Sparse model-based clustering of three-way data via lasso-type penalties

Andrea Cappozzo*
Department of Economics, Management and Quantitative Methods,
University of Milan

Alessandro Casa*
Faculty of Economics and Management, Free University of Bozen-Bolzano

Michael Fop
School of Mathematics & Statistics, University College Dublin

November 11, 2024

Abstract

Mixtures of matrix Gaussian distributions provide a probabilistic framework for clustering continuous matrix-variate data, which are increasingly common in various fields. Despite their widespread use and successful applications, these models suffer from over-parameterization, making them not suitable for even moderately sized matrix-variate data. To address this issue, we introduce a sparse model-based clustering approach for three-way data. Our approach assumes that the matrix mixture parameters are sparse and have different degrees of sparsity across clusters, enabling the induction of parsimony in a flexible manner. Estimation relies on the maximization of a penalized likelihood, with specifically tailored group and graphical lasso penalties. These penalties facilitate the selection of the most informative features for clustering three-way data where variables are recorded over multiple occasions, as well as allowing the identification of cluster-specific association structures. We conduct extensive testing of the proposed methodology on synthetic data and validate its effectiveness through an application to time-dependent crime patterns across multiple U.S. cities. Supplementary files for this article are available online.

Keywords: Group lasso, Matrix-variate data, Model-based clustering, Penalized likelihood, Sparse estimation

*These authors contributed equally to this work

1 Introduction

Matrix-variate data, where a matrix is observed for each statistical unit, are becoming more common in a large number of applications and data analysis routines. This data structure is often referred to as *three-way* and characterized by the presence of three different layers or modes; namely the units, the variables and the occasions. These data are nowadays often occurring in applications such as basketball analytics (Yin et al., 2023), export trade networks (Melnykov et al., 2021), image and brain scan data (Gao et al., 2021; Liu et al., 2022), multivariate time-dependent analysis (Anderlucci and Viroli, 2015). In this work, we focus on analyzing the evolution of crime trends across multiple U.S. cities (see, e.g., Melnykov and Zhu, 2019). In the dataset considered, each city serves as a statistical unit, with annual crime rates recorded for seven types of crimes (the variables) over a 13-year span from 2000 to 2012, representing distinct time occasions. Consequently, the data for each statistical unit can be naturally structured as a 7×13 matrix.

In spite of their potential in terms of informative content, matrix-variate data introduce several challenges which need to be dealt with in the modeling process. In fact, each of the three different layers induce specific peculiarities in terms of intricate dependency structures. In this landscape, clustering is often of interest to reduce the aforementioned complexities by proposing parsimonious summaries of the data and highlighting their most important patterns. To this extent, both distance-based (Vichi, 1999; Vichi et al., 2007) and nonparametric techniques (Ferraccioli and Menardi, 2023) have been proposed. Nevertheless, parametric or model-based approaches are undoubtedly the ones that have received the most attention: taking steps from Basford and McLachlan (1985) and building on mixtures of matrix-variate Gaussian distributions, the seminal papers by Viroli (2011a,b) have paved the way for a new and lively stream of research. Recently, several flexible approaches have been proposed for model-based clustering of matrix-variate data of different nature. These methodologies considered alternative distributional assumptions for skewed data (Chen and Gupta, 2005; Melnykov and Zhu, 2018; Gallaugher and McNicholas, 2018), transformations (see, among others, Chen and Gupta, 2005; Melnykov and Zhu, 2018; Gallaugher and McNicholas, 2018; Tomarchio et al., 2020, 2022; Tomarchio, 2022) and alternative models for

count data (Silva et al., 2023; Subedi, 2023).

Despite being practically useful, matrix-variate model-based clustering faces significant limitations in high-dimensional settings. These limitations are particularly pronounced in the three-way framework where the tendency to over-parameterization, inherited from the vector-valued setting (Bouveyron and Brunet-Saumard, 2014), becomes even more problematic. In fact, when employing matrix Gaussian distributions (Dawid, 1981) as component densities, two covariance matrices are employed to accommodate the data structure. Consequently, when dense parameterizations are assumed for these matrices, the number of parameters to be estimated grows quadratically with both row and column dimensions. This undermines the practical utility of the approach, even when a moderate number of variables and/or occasions are observed.

To address these limitations, in this work we introduce a novel approach where each parameter in the specification of the matrix Gaussian mixture model has its own cluster-specific degree of sparsity. This greatly increases the flexibility of the model, leads to a parsimonious modeling framework, and provides more interpretable insights regarding the clustering partition. The approach relies on the maximization of a penalized likelihood which automatically enforces sparsity. More specifically, we impose a graphical lasso penalty on row and column precision matrices, promoting a reduction in the number of association parameters while facilitating interpretation in terms of conditional dependencies, thanks to the connection with Gaussian graphical models. Additionally, we impose a group lasso penalty on the rows of the component mean matrices. In the common scenario where variables are observed over time for a set of statistical units, this penalization scheme allows to perform automatic variable selection in a three-way model-based clustering framework. As a supplementary contribution, we briefly generalize the applicability of the work by Heo and Baek (2021), where they consider a lasso entry-wise penalty for the elements of the mean matrices.

The remainder of the paper is structured as follows. Section 2 provides an overview of model-based clustering of matrix-variate data, with a specific focus on the issues arising in large dimensional scenarios. In Section 3, our proposal is introduced and motivated,

alongside with the description of the associated estimation and model selection methods. In Section 4 and 5, the performance of the proposed framework is tested on synthetic and real data, respectively. Conclusions and considerations about further improvements and future research directions end the paper in Section 6.

2 Model-based matrix-variate clustering

2.1 Mixture of matrix normal distributions

Model-based clustering (Fraley and Raftery, 2002; Bouveyron et al., 2019) assumes that the data are generated by a finite mixture distribution, which describes the presence of heterogeneous sub-populations. In this context, maximum likelihood estimation is usually implemented by means of the EM algorithm (Dempster et al., 1977), resorting to a data augmentation scheme where the latent group indicator variables are treated as missing data. Operationally, once the model is fitted, a partition is obtained by assuming a one-to-one correspondence between the groups and the mixture components, and assigning the i-th observation to a given cluster according to the maximum a posteriori (MAP) rule (see Fraley and Raftery, 2002; Bouveyron et al., 2019, for a detailed discussion).

When dealing with standard continuous vector-variate data, where multiple variables are measured for a set of units, it is routinely assumed that the mixture components correspond to multivariate Gaussian distributions (Fraley and Raftery, 2002). Nonetheless, nowadays it is becoming increasingly common to encounter three-way data structures, where multiple variables are measured over different occasions. This additional layer (or mode) introduces new modeling challenges that need to be taken into account when clustering samples is the final goal. Indeed, as noted by Anderlucci and Viroli (2015), models have to "*account simultaneously for three goals of the analysis, which arise from the three layers of the data structure; heterogeneous units, correlated occasions and dependent variables*". Matrix Gaussian mixture models have originally been proposed by Viroli (2011a,b) with the aim of concurrently accounting for these sources of complexity.

Consider the sample $\mathbf{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$, which in this framework is a collection of $p \times q$

matrices, with $\mathbf{X}_i \in \mathbb{R}^{p\times q}$, $i = 1, \ldots, n$. While the n matrices can generally contain any type of measurement, in the following we assume a temporal structure where p variables are observed in q different time occasions. This is suitable for most applications, including the one considered in Section 5, where crime time evolution is explored. Lastly, note that in this setting we assume that each unit has precisely the same p variables observed at the same q occasions. The natural Gaussian mixture model (GMM) extension for model-based clustering of three-way data is given by the matrix Gaussian mixture model (MGMM), expressed as follows:

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

where τ_k's are the mixing proportions with $\tau_k > 0$, $\forall k = 1, \ldots, K$ and $\sum_{k=1}^{K} \tau_k = 1$, K is the number of mixture components, while $\boldsymbol{\Theta}$ denotes the collection of all mixture parameters. Here, $\phi_{p\times q}(\cdot; \mathbf{M}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Psi}_k)$ denotes the density of a $p \times q$ matrix normal distribution (Dawid, 1981), reading as

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

where $\mathbf{M}_k$ is the $p \times q$ mean matrix, $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\Psi}_k$ are the row- and column-covariance matrices of the k-th component, with dimensions $p \times p$ and $q \times q$, respectively. Similar to the two-way scenario, the model in (1) can be estimated by means of the EM-algorithm, see for example Viroli (2011a); Glanz and Carvalho (2018); Gao et al. (2021). Alternatively, the model can also be formulated and estimated under a Bayesian framework, as proposed for example in Viroli (2011b); Yin et al. (2023).

An alternative specification of the matrix-variate Gaussian distribution may be given, since the following relation holds

$$\mathbf{X}_i \sim m\mathcal{N}_{p\times q}(\mathbf{M}, \boldsymbol{\Sigma}, \boldsymbol{\Psi}) \iff \mathrm{vec}(\mathbf{X}_i) \sim \mathcal{N}_{pq}(\mathrm{vec}(\mathbf{M}), \boldsymbol{\Psi} \otimes \boldsymbol{\Sigma}), \tag{2}$$

where $\mathrm{vec}(\cdot)$ and $\otimes$ denote the vectorization operator and the Kronecker product, respectively, while $m\mathcal{N}_{p\times q}$ denotes a matrix Normal distribution of dimensions p and q. From this relation, the matrix-variate Gaussian can be regarded as a direct generalization of

the normal distribution to the three-way matrix framework. For more details about the matrix Gaussian distribution, its properties, and its connection to the multivariate normal distribution, readers can refer to Gupta and Nagar (2018). The presence of the Kronecker product in (2) highlights an identifiability issue, since $\boldsymbol{\Psi} \otimes \boldsymbol{\Sigma} = c\boldsymbol{\Psi} \otimes c^{-1}\boldsymbol{\Sigma}$ for any $c \in \mathbb{R}^+$. Enforcing constraints on the trace or on the determinant of one of the two matrices is regarded as a viable solution to solve the problem (see e.g., Viroli, 2012; Melnykov and Zhu, 2018; Glanz and Carvalho, 2018); the latter approach will be considered in the rest of the paper. As a final point, note that, as is customary in penalized model-based clustering (see, for example Pan and Shen, 2007; Xie et al., 2008b,a; Zhou et al., 2009), we assume that the data matrix $\mathbf{X}$ is mean-centered element-wise. This enables the component means to be interpreted as weighted shifts around a global zero mean, clarifying which variables contribute to distinguishing the clusters. The reasoning behind this approach will become evident in the model specification discussed in Section 3.1.

2.2 Issues in matrix mixture models for high-dimensional data

Finite mixture models are routinely used for probabilistic cluster analysis. However, both in the two-way and in the three-way framework, they present a cumbersome issue which is related to their tendency to be over-parameterized even with a moderate number of variables. When dealing with vector-variate data, the cardinality of the parameter space $|\boldsymbol{\Theta}|$ scales quadratically with the number of variables. This problem is even more exacerbated in the matrix-variate scenario, where $|\boldsymbol{\Theta}|$ scales quadratically with both dimensions p and q of the component row and column covariance matrices. To deal with this challenge, different approaches have been proposed in the two-way setting (see e.g., Bouveyron and Brunet-Saumard, 2014; Fop and Murphy, 2018, for exhaustive reviews of the topic), which can be grouped into three distinct types: constrained modeling, variable selection, and sparse modeling; a brief overview is provided in Casa et al. (2022).

In line with this taxonomy, recent efforts have been devoted to addressing the issue of over-parameterization within the framework of matrix mixture modeling. Specifically, some of the existing approaches either adopt parsimonious parameterizations, or implement

variable selection to discard irrelevant variables and reduce the number of parameters. In Sarkar et al. (2020), the authors extend the family of covariance eigendecomposition models considered for vector-valued data (Banfield and Raftery, 1993; Celeux and Govaert, 1995) to the matrix-variate scenario. They introduced a collection of 98 constrained models and further enhanced parsimony by proposing an additive formulation for the mean matrices, resulting in a family of 196 matrix mixture models. On the other hand, Wang and Melnykov (2020) propose a variable selection approach where the work by Maugis et al. (2009) is extended to the matrix-variate framework. A stepwise variable selection procedure is proposed, which alternates variable inclusion and exclusion steps, where the resulting models are compared by means of the Bayesian Information Criterion (BIC, Schwarz, 1978). These two approaches present some relevant drawbacks: they can be computationally intensive, as they involve fitting and comparing a large number of models, and they implement a rigid way to induce parsimony, not allowing the association patterns among the variables and the structure of the mean matrices to vary across clusters.

For the above reasons, in this work we take a different perspective, based on the formulation of a sparse matrix mixture model, by extending the framework of sparse and penalized mixture models (Fop and Murphy, 2018; Fop et al., 2019; Casa et al., 2022, among others) to matrix-variate data. Building primarily upon the literature on sparse matrix graphical models (Leng and Tang, 2012; Chen and Liu, 2019, for example) and sparse model-based clustering (Pan and Shen, 2007; Zhou et al., 2009), sparse approaches for matrix-variate data clustering have been recently introduced and are gaining increasing attention. In application to brain imaging data, Gao et al. (2021) develop a penalized Gaussian matrix mixture model, where penalty functions on the entries of the component mean matrices are introduced to shrink the mean parameters. The method is shown to recover the low rank mean signal, however, it does not allow a flexible modeling of the association structure across the clusters. On a similar vein, Liu et al. (2022) presents a multi-step approach for clustering and sparse correlation estimation in application to functional magnetic resonance imaging data. Here, in contrast to Gao et al. (2021) and motivated by the application, the authors propose an optimization framework that focuses on recovering the

different association structures across the clusters, but covariance parameters rather than the means are employed to cluster the units, which could be a limitation if clusters are well separated in terms of mean signals. Additionally, the authors remark that the method suffers from the need to pre-specify the number of clusters beforehand and the lack of a principled method for its selection. Lastly, in Heo and Baek (2021), the authors describe a penalized matrix normal mixture model for clustering that employs penalty functions on both means and precision matrix parameters to induce sparsity. However, this approach relies on implicit restrictive independence assumptions during estimation that may hinder the correct identification of the actual clustering partition. In fact, neglecting the associations among variables and occasions often leads to a larger number of components needed to appropriately capture the shape of the clusters (Biernacki et al., 2000, for example). Moreover, the specific formulation of the penalty functions on the mean parameters does not allow for an effective variable selection in the context of three-way data where variables are measured over multiple occasions.

In what follows, we propose a sparse matrix Gaussian mixture model where we overcome the drawbacks of the aforementioned frameworks for three-way data clustering. Our proposed approach offers several advantages: it allows clusters to be characterized by different association structures, accommodating the estimation of sparse component inverse covariance matrices. Moreover, it allows mean parameters to have different sparsity patterns across clusters, and it implements variable selection in a matrix-variate context where the variables are observed over multiple time occasions. Lastly, it leverages a computationally efficient estimation procedure based on lasso-type penalties, and it considers a principled criterion to perform model selection. The proposed method is based on a penalized likelihood framework, presented in the next section.

3 Sparse matrix mixture models

3.1 Model specification

We hereafter introduce *Sparsemixmat*, a novel sparse matrix Gaussian mixture model for model-based clustering of matrix-variate data. Estimation of this model relies on the maximization of a penalized likelihood, which, following from the model in (1), is defined as follows:

$$\ell_P(\boldsymbol{\Theta}; \mathbf{X}) = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \tau_k \phi_{p\times q}(\mathbf{X}_i; \mathbf{M}_k, \boldsymbol{\Omega}_k, \boldsymbol{\Gamma}_k) \right\} - p_{\boldsymbol{\lambda}}(\boldsymbol{\Theta}). \tag{3}$$

The first term represents the standard MGMM log-likelihood, $p_{\boldsymbol{\lambda}}(\boldsymbol{\Theta})$ is a penalty term depending on a set of shrinkage hyperparameters generally denoted with $\boldsymbol{\lambda}$, while $\boldsymbol{\Omega}_k = \boldsymbol{\Sigma}_k^{-1}$ and $\boldsymbol{\Gamma}_k = \boldsymbol{\Psi}_k^{-1}$ for $k = 1, \ldots, K$ are the rows and column precision matrices, respectively. The collection of parameters is denoted with $\boldsymbol{\Theta} = \{\tau_k, \mathbf{M}_k, \boldsymbol{\Omega}_k, \boldsymbol{\Gamma}_k\}_{k=1}^{K}$. The choice of parameterizing the MGMM density in terms of precision matrices is motivated by their relation to Gaussian graphical models and their interpretation in terms of conditional dependencies (Whittaker, 1990; Leng and Tang, 2012). However, other options could be considered, and a discussion is reported in Section 6.

