Computational Statistics and Data Analysis 212 (2025) 108232

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Model-based clustering for covariance matrices via penalized Wishart mixture models

Andrea Cappozzo [a,1], Alessandro Casa [b,*,1]

[a] Department of Statistical Sciences, Università Cattolica del Sacro Cuore, Milan, Italy
[b] Faculty of Economics and Management, Free University of Bozen-Bolzano, Italy

ARTICLE INFO

Keywords:
Model-based clustering
Penalized likelihood
EM algorithm
Sparse covariance matrices
Sparse estimation
Covariance graph models

ABSTRACT

Covariance matrices provide a valuable source of information about complex interactions and dependencies within the data. However, from a clustering perspective, this information has often been underutilized and overlooked. Indeed, commonly adopted distance-based approaches tend to rely primarily on mean levels to characterize and differentiate between groups. Recently, there have been promising efforts to cluster covariance matrices directly, thereby distinguishing groups solely based on the relationships between variables. From a model-based perspective, a probabilistic formalization has been provided by considering a mixture model with component densities following a Wishart distribution. Notwithstanding, this approach faces challenges when dealing with a large number of variables, as the number of parameters to be estimated increases quadratically. To address this issue, a sparse Wishart mixture model is proposed, which assumes that the component scale matrices possess a cluster-dependent degree of sparsity. Model estimation is performed by maximizing a penalized log-likelihood, enforcing a covariance graphical lasso penalty on the component scale matrices. This penalty not only reduces the number of non-zero parameters, mitigating the challenges of high-dimensional settings, but also enhances the interpretability of results by emphasizing the most relevant relationships among variables. The proposed methodology is tested on both simulated and real data, demonstrating its ability to unravel the complexities of neuroimaging data and effectively cluster subjects based on the relational patterns among distinct brain regions.

1. Introduction

Covariance matrices play a pivotal role in many different scientific fields, as they encode linear relationships among observed variables, thus shedding light on how various complex and multidimensional phenomena are interrelated. In finance, they are essential for portfolio theory as they guide and inform investment strategies. In genomics, covariance matrices capture the dependencies among genes. As such, they are crucial for understanding genetic variation, playing a key role in identifying associations with traits and diseases, as well as in uncovering the underlying structure of complex genetic data. In neuroscience, they are often employed to investigate brain connectivity by examining correlations between different brain regions, providing insights into brain activities. Furthermore, covariance matrices serve as a cornerstone in many multivariate data analysis techniques such as principal component

* Corresponding author. Faculty of Economics and Management, Free University of Bozen-Bolzano, Piazzetta dell'Università, 1, 39031 Brunico, Italy.
E-mail address: alessandro.casa@unibz.it (A. Casa).
[1] Authors contributed equally to this work.

https://doi.org/10.1016/j.csda.2025.108232
Received 30 August 2024; Received in revised form 13 June 2025; Accepted 13 June 2025

Available online 20 June 2025

analysis and factor analysis, to name a few. Consequently, there is a substantial body of work in the literature dedicated to the precise estimation of covariance matrices, especially in those challenging situations where the number of observed variables is close or larger than the number of observations; readers can refer to Pourahmadi (2013) and references therein for a detailed review of the topic.

From a cluster analysis viewpoint, the informative content of the covariance matrices has often been somewhat overlooked. In fact, ubiquitous approaches, such as k-means and hierarchical clustering, primarily focus on differences in mean levels when attempting to group multivariate observations. Nonetheless, some recent scattered attempts have been made to propose clustering procedures aiming to identify groups of covariance matrices, therefore partitioning statistical units based on the intrinsic informative content of these objects. For example, in the seminal paper by Dryden et al. (2009), the statistical analysis of covariance matrices is thoroughly explored, leading to the introduction of non-Euclidean distances that are more appropriate in this context. Building on this, distance-based clustering approaches specifically designed for covariance matrices have been introduced (see e.g., Cabassi et al., 2018; Cappozzo et al., 2018). From a model-based perspective, Hidot and Saint-Jean (2010) have devised an approach to provide a probabilistically grounded formalization of this problem. More specifically, they consider a mixture model in which the component densities are assumed to follow a Wishart distribution, making it naturally suited for modeling covariance matrices. To obtain practical estimates of the parameters and the resulting partition, the authors introduce an EM-algorithm specifically designed to handle the characteristics of the proposed methodology. Although reasonable and successfully tested on real-world applications, this approach shows some limitations in high-dimensional situations. In fact, as stated by the authors, when the dimensionality of the covariance matrices to be clustered increases, the proposed method may encounter numerical and computational instabilities, jeopardizing the effectiveness of the procedure.

To address this limitation, in this work we introduce a sparse Wishart mixture model. Sparsity is induced in the component-specific scale matrices of the Wishart component densities by devising a penalized estimation procedure. More specifically, we develop a tailored EM-algorithm where covariance graphical lasso is embedded in the maximization step. By obtaining sparse estimates, our approach helps alleviate some of the challenges associated with covariance estimation when dealing with a large number of variables. On the one hand, it reduces estimation instabilities, leading to more reliable estimates. On the other hand, it simplifies the interpretation of the results by shrinking some of the entries of the component scale matrices toward zero, allowing us to focus on the most relevant relationships among variables. Moreover, it lends itself to a visual interpretation in terms of Gaussian covariance graph models (Chaudhuri et al., 2007). Our proposal is used to cluster subjects based solely on the information provided by their functional networks, obtained through functional magnetic resonance imaging (fMRI), and encoding the relationships in the activity of different regions of the brain. By highlighting different groups, the procedure serves as a stepping stone to unveil complex patterns in functional brain connectivity and its relation to anatomical connections and neurological diagnosis and diseases. Lastly, the induced sparsity allows focus on relevant differences in relationships among brain regions across groups, thus facilitating the interpretation of the results.

The remainder of the paper is structured as follows. In Section 2, we provide an overview of the model-based clustering framework, focusing specifically on approaches to partition covariance matrices. In Section 3, we motivate and present our proposal, covering both model estimation and selection. Sections 4 and 5 assess the performance and applicability of the proposed approach, testing it on simulated and real data. Lastly, Section 6 concludes the article with some final remarks and discusses potential future research directions.

2. Preliminaries and related work

Model-based clustering (Fraley and Raftery, 2002; McNicholas, 2016; Bouveyron et al., 2019) provides a widely employed probabilistic strategy to cluster analysis. This approach posits that data stem from a finite mixture distribution, describing the heterogeneity of the statistical units and reflecting the presence of diverse sub-populations or clusters. Conventionally, from a frequentist perspective, maximum likelihood estimation is performed by means of the EM-algorithm (Dempster et al., 1977), utilizing a data augmentation approach where latent group indicators are treated as missing data. From a practical standpoint, once the estimates of the parameters are obtained, each observation is assigned to the pertaining cluster according to the Maximum a Posteriori (MAP) rule, assuming a one-to-one correspondence between mixture components and groups.

When continuous data are observed, Gaussian densities are customarily considered as mixture components. However, there has also been growing attention towards more flexible modeling choices, potentially able to accommodate non-elliptical and asymmetric cluster shapes; readers can refer, for example, to McLachlan and Peel (1998); Lin (2009, 2010); Andrews et al. (2011); Gollini and Murphy (2014); Browne and McNicholas (2015); Tomarchio et al. (2024) for relevant methodological proposals on the topic. In recent years, the emergence of unique data structures has become increasingly common, presenting new challenges during the modeling phase. Consequently, innovative strategies have been developed to accommodate these scenarios, adapting model-based clustering to handle functional data (Bouveyron and Jacques, 2011; Bouveyron et al., 2015), network data (Snijders and Nowicki, 1997; Handcock et al., 2007), time-dependent data (De la Cruz-Mesía et al., 2008; McNicholas and Murphy, 2010), and matrix-variate data (Viroli, 2011), among others.

