POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

# A Model-Based Approach to Cluster Correlation Matrices from fMRI Signals via a Mixture of Sparse Wishart Distributions

Tesi di Laurea Magistrale in
Mathematical Engineering - Ingegneria Matematica

**Elisa Borrini, 993554**

**Advisor:**
Prof. Chiara Masci

**Co-advisors:**
Andrea Cappozzo
Alessandro Casa

**Academic year:**
2023-2024

**Abstract:** Functional Magnetic Resonance Imaging (fMRI) has revolutionized our ability to observe the human brain in action, capturing intricate spatiotemporal patterns of brain activity through rich, high-dimensional data. An essential challenge in fMRI analysis is unraveling the inherent structures within these complex datasets. This thesis is dedicated to addressing this critical need by exploring an innovative method for grouping patients based on the shared covariance structures inherent in their fMRI signals. Facing the challenge of high-dimensionality in fMRI data, this work introduces a novel framework: at the heart of this methodology lies the use of penalized estimation techniques, specifically using a mixture model of sparse Wishart distributions, to efficiently address the curse of dimensionality by promoting sparsity in covariance matrix estimation.
This thesis introduces a model-based approach that builds on penalized estimations to refine statistical analysis, specifically through developing a sparse covariance matrices framework. This advancement allows for clearer insights into complex data patterns and significantly improves both the interpretability of models and their computational execution.
By applying the proposed clustering framework to fMRI data, the study identifies distinct groups of subjects based on their brain activity patterns. The analysis reveals significant correlations between the derived clusters and subject-specific characteristics.
This paper bridges the gap between innovative statistical methodologies and their application in neuroscience research. By integrating model-based clustering with sparse covariance matrix estimation, it provides a powerful tool for dissecting the complex web of brain connectivity, paving the way for enhanced understanding and diagnosis of neurological disorders.

**Key-words:** Model-based clustering, Penalized likelihood, Sparse covariance matrices, Wishart mixture models, E-M algorithm

## 1. Introduction

The advent of functional magnetic resonance imaging (fMRI) has significantly enriched neuroscience research, offering unparalleled insights into brain activity and connectivity. This non-invasive imaging technology, along

with other methods like electroencephalography (EEG) and diffusion tensor imaging (DTI), has unveiled complex patterns of dynamic brain activity and anatomical connectivity. These high-dimensional datasets, characterized by a greater number of variables than observations, present a unique challenge, necessitating advanced statistical methodologies to accurately interpret the intricate relationships and dependencies within the brain. Novel statistical approaches are essential for dissecting these complex datasets, allowing researchers to explore how brain activity and connectivity vary across individuals and correlate with different traits.
In this context, clustering techniques emerge as a logical strategy to discern patterns among patients by examining correlation matrices derived from fMRI data. Such analyses can highlight potential relationships between neuronal areas, contributing to our understanding of how the brain's functional architecture varies across individuals and conditions. Model-based clustering presents a powerful approach for analyzing brain activity and connectivity patterns within high-dimensional datasets, assuming that observed data derives from a mixture of distinct probability distributions corresponding to unique clusters.
In the specific framework of this study, our focus is directed towards employing a mixture of Wishart distributions for modeling covariance matrices, an approach suggested by the work of Hidot and Saint-Jean ([16]), leveraging the Expectation-Maximization (E-M) algorithm for parameter estimation. However, their method encounters significant obstacles in high-dimensional scenarios due to the curse of dimensionality, where the exponential growth in the number of estimable parameters relative to data dimensionality leads to computational inefficiencies and the risk of model overfitting. The challenge is particularly pronounced in estimating covariance matrices, essential for defining and distinguishing clusters based on the variability and correlation structure among variables.
To mitigate these issues, sparsity has been identified as a crucial strategy within statistical analysis. High-dimensional data often reveal a sparse structure, suggesting that many variables contribute minimally to the overall pattern, thereby allowing for a focus on a smaller, more impactful subset of parameters. Penalized likelihood estimation, which introduces a penalty term to the likelihood function, emerges as a key method to promote sparsity, reducing model complexity while ensuring data fit. This approach, becoming more popular in many statistical fields, effectively reduces the tendency to create overly complex models.
Addressing the specific challenges in estimating high-dimensional covariance matrices in a model-based clustering framework, sparsity-inducing techniques like the Graphical Lasso (Glasso) and its extension, the Covariance Graphical Lasso, have been developed. These methods simplify estimating parameters by promoting a sparse approach, making it easier to understand and compute clustering for complex data. Glasso, introduced by Friedman et al. in [12], aims to estimate sparse inverse covariance matrices in Gaussian graphical models. By imposing an $L_1$ penalty on the elements of the precision matrix, Glasso ensures that many off-diagonal elements are exactly zero, leading to a sparse and interpretable representation of the conditional dependence structure among variables.
Expanding upon the concept of Glasso, the Covariance Graphical Lasso seeks to estimate sparse covariance matrices directly. This approach is particularly relevant when the goal is to understand the variability and correlation structure among variables in high-dimensional frameworks. By employing a similar $L_1$ penalization on the covariance matrix, this method facilitates the identification of significant correlations while reducing the complexity and enhancing the interpretability of the model (see [2] and [25]).
By incorporating sparsity, model-based clustering becomes adaptable to the complexities of high-dimensional data, enhancing result interpretability. This adjustment aligns with the broader trend in statistical methodologies, addressing the demands of increasingly complex and high-dimensional datasets.
Our method has demonstrated efficiency during the simulation phase, providing robust estimates. In its application to correlation matrices from fMRI, despite the limited number of available observations, the method showcased its capability to produce plausible estimates. It successfully highlighted the presence of two distinct clusters, revealing differences among the patients in the study, such as variations in age and the presence or absence of mental disorders. This outcome underscores the method's effectiveness in uncovering meaningful patterns and clustering characteristics within the data, even in scenarios with a restricted number of observations.
The rest of the thesis is structured as follows. Section 2 delves into previous work and background knowledge used to develop the study. In Section 3 the proposal is presented. In Sections 4 the performances and the applicability of the proposed approach are tested on simulated data. In Section 5, the procedure is applied on real data, the results are presented and analyzed. Lastly, the paper concludes in Section 6 with a brief discussion, drawing conclusions, and suggesting possible future directions for progress.