Different routes can be taken when specifying the penalty $p_{\boldsymbol{\lambda}}(\boldsymbol{\Theta})$ to obtain sparse estimates of the mixture component matrix parameters; readers may refer to the recent book by Hastie et al. (2019) for a detailed discussion. In this work, we consider the following penalty term:

$$p_{\boldsymbol{\lambda}}(\boldsymbol{\Theta}) = \lambda_1 \sum_{k=1}^{K} \sum_{r=1}^{p} \|\mathbf{m}_{r\cdot,k}\|_2 + \lambda_2 \sum_{k=1}^{K} \|\mathbf{P}_2 \odot \boldsymbol{\Omega}_k\|_1 + \lambda_3 \sum_{k=1}^{K} \|\mathbf{P}_3 \odot \boldsymbol{\Gamma}_k\|_1, \tag{4}$$

where $\mathbf{m}_{r\cdot,k}$ is the r-th row of matrix $\mathbf{M}_k$, while $\|\cdot\|_1$ and $\|\cdot\|_2$ are the L_1 and the L_2-norm respectively, with $\|A\|_1 = \sum_{jh} |A_{jh}|$. Moreover, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$ is a vector of positive shrinkage hyper-parameters controlling the strength of the penalization. Lastly, $\mathbf{P}_2$ and $\mathbf{P}_3$ are matrices with non-negative entries, and $\odot$ denotes the element-wise product.

The first term in (4) corresponds to a group lasso penalty (Yuan and Lin, 2006) imposed on the rows of the K mean matrices $\mathbf{M}_k$. Group lasso aims to simultaneously shrink to zero a set of grouped parameters. While being primarily used in regression frameworks

where some covariates may be structurally connected (see Ch.4.3 in Hastie et al., 2019, and references therein), this type of penalty has also been employed in model-based clustering of vector-variate data in Xie et al. (2008a) and Xie et al. (2008b). In these works, group lasso is used for joint penalization of the mixture parameters across clusters in the case where variables are grouped based on available prior information. Differently, in the present paper we propose to generalize the group lasso penalty to an unsupervised setting for matrix-variate clustering and variable selection. Specifically, we consider the parameters as being grouped according to the rows of $\mathbf{M}_k$. With the grouped lasso penalty on $\mathbf{M}_k$, for a given k, either the whole row $\mathbf{m}_{r\cdot,k} = (m_{r1,k}, \ldots, m_{rq,k})$ is estimated to be zero, or else its elements are shrunk towards zero by an amount depending on λ_1. If a row mean $\mathbf{m}_{r\cdot,k} = \mathbf{0}$ for all k, the r-th row of $\mathbf{M}_k$ is constant across all occasions and clusters. As a result, the variable is not useful for discriminating the mean levels of the clusters, since the cluster centers over that dimension and across time instants are indistinguishable from the zero center in the mean-centered data. Even when $\mathbf{m}_{r\cdot,k} = \mathbf{m}_{r\cdot,h} = \mathbf{0}$ for some components k and h, the r-th variable does not contain discriminating information to separate these clusters, resulting in overlap with the center of the data along that dimension. This penalization is therefore adopted to perform variable selection in model-based clustering of three-way data in the common scenario when p variables are observed over q time occasions. Note also that shrinking towards zero the row mean $\mathbf{m}_{r\cdot,k}$ of cluster k in the mean-centered data is equivalent to shrinking the component row mean towards the corresponding sample row mean in the original, un-centered data.

Going back to Equation (4), with the second and the third terms we impose graphical lasso penalties (see Banerjee et al., 2008; Friedman et al., 2008; Witten et al., 2011) on the group-specific precision matrices. This represents an extension of the work by Leng and Tang (2012) to the framework of mixture models. By shrinking to zero some parameters, the penalty terms allow to alleviate the over-parameterization problems outlined in Section 2.2 when dealing with matrix data of moderate-to-large dimensions. The resulting sparse representation of $\mathbf{\Omega}_k$ and $\mathbf{\Gamma}_k$, for $k = 1, \ldots, K$, provides a convenient interpretation of the dependencies among rows and columns of the observed matrices. In fact, zero entries in

the precision matrices imply that the corresponding variables are conditionally independent given the others, following the principles of Gaussian graphical models (Whittaker, 1990). The matrices $\mathbf{P}_2$ and $\mathbf{P}_3$ in the graphical lasso penalty terms provide an higher degree of flexibility, since particular specifications allow to introduce prior beliefs regarding the dependencies between the variables. Indications on how to choose these matrices can be found in Bien and Tibshirani (2011). Here the authors suggest to use all-ones matrices, ensuring homogeneous and uninformative penalization for all the terms. To prevent shrinkage of the diagonal entries, zeros can be placed on the main diagonal. Alternatively, $\mathbf{P}_2$ and $\mathbf{P}_3$ can be defined as adjacency matrices with user-defined patterns, thus allowing the a priori specification of the expected conditional dependence structures. More recently, Casa et al. (2022) introduced a data-driven method for specifying these matrices, which promotes cluster separation within the context of sparse model-based clustering and does not require initial knowledge of the association structure between the variables. In what follows we employ all-one matrices with zero diagonal entries for both $\mathbf{P}_2$ and $\mathbf{P}_3$, as this aspect is not the primary focus of the present paper.

The above-mentioned methodology is based on the assumption that all the parameter matrices in (3), namely $\{\mathbf{M}_k, \boldsymbol{\Omega}_k, \boldsymbol{\Gamma}_k\}_{k=1}^K$, have different component-specific levels of sparsity. This leads to a realistic and flexible modeling framework, where a variable may be relevant only for a subset of clusters, and where the conditional dependence patterns are allowed to vary across groups. Our proposal represents a natural extension to the three-way data scenario of the approach outlined by Zhou et al. (2009). Coherently with their work, the penalty on $\mathbf{M}_k$ aims to perform variable selection. On the other hand, the penalizations on $\boldsymbol{\Omega}_k$ and $\boldsymbol{\Gamma}_k$ are implemented to obtain sparse representations of the precision matrices and to reduce the number of free parameters to be estimated.

3.2 Model estimation

For a fixed number of components K and penalty vector $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$, the parameters are estimated by maximizing (3) with respect to $\boldsymbol{\Theta}$. The maximization is carried out by means of a tailored EM algorithm for maximum penalized likelihood estimation (Green,

1990; McLachlan and Krishnan, 2008), where the maximization step (M-step) is comprised of three partial optimization cycles. Let us firstly define the *penalized complete-data log-likelihood* associated with (3) as

$$\ell_C\left(\boldsymbol{\Theta}; \mathbf{X}, \mathbf{Z}\right) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\left[\log\tau_k - \frac{pq}{2}\log 2\pi + \frac{q}{2}\log\left|\boldsymbol{\Omega}_k\right| + \frac{p}{2}\log\left|\boldsymbol{\Gamma}_k\right| + \right.$$
$$\left. -\frac{1}{2}\operatorname{tr}\left\{\boldsymbol{\Omega}_k\left(\mathbf{X}_i - \mathbf{M}_k\right)\boldsymbol{\Gamma}_k\left(\mathbf{X}_i - \mathbf{M}_k\right)'\right\}\right] - p_{\boldsymbol{\lambda}}(\boldsymbol{\Theta}), \quad (5)$$

where z_{ik} is the realization of the latent group membership indicator variable Z_{ik}, with $z_{ik} = 1$ if matrix $\mathbf{X}_i$ belongs to the k-th component, and 0 otherwise. The posterior probability of Z_{ik} is updated at each expectation step (E-step), allowing to obtain the conditional expectation of (5), usually called Q-function, which defines the objective function to be maximized in the M-step. The estimation algorithm is described in detail in the following subsections.

3.2.1 Initialization strategy

Initialization plays a crucial role when resorting to EM-type algorithms to perform model estimation. In fact, whenever the likelihood surface has multiple modes, the convergence to the global maximum is not guaranteed and poorly chosen initial values may lead to sub-optimal solutions (McLachlan and Krishnan, 2008). Thanks to the correspondence between GMM and MGMM in Equation (2), initialization strategies developed for vector-variate data can be directly employed in the matrix-variate framework. In this regard, after the data have been vectorized, we resort to model-based agglomerative hierarchical clustering (Scrucca and Raftery, 2015). This initialization strategy, already employed in the popular `mclust` software (Scrucca et al., 2016), has been proven effective in partitioning the data into K initial groups.

Once the starting partition is obtained, the first iteration of the M-step also requires the initialization of the matrices $\boldsymbol{\Omega}_k$ and $\boldsymbol{\Gamma}_k$, $k = 1, \ldots, K$. For this purpose, identity matrices of dimensions respectively equal to $p \times p$ and $q \times q$ are employed as initial values.

3.2.2 E-step

At iteration t, the estimated a posteriori probabilities $\hat{z}_{ik}^{(t)} = \widehat{\Pr}\left(Z_{ik} = 1 \mid \mathbf{X}_i\right)$ are updated as follows:

$$\hat{z}_{ik}^{(t)} = \frac{\hat{\tau}_k^{(t-1)} \phi_{p\times q}\left(\mathbf{X}_i; \hat{\mathbf{M}}_k^{(t-1)}, \hat{\boldsymbol{\Omega}}_k^{(t-1)}, \hat{\boldsymbol{\Gamma}}_k^{(t-1)}\right)}{\sum_{v=1}^{K} \hat{\tau}_v^{(t-1)} \phi_{p\times q}\left(\mathbf{X}_i; \hat{\mathbf{M}}_v^{(t-1)}, \hat{\boldsymbol{\Omega}}_v^{(t-1)}, \hat{\boldsymbol{\Gamma}}_v^{(t-1)}\right)}, \quad i = 1, \ldots, n, \quad k = 1, \ldots, K,$$

where with the superscript $(t-1)$ we denote the parameter estimates obtained in the previous EM iteration.

3.2.3 M-step

The M-step requires the maximization of the penalized Q-function, defined as

$$\begin{aligned} Q\left(\boldsymbol{\tau}, \{\mathbf{M}_k, \boldsymbol{\Omega}_k, \boldsymbol{\Gamma}_k\}_{k=1}^{K}\right) = \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{z}_{ik}^{(t)} \Big[\log \tau_k + \frac{q}{2}\log|\boldsymbol{\Omega}_k| + \frac{p}{2}\log|\boldsymbol{\Gamma}_k| + \\ -\frac{1}{2}\operatorname{tr}\left\{\boldsymbol{\Omega}_k \left(\mathbf{X}_i - \mathbf{M}_k\right)\boldsymbol{\Gamma}_k \left(\mathbf{X}_i - \mathbf{M}_k\right)'\right\}\Big] + \\ - \lambda_1 \sum_{k=1}^{K}\sum_{r=1}^{p} \|\mathbf{m}_{r\cdot,k}\|_2 - \lambda_2 \sum_{k=1}^{K} \|\mathbf{P}_2 \odot \boldsymbol{\Omega}_k\|_1 - \lambda_3 \sum_{k=1}^{K} \|\mathbf{P}_3 \odot \boldsymbol{\Gamma}_k\|_1. \quad (6) \end{aligned}$$

The direct maximization of $Q(\cdot)$ with respect to all parameters at once is an unfeasible task, so a partial optimization strategy is required. The closed-form expression for the mixing proportions $\boldsymbol{\tau}$ is readily available:

$$\hat{\tau}_k^{(t)} = \frac{\hat{n}_k^{(t)}}{n}, \quad \hat{n}_k^{(t)} = \sum_{i=1}^{n} \hat{z}_{ik}^{(t)}, \quad k = 1, \ldots, K.$$

Updates for $\mathbf{M}_k$, $\boldsymbol{\Omega}_k$, and $\boldsymbol{\Gamma}_k$, $k = 1, \ldots, K$ are implemented using custom procedures described in what follows.

Sparse estimation of the mean matrices $\mathbf{M}_k$

When maximization of (6) is performed with respect to $\mathbf{M}_k$, given current estimates of the precision matrices $\hat{\mathbf{\Omega}}_k^{(t-1)}$ and $\hat{\mathbf{\Gamma}}_k^{(t-1)}$, the Q-function simplifies as follows

$$\begin{aligned} Q_M(\mathbf{M}_k) &= \sum_{i=1}^{n} \hat{z}_{ik}^{(t)} \left[\operatorname{tr}\left\{ \hat{\mathbf{\Omega}}_k^{(t-1)} \mathbf{X}_i \hat{\mathbf{\Gamma}}_k^{(t-1)} \mathbf{M}_k' \right\} - \frac{1}{2} \operatorname{tr}\left\{ \hat{\mathbf{\Omega}}_k^{(t-1)} \mathbf{M}_k \hat{\mathbf{\Gamma}}_k^{(t-1)} \mathbf{M}_k' \right\} \right] - \lambda_1 \sum_{r=1}^{p} \|\mathbf{m}_{r\cdot,k}\|_2 \\ &= \operatorname{tr}\left\{ \hat{\mathbf{\Omega}}_k^{(t-1)} \boldsymbol{S}_M \hat{\mathbf{\Gamma}}_k^{(t-1)} \mathbf{M}_k' \right\} - \frac{\hat{n}_k^{(t)}}{2} \operatorname{tr}\left\{ \hat{\mathbf{\Omega}}_k^{(t-1)} \mathbf{M}_k \hat{\mathbf{\Gamma}}_k^{(t-1)} \mathbf{M}_k' \right\} - \lambda_1 \sum_{r=1}^{p} \|\mathbf{m}_{r\cdot,k}\|_2, \end{aligned} \tag{7}$$

where $\boldsymbol{S}_M$ is the sum of the matrix-variate observations weighted by $\hat{z}_{ik}^{(t)}$:

$$\boldsymbol{S}_M = \sum_{i=1}^{n} \hat{z}_{ik}^{(t)} \mathbf{X}_i.$$

The optimization of (7) with respect to $\mathbf{M}_k$ is solved via a proximal gradient descent algorithm (Parikh and Boyd, 2014). Briefly, proximal gradient methods address a general class of convex problems where the objective function may be decomposed into two terms: the first, generally denoted with $f(\cdot)$, is convex and differentiable, while the other, $g(\cdot)$, may not be everywhere differentiable. On that account, proximal gradient methods, also known as forward backward splitting procedures, can be seen as an extension of gradient descent for optimization problems whose gradient is not available for the entire objective function. In recent years, such approaches have gained increasing popularity in the field of statistics and machine learning, as they provide reliable and numerically efficient solutions to regularized models with non-differentiable penalties (Mosci et al., 2010; Klosa et al., 2020). In our case, the maximization of (7) can be recast as follows:

$$\underset{\mathbf{M}_k}{\text{minimize}}\, f(\mathbf{M}_k) + g(\mathbf{M}_k),$$

where

$$f(\mathbf{M}_k) = \frac{\hat{n}_k^{(t)}}{2} \operatorname{tr}\left\{ \hat{\mathbf{\Omega}}_k^{(t-1)} \mathbf{M}_k \hat{\mathbf{\Gamma}}_k^{(t-1)} \mathbf{M}_k' \right\} - \operatorname{tr}\left\{ \hat{\mathbf{\Omega}}_k^{(t-1)} \boldsymbol{S}_M \hat{\mathbf{\Gamma}}_k^{(t-1)} \mathbf{M}_k' \right\} \quad \text{and} \quad g(\mathbf{M}_k) = \lambda_1 \sum_{r=1}^{p} \|\mathbf{m}_{r\cdot,k}\|_2.$$

Define $\nabla \mathbf{m}_{l\cdot,k}$ to be the l-th row, $l = 1, \ldots, p$, of

$$\frac{\partial f(\mathbf{M}_k)}{\partial \mathbf{M}_k} = \hat{n}_k^{(t)} \hat{\mathbf{\Omega}}_k^{(t-1)} \mathbf{M}_k \hat{\mathbf{\Gamma}}_k^{(t-1)} - \hat{\mathbf{\Omega}}_k^{(t-1)} \boldsymbol{S}_M \hat{\mathbf{\Gamma}}_k^{(t-1)}, \tag{8}$$

where (8) is the $p \times q$ matrix of first-order partial derivatives of $f(\cdot)$ with respect to $\mathbf{M}_k$. A proximal gradient update for the l-th row of matrix $\mathbf{M}_k$ is constructed as follows:

$$\boldsymbol{b} = \mathbf{m}_{l\cdot,k} - \nu \nabla \mathbf{m}_{l\cdot,k}, \tag{9a}$$

$$\hat{\mathbf{m}}_{l\cdot,k} = \text{prox}_{\nu\lambda_1}(\boldsymbol{b}), \tag{9b}$$

where ν is a step-size parameter and $\text{prox}_{\nu\lambda_1}(\cdot)$ is the proximity operator of the considered group lasso penalty, namely the row-wise soft thresholding operator:

$$\text{prox}_{\nu\lambda_1}(\boldsymbol{b}) = \begin{cases} \boldsymbol{b}\left(1 - \frac{\lambda_1 \nu}{\|\boldsymbol{b}\|_2}\right) & \text{if } \|\boldsymbol{b}\|_2 > \lambda_1 \nu, \\ \mathbf{0} & \text{if } \|\boldsymbol{b}\|_2 \leq \lambda_1 \nu. \end{cases} \tag{10}$$

Iterating equations (9a) and (9b) until convergence sequentially along the p rows yields $\hat{\mathbf{M}}_k^{(t)}$, the estimate of the k-th component mean matrix for the t-th iteration of the EM algorithm. This estimate is the proximal gradient solution to the maximization problem in (7). When λ_1 is sufficiently large, the rows of $\hat{\mathbf{M}}_k^{(t)}$ are set to zero as a result of the proximity operator. In practice, the weighted sample mean matrix $\sum_{i=1}^n \hat{z}_{ik}^{(t)} \mathbf{X}_i / \hat{n}_k^{(t)}$ is employed as an initial guess for starting the proximal gradient search, while the step-size parameter ν is kept fixed at 10^{-4}.