Along the same lines, Hidot and Saint-Jean (2010) have recently proposed a Wishart mixture model, which extends the model-based approach to partition cross-product matrices. Formally, let $\mathbf{\Gamma} = \{\mathbf{\Gamma}_1, \dots, \mathbf{\Gamma}_n\}$ be a sample of n square, symmetric, and positive semi-definite matrices, with $\mathbf{\Gamma}_i \in \mathbb{R}^{p \times p}$, $i = 1, \dots, n$. These matrices clearly share a strong connection with sample covariance matrices, differing from them only by a normalization term; consequently, they encode linear relationships among the p observed variables. Implicitly, it is assumed that $\mathbf{\Gamma}_i$ has been generated as the cross-product $\mathbf{X}_{(i)}^T \mathbf{X}_{(i)}$, where $\mathbf{X}_{(i)} = \{\mathbf{x}_1^{(i)}, \dots, \mathbf{x}_{n_i}^{(i)}\}$ is a random sample with $\mathbf{x}_l^{(i)} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{\Sigma})$, for $l = 1, \dots, n_i$. According to this assumption, $\mathbf{\Gamma}_i$ is distributed as a central Wishart with scale matrix $\mathbf{\Sigma}$ and degrees

of freedom n_i, respectively identifying the covariance matrix and the size of the underlying sample $\mathbf{X}_{(i)}$. Note that, in the following and in line with Section 2.1 and 2.2 in Hidot and Saint-Jean (2010), we assume to have observed directly $\mathbf{\Gamma}_i$, without having access to $\mathbf{X}_{(i)}$ and considering the sample size n_i as unknown, for $i = 1, \dots, n$. To cluster the observed matrices, the authors posit that $\mathbf{\Gamma}_i$ arises from a Wishart mixture model, with density given by

$$f(\mathbf{\Gamma}_i; \mathbf{\Theta}) = \sum_{k=1}^{K} \tau_k f_{\mathcal{W}}(\mathbf{\Gamma}_i; \mathbf{\Sigma}_k, \nu_k), \tag{1}$$

where K is the number of mixture components, τ_k's are the mixing proportions with $\tau_k > 0$ for all $k = 1, \dots, K$, $\sum_{k=1}^{K} \tau_k = 1$, and $\mathbf{\Theta} = \{\mathbf{\Theta}_k\}_{k=1}^{K}$ with $\mathbf{\Theta}_k = \{\tau_k, \mathbf{\Sigma}_k, \nu_k\}$ the set of model parameters. Lastly, $f_{\mathcal{W}}(\cdot; \mathbf{\Sigma}_k, \nu_k)$ denotes the density of a Wishart distribution with component-specific scale matrix $\mathbf{\Sigma}_k$ and degrees of freedom ν_k, given by:

$$f_{\mathcal{W}}(\mathbf{\Gamma}_i; \mathbf{\Sigma}_k, \nu_k) = \frac{|\mathbf{\Gamma}_i|^{\frac{\nu_k - p - 1}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}_k^{-1}\mathbf{\Gamma}_i\right)\right\}}{2^{\frac{\nu_k p}{2}} |\mathbf{\Sigma}_k|^{\nu_k/2} \gamma_p\left(\nu_k/2\right)},$$

with $\gamma_p(\cdot)$ denoting the multivariate Gamma function:

$$\gamma_p(a) = \pi^{p(p-1)/4} \prod_{j=1}^{p} \gamma\left(a + (1-j)/2\right),$$

where $\gamma(b) = \int_0^{+\infty} t^{b-1} e^{-t} dt, b > 0$, is the Gamma function.

In the work by Hidot and Saint-Jean (2010), parameters are estimated by means of a tailored EM-algorithm and, coherently with the model-based clustering framework, the final partition of the matrices $\mathbf{\Gamma}_1, \dots, \mathbf{\Gamma}_n$ in K clusters is obtained by resorting to the MAP rule. The authors demonstrate the effectiveness of their clustering method, which implicitly groups the underlying unobserved samples $\mathbf{X}_{(i)}$ based solely on the relationships among the observed variables, as encoded in the estimated group-specific scale matrices $\mathbf{\Sigma}_k$.

As previously stated, Hidot and Saint-Jean (2010) consider a scenario where $\mathbf{\Gamma}_i$ is observed, while the associated sample size n_i, for $i = 1, \dots, n$, is not directly accessible. The unobservable ratio p/n_i critically impacts their approach, leading to relevant drawbacks when this ratio is relatively large. Specifically, when $p/n_i > 1$ the cross-product matrix and consequently the sample covariance become singular, inherently violating the standard assumption that $\mathbf{\Gamma}_i$ follows a Wishart distribution and precluding reliable inference on the relationships among observed features. To address this issue, they assume – an assumption we also adopt throughout the remainder of this paper – that $p \leq n_i$ for all $i = 1, \dots, n$. Nonetheless, even when $p/n_i < 1$ but remains moderately large, the sample covariance matrix is known to be highly unstable, accumulating substantial estimation errors due to the large dimensionality of the parameter space, which scales quadratically with p (see, e.g., Pourahmadi, 2013, for a thorough discussion). In this context, $\hat{\mathbf{\Sigma}}_k$, for $k = 1, \dots, K$, may not reliably characterize the relationships among the observed features, potentially compromising the soundness of the resulting partition. Additionally, in a clustering framework, it is often of interest to conduct post-hoc analyses to explore cluster-specific patterns, primarily captured by $\hat{\mathbf{\Sigma}}_k$. Such analyses can yield valuable insights into the underlying population and the factors contributing to its heterogeneity. However, when the dimensionality p is large, interpreting the results becomes challenging, as it is difficult to discern which relationships among variables are driving the clustering.

For these reasons, in this work we extend the approach by Hidot and Saint-Jean (2010) by proposing a sparse mixture of Wishart distributions. The proposal builds upon the ever-expanding literature on sparse model-based clustering (see Pan and Shen, 2007; Zhou et al., 2009; Fop et al., 2019; Casa et al., 2022, among others) by extending its rationale to clustering covariance matrices. In a nutshell, our proposal addresses some of the aforementioned issues by providing sparse estimates of the matrices $\mathbf{\Sigma}_k$ for $k = 1, \dots, K$. On the one hand, this yields more reliable results in large-dimensional settings, potentially refining the final partitions. On the other hand, it simplifies the interpretation of the results by capturing group-specific marginal dependency structures. The proposed method is based on a penalized likelihood framework, which is detailed in the following section.

3. Sparse Wishart mixture models

3.1. Model specification

As previously highlighted, the method proposed by Hidot and Saint-Jean (2010) possesses some weaknesses regarding its applicability in high-dimensional settings and its interpretability. To overcome these limitations, we propose an alternative method referred to as *Sparsemixwishart*. Our method estimates parameters and subsequently performs clustering by maximizing a penalized log-likelihood defined as:

$$\ell_P(\mathbf{\Theta}; \mathbf{\Gamma}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \tau_k f_{\mathcal{W}}(\mathbf{\Gamma}_i; \mathbf{\Sigma}_k, \nu_k) - p_\lambda(\mathbf{\Theta}), \tag{2}$$

where the first term is the log-likelihood of a mixture of Wishart distributions, while the second one is a penalty term on the model parameters. As it is customary in the statistical learning with sparsity literature, λ denotes the positive hyperparameter that controls the strength of the penalization.

Taking our steps from the covariance graphical lasso introduced by Bien and Tibshirani (2011), in this work we consider the following penalty term:

$$p_\lambda(\boldsymbol{\Theta}) = \lambda \sum_{k=1}^{K} ||\mathbf{P} * \boldsymbol{\Sigma}_k||_1 \,, \tag{3}$$

where $*$ denotes the element-wise multiplication and $||\cdot||_1$ is the L_1-norm with $||A||_1 = \sum_{jh}|A_{jh}|$. Lastly, $\mathbf{P}$ is a predefined matrix with non-negative entries, whose specifics will be discussed shortly.

The penalty in (3), thanks to the geometry of the L_1-norm, induces a certain degree of sparsity by shrinking to zero some of the entries of the matrices $\boldsymbol{\Sigma}_k, k = 1, \dots, K$, with the amount of sparsity depending on λ. This allows us to alleviate potential issues arising when estimating association matrices in large-dimensional scenarios. Sparse representations of the component scale matrices introduce a convenient connection with *covariance graph models* (Chaudhuri et al., 2007). Within this framework, the observed variables correspond to distinct nodes in the graph, with edges representing marginal dependencies between them. Consequently, two variables are not connected by an edge when they are marginally independent. This not only facilitates convenient graphical visualizations of the results, but also simplifies interpretation by enabling a better characterization of the obtained partition by exploring how marginal dependencies among features vary across clusters.

The inclusion of $\mathbf{P}$ in Equation (3) has the potential to enhance the flexibility of the method, as thoughtful specifications may incorporate user-defined constraints or prior beliefs about relationships among variables. Some guidance on this choice is provided in the seminal paper by Bien and Tibshirani (2011). Specifically, common approaches include using an all-ones matrix with zeros on the main diagonal, thereby avoiding shrinkage of the diagonal entries of $\boldsymbol{\Sigma}_k$; or defining $\mathbf{P}$ as an adjacency matrix with predefined patterns. For example, in neuroscience, the number of white matter fibers connecting different brain regions can offer prior insights into the association structure within functional Magnetic Resonance Imaging; readers can refer to Section 5 for further details.

3.2. Model estimation

Considering for the moment K and λ as fixed, parameter estimates are obtained by maximizing (2) with respect to $\boldsymbol{\Theta}$. In this work, we rely on an EM-algorithm specifically tailored for maximum penalized likelihood estimation. To this end, we first define the *penalized complete-data log-likelihood* related to (2) as

$$\ell_C(\boldsymbol{\Theta}) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \log \tau_k f_{\mathcal{W}}(\boldsymbol{\Gamma}_i; \boldsymbol{\Sigma}_k, \nu_k) - \lambda \sum_{k=1}^{K} \|\mathbf{P} * \boldsymbol{\Sigma}_k\|_1 \,, \tag{4}$$

where, as usual, $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})$ is the realization of the latent group membership indicator variable $\mathbf{Z}_i$, with $z_{ik} = 1$ if the covariance matrix $\boldsymbol{\Gamma}_i$ belongs to the k-th cluster and zero otherwise.

