CLADAG 2021

BOOK OF ABSTRACTS AND SHORT PAPERS

13th Scientific Meeting of the Classification and Data Analysis Group
Firenze, September 9-11, 2021

edited by
Giovanni C. Porzio
Carla Rampichini
Chiara Bocci

FIRENZE
UNIVERSITY
PRESS

PROCEEDINGS E REPORT

ISSN 2704-601X (PRINT) - ISSN 2704-5846 (ONLINE)

– 128 –

SCIENTIFIC PROGRAM COMMITTEE

Giovanni C. Porzio (chair) (University of Cassino and Southern Lazio - Italy)

Silvia Bianconcini (University of Bologna - Italy)
Christophe Biernacki (University of Lille - France)
Paula Brito (University of Porto - Portugal)
Francesca Marta Lilja Di Lascio (Free University of Bozen-Bolzano - Italy)
Marco Di Marzio ("Gabriele d'Annunzio" University of Chieti-Pescara - Italy)
Alessio Farcomeni ("Tor Vergata" University of Rome - Italy)
Luca Frigau (University of Cagliari - Italy)
Luis Ángel García Escudero (University of Valladolid - Spain)
Bettina Grün (Vienna University of Economics and Business - Austria)
Salvatore Ingrassia (University of Catania - Italy)
Volodymyr Melnykov (University of Alabama - USA)
Brendan Murphy (University College Dublin -Ireland)
Maria Lucia Parrella (University of Salerno - Italy)
Carla Rampichini (University of Florence - Italy)
Monia Ranalli (Sapienza University of Rome - Italy)
J. Sunil Rao (University of Miami - USA)
Marco Riani (University of di Parma - Italy)
Nicola Salvati (University of Pisa - Italy)
Laura Maria Sangalli (Polytechnic University of Milan - Italy)
Bruno Scarpa (University of Padua - Italy)
Mariangela Sciandra (University of Palermo - Italy)
Luca Scrucca (University of Perugia - Italy)
Domenico Vistocco (Federico II University of Naples - Italy)
Mariangela Zenga (University of Milan-Bicocca - Italy)

LOCAL PROGRAM COMMITTEE

Carla Rampichini (chair) (University of Florence - Italy)

Chiara Bocci (University of Florence - Italy)
Anna Gottard (University of Florence - Italy)
Leonardo Grilli (University of Florence - Italy)
Monia Lupparelli (University of Florence - Italy)
Maria Francesca Marino (University of Florence - Italy)
Agnese Panzera (University of Florence - Italy)
Emilia Rocco (University of Florence - Italy)
Domenico Vistocco (Federico II University of Naples - Italy)

CLADAG 2021
BOOK OF ABSTRACTS AND SHORT PAPERS

13th Scientific Meeting of the Classification
and Data Analysis Group
Firenze, September 9-11, 2021

edited by
Giovanni C. Porzio
Carla Rampichini
Chiara Bocci

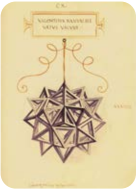
FIRENZE UNIVERSITY PRESS
2021

CLADAG 2021 BOOK OF ABSTRACTS AND SHORT PAPERS : 13th Scientific Meeting of the Classification and Data Analysis Group Firenze, September 9-11, 2021/ edited by Giovanni C. Porzio, Carla Rampichini, Chiara Bocci. — Firenze : Firenze University Press, 2021.
(Proceedings e report ; 128)

https://www.fupress.com/isbn/9788855183406

ISSN 2704-601X (print)
ISSN 2704-5846 (online)
ISBN 978-88-5518-340-6 (PDF)
ISBN 978-88-5518-341-3 (XML)
DOI 10.36253/978-88-5518-340-6

Graphic design: Alberto Pizarro Fernández, Lettera Meccanica SRLs
Front cover: Illustration of the statue by Giambologna, *Appennino* (1579-1580) by Anna Gottard



CLAssification and Data
Analysis Group (CLADAG)
of the Italian Statistical
Society (SIS)

Published by Firenze University Press
Firenze University Press
Università degli Studi di Firenze
via Cittadella, 7, 50144 Firenze, Italy
www.fupress.com

This book is printed on acid-free paper
Printed in Italy

MODEL-BASED CLUSTERING WITH SPARSE MATRIX MIXTURE MODELS

Andrea Cappozzo [1], Alessandro Casa[2] and Michael Fop[2]

[1] Deparment of Mathematics, Politecnico di Milano
(e-mail: andrea.cappozzo@polimi.it)

[2] School of Mathematics and Statistics, University College Dublin
(e-mail: alessandro.casa@ucd.ie, michael.fop@ucd.ie)

ABSTRACT: In recent years we are witnessing to an increased attention towards methods for clustering matrix-valued data. In this framework, matrix Gaussian mixture models constitute a natural extension of the model-based clustering strategies. Regrettably, the overparametrization issues, already affecting the vector-valued framework in high-dimensional scenarios, are even more troublesome for matrix mixtures. In this work we introduce a sparse model-based clustering procedure conceived for the matrix-variate context. We introduce a penalized estimation scheme which, by shrinking some of the parameters towards zero, produces parsimonious solutions when the dimensions increase. Moreover it allows cluster-wise sparsity, possibly easing the interpretation and providing richer insights on the analyzed dataset.