Sparse estimation of the row-precision matrices $\boldsymbol{\Omega}_k$

When (6) is to be maximized with respect to $\boldsymbol{\Omega}_k$, given the current estimates of the precision matrices $\hat{\boldsymbol{\Gamma}}_k^{(t-1)}$ and of the mean parameters $\hat{\mathbf{M}}_k^{(t)}$, the Q-function simplifies as follows:

$$Q_\Omega(\boldsymbol{\Omega}_k) = \sum_{i=1}^n \hat{z}_{ik}^{(t)} \left[\frac{q}{2} \log|\boldsymbol{\Omega}_k| - \frac{1}{2} \text{tr}\left\{\boldsymbol{\Omega}_k \left(\mathbf{X}_i - \hat{\mathbf{M}}_k^{(t)}\right) \hat{\boldsymbol{\Gamma}}_k^{(t-1)} \left(\mathbf{X}_i - \hat{\mathbf{M}}_k^{(t)}\right)'\right\}\right] - \lambda_2 ||\mathbf{P}_2 \odot \boldsymbol{\Omega}_k||_1. \tag{11}$$

By rearranging terms in (11), we obtain:

$$Q_\Omega(\boldsymbol{\Omega}_k) = \log|\boldsymbol{\Omega}_k| - \text{tr}\{\boldsymbol{\Omega}_k \boldsymbol{S}_\Omega\} - \frac{2}{\hat{n}_k q} \lambda_2 ||\mathbf{P}_2 \odot \boldsymbol{\Omega}_k||_1, \tag{12}$$

where

$$\boldsymbol{S}_\Omega = \sum_{i=1}^n \hat{z}_{ik}^{(t)} \frac{\left(\mathbf{X}_i - \hat{\mathbf{M}}_k^{(t)}\right) \hat{\boldsymbol{\Gamma}}_k^{(t-1)} \left(\mathbf{X}_i - \hat{\mathbf{M}}_k^{(t)}\right)'}{\hat{n}_k^{(t)} q}.$$

Maximization of (12) with respect to $\boldsymbol{\Omega}_k$ corresponds a graphical lasso problem, which is solved using the coordinate descent algorithm by Friedman et al. (2008), where in our context their penalty coefficient is equal to $\frac{2}{\hat{n}_k q}\lambda_2 \mathbf{P}_2$. The algorithm is implemented in the `R` (R Core Team, 2023) package `glassoFast` (Sustik et al., 2018) and returns the estimates of the row precision matrices $\hat{\boldsymbol{\Omega}}_k^{(t)}$, for $k = 1, \ldots, K$.

Sparse estimation of the column-precision matrices $\boldsymbol{\Gamma}_k$

In the maximization of (6) with respect to $\boldsymbol{\Gamma}_k$, given the current estimates $\hat{\boldsymbol{\Omega}}_k^{(t)}$ and $\hat{\mathbf{M}}_k^{(t)}$, the Q-function simplifies to:

$$Q_\Gamma(\boldsymbol{\Gamma}_k) = \sum_{i=1}^{n} \hat{z}_{ik}^{(t)} \left[\frac{p}{2} \log |\boldsymbol{\Gamma}_k| - \frac{1}{2} \operatorname{tr} \left\{ \boldsymbol{\Gamma}_k \left(\mathbf{X}_i - \hat{\mathbf{M}}_k^{(t)} \right)' \hat{\boldsymbol{\Omega}}_k^{(t)} \left(\mathbf{X}_i - \hat{\mathbf{M}}_k^{(t)} \right) \right\} \right] - \lambda_3 ||\mathbf{P}_3 \odot \boldsymbol{\Gamma}_k||_1. \tag{13}$$

By rearranging terms in (13), we obtain the following objective function:

$$Q_\Gamma(\boldsymbol{\Gamma}_k) = \log |\boldsymbol{\Gamma}_k| - \operatorname{tr} \{\boldsymbol{\Gamma}_k \boldsymbol{S}_\Gamma\} - \frac{2}{\hat{n}_k p} \lambda_3 ||\mathbf{P}_3 \odot \boldsymbol{\Gamma}_k||_1, \tag{14}$$

where

$$\boldsymbol{S}_\Gamma = \sum_{i=1}^{n} \hat{z}_{ik}^{(t)} \frac{\left(\mathbf{X}_i - \hat{\mathbf{M}}_k^{(t)} \right)' \hat{\boldsymbol{\Omega}}_k^{(t)} \left(\mathbf{X}_i - \hat{\mathbf{M}}_k^{(t)} \right)}{\hat{n}_k^{(t)} p}.$$

Maximization of (14) with respect to $\boldsymbol{\Gamma}_k$ corresponds again to the graphical lasso, where in this case the original penalty coefficient is equal to $\frac{2}{\hat{n}_k p}\lambda_3 \mathbf{P}_3$. Also in this case the estimation is performed using the algorithm implemented in the package `glassoFast`, retrieving $\tilde{\boldsymbol{\Gamma}}_k^{(t)}$ for $k = 1, \ldots, K$. Note that, to address the non-identifiability issue highlighted at the end of Section 2.1, we require the estimates of the column precision matrices $\hat{\boldsymbol{\Gamma}}_k^{(t)}$ to have determinant equal to 1. In practice, this is achieved by rescaling $\tilde{\boldsymbol{\Gamma}}_k^{(t)}$ as follows:

$$\hat{\boldsymbol{\Gamma}}_k^{(t)} = \tilde{\boldsymbol{\Gamma}}_k^{(t)} / |\tilde{\boldsymbol{\Gamma}}_k^{(t)}|^{1/q}, \tag{15}$$

where the scaling can be applied either at each iteration or at the conclusion of the algorithm upon convergence, with virtually no impact on the final solution (Tomarchio et al., 2022); in our case, the scaling is applied at each iteration of the algorithm.

The updates based on the graphical lasso expressions (12) and (14-15) are iterated sequentially within the M-step at each cycle of the EM algorithm until convergence is

reached, returning sparse estimates of the precision matrices $\boldsymbol{\Omega}_k$ and $\boldsymbol{\Gamma}_k$. The global convergence is evaluated by monitoring the increase in the penalized log-likelihood at each full EM iteration. The algorithm is considered to have reached convergence when $\ell_P(\hat{\boldsymbol{\Theta}}^{(t)}; \mathbf{X}) - \ell_P(\hat{\boldsymbol{\Theta}}^{(t-1)}; \mathbf{X}) < \varepsilon$ for a given $\varepsilon > 0$. In our analyses, ε is set equal to 10^{-5}.

The procedure described in this section is available within an `R` package at `https://github.com/AndreaCappozzo/sparsemixmat`, where some of the routines have been implemented in `C++` to reduce the overall computing time.

3.3 A note on related penalty specifications

As briefly mentioned in Section 3.1, several options can be considered when specifying the penalty term in (3). An alternative approach to our proposal could involve applying a standard lasso penalty to the matrices $\mathbf{M}_k$, which would be consistent with the penalty used for the precision matrices. In this case, the penalty term would read as follows

$$p_{\boldsymbol{\lambda}}(\boldsymbol{\Theta}) = \lambda_1 \sum_{k=1}^{K} \|\mathbf{P}_1 \odot \mathbf{M}_k\|_1 + \lambda_2 \sum_{k=1}^{K} \|\mathbf{P}_2 \odot \boldsymbol{\Omega}_k\|_1 + \lambda_3 \sum_{k=1}^{K} \|\mathbf{P}_3 \odot \boldsymbol{\Gamma}_k\|_1, \tag{16}$$

where $\mathbf{P}_1$ is a $p \times q$ matrix with non-negative entries, while the other quantities are defined as in the previous sections. Compared to the formulation in (4), this penalty offers a less-structured way to inducing sparsity in the mean matrices. In general, it does not facilitate proper variable selection, as the dimensions of the mean matrices are not jointly shrunk to zero. Nevertheless, the sparsity patterns could provide relevant insights and the method can be useful in some specific applications, as for example when no temporal dimension is present in the data. As highlighted in Section 2.2, Gao et al. (2021) consider lasso cell-wise penalization of matrix mixture mean parameters. However, the authors do not consider penalization of the component covariance matrices. Consequently, the method may still necessitate the estimation of a large number of parameters and does not offer a flexible model for the association structures within row and column variables. To overcome these limitations, recently Heo and Baek (2021) derive a penalized matrix normal mixture model where sparsity is also induced on the precision matrices, by using a penalty function similar to (16). Nonetheless, in their proposed estimation procedure,

and in particular in the M-step update for $\mathbf{M}_k$, the authors implicitly assume that both the precision matrices corresponding to rows and columns, respectively, are diagonal. This assumption can lead to inaccurate estimates, especially in those applications where complex conditional dependency patterns exist. For these reasons, in the following we derive an estimation scheme where the independence assumption is not required. We refer to this approach as *Sparsemixmat-lasso*, to remark that the approach implements a sparse matrix Gaussian mixture, but where a cell-wise lasso penalty is placed on the entries of the mean parameters $\mathbf{M}_k$. Note that $\mathbf{\Omega}_k$ and $\mathbf{\Gamma}_k$ are estimated as in Section 3.2.3, therefore in what follows we only outline the updating formula for $\mathbf{M}_k$. Furthermore, the E-step and the considerations about the initialization strategy and the convergence criterion remain unchanged.

Consider the current estimates of the precision matrices $\hat{\mathbf{\Omega}}_k$ and $\hat{\mathbf{\Gamma}}_k$, where we omit the iteration superscript for ease of notation. When the penalty term is defined as in (16), in the maximization step with respect to $\mathbf{M}_k$, the Q-function can be expressed as follows:

$$Q_M(\mathbf{M}_k) = \sum_{i=1}^{n} \hat{z}_{ik}^{(t)} \left[\operatorname{tr}\left\{ \hat{\mathbf{\Omega}}_k \mathbf{X}_i \hat{\mathbf{\Gamma}}_k \mathbf{M}_k' \right\} - \frac{1}{2} \operatorname{tr}\left\{ \hat{\mathbf{\Omega}}_k \mathbf{M}_k \hat{\mathbf{\Gamma}}_k \mathbf{M}_k' \right\} \right] - \lambda_1 ||\mathbf{P}_1 \odot \mathbf{M}_k||_1. \tag{17}$$

We propose a cell-wise coordinate ascent estimation for $m_{ls,k}$, where $m_{ls,k}$ denotes the element in the l-th row and s-th column of matrix $\mathbf{M}_k$. Similarly, let $\hat{\omega}_{ls,k}$, $\hat{\gamma}_{ls,k}$ and $p_{ls,1}$ denote the elements in the l-th row and s-th column of matrices $\hat{\mathbf{\Omega}}_k$, $\hat{\mathbf{\Gamma}}_k$ and $\mathbf{P}_1$, respectively. Lastly, $x_{ls,i}$ is similarly defined in relation to a matrix observation $\mathbf{X}_i$. The following proposition characterizes the updating formula:

Proposition 1: *The sufficient and necessary conditions for $\hat{m}_{ls,k}$ to be a (global) maximizer of (17) (for fixed l, s, and k) are*

$$\sum_{i=1}^{N} \hat{z}_{ik} \sum_{r=1}^{p} \sum_{c=1}^{q} \hat{\omega}_{lr,k} x_{rc,i} \hat{\gamma}_{cs,k} - \hat{n}_k \sum_{r=1}^{p} \sum_{c=1}^{q} \hat{\omega}_{lr,k} \hat{m}_{rc,k} \hat{\gamma}_{cs,k} = \lambda_1 p_{ls,1} \operatorname{sign}(\hat{m}_{ls,k}), \quad \text{if } \hat{m}_{ls,k} \neq 0 \tag{18}$$

and

$$\left| \sum_{i=1}^{n} \hat{z}_{ik} \left[\sum_{\substack{r=1 \\ r \neq l}}^{p} \hat{\omega}_{lr,k} \left(\sum_{c=1}^{q} \left(x_{rc,i} - \hat{m}_{rc,k} \right) \hat{\gamma}_{cs,k} \right) + \right. \right.$$
$$\left. \left. + \hat{\omega}_{ll,k} \left(\sum_{\substack{c=1 \\ c \neq s}}^{q} \left(x_{lc,i} - \hat{m}_{lc,k} \right) \hat{\gamma}_{cs,k} \right) + \hat{\omega}_{ll,k} x_{ls,i} \hat{\gamma}_{ss,k} \right] \right| \leq \lambda_1 p_{ls,1}, \quad \textit{if } \hat{m}_{ls,k} = 0. \quad (19)$$

Thus, at the t-th iteration of the EM algorithm $\hat{m}_{lsk}^{(t)} = 0$ if

$$\left| \sum_{i=1}^{n} \hat{z}_{ik}^{(t)} \left[\sum_{\substack{r=1 \\ r \neq l}}^{p} \hat{\omega}_{lr,k}^{(t-1)} \left(\sum_{c=1}^{q} \left(x_{rc,i} - \hat{m}_{rc,k}^{(t)} \right) \hat{\gamma}_{cs,k}^{(t-1)} \right) + \right. \right.$$
$$\left. \left. + \hat{\omega}_{ll,k}^{(t-1)} \left(\sum_{\substack{c=1 \\ c \neq s}}^{q} \left(x_{lc,i} - \hat{m}_{lc,k}^{(t)} \right) \hat{\gamma}_{cs,k}^{(t-1)} \right) + \hat{\omega}_{ll,k}^{(t-1)} x_{ls,i} \hat{\gamma}_{ss,k}^{(t-1)} \right] \right| \leq \lambda_1 p_{ls,1}, \quad (20)$$

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

$$\hat{n}_k^{(t)} \hat{\omega}_{ll,k}^{(t-1)} \hat{m}_{ls,k}^{(t)} \hat{\gamma}_{ss,k}^{(t-1)} + \lambda_1 p_{ls,1} \operatorname{sign}\left(\hat{m}_{ls,k}^{(t)} \right) = \sum_{i=1}^{n} \hat{z}_{ik}^{(t)} \sum_{r=1}^{p} \sum_{c=1}^{q} \hat{\omega}_{lr,k}^{(t-1)} x_{rc,i} \hat{\gamma}_{cs,k}^{(t-1)} +$$
$$- \hat{n}_k^{(t)} \left(\sum_{\substack{r=1 \\ r \neq l}}^{p} \sum_{\substack{c=1 \\ c \neq s}}^{q} \hat{\omega}_{lr,k}^{(t-1)} \hat{m}_{rc,k}^{(t)} \hat{\gamma}_{cs,k}^{(t-1)} \right) \quad (21)$$

with respect to $\hat{m}_{ls,k}^{(t)}$.