During each expectation step (E-step), the posterior probability of $\mathbf{Z}_i$ is updated, allowing for the computation of the conditional expectation of Equation (4), commonly referred to as the Q-function. Subsequently, the Q-function serves as the objective function to be maximized in the maximization step (M-step) to obtain updated parameter estimates. A detailed description of the algorithm follows in the next subsections.

3.2.1. E-step

At the t-th iteration, the E-step involves computing the posterior probability that $\boldsymbol{\Gamma}_i$ belongs to the k-th component conditionally on the parameter estimates $\hat{\boldsymbol{\Theta}}^{(t-1)}$ obtained at the previous iteration. More specifically, $\hat{z}_{ik}^{(t)}$ are ordinarily updated as

$$\hat{z}_{ik}^{(t)} = \frac{\hat{\tau}_k^{(t-1)} f_{\mathcal{W}}\left(\boldsymbol{\Gamma}_i; \hat{\boldsymbol{\Sigma}}_k^{(t-1)}, \hat{\nu}_k^{(t-1)}\right)}{\sum_{l=1}^{K} \hat{\tau}_l^{(t-1)} f_{\mathcal{W}}\left(\boldsymbol{\Gamma}_i; \hat{\boldsymbol{\Sigma}}_l^{(t-1)}, \hat{\nu}_l^{(t-1)}\right)} \,,$$

for $i = 1, \dots, n$ and $k = 1, \dots, K$.

3.2.2. M-step

In the M-step, the updates for the parameter estimates are obtained by maximizing the penalized Q-function, which in this context is defined as follows:

$$\begin{aligned} Q(\boldsymbol{\Theta}_1, \dots, \boldsymbol{\Theta}_K) = \sum_{i=1}^{n}\sum_{k=1}^{K} \hat{z}_{ik}^{(t)} \Bigg[\log \tau_k - \frac{\nu_k p}{2} \log 2 - \frac{\nu_k}{2} \log |\boldsymbol{\Sigma}_k| - \log \gamma_p\left(\frac{\nu_k}{2}\right) + \\ \left(\frac{\nu_k - p - 1}{2}\right) \log |\boldsymbol{\Gamma}_i| - \frac{\operatorname{tr}(\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\Gamma}_i)}{2} \Bigg] - \lambda \sum_{k=1}^{K} \|\mathbf{P} * \boldsymbol{\Sigma}_k\|_1 \,. \end{aligned} \tag{5}$$

The simultaneous maximization of (5) with respect to all the parameters is unfeasible. Consequently, we resort to a partial optimization strategy cycling over three distinct steps. First, note that closed-form updates for τ_k are available and are given by

$$\hat{\tau}_k^{(t)} = \frac{\hat{n}_k^{(t)}}{n}, \tag{6}$$

where $\hat{n}_k^{(t)} = \sum_{i=1}^{n} \hat{z}_{ik}^{(t)}$ is the estimated sample size of cluster k, for $k = 1, \dots, K$, at the t-th iteration of the EM-algorithm. Subsequently, the update for the degrees of freedom is obtained by maximizing (5) with respect to ν_k. Following the rationale outlined in Hidot and Saint-Jean (2010), this corresponds to solving the following equation for each mixture component separately:

$$\sum_{i=1}^{n} \hat{z}_{ik}^{(t)} \log \left| \frac{\boldsymbol{\Gamma}_i (\hat{\boldsymbol{\Sigma}}_k^{(t-1)})^{-1}}{2} \right| = \sum_{i=1}^{n} \hat{z}_{ik}^{(t)} \sum_{j=1}^{p} \psi \left(\frac{\nu_k - j + 1}{2} \right), \tag{7}$$

where $\psi(\cdot)$ is the digamma function. No closed-form solution exists; nonetheless, (7) defines a simple one-dimensional root-finding problem with a continuous and monotonic function in ν_k. It is therefore solved numerically to obtain $\hat{\nu}_k^{(t)}$ for $k = 1, \dots, K$.

Updates for component scale matrices are more involved and they differ from those in Hidot and Saint-Jean (2010), as our proposal aims to induce sparsity in $\boldsymbol{\Sigma}_k$. More specifically, it is easy to see that maximization of (5) with respect to $\boldsymbol{\Sigma}_k$ boils down to minimizing the following objective function:

$$Q(\boldsymbol{\Sigma}_k) = \log|\boldsymbol{\Sigma}_k| + \operatorname{tr}\left(\boldsymbol{\Sigma}_k^{-1} \tilde{\mathbf{S}}_k\right) + \frac{2\lambda}{\hat{n}_k^{(t)} \hat{\nu}_k^{(t)}} \|\mathbf{P} * \boldsymbol{\Sigma}_k\|_1, \tag{8}$$

where $\tilde{\mathbf{S}}_k = (\sum_{i=1}^{n} \hat{z}_{ik}^{(t)} \boldsymbol{\Gamma}_i) / (\hat{n}_k^{(t)} \hat{\nu}_k^{(t)})$. Minimizing (8) with respect to $\boldsymbol{\Sigma}_k$ is equivalent to solving a covariance graphical lasso problem (see Equation (1) in Bien and Tibshirani, 2011) with a modified penalty coefficient of $2\lambda/(\hat{n}_k^{(t)} \hat{\nu}_k^{(t)})$. The L_1-norm naturally induces sparse and positive definite estimates, forcing some entries of $\hat{\boldsymbol{\Sigma}}_k$ to be exactly zero. The covariance graphical lasso in (8) poses a non-convex optimization problem, for which a majorize-minimize approach was initially developed as a solution. More recently, Wang (2014) proposed a coordinate descent method to minimize (8), iteratively updating one column and one row of $\hat{\boldsymbol{\Sigma}}_k$ at a time while keeping all other elements fixed. Since the latter algorithm has been shown to be both faster and numerically more stable, we adopted it for our estimation process, as implemented in the `covglasso` `R` package (Fop, 2021). Regarding the choice of $\mathbf{P}$ in what follows, unless otherwise stated (see Section 5), we assume it to be an all-one matrix with zeros on the main diagonal.

The assessment of the overall convergence of the EM-algorithm involves monitoring increases in the penalized log-likelihood throughout each complete iteration. More specifically, the implemented EM-algorithm is considered to have achieved convergence when

$$\left| \ell_P(\boldsymbol{\Theta}^{(t)}; \boldsymbol{\Gamma}) - \ell_P(\boldsymbol{\Theta}^{(t-1)}; \boldsymbol{\Gamma}) \right| \leq \varepsilon, \quad \varepsilon > 0.$$

In the following analyses, ε is set to 10^{-6}.

One final computational aspect worth discussing is the initialization process, which is crucial for any deterministic algorithm. While multiple random initializations are a consistently viable option, they come with significant computational cost. Despite their undeniable effectiveness, this approach can become impractical when clustering high-dimensional covariance objects. To address this issue, we propose using hierarchical clustering with a non-Euclidean Riemannian distance (Dryden et al., 2009) as a measure of dissimilarity to provide an initial partition of the data. This distance-based initialization is not only computationally efficient, but also applicable when exploring different numbers of mixture components, as discussed in the upcoming section on model selection.

The whole procedure has been implemented mainly using the `R` software (R Core Team, 2024), with some routines written in `C++` to reduce total computational time. The code is freely available in the form of an `R` package at https://github.com/AndreaCappozzo/sparsemixwishart.

3.3. Model selection

In the previous sections, the proposed methodology has been presented considering both the number of clusters K and the hyperparameter λ as fixed. However, in practical applications, these values are typically unknown and must be determined to obtain a valid partition. In this work, we select values of K and λ which maximize a modified version of the Bayesian Information Criterion (BIC, Schwarz, 1978) defined as

$$\text{BIC} = 2 \log L(\hat{\boldsymbol{\Theta}}) - d_0 \log(n), \tag{9}$$

where $\log L(\hat{\boldsymbol{\Theta}})$ is the log-likelihood evaluated in $\hat{\boldsymbol{\Theta}}$. Here, d_0 denotes the number of parameters not shrunk to zero by the penalized estimation procedure outlined in Section 3.2. The criterion in (9) has already proven useful for simultaneously selecting K and λ within penalized model-based clustering (see, e.g., Pan and Shen, 2007; Fop et al., 2019), supported theoretically by Zou et al. (2007) and following conjectures by Efron et al. (2004). However, while the consistency of the BIC selection has been demonstrated (see, e.g., Keribin, 2000) and its empirical effectiveness has been observed in numerous clustering applications (see Bouveyron et al., 2019, and references therein), its adoption deserves careful consideration. Selecting the number of groups and recovering the association structure present distinct challenges with potentially different requirements. Although cross-validation is a common strategy for tuning penalization strength in covariance regularization (see, e.g., Warton, 2008; Bien and Tibshirani, 2011; Cai and Liu, 2011; Cibinel et al., 2024), information criteria offer a computationally more efficient approach for model selection, as noted, for example, by Yuan and Lin (2007). This computational advantage, combined with the inherent challenges of using cross-validation for the unsupervised selection of K, motivates our adoption of the criterion in (9). However, as recently highlighted by Tomarchio and Punzo (2025), the selection of an information criterion should not be made without careful consideration, as blind choice can lead to underfitting or overfitting.