KEYWORDS: model-based clustering, penalized likelihood, sparse matrix estimation, EM-algorithm

1 Introduction

Model-based clustering represents a well established framework to cluster multivariate data. When dealing with continuous data, the generative mechanism is routinely described by means of Gaussian Mixture Models (GMMs). Partitions are obtained by exploiting the one-to-one correspondence between the groups and the components of the mixture. This approach has been used in many different applications; nonetheless GMMs tend to be over-parameterized in high-dimensional settings where their usefulness might be jeopardized.

This problem complicates even further in three-way data scenarios, where multiple variables are measured on different occasions for the considered units. Here matrix-variate distributions have often been used and embedded in the mixtures framework, thus providing a valid solution when partitions of matrices are required (Viroli, 2011). In spite of its strenght points, this approach is dramatically over-parameterized even in moderate dimensions. Therefore, we propose a penalized model-based clustering strategy in the matrix-variate framework. Our approach reduces the number of parameters to be estimated, by shrinking some of them towards zero, and possibly leads to a gain in terms of interpretability. The rest of the paper is organized as follows. In Section 2 we introduce matrix Gaussian mixture models (MGMMs) and we outline our proposal. An application to real world data is reported in Section 3 alongside with some concluding remarks and possible future research directions.

2 Penalized matrix-variate mixture model

Let $\mathbf{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}$ be a set of n matrices with $\mathbf{X}_i \in \mathbb{R}^{p \times q}$. MGMM provides an extension of the GMM when clustering of matrices are needed. The density of $\mathbf{X}_i$ is then expressed as follows

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

where $\Theta = \{\tau_k, M_k, \Omega_k, \Gamma_k\}_{k=1}^{K}$, τ_k's are the mixing proportions, with $\tau_k > 0$ and $\sum_k \tau_k = 1$. On the other hand $\phi_{(p\times q)}(\mathbf{X}_i; M_k, \Omega_k, \Gamma_k)$ denotes the density of a $p \times q$ matrix normal distribution where $M_k \in \mathbb{R}^{p\times q}$ is the mean of the *k-th* component while $\Omega_k \in \mathbb{R}^{p\times p}$ and $\Gamma_k \in \mathbb{R}^{q\times q}$ represent respectively the rows and the columns component precision matrices.

In (1) the number of parameters to estimate scales quadratically with both p and q, endangering the pratical usefulness of the model. Recently some solutions have been proposed, trying to overcome this issue (see Wang & Melnykov, 2020 and Sarkar *et al.* , 2020). These approaches present some drawbacks as they are computationally intensive and as they implement a rigid way to induce parsimony. Therefore in this work we take a different path, adopting a penalized estimation approach which implicitly assumes that M_k, Ω_k, Γ_k, for $k = 1, \dots, K$, possess some degree of sparsity.

To this aim, we introduce a penalized likelihood strategy to obtain $\hat{\Theta}$. The log-likelihood function to be maximized is defined as

$$\ell(\Theta;\mathbf{X}) = \sum_{i=1}^{n} \log\left\{\sum_{k=1}^{K} \tau_k \phi_{p\times q}(\mathbf{X}_i; M_k, \Omega_k, \Gamma_k)\right\} - p_{\lambda_1,\lambda_2,\lambda_3}(M_k, \Omega_k, \Gamma_k) \tag{2}$$

with the penalization term $p_{\lambda_1,\lambda_2,\lambda_3}(M_k, \Omega_k, \Gamma_k)$ equals to

$$p_{\lambda_1,\lambda_2,\lambda_3}(M_k, \Omega_k, \Gamma_k) = \sum_{k=1}^{K} \lambda_1 ||P_1 * M_k||_1 + \sum_{k=1}^{K} \lambda_2 ||P_2 * \Omega_k||_1 + \sum_{k=1}^{K} \lambda_3 ||P_3 * \Gamma_k||_1$$

Table 1. *Adjusted Rand Index (ARI) and number of free estimated parameters for three clustering procedures.*