The proof of Proposition 1 is reported in Section S1 of the Supplementary Material. This result corrects an inaccuracy introduced in Equation (5) of Heo and Baek (2021), and it can be seen as the matrix-variate extension of Theorem 1 of Zhou et al. (2009). Convergence to the global maximum is assured due to the theoretical properties of coordinate descent algorithms (see e.g., Wright, 2015). The described procedure, for sufficiently large λ_1, forces some $\hat{m}_{lsk}^{(t)}$ to be shrunk to 0, ultimately inducing sparsity in $\mathbf{M}_k$, $k = 1, \ldots, K$. However, as previously noted, this type of penalty does not allow for direct variable selection within a matrix-variate data framework. Direct variable selection can only be achieved by employing a group-lasso penalization scheme, as highlighted in Section 3.1.

3.4 Model selection

The model estimation strategy in Section 3.2 has been outlined by considering K and $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3)$ fixed. However, in practical applications, the number of clusters and the penalty hyperparameters are not known a priori and need to be chosen using model selection strategies. In this work, we select K and $\boldsymbol{\lambda}$ to maximize a modified version of the Bayesian information criterion (BIC, Schwarz, 1978), already considered in Pan and Shen (2007) and Casa et al. (2022). In detail, we use the following criterion:

$$BIC = 2 \log L(\hat{\boldsymbol{\Theta}}) - d_0 \log(n), \tag{22}$$

where d_0 is the number of non-zero estimated parameters and $\log L(\hat{\boldsymbol{\Theta}})$ is the log-likelihood evaluated at $\hat{\boldsymbol{\Theta}}$, the estimate of the parameters obtained by the algorithm described in Sections 3.2 and 3.3.

The adequacy of the BIC for selecting the number of mixture components has been thoroughly studied (see e.g., Roeder and Wasserman, 1997; Keribin, 2000), and the criterion has been widely used both in the two-way, and, more recently, the three-way model-based clustering frameworks (Sarkar et al., 2020; Tomarchio et al., 2022; Sharp et al., 2022). Moreover, the formulation in (22) has been proven useful also to tune the intensity of the penalization in both lasso (Zou et al., 2007) and sparse precision matrix estimation contexts (Lian, 2011). Nevertheless, other model selection strategies may be pursued, especially in situations where exhaustive grid searches are considered too computationally demanding. Possible alternatives include stochastic optimization algorithms, such as genetic algorithms (Holland, 1992), or conditional search schemes. Another interesting approach is outlined in Jiang et al. (2015), where the authors develop the E-MS algorithm, in which model selection is performed within each iteration of the standard EM algorithm.

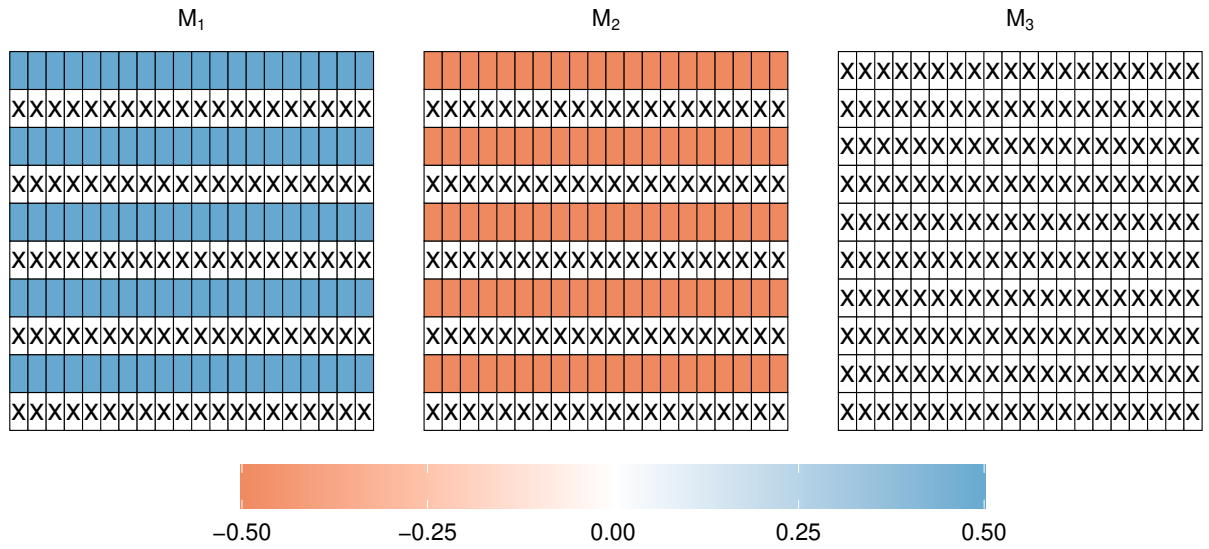

Figure 1: Heatmaps of the true 10×20 *mean matrices* $\mathbf{M}_k$, $k = 1, 2, 3$, *considered in the simulated data experiment. A zero entry in the matrices is indicated with the symbol* $\times$.

4 Simulation study

4.1 Experimental Setup

In this section, we assess the performance of the proposed method on synthetic data, evaluating its ability in recovering the underlying sparse patterns and the clustering structure. For each replication of the simulation experiment, we generate $n = 1000$ samples from a 3-component matrix Gaussian mixture model, in which mean matrices and both row and column precision matrices have some degree of sparsity. The row and column precision matrices have dimensions $p \times p$ and $q \times q$, with p and q equal to 10 and 20, respectively. The 10×20 mean matrices $\mathbf{M}_k$, $k = 1, 2, 3$, have a row-wise sparse structure, displayed in Figure 1. The data-generating process purposely reproduces a situation in which some of the p variables measured on q occasions are irrelevant for clustering. In this specific context, the second, fourth, sixth, eighth, and tenth rows do not convey any grouping information, being identically equal to 0 in all clusters. We consider two distinct scenarios according to the sparsity structure enforced for the row precision matrices $\mathbf{\Omega}_k$:

- *Alternated-blocks row precision matrices:* the 10×10 row precision matrices $\mathbf{\Omega}_k$,

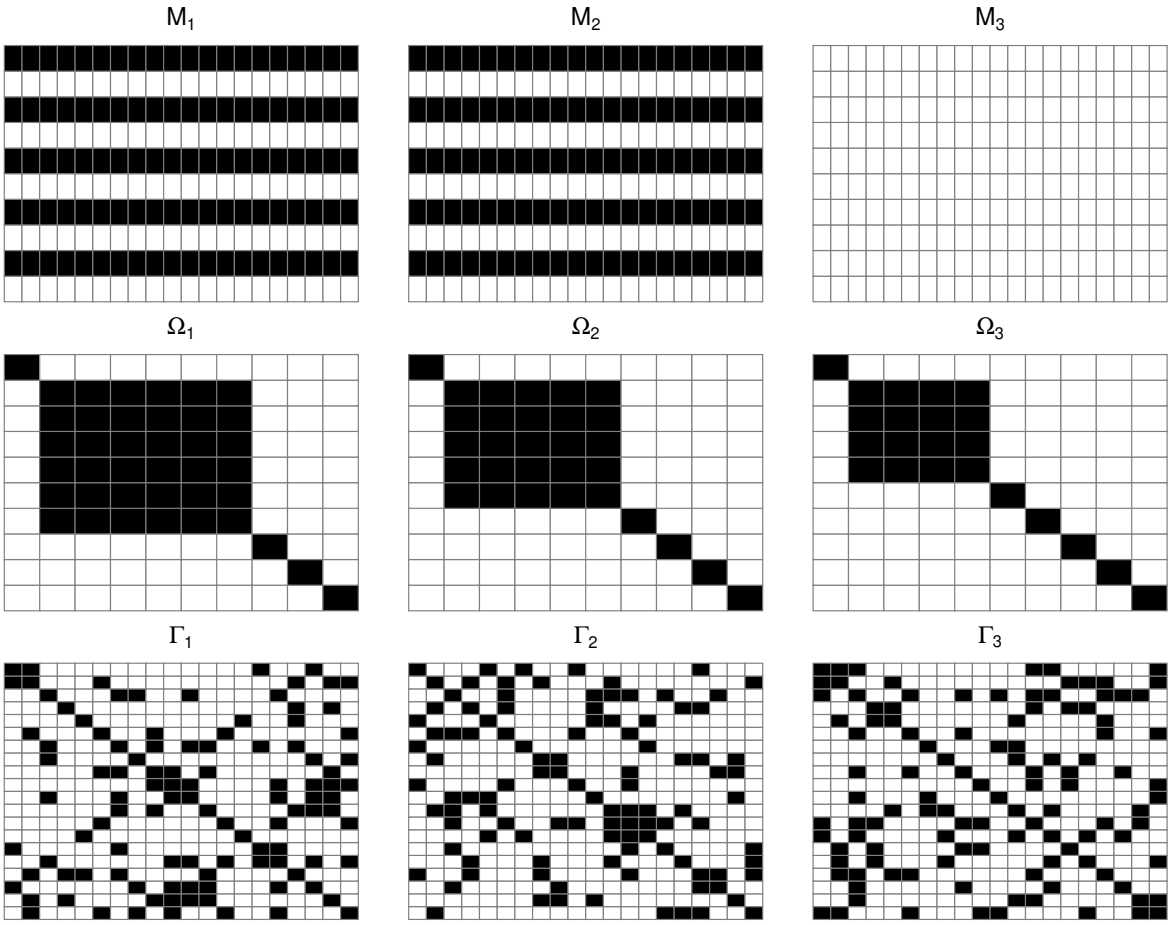

Figure 2: True sparsity patterns of the parameters associated to the Alternated-blocks row precision matrices *simulated data scenario.*

$k = 1, 2, 3$, have a block-wise sparse structure, as shown in Figure 2.

- *Sparse-at-random row precision matrices:* the row precision matrices have a sparse-at-random Erdős-Rényi graph structure (Erdős and Rényi, 1960) with probabilities of connection equal to 0.2, 0.5 and 0.8 for $\mathbf{\Omega}_1$, $\mathbf{\Omega}_2$ and $\mathbf{\Omega}_3$, respectively. These are displayed in Figure 3.

In both scenarios, the column precision matrices $\mathbf{\Gamma}_k$ are generated according to a sparse-at-random Erdős-Rényi graph structure, while the mixing proportions τ_k are assumed equal to $1/K$, $K = 3$. The experiment is repeated 100 times, and for each replication the following models are fitted to the synthetic data samples:

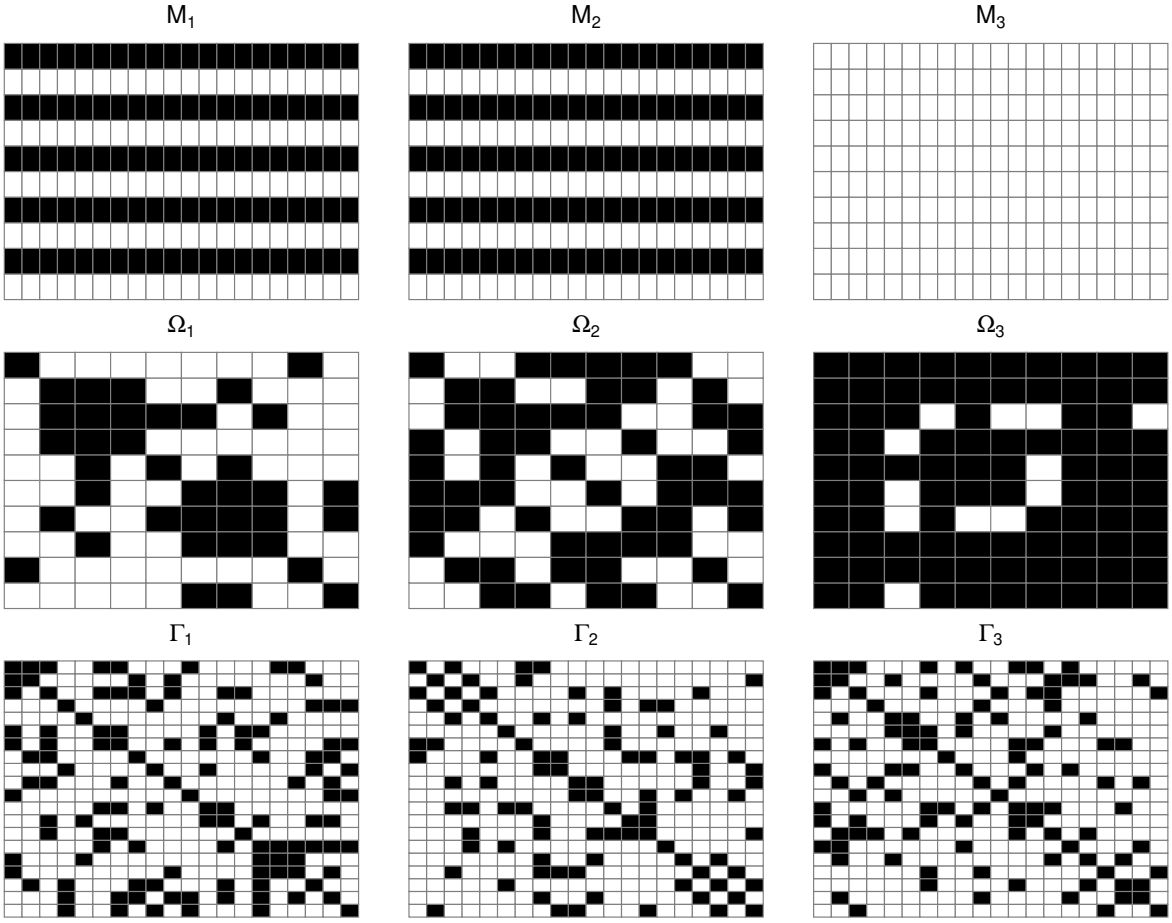

Figure 3: True sparsity patterns of the parameters associated to the Sparse-at-random row precision matrices *simulated data scenario.*

- *Full MGMM:* the finite mixtures of matrix normal distributions originally introduced in Viroli (2011a), where full matrix parameters are estimated for each component. This model specification corresponds to a G-VVV-VV model following the nomenclature introduced in Sarkar et al. (2020).

- *Sparsemixmat:* the penalized MGMM method introduced in this paper, with a group-lasso penalization imposed on the rows of the mean matrices according to the penalty term in (4).

- *Sparsemixmat-lasso:* the modification of the penalized MGMM methodology introduced in Heo and Baek (2021), which uses entry-wise lasso penalization on the mean matrices as in (16). Unlike Heo and Baek (2021), the precision matrices are not assumed to be diagonal, as outlined in Section 3.3.

In addition, we tested also procedures for clustering two-way data on the vectorized matrices; however, due to their poorer performance compared to methods tailored specifically for matrix-variate data, these results are omitted from the main text and are provided instead in Section S2 of the Supplementary Material.

For *Sparsemixmat* and *Sparsemixmat-lasso*, a search over an equispaced grid of elements for each penalty term λ_1, λ_2 and λ_3 is conducted, and the best model according to the BIC in (22) is retained. All competing methods are initialized via model-based agglomerative hierarchical clustering, as discussed in Section 3.2.1. The methods are evaluated based on their ability in performing variable selection, recovering the true sparsity structure, and correctly retrieving the cluster allocations. The issue of matching the estimated clustering with the actual classification is addressed using the `matchClasses` function from the `e1071` `R` package (Meyer et al., 2020). Simulation results are reported in the next subsection.