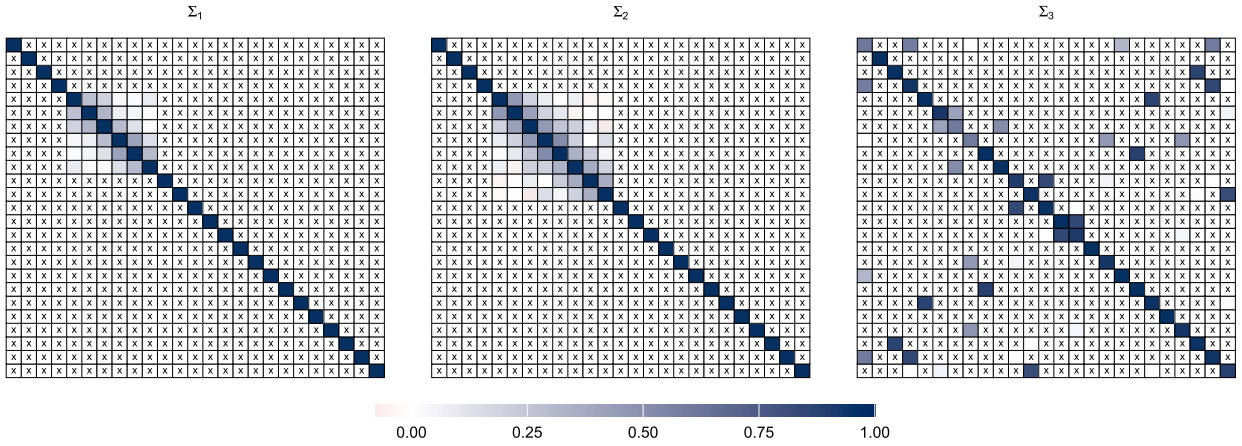

Fig. 1. Heatmaps of the true 25×25 scale matrices $\mathbf{\Sigma}_k$, $k = 1, 2, 3$, considered in the simulated data experiment. A zero entry in the matrices is indicated with the symbol $\times$.

Given that a comprehensive solution to the problem of model selection in penalized model-based clustering lies beyond the scope of this paper, we adopt hereafter the BIC in (9) as a working choice. However, caution is advised, especially in high-dimensional contexts and with complex data structures such as those considered in this study, as BIC can tend to favor models with fewer clusters and smaller values of the shrinkage parameter λ. Specific considerations related to this issue are discussed in the simulation study and in the real data application presented in the following sections.

4. Simulation study

4.1. Experimental setup

In this section, we evaluate the performance of the proposed method on synthetic data, examining its capability to recover the underlying data partition while effectively identifying the sparsity structure in the component-specific scale matrices. For each replication of the experiment, we draw $n = 200$ random matrices with dimension $p \times p$, with $p = 25$, from a Wishart mixture model with $K = 3$ components. We consider equal mixing proportions $\tau_k = \frac{1}{3}$ for all $k = 1, 2, 3$; with degrees of freedom of 30, 30, and 40 for the first, second, and third component, respectively. The scale matrices are generated using two different mechanisms: an alternated-blocks structure for $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, and a sparse-at-random Erdős-Rényi graph structure (Erdős and Rényi, 1960) for $\mathbf{\Sigma}_3$. This procedure results in component-wise different sparsity patterns, as displayed in Fig. 1. We repeat the experiment $B = 500$ times and, alongside the methodology introduced in Section 3, we consider the following competing models:

- *Hidot Saint-Jean:* the original Wishart mixture model presented by Hidot and Saint-Jean (2010), where maximum likelihood estimation is used for model fitting, resulting in non-sparse estimates of the scale matrices $\mathbf{\Sigma}_k$ for $k = 1, 2, 3$.
- *Hclust-Euclidean*: hierarchical clustering using Frobenius distance as the measure of dissimilarity between random matrices, employing Ward's linkage method (Murtagh and Legendre, 2014). Specifically, for each pair of observations $\mathbf{\Gamma}_i$ and $\mathbf{\Gamma}_j$, we compute the following quantity to construct the dendrogram:

$$d_E(\mathbf{\Gamma}_i, \mathbf{\Gamma}_j) = \|\mathbf{\Gamma}_i - \mathbf{\Gamma}_j\|_F = \sqrt{\mathrm{tr}\{(\mathbf{\Gamma}_i - \mathbf{\Gamma}_j)^{\mathrm{T}}(\mathbf{\Gamma}_i - \mathbf{\Gamma}_j)\}}.$$

- *Hclust-Riemannian*: hierarchical clustering using Non-Euclidean Riemannian distance (Dryden et al., 2009) as the measure of dissimilarity between random matrices, employing Ward's linkage method. Specifically, for each pair of observations $\mathbf{\Gamma}_i$ and $\mathbf{\Gamma}_j$, we compute the following quantity to construct the dendrogram:

$$d_R(\mathbf{\Gamma}_i, \mathbf{\Gamma}_j) = \|\log(\mathbf{\Gamma}_i^{-1/2}\mathbf{\Gamma}_j\mathbf{\Gamma}_i^{-1/2})\|_F,$$

  where both the logarithm and square root are taken in the sense of matrix operations (Arsigny et al., 2007).

The simulation study has three main objectives. First, we aim to evaluate the ability to recover the true underlying group structure. To this end, all competing clustering methods are evaluated using the adjusted Rand index (ARI, Hubert and Arabie, 1985). Secondly, for the model-based methods only, namely our proposal *Sparsemixwishart* and the original *Hidot Saint-Jean* procedure, we assess the quality of the parameter estimation by computing, for each Wishart component, the Kullback–Leibler divergence between true and estimated densities:

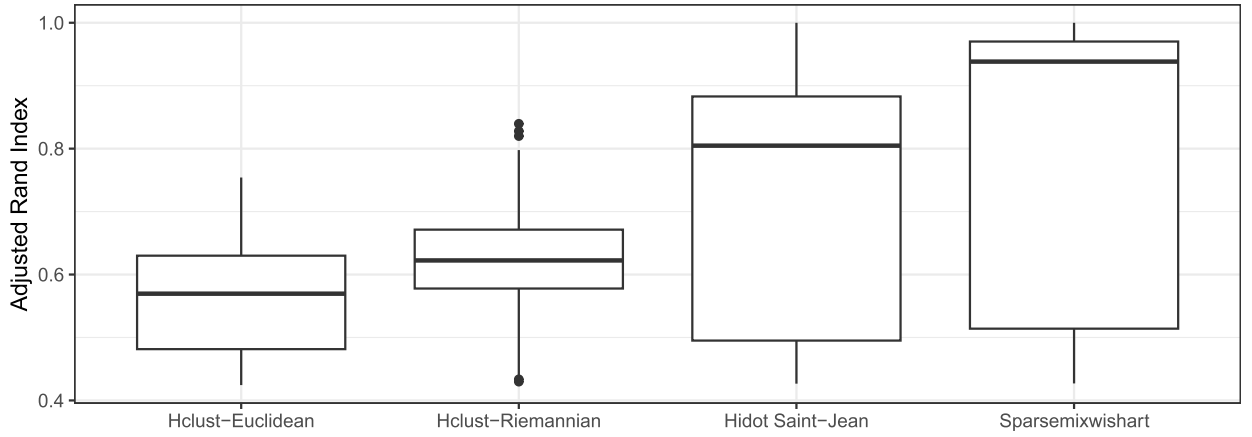

Fig. 2. Boxplots of the ARI for the $B = 500$ repetitions of the simulated experiment. For the *Sparsemixwishart* method, the shrinkage parameter λ is selected according to the BIC defined in (9).

$$D_{\mathrm{KL}}(\mathcal{W}_k \,\|\, \hat{\mathcal{W}}_k) = \frac{1}{2}\left[\hat{\nu}_k\left(\log|\hat{\boldsymbol{\Sigma}}_k| - \log|\boldsymbol{\Sigma}_k|\right) + \nu_k \mathrm{tr}\left(\hat{\boldsymbol{\Sigma}}_k^{-1}\boldsymbol{\Sigma}_k\right) + 2\log\left(\frac{\gamma_p\left(\hat{\nu}_k/2\right)}{\gamma_p\left(\nu_k/2\right)}\right) + \left(\nu_k - \hat{\nu}_k\right)\psi\left(\frac{\nu_k}{2}\right) - \frac{\nu_k}{p}\right], \tag{10}$$

where $\mathcal{W}_k$ and $\hat{\mathcal{W}}_k$ denote the true and estimated Wishart distributions, respectively, for the k-th component, with parameters ν_k, $\boldsymbol{\Sigma}_k$ and $\hat{\nu}_k$, $\hat{\boldsymbol{\Sigma}}_k$. Lastly, we study the support recovery performance of our approach by determining whether the penalized estimation introduced in *Sparsemixwishart* can accurately identify the true sparsity patterns in $\boldsymbol{\Sigma}_k$. To do so, we consider the F_1 score defined as follows:

$$F_1 = \frac{\mathtt{tp}}{\mathtt{tp} + 0.5(\mathtt{fp} + \mathtt{fn})}. \tag{11}$$

In Equation (11), `tp` denotes the number of entries different from zero in $\boldsymbol{\Sigma}_k$ that are correctly estimated as such; `fp` represents the number of zero entries incorrectly identified as non-zero, and `fn` is the number of non-zero elements which are wrongly shrunk to zero.