## 2. Preliminaries and related works

Model-based clustering is a statistical approach that groups observations into distinct clusters based on probabilistic models. The fundamental idea is to assume that data are generated from a mixture of probability distributions, each associated with a particular cluster. The likelihood of multivariate independent observations

$x_1, \ldots, x_N$, with $x_i \in \mathbb{R}^p$, for a model with $g$ components, is defined as:

$$L_{mix}(\Psi_1, \ldots, \Psi_g; \beta_1, \ldots, \beta_g | x) = \prod_{i=1}^{N} \sum_{k=1}^{g} \beta_k f_k(x_i | \Psi_k) \tag{1}$$

where $f_k$ is the k-th component density, $\Psi_k$ is the vector of the parameters of the $k$-th component of the mixture, and $\beta_1, \ldots, \beta_g$ are the mixture proportions. Typically, $f_k$ is the multivariate Gaussian with $\Psi_k = (\mu_k, \Sigma_k)$, for k = 1, ..., g.
Recently, there has been a growing interest in studying extensions beyond the Gaussian model. More flexible approaches, such as the use of mixture of Student's t-distributions [20], Weibull distribution or Exponential distribution [8], asymmetric distributions [18] [24] and mixtures with matrix-variate distributions [13], have been explored. In the case in which data corresponds to covariance matrices, one can refer to a mixture of Wishart distributions. This choice is motivated by the need to capture the intricate structure of covariances between variables within each cluster. By incorporating the Wishart distribution, the model can effectively account for varying degrees of correlation and dispersion in different dimensions.
One of the major limitations of model-based clustering lies in their tendency to be over-parameterized in high-dimensional scenarios. The expression "high-dimensional clustering" denotes situations in which the number of variables is significantly higher than the available observations in the dataset. In such cases, traditional statistical model estimation confronts difficulties, leading to unstable estimates and overparameterized models that compromise the accuracy of statistical analyses.
In the context of this study, our primary objective is to address these limitations by sparsely estimating covariance matrices which are the parameters of a mixture of Wishart distributions within the E-M framework.

This study builds upon the foundations laid by a prior work of Hidot and Saint-Jean in [16], where they performed parameter estimation for a mixture of Wishart distributions, without considering sparsity in the covariance matrix parameters, the limitation addressed in our study. The method they proposed has the following structure.

Let $x = (x_1, \ldots, x_N)$ be such that $x_i \sim \mathcal{N}_p(\mu, \Sigma)$. Then $\Gamma = x^t x$ follows a central Wishart distribution with covariance matrix $\Sigma$ and degrees of freedom $\nu$, with $\nu = N$, denoted as $\Gamma \sim \mathcal{W}_p(\Sigma, \nu)$. Let $\mathbf{\Gamma} = (\Gamma_1, \ldots, \Gamma_N)$ be a set of $N$ real squared matrices with order $p$.
It is assumed that each $\Gamma_i$ arises from a mixture of $g$ Wishart distribution with parameters $\Sigma_k$ and $\nu_k$, for k in 1, ..., g. The corresponding density function for each $\Gamma_i$ is:

$$f(\Gamma_i, \Psi) = \sum_{k=1}^{g} \beta_k f_W(\Gamma_i; \Psi_k) \tag{2}$$

where $\Psi_k = (\Sigma_k, \nu_k)$ is the vector of the parameters of the k-th component of the mixture and $f_W(\Gamma_i; \Psi_k)$ is the density function of a Wishart distribution.
The parameters' estimation is computed using an E-M algorithm, a powerful iterative optimization method. The E-M algorithm is an optimization technique commonly employed for parameter estimation in statistical models. It is particularly useful when dealing with incomplete or latent data. Latent variables are introduced to simplify the expression of the likelihood function and to make the optimization more tractable. These variables capture the hidden information in the data that are not directly observable.
Fraley and Raftery in [11] exhaustively explain how E-M algorithm is applied to estimate finite mixture models. In the framework presented by Hidot and Saint-Jean ([16]), consider $N$ observations $t_i = (\Gamma_i, z_i)$ being the complete data, with $\Gamma_i$ being the observed one as defined before, and $z_i$ the missing data, where $z_i = (z_{i1}, \ldots, z_{ig})$ and $z_{ik} = 1$ if the observation $\Gamma_i$ belongs to the k-th cluster. In the case of mixture models, $z_i = (z_{i1}, \ldots, z_{ig})$ is *iid* according to a multinomial distribution on the $g$ groups with probabilities $\beta_1, \ldots, \beta_g$. Therefore, the complete-data log-likelihood is:

$$L(\Psi_k, \beta_k, z_{ik} \mid t) = \sum_{i=1}^{N} \sum_{k=1}^{g} z_{ik} \log[\beta_k f_W(\Gamma_i \mid \Psi_k)]$$

The E-M algorithm operates in two steps as follows. The E-step involves estimating the expected value of the latent variables given the current parameters. In particular, at iteration $t$, a matrix $Z$ with dimensions $(N, g)$ is updated according to the formula:

$$z_{ik}^{(t)} = \beta_k^{(t)} \cdot \frac{f_w(\Gamma_i, \Sigma_k^{(t)}, \nu_k^{(t)})}{\sum_{l=1}^{g} \beta_l^{(t)} \cdot f_w(\Gamma_i, \Sigma_l^{(t)}, \nu_l^{(t)})} \tag{3}$$

The matrix Z is composed by N g-dimensional vectors $z_i$ , where $z_{ik}$ is a scalar between 0 and 1 representing the probability that $\Gamma_i$ arises from the k-th component.
In the M-step, the algorithm maximizes the complete-data log-likelihood by updating the model parameters $\Psi_k$ and the mixture proportions $\beta_k$ based on the observed data and the expected values of the latent variables obtained from the E-step, with respect to the following formulas:

$$\beta_k^{(t)} = \frac{1}{N} \sum_{i=1}^{N} z_{ik}^{(t-1)} \tag{4}$$

$$\sum_{i=1}^{N} z_{ik}^{(t-1)} \log \left( \frac{|\Gamma_i (\Sigma_k^{(t-1)})^{-1}|}{2} \right) = \sum_{i=1}^{N} z_{ik}^{(t-1)} \sum_{j=1}^{p} \psi \left( \frac{1}{2} (\nu_k^{(t)} - j + 1) \right) \tag{5}$$

$$\Sigma_k^{(t)} = \frac{\sum_{i=1}^{N} z_{ik}^{(t-1)} \Gamma_i}{\sum_{j=1}^{N} z_{jk}^{(t-1)} \nu_k^{(t)}} \tag{6}$$

where the superscript $(t-1)$ indicates the corresponding values obtained at the previous iteration.

$\psi$ represents the digamma function:

$$\psi(x) = \frac{d}{dx} \log(\Gamma(x))$$

where $\Gamma(x)$ is the gamma function, defined for $x > 0$ as:

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$$

This method returns an estimate of the specific parameters of each Wishart distribution of the mixture, $\Psi_k = (\Sigma_k, \nu_k)$, with $k = 1, \ldots, g$, together with a partition P of the initial data $(\Gamma_1, \ldots, \Gamma_N)$ into g clusters. This last procedure is carried on by the maximum a posteriori (MAP) rule on the vectors $z_i$. The MAP rule looks for the parameter value that makes the observed data most likely, taking into account both the likelihood of the data under different parameter values and any prior beliefs about the parameters. This contrasts with the maximum likelihood estimation (MLE), which only considers the likelihood of the observed data without incorporating prior information. The MAP rule, therefore, provides a more comprehensive approach by integrating both prior knowledge and the evidence from the data to make statistical inferences or decisions.

Starting from this foundation, the primary focus and novelty of this paper lies in the sparse estimation of the covariance matrix parameter, $\Sigma_k$. This innovative approach enhances the interpretability of the results and addresses the challenges associated with high-dimensional data. Our contribution builds upon the existing framework, extending its applicability to scenarios where sparse covariance estimation is essential for an accurate representation of underlying patterns. Specifically, we focus on covariance objects coming from a mixture of sparse Wishart distributions, a specialized variant allowing for sparsity in covariance matrices.
As described in detail in the next section, retaining the core principles of Hidot and Saint-Jean, we developed a tailored approach to suit our specific needs.