4.2 Simulation study results

4.2.1 Alternated-blocks row precision matrices

In Figure 4, we present the heatmap plots associated to the 10×10 row precision matrices $\boldsymbol{\Omega}_k$, $k = 1, 2, 3$ for the *alternated-blocks row precision matrices* scenario. The top row

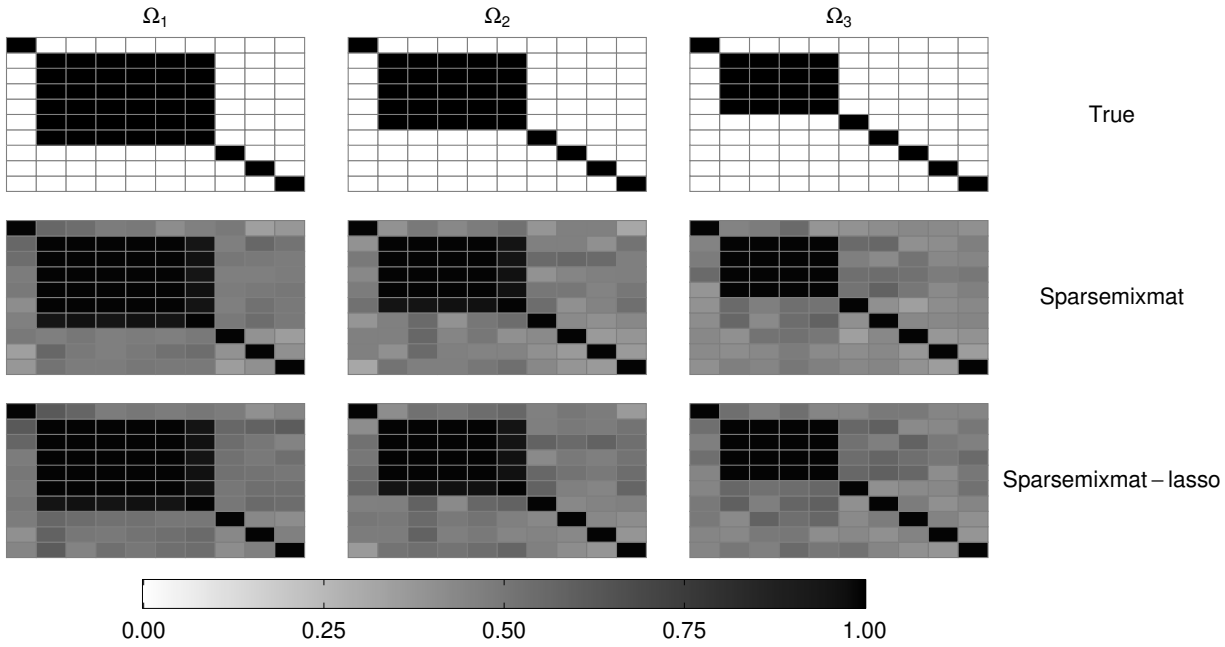

Figure 4: Alternated-blocks row precision matrices *scenario. In the top row, the true sparsity patterns of the row precision matrices* $\mathbf{\Omega}_k$*,* $k = 1, 2, 3$*. In the middle and bottom rows, heatmap plots displaying the proportion of times parameters in the row precision matrices have been estimated as non-zero across* 100 *replications of the simulated experiment.*

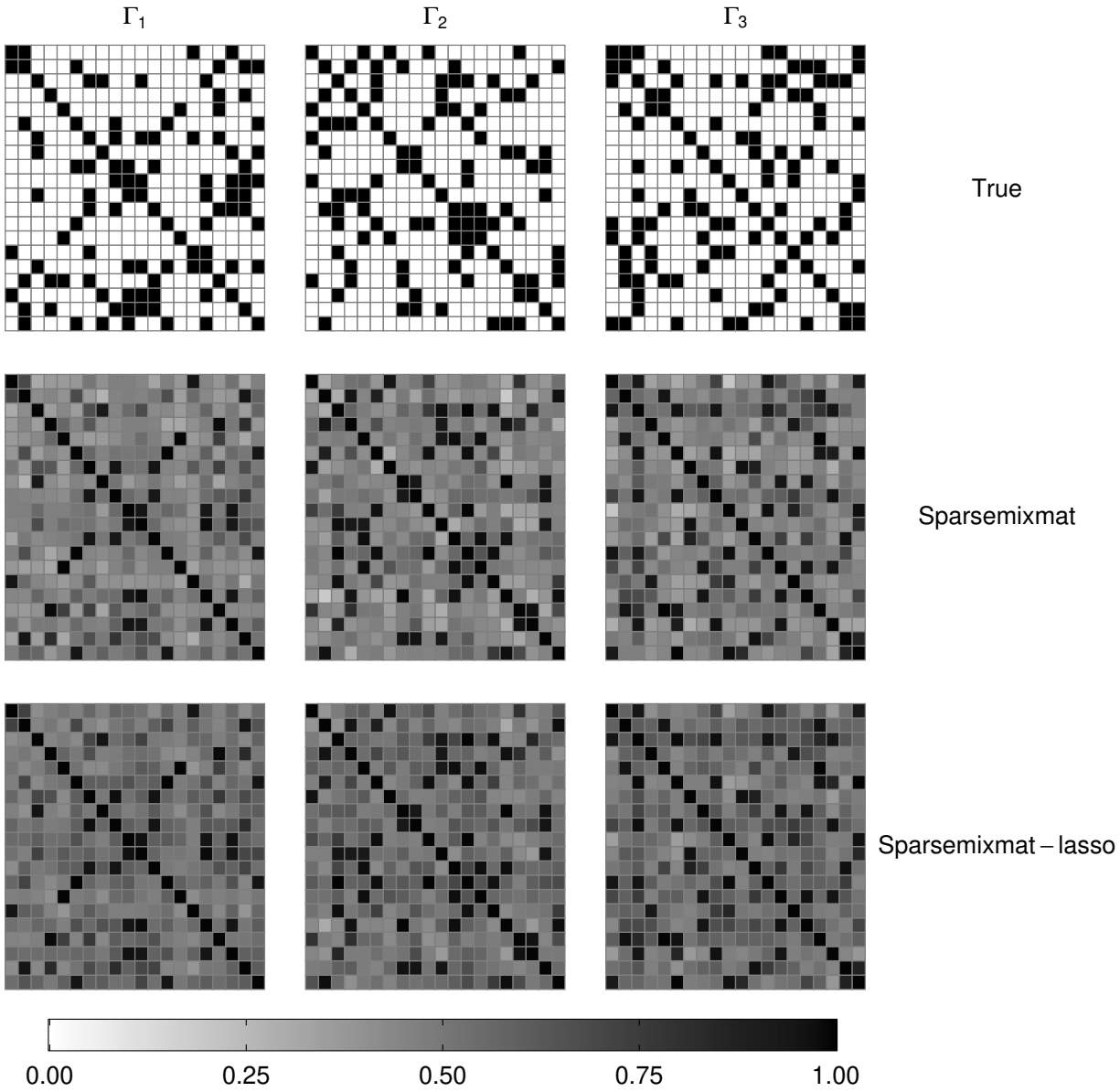

Figure 5: Alternated-blocks row precision matrices *scenario. In the top row, the true sparsity patterns of the column precision matrices* $\mathbf{\Gamma}_k$*,* $k = 1, 2, 3$*. In the middle and bottom rows, heatmap plots displaying the proportion of times parameters in the column precision matrices have been estimated as non-zero across* 100 *replications of the simulated experiment.*

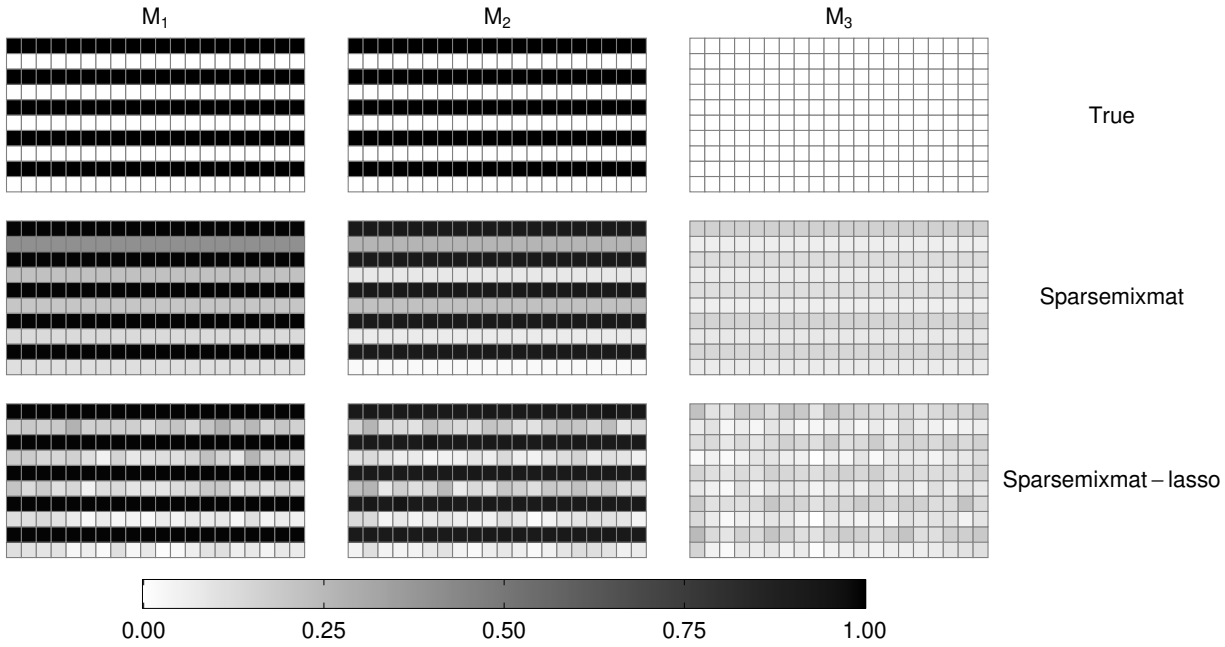

Figure 6: Alternated-blocks row precision matrices *scenario. In the top row, the true sparsity patterns of the mean matrices* $\mathbf{M}_k$*,* $k = 1, 2, 3$*. In the middle and bottom rows, heatmap plots displaying the proportion of times parameters in the mean matrices have been estimated as non-zero across* 100 *replications of the simulated experiment.*

displays each heatmap representing the association structure corresponding to a component row precision matrix, where each black square denotes the presence of a non-zero parameter, and hence an association between a pair of variables. The second and third rows show the heatmaps indicating the proportion of times a non-zero precision parameter has been estimated between pairs of variables. As illustrated in the graphs, both *Sparsemixmat* and *Sparsemixmat-lasso* satisfactorily recover the true underlying sparsity structure. A moderate penalty on the row-precision matrices induces shrinkage to zero of some of the elements of $\mathbf{\Omega}_k$, enabling the correct identification of the conditional association structures among the variables in the clusters. Figure 5 displays similar heatmaps for the 20×20 column precision matrices, where the association structure is again well recovered by both methods.

Different results emerge when examining the estimates of the cluster mean matrices $\mathbf{M}_k$, reported in Figure 6. In the figure, the heatmaps report the non-zero entries of the data-generating mean matrices and the proportion of zero entries in the estimated matrices, averaged over 100 replications. The row-wise shrinkage of *Sparsemixmat*, enforced by the group-lasso penalty, enables more accurate recovery of the mean matrices compared to the entry-wise lasso shrinkage of *Sparsemixmat-lasso*. This is further supported by the metrics displayed in Table 1 where we report the average Frobenius distance between true and estimated parameters for each mixture component. Notably, *Sparsemixmat* outperforms the competing methods, exhibiting the lowest average distance for every mean matrix across all three clusters. While *Sparsemixmat-lasso* and *Full MGMM* seem to perform slightly better when examining row and column precision matrices, the difference is often negligible. Moreover, *Sparsemixmat* achieves superior results in terms of recovery of the underlying cluster partition, as measured by the adjusted Rand index (ARI, Hubert and Arabie, 1985), as well as in overall model parsimony, quantified by the number of estimated parameters. *Sparsemixmat* shows a higher ARI and a lower number of non-zero parameters compared to *Full MGMM* and *Sparsemixmat-lasso*. It is important to note that *Full MGMM* does not employ any shrinkage, resulting in a total of $(K-1)+K(pq+p(p+1)/2+q(q+1)/2)$ estimated parameters in all cases. Lastly, it is worth noting that while the proximal gradient

Table 1: Alternated-blocks row precision matrices *scenario. Frobenius distance between true and estimated parameters, adjusted Rand index (ARI), number of non-zero parameters (d_0), and computing time (in seconds) per iteration averaged over* 100 *repetitions of the simulated experiment. Bold font indicates the best performing method according to the considered metric. Standard errors are reported in brackets.*

	Full MGMM	*Sparsemixmat*	*Sparsemixmat-lasso*
$\|\|\mathbf{M}_1 - \hat{\mathbf{M}}_1\|\|_F$	10.032 (27.196)	**8.015 (26.333)**	8.54 (25.808)
$\|\|\mathbf{M}_2 - \hat{\mathbf{M}}_2\|\|_F$	4.709 (10.43)	**2.252 (6.781)**	2.751 (6.643)
$\|\|\mathbf{M}_3 - \hat{\mathbf{M}}_3\|\|_F$	4.716 (12.375)	**1.939 (6.572)**	1.952 (6.597)
$\|\|\boldsymbol{\Omega}_1 - \hat{\boldsymbol{\Omega}}_1\|\|_F$	**0.334 (1.084)**	0.577 (0.809)	0.527 (0.813)
$\|\|\boldsymbol{\Omega}_2 - \hat{\boldsymbol{\Omega}}_2\|\|_F$	**0.241 (0.729)**	0.325 (0.693)	0.311 (0.697)
$\|\|\boldsymbol{\Omega}_3 - \hat{\boldsymbol{\Omega}}_3\|\|_F$	0.436 (1.342)	**0.303 (0.901)**	0.305 (0.928)
$\|\|\boldsymbol{\Gamma}_1 - \hat{\boldsymbol{\Gamma}}_1\|\|_F$	**28.648 (88.923)**	32.541 (81.452)	29.83 (83.172)
$\|\|\boldsymbol{\Gamma}_2 - \hat{\boldsymbol{\Gamma}}_2\|\|_F$	22.428 (72.793)	21.871 (67.351)	**21.544 (68.748)**
$\|\|\boldsymbol{\Gamma}_3 - \hat{\boldsymbol{\Gamma}}_3\|\|_F$	31.787 (99.288)	**26.026 (82.058)**	26.168 (84.907)
ARI	0.995 (0.047)	**1 (<0.01)**	0.997 (0.058)
d_0	1397 (0)	**720.051 (34.151)**	750.051 (24.829)
Average time per iteration (s)	**1.206 (0.07)**	2.482 (0.078)	143.907 (9.422)

algorithm of *Sparsemixmat* is slower than the model with no penalty, as expected, it is considerably faster than the *Sparsemixmat-lasso* approach.