For tuning the hyperparameter λ, a grid of 150 equispaced values is considered for each repetition of the simulated experiment, with the best shrinkage factor selected using the modified BIC as suggested in Section 3.2. Lastly, the issue of matching the estimated clusters with the original mixture components is addressed using the `matchClasses` function from the `e1071 R` package (Meyer et al., 2023). Simulation results are reported in the next subsection.

4.2. Simulation results

Fig. 2 displays the empirical distribution of the ARI metric across the $B = 500$ repetitions of the simulated experiment for all the competing models introduced in the previous section. We immediately recognize that model-based methods outperform distance-based procedures in recovering the true underlying partition. It is noteworthy that hierarchical clustering based on Riemannian distance exhibits a higher median ARI and a lower standard error compared to its Frobenius counterpart. This advantage is likely due to the non-Euclidean nature of the space of positive semi-definite symmetric matrices, where a Riemannian metric may be more suitable for calculating distances between covariance objects (Dryden et al., 2009). Overall, our proposal demonstrates the best performance in terms of ARI, greatly surpassing the quality of the partition achieved by Hidot and Saint-Jean (2010). The reason for this improvement stems from the covariance graphical lasso penalty introduced in *Sparsemixwishart*, which allows for the sparse and more numerically stable estimation of the model parameters. This result is reflected in the median Kullback–Leibler divergence between the true and estimated components, computed across the $B = 500$ repetitions of the simulated experiment, as reported in Table 1. In detail, the median value of $D_{\mathrm{KL}}(\mathcal{W}_k \,\|\, \hat{\mathcal{W}}_k)$ is lower for every component when the model is estimated using our *Sparsemixwishart* approach. Turning to the support recovery performance of the proposed methodology, Fig. 3 presents smoothed line plots of the component-wise different F_1 scores calculated according to Equation (11), shown as functions of the shrinkage factor λ. As the figure displays, a λ value close to 100 results in good recovery of the sparsity structure for the first two components, whereas a stronger penalty appears necessary for the accurate recovery of $\boldsymbol{\Sigma}_3$. Note that, across the 500 replications of the simulated experiment, an average λ of 34.5 (SD = 14) was selected by the modified BIC in (9). However, the visual inspection of Fig. 3 suggests that a higher λ might provide better overall F_1 scores across the three components. As such, the selection of λ using the criterion in (9) seems to be sub-optimal, if the recovery of the true structures in the scale matrices is the final aim. Needless to say the BIC, here used to select the shrinkage factor λ, is not specifically designed to maximize support recovery metrics. Therefore, when support recovery is the primary

Table 1
Median Kullback–Leibler divergence between true and estimated Wishart components over 500 repetitions of the simulated experiment. Standard errors are reported in brackets.

	Hidot Saint-Jean	Sparsemixwishart
$D_{\mathrm{KL}}(\mathcal{W}_1 \parallel \hat{\mathcal{W}}_1)$	2.977 (386.3)	2.065 (407.807)
$D_{\mathrm{KL}}(\mathcal{W}_2 \parallel \hat{\mathcal{W}}_2)$	3.596 (498.709)	2.535 (543.881)
$D_{\mathrm{KL}}(\mathcal{W}_3 \parallel \hat{\mathcal{W}}_3)$	2.943 (10.808)	2.258 (10.996)

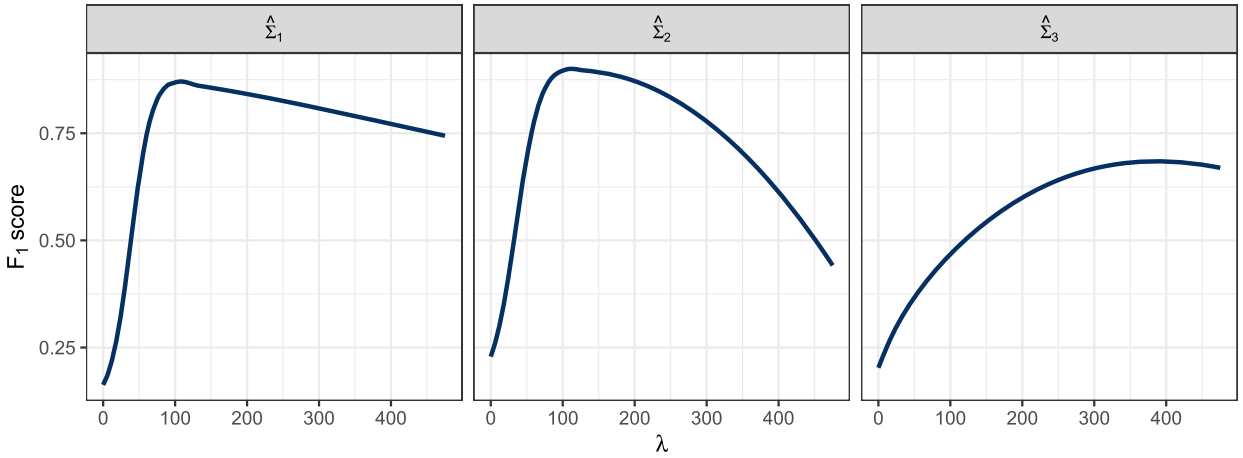

Fig. 3. Smoothed lines plots of the component-wise different F_1 score for $B = 500$ repetitions of the simulated experiment, varying shrinkage factor λ.

goal, a more appropriate criterion should be considered. For instance, in a different framework Casa et al. (2024) recently re-framed hyperparameter selection for sparse covariance matrix estimation as a sequential hypothesis testing procedure, specifically designed to recover the underlying structure. This might serve as a stepping stone for further generalizations to the covariance graphical lasso in the model-based clustering framework. Nonetheless, when dealing with mixture models, an additional complexity potentially influencing the results is given by the possibly different sparsity patterns and magnitudes across mixture components. Such difficulties are known in the penalized model-based clustering literature. As a matter of example, empirical results have demonstrated that an alternated-blocks structure is more easily recovered in this context (Fop et al., 2019). Recently, a solution has been proposed in Casa et al. (2022) in the framework of Gaussian mixture models. While adapting the methodology developed in Casa et al. (2022) to sparse mixtures of Wishart is beyond the scope of this paper, it could be a worthwhile avenue for future research. Lastly, the computational costs of the considered methods are compared in terms of computing times required for performing the fitting. Fig. 4 displays boxplots of computing time (in seconds) over $B = 500$ runs of the simulated experiment. As expected, the penalized estimation embedded in our approach results in a slightly slower methodology compared to competitors, with the covariance graphical lasso representing the primary bottleneck in the EM algorithm described in Section 3.2. Nonetheless, the computational burden for implementing our proposed method remains manageable, especially considering the significant improvements it yields in both partition quality and structure recovery. The simulation was run on a computer cluster with 20 AMD EPYC Rome cores (2.2 GHz).

The simulation study demonstrates that the flexibility and parsimonious nature of the covariance graphical lasso penalty not only lead to improved estimation compared to the *Hidot Saint-Jean* method, when dealing with sparse scale matrices, but also facilitate more accurate recovery of the true underlying partition in the clustering of covariance objects. These improvements are also observed in real data analyses in the neuroimaging context, as detailed in the next section.

5. Application to fMRI functional networks

The dataset analyzed in this study was collected as part of a pilot project for the Enhanced Nathan Kline Institute-Rockland Sample. A detailed description of the project's objectives is available at http://fcon_1000.projects.nitrc.org/indi/enhanced/. Specifically, brain imaging data were obtained from 24 subjects, covering 70 brain regions defined using anatomical segmentation based on the Desikan atlas (Desikan et al., 2006). The study incorporates multiple sources of information, including:

- *Structural networks*, collected by means of Diffusion Tensor Imaging (DTI), measuring anatomical interconnections based on the amount of white matter fibers connecting different brain regions.
- *Dynamic functional activity*, collected by means of resting-state fMRI (R-fMRI), measuring the activity of the 70 brain regions over 404 equally spaced time points, recorded when the subjects are not performing specific tasks.