	Sparsemixmat	Sarkar *et al.* , 2020	GMM
ARI	0.7883	0.7772	0.3841
# of parameters	218	275	850

P_1, P_2, P_3 are matrices with non-negative entries, $||A||_1 = \sum_{jh} |A_{jh}|$, $\lambda_1, \lambda_2, \lambda_3$ are the penalization parameters while $*$ denotes the element-wise product.
To estimate Θ, we devise an ad-hoc EM-algorithm which maximizes the *penalized complete data log-likelihood* associated with (2). The E-step computes class membership a posteriori probabilities via the standard updating formula. On the other hand the M-step consists of three partial optimization cycles. An estimate for M_k is obtained by means of a cell-wise coordinate ascent algorithm while, to estimate Ω_k and Γ_k, we propose a suitable modification of the graphical LASSO (Friedman *et al.* , 2008). The resulting model, inducing sparsity in the precision matrices, accounts for cluster-wise conditional independence patterns, which might ease the interpretation of the results, and possibly provides indications about irrelevant variables. Moreover the number of parameters is reduced without imposing rigid structures.

3 Application and concluding remarks

We employ the procedure outlined in Section 2 to obtain a partition of the Landsat satellite data, where $n = 845$ matrices, with dimensions 4×9, coming from three different classes are available (see Viroli, 2011 for a detailed description). In Table 1 we report the results obtained with the proposed procedure (Sparsemixmat) and with two plausible competitors being the approach by Sarkar *et al.* , 2020 and the standard GMM applied to the unfolded two-way representation of the data. Our model outperforms the competitors, when recovering the true clustering structure is the aim. Furthermore, we provide the most parsimonious solution, displaying the lowest number of non zero estimated parameters. The retrieved sparse matrix structures are graphically displayed, for the three classes, in Figure 1. While the clustering is mainly driven by the different patterns in M_k's, the Γ_k's are the ones showing the highest degree of sparsity, with different intensities for the three classes.

The promising results obtained in the application demonstrate how the penalized matrix-variate mixture model proposed in this work might alleviate the flaws of standard three-way data clustering in high-dimensional scenarios.

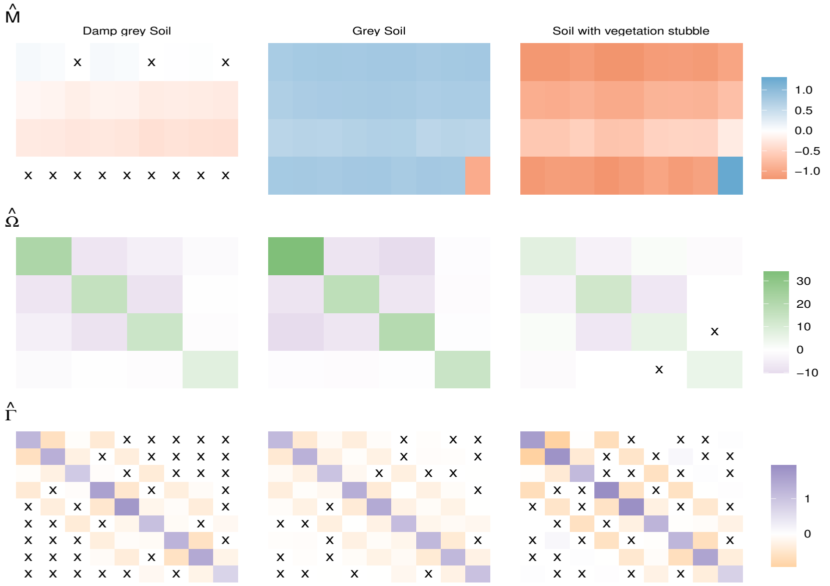


Figure 1. *Sparsely estimated M_k (upper plots), Ω_k (middle plots) and Γ_k (lower plots) for $k = 1,2,3$. Entries that are shrunk to 0 by the estimator are highlighted with an $\times$.*

Our proposal is able to effectively reduce the number of parameters to estimate while, at the same time, flexibly accounting for different relationships among the variables and for different level of sparsity across the groups. Future research directions would focus on the derivation of an appropriate model selection procedure, determining jointly reasonable values for the penalty coefficients as well as for the number of mixture components.

References

FRIEDMAN, J., HASTIE, T., & TIBSHIRANI, R. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**(3), 432–441.

SARKAR, S., ZHU, X., MELNYKOV, V., & INGRASSIA, S. 2020. On parsimonious models for modeling matrix data. *Computational Statistics & Data Analysis*, **142**, 106822.

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

WANG, Y., & MELNYKOV, V. 2020. On variable selection in matrix mixture modelling. *Stat*, **9**(1), e278.