Another aspect to examine is the performance of the proposed methods in terms of variable selection. Given the matrix-variate nature of the data, we are interested in monitoring a method's ability to correctly identify the zero rows of the mean matrices, and hence detect those variables that have constant means across occasions and clusters. To measure this, we make use of the F_1 score, defined as follows:

$$F_1 = \frac{\texttt{tp}}{\texttt{tp} + 0.5(\texttt{fp} + \texttt{fn})}, \tag{23}$$

where `tp` denotes the number of zero rows in $\mathbf{M}_k$ correctly estimated as such, while `fp` and `fn` denote the number of non-zero rows wrongly shrunk to zero and the number of

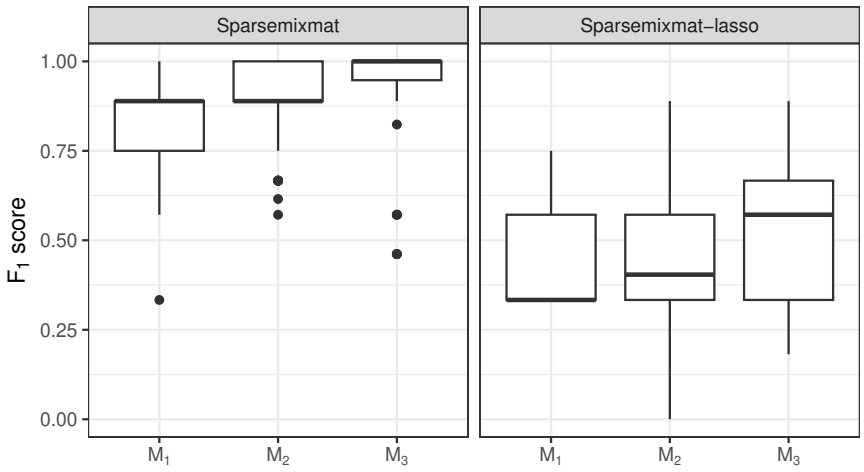

Figure 7: Alternated-blocks row precision matrices *scenario. Boxplots of the* F_1 *score for* 100 *replications of the simulated experiment.*

zero rows not shrunk to zero, respectively. Figure 7 presents boxplots of the F_1 score for the *Sparsemixmat* and *Sparsemixmat-lasso* methods. By enforcing entire rows of $\hat{\mathbf{M}}_k$ to be shrunk to zero via the group-lasso penalty, *Sparsemixmat* achieves better variable selection performance. Conversely, for the *Sparsemixmat-lasso*, which applies entry-wise lasso shrinkage, there is no guarantee that entire rows will be ultimately set to zero. Therefore, when the primary aim is multivariate variable selection within a matrix mixture context, our proposed *Sparsemixmat* is preferable.

Similar results are observed when more complex dependence structures between the p variables are considered, as will be reported in the next subsection.

4.2.2 Sparse-at-random row precision matrices

In the second scenario, the row precision matrices have a sparse at random Erdős-Rényi graph structure. By inspecting Figure 8, we notice how the less trivial dependence patterns among the variables affect the performance of the penalized models. Regardless of the methods considered, the number of non-zero edges is greatly overestimated, resulting in solutions where the sparsity in $\mathbf{\Omega}_k$ is underrated. Similar results are observed, albeit less remarkably, for the column precision matrices and the mean matrices reported in Fig-

ure 9 and 10, respectively. Nonetheless, *Sparsemixmat* displays the best results also in this more challenging scenario, as it is indicated in Table 2. Particularly, our proposal outperforms the competitors in terms of Frobenius distance, parsimony, and recovering of the true clustering, demonstrating only a slightly longer execution time than *Full MGMM* for estimation. Similarly to the previous scenario, with regard to the ability to perform variable selection, a group-lasso penalty on the rows of $\mathbf{M}_k$ is preferred. This is highlighted in the boxplots of Figure 11, where *Sparsemixmat* shows consistently higher F_1 score values compared to *Sparsemixmat-lasso*. Interestingly, the variable selection performance of both methods in terms of the F_1 score is slightly lower in this scenario compared to the previous one. This finding suggests that variable selection performance depends not only on the penalty imposed on the mean matrices, but also on how accurately the dependence structure among the p variables in the K clusters is recovered.

In summary, the proposed *Sparsemixmat* adequately tackles the problem of clustering matrix-variate data with sparse model parameters. The method is flexible, capable of capturing cluster-wise different dependence structures in both variables and occasions, it enables row-wise variable selection when variables are recorded over multiple occasions, and it effectively detects the clustering structure in matrix data. These considerations hold true not only in an experimental setup but also in the analysis of real-world data, as reported in the next section.

5 Application: criminal trends in the US

5.1 Data description

We analyze data from the United States Department of Justice Federal Bureau of Investigation concerning violent and property crimes of 236 American cities. The aim of the analysis is to cluster cities with similar crime trends and to identify which crime types exhibit relevant differences in the time patterns across clusters. In the data, for each city, $p = 7$ variables corresponding to the rates of murder, rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft are measured over $q = 13$ years in the pe-

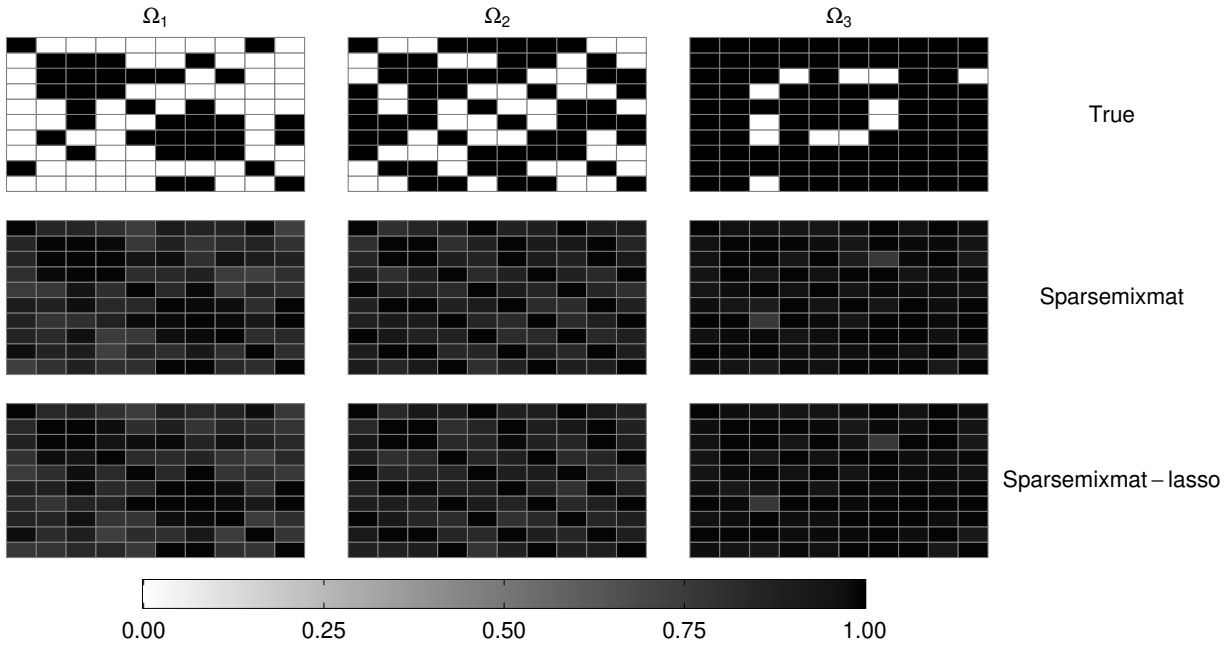

Figure 8: Sparse-at-random row precision matrices *scenario. In the top row, the true sparsity patterns of the row precision matrices* $\mathbf{\Omega}_k$*,* $k = 1, 2, 3$*. In the middle and bottom rows, heatmap plots displaying the proportion of times parameters in the row precision matrices have been estimated as non-zero across* 100 *replications of the simulated experiment.*

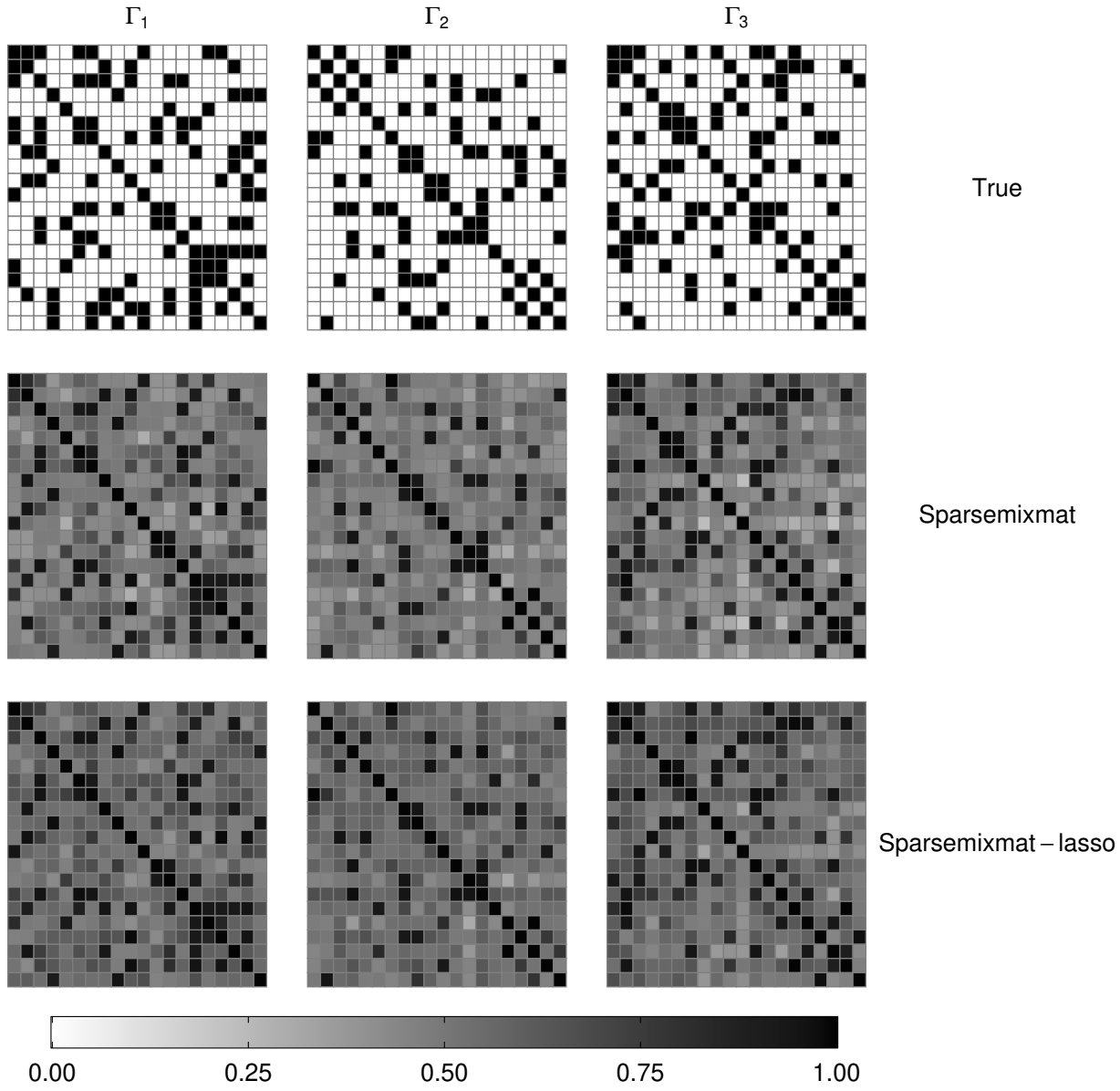

Figure 9: Sparse-at-random row precision matrices *scenario. In the top row, the true sparsity patterns of the column precision matrices* $\mathbf{\Gamma}_k$*,* $k = 1, 2, 3$*. In the middle and bottom rows, heatmap plots displaying the proportion of times parameters in the column precision matrices have been estimated as non-zero across* 100 *replications of the simulated experiment.*

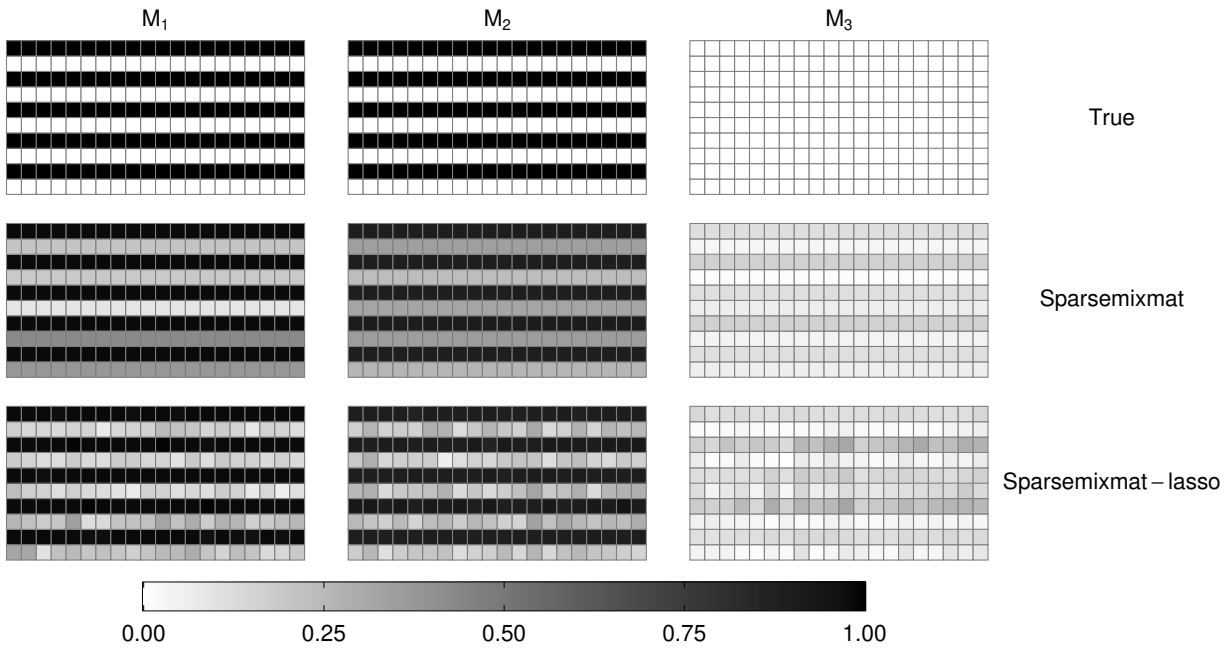

Figure 10: Sparse-at-random row precision matrices *scenario. In the top row, the true sparsity patterns of the mean matrices* $\mathbf{M}_k$*,* $k = 1, 2, 3$*. In the middle and bottom rows, heatmap plots displaying the proportion of times parameters in the mean matrices have been estimated as non-zero across* 100 *replications of the simulated experiment.*

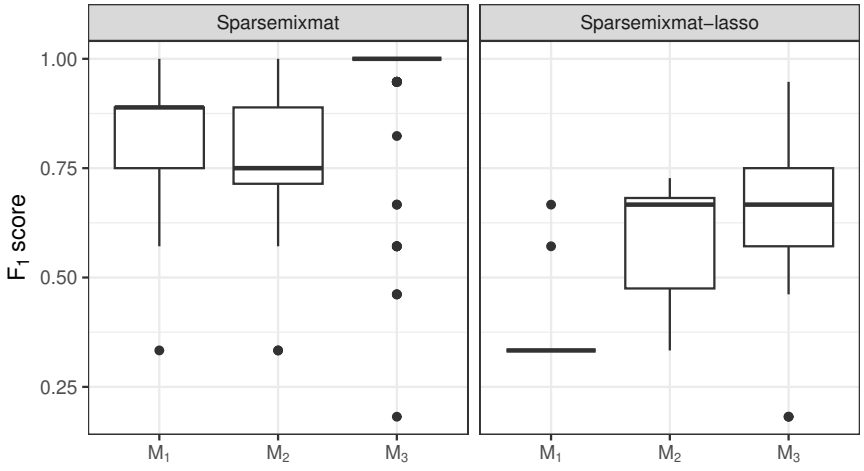

Figure 11: Sparse-at-random row precision matrices *scenario. Boxplots of the* F_1 *score for* 100 *replications of the simulated experiment.*

Table 2: Sparse-at-random row precision matrices *scenario. Frobenius distance between true and estimated parameters, adjusted Rand index (ARI), number of non-zero parameters (d_0) and computing time (in seconds) per iteration averaged over* 100 *repetitions of the simulated experiment. Bold numbers indicate the best performing method according to the considered metric. Standard errors are reported in brackets.*

	Full MGMM	*Sparsemixmat*	*Sparsemixmat-lasso*
$\|\|\mathbf{M}_1 - \hat{\mathbf{M}}_1\|\|_F$	19.492 (33.406)	**10.29 (29.308)**	12.04 (29.127)
$\|\|\mathbf{M}_2 - \hat{\mathbf{M}}_2\|\|_F$	19.062 (40.251)	**3.729 (12.083)**	6.213 (11.418)
$\|\|\mathbf{M}_3 - \hat{\mathbf{M}}_3\|\|_F$	18.122 (42.517)	**2.64 (7.504)**	2.792 (7.925)
$\|\|\mathbf{\Omega}_1 - \hat{\mathbf{\Omega}}_1\|\|_F$	3.116 (7.586)	**2.31 (6.597)**	2.321 (6.63)
$\|\|\mathbf{\Omega}_2 - \hat{\mathbf{\Omega}}_2\|\|_F$	2.311 (5.814)	**1.904 (5.355)**	1.908 (5.357)
$\|\|\mathbf{\Omega}_3 - \hat{\mathbf{\Omega}}_3\|\|_F$	2.314 (5.679)	**1.572 (4.251)**	1.583 (4.248)
$\|\|\mathbf{\Gamma}_1 - \hat{\mathbf{\Gamma}}_1\|\|_F$	50.2 (118.507)	43.46 (101.632)	**41.5 (103.282)**
$\|\|\mathbf{\Gamma}_2 - \hat{\mathbf{\Gamma}}_2\|\|_F$	36.462 (90.953)	**27.947 (75.949)**	28.065 (77.504)
$\|\|\mathbf{\Gamma}_3 - \hat{\mathbf{\Gamma}}_3\|\|_F$	49.875 (117.018)	34.301 (88.208)	**34.095 (90.846)**
ARI	0.991 (0.061)	**1 (<0.01)**	**1 (<0.01)**
d_0	1397 (0)	**794.97 (40.029)**	813.737 (13.536)
Average time per iteration (s)	**1.415 (0.127)**	7.705 (0.859)	266.736 (26.616)

riod between 2000 and 2012. Thus, the data can be conveniently arranged in a $7 \times 13 \times 236$ array, where each statistical unit $\mathbf{X}_i$, $i = 1, \ldots, 236$, takes the form of a 7×13 matrix. The dataset is publicly available in the `MatTransMix R` package (Zhu et al., 2022) and has been previously analyzed in Melnykov and Zhu (2019), where the authors introduced a method based on mixture of matrix transformation regression time series. In the next subsection, we discuss the results of our modeling approach and compare them with the findings from Melnykov and Zhu (2019).