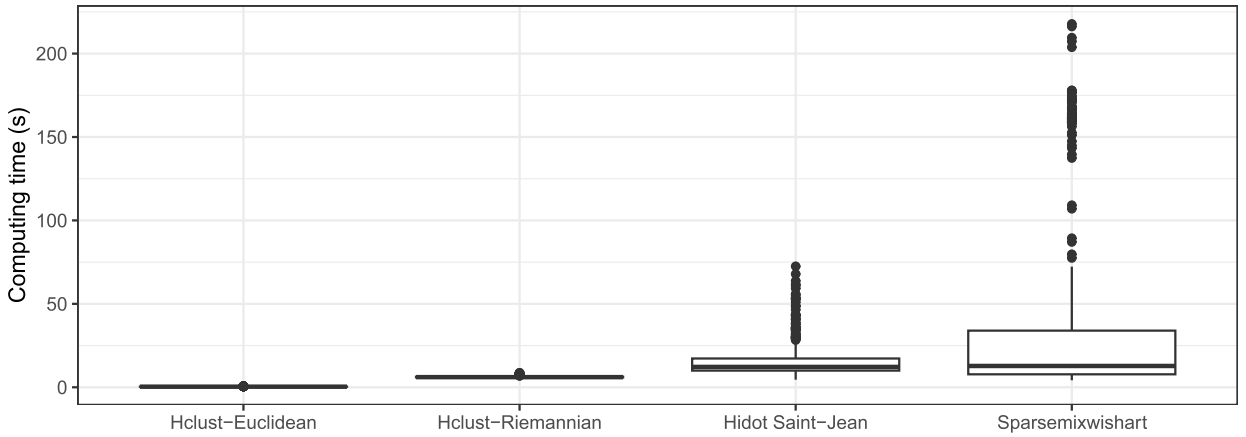

Fig. 4. Boxplots of computing time (in seconds) over $B = 500$ runs of the simulated experiment. For the *sparsemixwishart* methodology, the metric refers to the time required in fitting a model with a single shrinkage term λ.

- *Functional networks*, representing the synchronization in brain activity between different areas of the brain, obtained by computing the correlation of dynamic functional activity over time for the 70 brain regions.

Additionally, both subject-specific and region-specific covariates are available for each statistical unit, aiding in the interpretation of the results.

In this work, we primarily focus on *Functional Networks*, where information is encoded in two arrays, each corresponding to a different scan of the same set of subjects. Both arrays have dimensions of $70 \times 70 \times 24$, with the first two representing the 70 brain regions defined by the Desikan atlas and the third indicating the sample size. Given the way functional networks are constructed, they essentially encode information about how the activities of different brain regions are connected. Consequently, the goal of the analysis is to apply the developed methodology to cluster subjects based on the information contained in their functional networks. This approach not only facilitates the identification of distinct groups of subjects based on their brain activity patterns but also leverages the proposed penalized estimation scheme to highlight the connections between regions that are most important for distinguishing different clusters. When combined with subject-specific information, this approach may provide insights into brain activity mechanisms and their relationships with factors such as age, neurological diseases, or prior diagnoses. Lastly, it is important to note that this application represents clustering in the truest sense, as no ground-truth labels are available for direct comparison. Therefore, the clustering results will be interpreted and evaluated alongside the existing information about brain regions and subjects.

After conducting preliminary analyses, we decided to focus exclusively on the first scan, as the second scan was available for only a limited number of subjects. Additionally, two subjects were excluded due to the unavailability of their first-scan functional networks. Furthermore, two brain regions labeled as unknown were discarded, as their inclusion would hinder meaningful post-hoc interpretation of the results. Consequently, the analysis was conducted on a sample of correlation matrices $\{\mathbf{\Gamma}_i\}_{i=1}^n$, where $n = 22$ and $\mathbf{\Gamma}_i \in \mathbb{R}^{p \times p}$ with $p = 68$. The *Sparsemixwishart* procedure described in Section 3.1 was applied for $K \in \{2, 3, 4\}$, with the shrinkage parameter λ varying over a predefined grid of values. Note that $\lambda = 0$ is included in the grid to allow comparison with the work of Hidot and Saint-Jean (2010).

Initially, $\mathbf{P}$ was fixed to an all-one matrix with zero entries on the diagonal. In this scenario, the BIC in Equation (9) selects a model with $K = 2$ components and a λ value equal to 194, suggesting that our approach, which induces sparsity in the matrices $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, is appropriate in this context. Fig. 5 presents the estimated scale matrices $\hat{\mathbf{\Sigma}}_1$ and $\hat{\mathbf{\Sigma}}_2$, with the resulting sparsity pattern clearly highlighted. We immediately notice that both matrices exhibit an alternating-blocks structure that mirrors the neuro-anatomic division of the brain into two hemispheres, with regions within the same hemisphere being more connected to each other than to regions in the opposite hemisphere. Secondly, despite the overall low level of estimated sparsity, the first scale matrix has more than three times as many entries shrunk to zero compared to the scale matrix of the second component. From a clustering perspective, this results in a partition of subjects into two groups, with sizes of 12 and 10, respectively. Analyzing additional subject information reveals significant differences between the groups. The average ages of the two clusters are 40 and 26.5, respectively. Additionally, the first group is predominantly right-handed, while the second group has a higher proportion of left-handed and ambidextrous individuals. Notably, the first group also has a higher number of subjects with a current diagnosis of mental disorder compared to the second group. Specifically, the diagnoses for the first group are mainly related to major depressive status.

Alongside the analysis described in the previous paragraph, we applied the proposed methodology to the same data using a data-driven parameterization of the matrix $\mathbf{P}$ instead. As mentioned in Section 3.1, clever specifications of this matrix allow us to include additional available information on the phenomenon under study, thus enhancing the flexibility of the model. In the considered context, we leverage the knowledge of structural anatomic interconnections contained in the *structural networks* dataset. In neuroscience, it is often assumed that brain activity reflects the underlying structural network and the connections of white matter fibers (Rykhlevskaia et al., 2008). Based on this rationale, we devise a heuristic strategy to specify $\mathbf{P}$, ensuring that correlations between brain regions with fewer connections in terms of white matter fibers are more heavily penalized. Specifically, we first

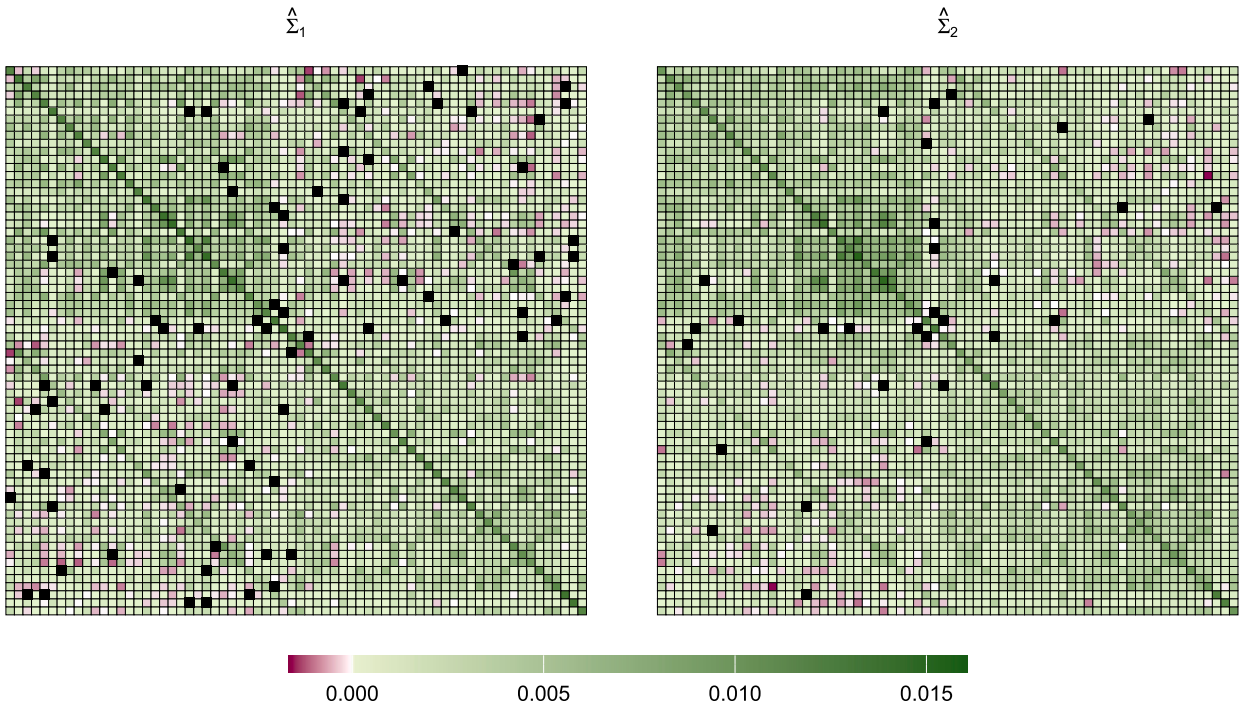

Fig. 5. fMRI functional networks data. Estimated scale matrices $\hat{\mathbf{\Sigma}}_1$ and $\hat{\mathbf{\Sigma}}_2$ for the *Sparsemixwishart* model, with $\mathbf{P}$ set to an all-ones matrix with zeros on the diagonal. Colors represent the values of the entries, with a 0 entry denoted by the symbol ■.