5.2 Results

We implement an initial pre-processing step in which the statistical units are log-transformed to alleviate skewness, and subsequently centered cell-wise for the reasons described in Section 3.1. Specifically, if the row mean parameters of certain clusters are shrunk to zero, the average log rate of the crimes associated with those zero rows align with the center of the data. As a result, when a given cluster has a row shrunk toward zero, this suggests that the cluster-specific trend associated with that variable is not appreciably different from the global trend for that variable. After this pre-processing step, the *Sparsemixmat* model introduced in Section 3 is fitted to the transformed crime data. The shrinkage parameters are varied within a pre-specified grid of values, and we consider $K \in \{3, 4, 5, 6\}$.

The BIC as introduced in Equation (22) selects $K = 3$ clusters, with corresponding shrinkage hyperparameters λ_1, λ_2, and λ_3 equal to 3.81, 0, and 14.3, respectively. The penalty coefficient $\lambda_2 = 0$ implies that the estimated row precision matrices $\hat{\boldsymbol{\Omega}}_k$ for the selected model are non-sparse, indicating relevant associations between the crime types across clusters. On the other hand, with both the selected λ_1 and λ_3 greater than zero, the estimated mean matrices and column precision matrices measuring the dependence between the time occasions exhibit certain degrees of sparsity. Visual representations of these estimated parameters are shown in Figures 12 and 13. From Figure 12, we observe that no crime type presents estimated cluster mean log rates equal to zero across all clusters, indicating that all variables contain some discriminating information. However, in light of the considerations outlined in Section 3.1, clusters tend to be differentiated over the log rates of certain crimes across the years. For example, all clusters have dissimilar burglary and larceny-theft log rates, while cluster 1 and 3 tend to overlap in terms of murder, rape, and motor vehicle theft log rates. In addition, robbery and assault crime log rates tend to stay constant over time for the cities in cluster 3, while they vary for those in clusters 1 and 2. Figure 13 shows that the estimated column precision matrices, which embed the conditional association structure of the crime log rates between years, tend to have a banded structure. The entries along the diagonal are generally non-zero, while entries between occasions far in time are generally shrunk to zero, indicating higher levels

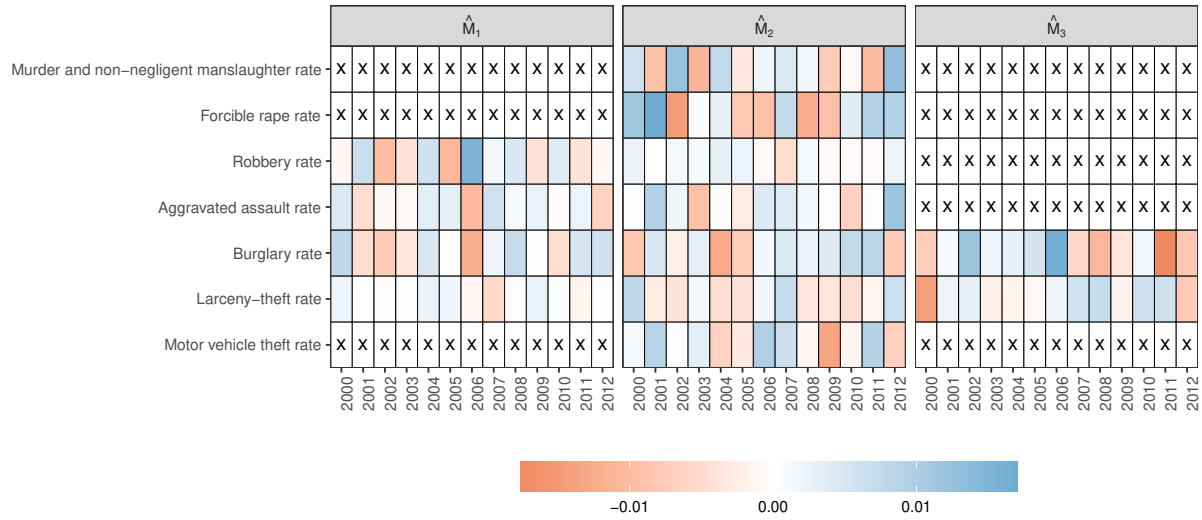

Figure 12: Crime data. *Estimated mean matrices* $\hat{\mathbf{M}}_k$, $k = 1, 2, 3$, *for the Sparsemixmat model. Colors denote the values of the estimates; a* 0 *entry in the matrices is indicated by the symbol* $\times$.

of association among consecutive years.

The clustering of the cities in the data is displayed in the map of Figure 14. To aid interpretation of the results, Figure 15 reports the average crime rate profiles computed on the original non-transformed data, obtained by calculating the average rate of the cities assigned to each cluster. In more detail, cluster 3 (blue color) identifies the safest cities in the country, which tend also to be the smallest in size. Higher concentration of safe cities can be observed in Northern Texas, the Los Angeles-San Diego area, and parts of the northern states, together with a few coastal areas in Florida and in the south of Indiana. Cluster 2 (red color) includes the cities with the highest crime rates of the considered types. From the map, it appears that these cities tend to be unevenly distributed across the US, with a concentration in the eastern part of the country, which is also the most densely populated. Lastly, cluster 1 (orange color) comprises cities that are slightly less safe, for which the mean crime rates over time tend to be higher than those in cluster 3. However, as remarked previously, the cities in these clusters tend to have overlapping mean rates of murder, rape, and motor vehicle theft over time (see Figure 12).

We highlight several similarities in the results discussed here and those of the analyses

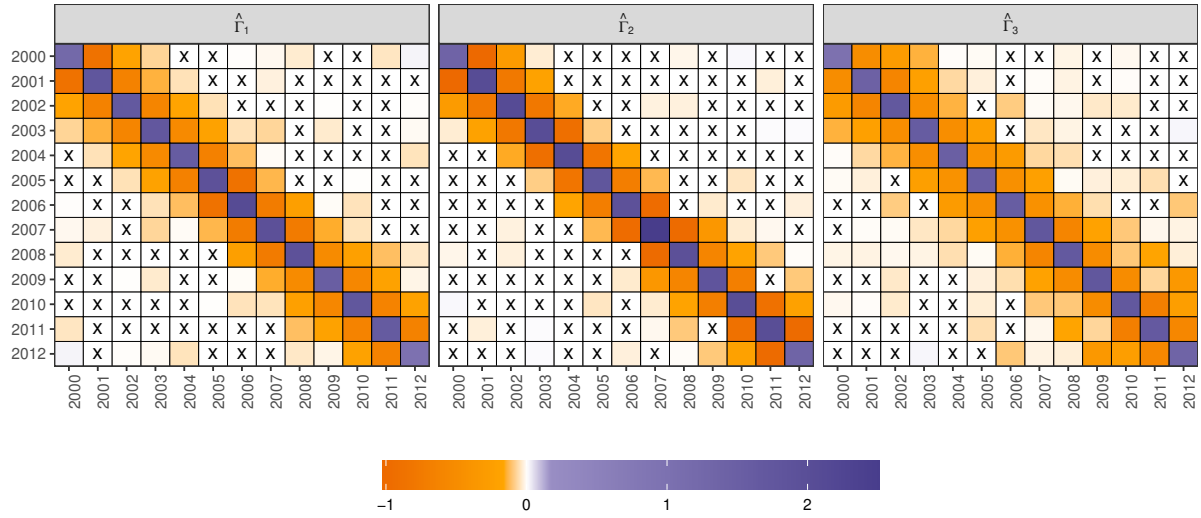

Figure 13: Crime data. *Estimated column precision matrices* $\hat{\mathbf{\Gamma}}_k$, $k = 1, 2, 3$, *for the Sparsemixmat model. Colors denote the values of the entries; a* 0 *entry in the matrices is indicated by the symbol* $\times$.

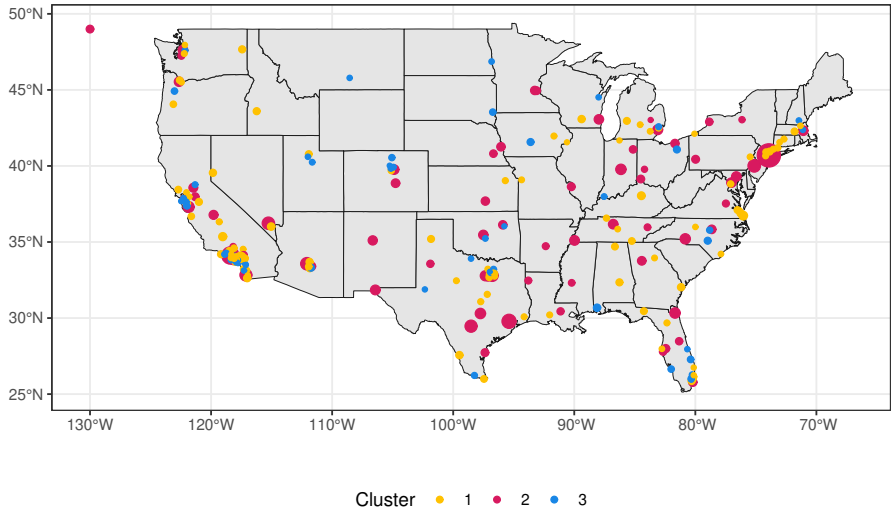

Figure 14: Crime data. Map of the USA showing the clustering of the cities obtained from the sparsemixmat model. The sizes of the circles are proportional to the city population. Colors indicate different clusters.

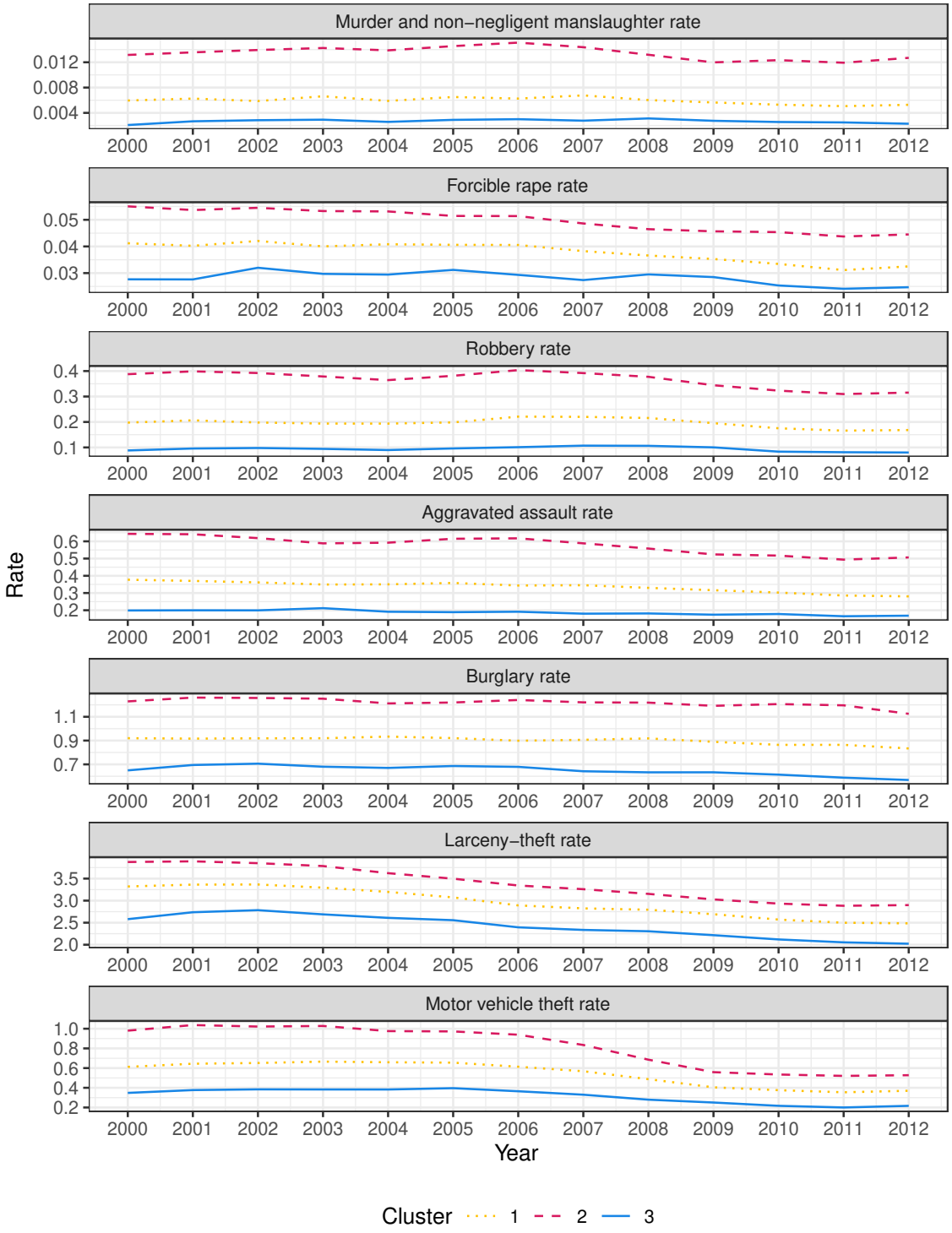

Figure 15: Crime data. Mean profiles for the sparsemixmat model. The mean profiles are computed for the variables in original scale. Colors and line types illustrate different clusters.