compute the matrix $\mathbf{D}$, which encodes the average white matter fiber counts between brain regions across all subjects. Subsequently, we categorize these connections among brain regions into five groups based on their anatomical connectivity strength: *zero, low, mild, moderate* and *high connectivity*. Finally, we construct the matrix $\mathbf{P}$ by assigning values ranging from 1 to 0.2 to its entries, reflecting this connectivity-based grouping. This specific assignment in $\mathbf{P}$ might seem somewhat arbitrary. However, within our framework, the hyperparameter λ scales the contribution of $\mathbf{P}$, thus calibrating the impact of the magnitude of its entries. Also in this scenario, the BIC selects a model with $K = 2$ components and a positive shrinkage factor λ equal to 204, providing further indications about the soundness of the proposal when compared to the one by Hidot and Saint-Jean (2010). The estimates $\hat{\mathbf{\Sigma}}_1$ and $\hat{\mathbf{\Sigma}}_2$ are reported in Fig. 6. Compared to using $\mathbf{P}$ as an all-ones matrix with zeros on the diagonal, the data-driven specification of $\mathbf{P}$ yields a higher overall level of estimated sparsity. Although $\hat{\mathbf{\Sigma}}_1$ remains sparser than $\hat{\mathbf{\Sigma}}_2$, the difference in sparsity magnitude is smaller compared to the previous scenario, with the separation between the right and left hemispheres remaining clearly visible. From a clustering perspective, incorporating information from structural networks does not alter the resulting partition. In fact, the resulting clustering exactly matches the previously discussed solution, preserving the meaningful distinctions between the two groups in terms of age, handedness, and mental disorder diagnoses.

The proposed approach has shown promising results in effectively distinguishing subjects based on their functional networks, while also integrating external information on structural anatomical interconnections. However, given the very limited sample size, we acknowledge that the analysis presented above is primarily illustrative. Further investigation, supported by larger datasets and deeper domain expertise, is necessary to better understand the determinants of specific mental disorder diagnoses and their relationship to brain activity and the co-regulation of distinct brain regions.

6. Discussion and conclusion

As the complexity of routinely collected data continues to grow, effectively analyzing covariance matrices is becoming increasingly important. These objects play a crucial role in capturing intricate linear relationships among variables, making them essential in a variety of fields like finance, genomics, and neuroscience. When it comes to clustering, both distance-based and model-based methods can encounter challenges in high-dimensional settings, particularly when the sample size is limited.

To address this limitation, in the present paper we have introduced a novel sparse Wishart mixture model. By resorting to a penalized likelihood approach with a covariance graphical lasso penalty, we have enforced sparsity in the component-specific scale matrices. This solution not only reduces the number of parameters that need to be estimated but also enhances interpretability by shrinking negligible relationships among variables to zero. A tailored EM-algorithm to maximize the objective function has been developed, and specific strategies for initialization, convergence, and model selection have been proposed.

A simulation study has been conducted to assess the effectiveness of our proposed method in clustering high-dimensional and sparse covariance objects. The results have demonstrated that our novel procedure outperforms state-of-the-art competitors in accurately

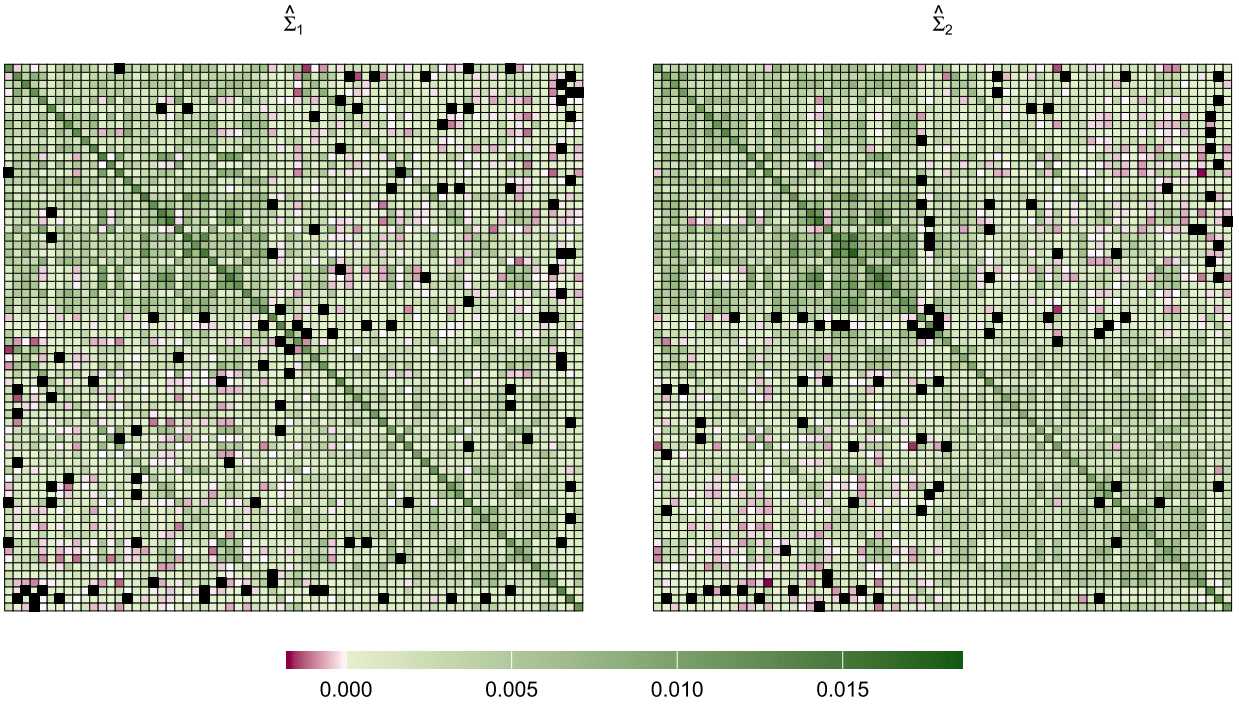

Fig. 6. fMRI functional networks data. Estimated scale matrices $\hat{\boldsymbol{\Sigma}}_1$ and $\hat{\boldsymbol{\Sigma}}_2$ for the *Sparsemixwishart* model, with a data-driven specification of $\mathbf{P}$ reflecting the structural anatomic interconnections among brain regions. Colors represent the values of the entries, with a 0 entry denoted by the symbol ■.

detecting the underlying partition of the data. In addition, our approach has successfully recovered the true sparsity patterns within the component-specific scale matrices.

Finally, we have presented a pilot application in which our procedure has been employed to cluster subjects based on their fMRI functional networks. The resulting groups have revealed distinct characteristics with respect to age, handedness, and mental disorder diagnoses. These encouraging findings, achieved in a complex real-world setting, highlight the potential of our method to contribute meaningfully to neuroscience by offering clearer insights into functional brain connectivity and its associations with individual differences and clinical conditions.

Lastly, it is worth noting that sparse estimation of covariance matrices has received considerable attention in recent years, especially in high-dimensional settings (Pourahmadi, 2013; Lam, 2020). In this work, we have focused on the covariance graphical lasso due to its seamless integration with the Wishart mixture model framework. However, exploring different penalization strategies or alternatives with different rationales such as thresholding (Bickel and Levina, 2008; Cai and Liu, 2011) or banding and tapering (Bien et al., 2016; Bien, 2019) could further improve our methodology, as they represent promising avenues for investigation beyond the covariance graphical lasso approach. Additionally, future research could explore the incorporation of more flexible distributions for positive definite symmetric matrices. Potential candidates include the Riesz distribution (Hassairi and Lajmi, 2001), the inverse Riesz distribution (Tounsi and Zine, 2012), and the F-Riesz distribution (Blasques et al., 2021). These distributions offer alternative approaches to modeling covariance objects, which may enhance the adaptability and performance of clustering methods in complex and high-dimensional settings. Some of these options are currently under examination, and their potential contributions will be addressed in future research.

Acknowledgements

We are grateful to the two anonymous reviewers, the Editor, and the Associate Editor for their careful evaluation of this manuscript and their helpful recommendations. Their contributions have substantially strengthened this work. We acknowledge Greg Kiar and Eric Bridgeford from NeuroData at Johns Hopkins University, who pre-processed the raw DTI and R-fMRI imaging data discussed in Section 5. We also express our gratitude to Elisa Borrini, whose master's thesis contains a preliminary version of this work.

References

Andrews, J.L., McNicholas, P.D., Subedi, S., 2011. Model-based classification via mixtures of multivariate t-distributions. Comput. Stat. Data Anal. 55 (1), 520–529.
Arsigny, V., Fillard, P., Pennec, X., Ayache, N., 2007. Geometric means in a novel vector space structure on symmetric positive-definite matrices. SIAM J. Matrix Anal. Appl. 29 (1), 328–347.
Bickel, P.J., Levina, E., 2008. Covariance regularization by thresholding. Ann. Stat. 36 (6), 2577–2604.
Bien, J., 2019. Graph-guided banding of the covariance matrix. J. Am. Stat. Assoc. 114 (526), 782–792.

Bien, J., Bunea, F., Xiao, L., 2016. Convex banding of the covariance matrix. J. Am. Stat. Assoc. 111 (514), 834–845.
Bien, J., Tibshirani, R.J., 2011. Sparse estimation of a covariance matrix. Biometrika 98 (4), 807–820.
Blasques, F., Lucas, A., Opschoor, A., Rossini, L., 2021. Tail heterogeneity for dynamic covariance-matrix-valued random variables: the F-Riesz distribution. Tinbergen Institute Discussion Paper 2021-010/III.
Bouveyron, C., Celeux, G., Murphy, T.B., Raftery, A.E., 2019. Model-Based Clustering and Classification for Data Science: with Applications in R. Cambridge University Press.
Bouveyron, C., Côme, E., Jacques, J., 2015. The discriminative functional mixture model for a comparative analysis of bike sharing systems. Ann. Appl. Stat. 9 (4), 1726–1760.
Bouveyron, C., Jacques, J., 2011. Model-based clustering of time series in group-specific functional subspaces. Adv. Data Anal. Classif. 5, 281–300.
Browne, R.P., McNicholas, P.D., 2015. A mixture of generalized hyperbolic distributions. Can. J. Stat. 43 (2), 176–198.
Cabassi, A., Casa, A., Fontana, M., Russo, M., Farcomeni, A., 2018. Three testing perspectives on connectome data. In: Studies in Neural Data Science: StartUp Research 2017. Siena, Italy, June 25–27. Springer, pp. 37–55.
Cai, T., Liu, W., 2011. Adaptive thresholding for sparse covariance matrix estimation. J. Am. Stat. Assoc. 106 (494), 672–684.
Cappozzo, A., Ferraccioli, F., Stefanucci, M., Secchi, P., 2018. An object oriented approach to multimodal imaging data in neuroscience. In: Studies in Neural Data Science: StartUp Research 2017. Siena, Italy, June 25–27. Springer, pp. 57–73.
Casa, A., Cappozzo, A., Fop, M., 2022. Group-wise shrinkage estimation in penalized model-based clustering. J. Classif. 39 (3), 648–674.
Casa, A., Ferrari, D., Huang, Z., 2024. High-dimensional covariance estimation by pairwise likelihood truncation. arXiv preprint arXiv:2407.07717.
Chaudhuri, S., Drton, M., Richardson, T.S., 2007. Estimation of a covariance matrix with zeros. Biometrika 94 (1), 199–216.
Cibinel, L., Roverato, A., Vinciotti, V., 2024. A unified approach to penalized likelihood estimation of covariance matrices in high dimensions. arXiv preprint arXiv:2410.02403.
De la Cruz-Mesía, R., Quintana, F.A., Marshall, G., 2008. Model-based clustering for longitudinal data. Comput. Stat. Data Anal. 52 (3), 1441–1457.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. Ser. B 39 (1), 1–22.
Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. NeuroImage 31 (3), 968–980.
Dryden, I.L., Koloydenko, A., Zhou, D., 2009. Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. Ann. Appl. Stat. 3 (3), 1102–1123.
Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Stat. 32, 407–499.
Erdős, P., Rényi, A., 1960. On the evolution of random graphs. Publ. Math. Inst. Hung. Acad. Sci. 5 (1), 17–60.
Fop, M., 2021. covglasso: Sparse Covariance Matrix Estimation. R package version 2.0.
Fop, M., Murphy, T.B., Scrucca, L., 2019. Model-based clustering with sparse covariance matrices. Stat. Comput. 29 (4), 791–819.
Fraley, C., Raftery, A.E., 2002. Model-based clustering, discriminant analysis, and density estimation. J. Am. Stat. Assoc. 97 (458), 611–631.
Gollini, I., Murphy, T.B., 2014. Mixture of latent trait analyzers for model-based clustering of categorical data. Stat. Comput. 24, 569–588.
Handcock, M.S., Raftery, A.E., Tantrum, J.M., 2007. Model-based clustering for social networks. J. R. Stat. Soc. A 170 (2), 301–354.
Hassairi, A., Lajmi, S., 2001. Riesz exponential families on symmetric cones. J. Theor. Probab. 14 (4), 927–948.
Hidot, S., Saint-Jean, C., 2010. An expectation–maximization algorithm for the Wishart mixture model: application to movement clustering. Pattern Recognit. Lett. 31 (14), 2318–2324.
Hubert, L., Arabie, P., 1985. Comparing partitions. J. Classif. 2, 193–218.
Keribin, C., 2000. Consistent estimation of the order of mixture models. Sankhya, Ser. A, 49–66.
Lam, C., 2020. High-dimensional covariance matrix estimation. Wiley interdisciplinary reviews. Comput. Stat. 12 (2), e1485.
Lin, T.I., 2009. Maximum likelihood estimation for multivariate skew normal mixture models. J. Multivar. Anal. 100 (2), 257–265.
Lin, T.-I., 2010. Robust mixture modeling using multivariate skew t distributions. Stat. Comput. 20, 343–356.
McLachlan, G.J., Peel, D., 1998. Robust cluster analysis via mixtures of multivariate t-distributions. In: Advances in Pattern Recognition: Joint IAPR International Workshops SSPR'98 and SPR'98. Sydney, Australia, August 11–13. Proceedings. Springer, 1998, pp. 658–666.
McNicholas, P.D., 2016. Mixture Model-Based Classification. Chapman and Hall/CRC.
McNicholas, P.D., Murphy, T.B., 2010. Model-based clustering of longitudinal data. Can. J. Stat. 38 (1), 153–168.
Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., 2023. e1071: Misc Functions of the Department of Statistics, Probability Theory Group. (Formerly: E1071), TU Wien. R package version 1.7-14.
Murtagh, F., Legendre, P., 2014. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? J. Classif. 31 (3), 274–295.
Pan, W., Shen, X., 2007. Penalized model-based clustering with application to variable selection. J. Mach. Learn. Res. 8 (5).
Pourahmadi, M., 2013. High-Dimensional Covariance Estimation. John Wiley & Sons.
R Core Team, 2024. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Rykhlevskaia, E., Gratton, G., Fabiani, M., 2008. Combining structural and functional neuroimaging data for studying brain connectivity: a review. Psychophysiology 45 (2), 173–187.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6 (2), 461–464.
Snijders, T.A., Nowicki, K., 1997. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. J. Classif. 14 (1), 75–100.
Tomarchio, S.D., Bagnato, L., Punzo, A., 2024. Model-based clustering using a new multivariate skew distribution. Adv. Data Anal. Classif. 18 (1), 61–83.
Tomarchio, S.D., Punzo, A., 2025. On the number of components for matrix-variate mixtures: a comparison among information criteria. Int. Stat. Rev.
Tounsi, M., Zine, R., 2012. The inverse Riesz probability distribution on symmetric matrices. J. Multivar. Anal. 111, 174–182.
Viroli, C., 2011. Finite mixtures of matrix normal distributions for classifying three-way data. Stat. Comput. 21, 511–522.
Wang, H., 2014. Coordinate descent algorithm for covariance graphical lasso. Stat. Comput. 24, 521–529.
Warton, D.I., 2008. Penalized normal likelihood and ridge regularization of correlation and covariance matrices. J. Am. Stat. Assoc. 103 (481), 340–349.
Yuan, M., Lin, Y., 2007. Model selection and estimation in the Gaussian graphical model. Biometrika 94 (1), 19–35.
Zhou, H., Pan, W., Shen, X., 2009. Penalized model-based clustering with unconstrained covariance matrices. Electron. J. Stat. 3, 1473.
Zou, H., Hastie, T., Tibshirani, R., 2007. On the "degrees of freedom" of the lasso. Ann. Stat. 35 (5), 2173–2192.