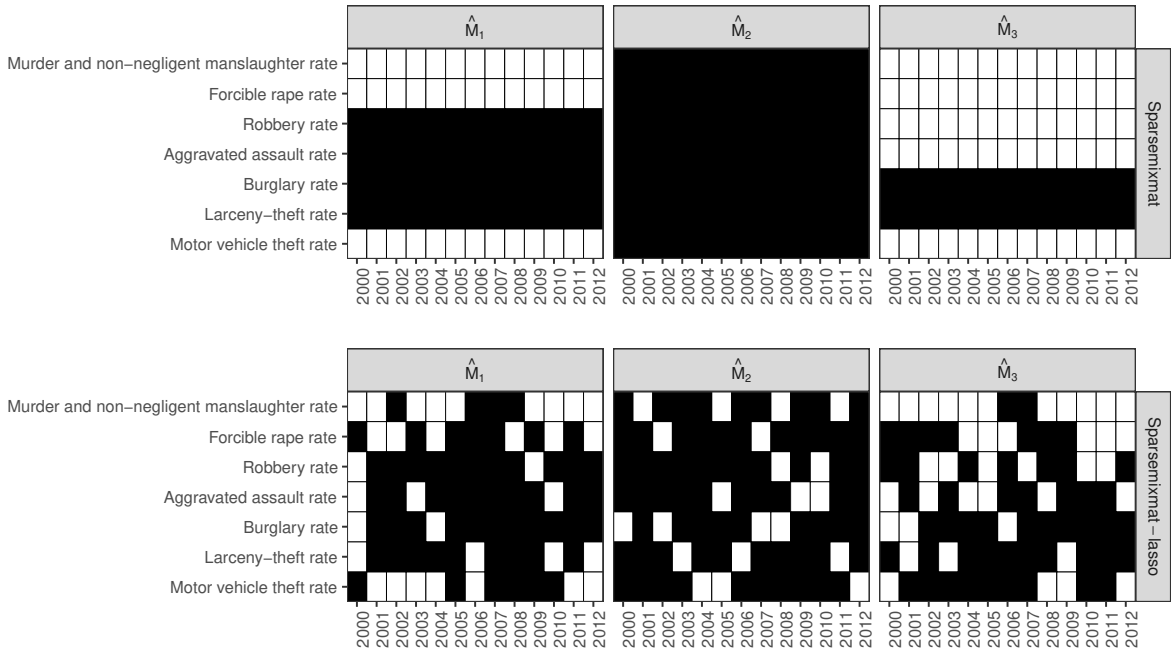

Figure 16: Crime data. *Sparse structure associated to the mean matrices* $\mathbf{M}_k$, $k = 1, 2, 3$, *for the Sparsemixmat and Sparsemixmat-lasso models. Black squares denote a non-zero entry.*

reported in Melnykov and Zhu (2019). First off, compared to the partition obtained in their 3-cluster model, we observe an agreement of approximately 75% of cases, along with a similar interpretation of the resulting clusters. Dependence patterns across time, similar to those in Figure 13, were also observed in Melnykov and Zhu (2019), in which a first order autoregressive model was employed to reduce the number of parameters and model the time dependence. While this is indeed a sensible modeling choice given the temporal nature of the crime data, this highlights the flexibility of our procedure in automatically capturing an autoregressive-like structure in the time occasions. This is achieved through the penalization on the column precision matrices, without the need to pre-specify any pattern or dependence structure. In addition, to overcome the over-parametrization issue associated with the mean matrices, Melnykov and Zhu (2019) consider regressing crime rates on years. In contrast, our proposed method employs a group-lasso penalty which effectively serves the same purpose, without the specification of a regression model.

We conclude this section by comparing the results obtained with our *Sparsemixmat* procedure with those of *Sparsemixmat-lasso* of Heo and Baek (2021) as modified in Section 3.3. The two models yield very similar partitions, with almost perfect agreement and only 9 cities assigned to different clusters. Figure 16 shows the sparse structures of the estimated cluster mean matrices, obtained with the two different penalties. For *Sparsemixmat-lasso*, all the crime types have non-zero mean log rates for some of the years and clusters, making difficult to differentiate the clusters based on overall mean crime log rate patterns across years. By contrast, the group-lasso penalty in *Sparsemixmat* facilitates easier interpretation, making it a more favorable option for clustering matrix-variate data with variables recorded over multiple time occasions.

6 Conclusion

The complex structure of three-way data makes clustering matrices a particularly difficult task. By framing the problem within a well-defined probabilistic framework, model-based methods are widely adopted to address these challenge. However, these approaches have to face severe over-parameterization issues, even with three-way data of moderate dimensions. In this work, we propose a methodology that alleviates these drawbacks, enabling the clustering of matrix-variate data even when the number of variables p or the number of occasions q is not small. Specifically, the proposed method relies on a penalized likelihood approach that allows to induce sparsity in the model parameters. The penalties on the row and column precision matrices greatly reduce the number of parameters to estimate, while also simplifying the interpretation of the dependence patterns through their connection to Gaussian graphical models. Furthermore, the group lasso penalty on the rows of the component mean matrices enables variable selection when three-way data arise from variables recorded over multiple occasions. This enhances model parsimony and provides valuable indications regarding which variables are useful in separating clusters across occasions. Assessments on both synthetic data and US crime rate data have shown the validity and effectiveness of our proposed method, overcoming several drawbacks of the approaches currently existing in the literature.

The paper opens several paths for future research. Firstly, while the group lasso penalty effectively performs variable selection, it may be too restrictive in some applications. In fact, as highlighted in Section 3.1, this specification sets to zero entire rows of the mixture component mean matrices. However, in some cases, sparsity within rows may be desirable, thus requiring only some occasions and not the entire variable to be shrunk to zero. This could be achieved by adapting the sparse group lasso (Simon et al., 2013) to the framework considered in our work. In fact, this penalty is a convex combination of the group-lasso and the entry-wise lasso penalty described in Section 3.3, potentially extending the application of the proposed method to other contexts. Another alternative approach could involve the use of a fused lasso penalty, as for example, in Guo et al. (2010) and Ren et al. (2022). This penalty could induce shrinkage of the rows of the matrix means toward each other, both within and between clusters, thereby facilitating the grouping of variables and the identification of those capable of differentiating some, but not all, clusters.

Throughout the manuscript, matrix Gaussian mixture models have been parameterized in terms of precision matrices. Nonetheless, the penalized approach can be adapted to a setting where sparsity is imposed on the covariance matrices, extending the method proposed by Fop et al. (2019) to the matrix-variate case. This extension would still lend itself to a convenient representation in terms of *covariance graphs*, where a missing edge between two nodes implies that the corresponding variables are marginally independent (Chaudhuri et al., 2007). Furthermore, in this work we focused on matrix Gaussian distributions, as they are commonly used for modeling continuous data. However, it would be interesting to explore whether the proposed penalized method could be employed in conjunction with alternative choices for the component densities, potentially encompassing situations with heavy-tails or skewness (see e.g., Melnykov and Zhu, 2018; Tomarchio et al., 2020). Lastly, alternative model selection strategies could be developed. While the grid search adopted in our numerical assessments produced good results, it may be computationally demanding in some applications, as discussed in Section 3.4. Therefore, stochastic optimization techniques could be adapted to our setting, as well as the E-MS algorithm introduced by Jiang et al. (2015).

As a final observation, we note that even in the matrix-variate scenario, the works focusing on precision matrices estimation in multi-class settings often enforce similarities between the underlying graphical models (Huang and Chen, 2015). While this assumption may be reasonable in various applications, it could compromise the quality of the results when clustering is the primary aim. Therefore, we believe that the strategy adopted in Casa et al. (2022) could be combined with the method proposed in this paper to encompass situations where different component precision matrices exhibit markedly different degrees of sparsity.

Supplementary materials

README: The supplemental files include a README file that describes the contents of the supplementary materials.

Appendix: The supplemental files contain the proof of Proposition 1 and additional results from the simulation study.

R code: The supplemental files include the source code for the `sparsemixmat R` package, which implements the clustering procedure described in the paper (also available at github.com/AndreaCappozzo/sparsemixmat).

Acknowledgments

The authors wish to thank two anonymous reviewers and the Associate Editor for their helpful comments. Andrea Cappozzo acknowledges financial support from the Italian Ministry of University and Research (MUR) under the Department of Excellence 2023-2027 grant agreement "Centre of Excellence in Economics and Data Science" (CEEDS).

Conflicts of interest

The authors report there are no competing interests to declare.

References

Anderlucci, L. and Viroli, C. (2015). Covariance pattern mixture models for the analysis of multivariate heterogeneous longitudinal data. *The Annals of Applied Statistics*, 9(2):777–800.

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *The Journal of Machine Learning Research*, 9:485–516.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 49(3):803.

Basford, K. E. and McLachlan, G. J. (1985). The mixture method of clustering applied to three-way data. *Journal of Classification*, 2:109–125.

Bien, J. and Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7):719–725.

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

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). *Model-based clustering and classification for data science: with applications in R*. Cambridge University Press.

Casa, A., Cappozzo, A., and Fop, M. (2022). Group-wise shrinkage estimation in penalized model-based clustering. *Journal of Classification*, 39(3):648–674.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, 28(5):781–793.

Chaudhuri, S., Drton, M., and Richardson, T. S. (2007). Estimation of a covariance matrix with zeros. *Biometrika*, 94(1):199–216.

Chen, J. T. and Gupta, A. K. (2005). Matrix variate skew normal distributions. *Statistics*, 39(3):247–253.

Chen, X. and Liu, W. (2019). Graph estimation for matrix-variate Gaussian data. *Statistica Sinica*, 29:479–504.

Dawid, A. P. (1981). Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika*, 68(1):265–274.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publication of the Mathematical Institute of the Hungarian Academy of Sciences*, 5(1):17–60.

Ferraccioli, F. and Menardi, G. (2023). Modal clustering of matrix-variate data. *Advances in Data Analysis and Classification*, 17:323–345.

Fop, M. and Murphy, T. B. (2018). Variable selection methods for model-based clustering. *Statistics Surveys*, 12:18–65.

Fop, M., Murphy, T. B., and Scrucca, L. (2019). Model-based clustering with sparse covariance matrices. *Statistics and Computing*, 29(4):791–819.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis and density estimation. *Journal of the American Statistical Association*, 97:611–631.

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

Gallaugher, M. P. and McNicholas, P. D. (2018). Finite mixtures of skewed matrix variate distributions. *Pattern Recognition*, 80:83–93.

Gao, X., Shen, W., Zhang, L., Hu, J., Fortin, N. J., Frostig, R. D., and Ombao, H. (2021). Regularized matrix data clustering and its application to image analysis. *Biometrics*, 77(3):890–902.

Glanz, H. and Carvalho, L. (2018). An expectation–maximization algorithm for the matrix normal distribution with an application in remote sensing. *Journal of Multivariate Analysis*, 167:31–48.

Green, P. J. (1990). On use of the EM for penalized likelihood estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 443–452.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2010). Pairwise variable selection for high-dimensional model-based clustering. *Biometrics*, 66(3):793–804.

Gupta, A. K. and Nagar, D. K. (2018). *Matrix variate distributions*, volume 104. CRC Press.

Hastie, T., Tibshirani, R., and Wainwright, M. (2019). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

Heo, J. and Baek, J. (2021). A penalized matrix normal mixture model for clustering matrix data. *Entropy*, 23(10):1249.

Holland, J. H. (1992). Genetic algorithms. *Scientific American*, 267(1):66–73.

Huang, F. and Chen, S. (2015). Joint learning of multiple sparse matrix Gaussian graphical models. *IEEE Transactions on Neural Networks and Learning Systems*, 26(11):2606–2620.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2:193–218.

Jiang, J., Nguyen, T., and Rao, J. S. (2015). The E-MS algorithm: model selection with incomplete data. *Journal of the American Statistical Association*, 110(511):1136–1147.

Keribin, C. (2000). Consistent estimation of the order of mixture models. *Sankhyā: The Indian Journal of Statistics, Series A*, 62(1):49–66.

Klosa, J., Simon, N., Westermark, P. O., Liebscher, V., and Wittenburg, D. (2020). Seagull: lasso, group lasso and sparse-group lasso regularization for linear regression models via proximal gradient descent. *BMC Bioinformatics*, 21(1):407.

Leng, C. and Tang, C. (2012). Sparse matrix graphical models. *Journal of the American Statistical Association*, 107(499):1187–1200.

Lian, H. (2011). Shrinkage tuning parameter selection in precision matrices estimation. *Journal of Statistical Planning and Inference*, 141(8):2839–2848.

Liu, D., Zhao, C., He, Y., Liu, L., Guo, Y., and Zhang, X. (2022). Simultaneous cluster structure learning and estimation of heterogeneous graphs for matrix-variate fMRI data. *Biometrics*, Online.

Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2009). Variable selection in model-based clustering: A general variable role modeling. *Computational Statistics & Data Analysis*, 53(11):3872–3882.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. Wiley.

Melnykov, V., Sarkar, S., and Melnykov, Y. (2021). On finite mixture modeling and model-based clustering of directed weighted multilayer networks. *Pattern Recognition*, 112:107641.

Melnykov, V. and Zhu, X. (2018). On model-based clustering of skewed matrix data. *Journal of Multivariate Analysis*, 167:181–194.

Melnykov, V. and Zhu, X. (2019). Studying crime trends in the USA over the years 2000–2012. *Advances in Data Analysis and Classification*, 13(1):325–341.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2020). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*. R package version 1.7-4.

Mosci, S., Rosasco, L., Santoro, M., Verri, A., and Villa, S. (2010). Solving Structured Sparsity Regularization with Proximal Methods. In Balcázar, J. L., Bonchi, F., Gionis, A., and Sebag, M., editors, *Machine Learning and Knowledge Discovery in Databases*, pages 418–433, Berlin, Heidelberg. Springer Berlin Heidelberg.

Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, 8:1145–1164.

Parikh, N. and Boyd, S. (2014). Proximal Algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ren, M., Zhang, S., and Wang, J. (2022). Consistent estimation of the number of communities via regularized network embedding. *Biometrics*, 79(3):2404–2416.

Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92(439):894–902.

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

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, 8(1):289–317.

Scrucca, L. and Raftery, A. E. (2015). Improved initialisation of model-based clustering using Gaussian hierarchical partitions. *Advances in Data Analysis and Classification*, 9(4):447–460.

Sharp, A., Chalatov, G., and Browne, R. P. (2022). A dual subspace parsimonious mixture of matrix normal distributions. *Advances in Data Analysis and Classification*.

Silva, A., Qin, X., Rothstein, S. J., McNicholas, P. D., and Subedi, S. (2023). Finite mixtures of matrix variate Poisson-log normal distributions for three-way count data. *Bioinformatics*, 39(5):btad167.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of computational and graphical statistics*, 22(2):231–245.

Subedi, S. (2023). Clustering matrix variate longitudinal count data. *Analytics*, 2(2):426–437.

Sustik, M. A., Calderhead, B., and Clavel, J. (2018). *glassoFast: Fast Graphical LASSO*. R package version 1.0.

Tomarchio, S. D. (2022). Matrix-variate normal mean-variance Birnbaum–Saunders distributions and related mixture models. *Computational Statistics*, pages 1–28.

Tomarchio, S. D., Gallaugher, M. P., Punzo, A., and McNicholas, P. D. (2022). Mixtures of matrix-variate contaminated normal distributions. *Journal of Computational and Graphical Statistics*, 31(2):413–421.

Tomarchio, S. D., Punzo, A., and Bagnato, L. (2020). Two new matrix-variate distributions with application in model-based clustering. *Computational Statistics & Data Analysis*, 152:107050.

Vichi, M. (1999). One-mode classification of a three-way data matrix. *Journal of Classification*, 16(1):27–44.

Vichi, M., Rocci, R., and Kiers, H. A. (2007). Simultaneous component and clustering models for three-way data: Within and between approaches. *Journal of Classification*, 24(1):71–98.

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

Viroli, C. (2011b). Model based clustering for three-way data structures. *Bayesian Analysis*, 6(4):573–602.

Viroli, C. (2012). On matrix-variate regression analysis. *Journal of Multivariate Analysis*, 111:296–309.

Wang, Y. and Melnykov, V. (2020). On variable selection in matrix mixture modelling. *Stat*, 9(1):e278.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley.

Witten, D. M., Friedman, J. H., and Simon, N. (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900.

Wright, S. J. (2015). Coordinate descent algorithms. *Mathematical Programming*, 151(1):3–34.

Xie, B., Pan, W., and Shen, X. (2008a). Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables. *Electronic Journal of Statistics*, 2:168–212.

Xie, B., Pan, W., and Shen, X. (2008b). Variable selection in penalized model-based clustering via regularization on grouped parameters. *Biometrics*, 64(3):921–930.

Yin, F., Hu, G., and Shen, W. (2023). Analysis of professional basketball field goal attempts via a bayesian matrix clustering approach. *Journal of Computational and Graphical Statistics*, 32(1):49–60.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68(1):49–67.

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

Zhu, X., Sarkar, S., and Melnykov, V. (2022). MatTransMix: An R package for Matrix Model-Based Clustering and Parsimonious Mixture Modeling. *Journal of Classification*, 39(1):147–170.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics*, 35(5):2173–2192.