## 3. Proposal

The approach by Hidot and Saint-Jean, as outlined in the previous section, encounters limitations in high-dimensional scenarios. In this framework, the challenge lies in the overwhelming number of free parameters to estimate and the complexity of interpreting dense covariance matrices. The quadratic scaling of parameters with variables leads to computational burdens and model overfitting. Dense covariance matrices, in particular, are hard to interpret because they suggest that every variable is related to every other variable, a scenario that is often unrealistic in practical applications. To address these issues, in this work, we propose a modification of the method by Hidot and Saint-Jean that is better suited to cluster high-dimensional covariance matrices. By doing so, we enhance both the interpretability and computational efficiency of the model, making it feasible to uncover the underlying structure in complex datasets where traditional methods vacillate. This modification not only simplifies the estimation process but also provides clearer insights into the data, allowing us to identify significant relationships between variables without being overwhelmed by unnecessary complexity.

## 3.1. Model Specification

Adapting the methodology proposed by Hidot and Saint-Jean ([16]) to accommodate a sparse estimate of $\Sigma_k$, considering the same scenario presented earlier, the objective function to maximize becomes a penalized likelihood:

$$L(\Psi; \mathbf{\Gamma}) = \prod_{i=1}^{N} \sum_{k=1}^{g} \beta_k f_W(\Gamma_i; \Psi_k) - Pen(\Sigma_k; \lambda) \tag{7}$$

Here, $L(\Psi; \mathbf{\Gamma})$ represents the penalized likelihood of the parameters $\Psi$ given the data $\mathbf{\Gamma}$, encapsulating our aim to balance data fit with model complexity. The first term denotes the likelihood contribution of observing the $i$-th data point $\Gamma_i$ under the $k$-th component of the mixture model, weighted by $\beta_k$. The term $Pen(\Sigma_k; \lambda)$ introduces a general penalty on the covariance matrices $\Sigma_k$, where $\lambda$ acts as a regularization parameter. By penalizing complexity, we aim to encourage simpler, more interpretable models.

The overall objective is to find the optimal parameters $\Psi_k = (\Sigma_k, \nu_k)$, in particular the sparse covariance matrices for each model k in the mixture, that maximize this penalized likelihood, thereby achieving a balance between model fit and sparsity. Our method aligns with the structure of Hidot and Saint-Jean's study presented in the previous section, leveraging the Expectation-Maximization (E-M) algorithm for iterative updates and maximizing the penalized likelihood. In the E-step, each component of the matrix $Z$, indicating the cluster membership of each observation, is updated as described in Equation (3). Additionally, in the M-step, the mixture proportions $\beta_k$ , for k in 1, . . . , g, are updated using the formula in Equation (4), ensuring a comprehensive adjustment of the model parameters at each iteration. However, departing from the method by Hidot and Saint-Jean, the model parameters $\Psi_k = (\Sigma_k, \nu_k)$ are updated following the strategy presented hereafter.

## 3.2. Covariance Graphical Lasso

The covariance graphical lasso is a statistical technique employed for estimating sparse covariance matrices in high-dimensional datasets. The primary goal of covariance graphical lasso is to identify and quantify the marginal dependencies among variables in multivariate data, emphasizing the underlying network structure. The method introduces sparsity by incorporating a regularization term that penalizes the absolute values of the elements in the covariance matrix. This penalization encourages many entries in the matrix to be exactly zero, effectively leading to a sparse representation.
A relevant study in this context is the one by Bien and Tibshirani described in [2]. The paper suggests a method for sparsely estimating a covariance matrix $\Sigma_k$, augmenting the likelihood with a lasso penalty applied to P * $\Sigma_k$, where P is an arbitrary matrix with non-negative elements and * denotes elementwise multiplication. The proposed estimator is the one satisfying:

$$\underset{\Sigma_k}{\operatorname{argmin}} \log(\det(\Sigma_k)) + \operatorname{trace}(\Sigma_k^{-1} S_k) + \lambda \|P * \Sigma_k\|_1 \tag{8}$$

where $S_k$ is the sample covariance matrix.
In order to perform the optimization, since the objective function is not convex, the paper presents a majorization-minimization approach that leads to the following:

$$\hat{\Sigma}_k^{(t)} = \underset{\Sigma_k}{\operatorname{argmin}} \left[ \operatorname{trace} \left[ \left( \hat{\Sigma}_k^{(t-1)} \right)^{-1} \Sigma_k \right] + \operatorname{trace} \left[ (\Sigma_k^{-1} S_k) \right] + \lambda \|P * \Sigma_k\|_1 \right] \tag{9}$$

where $\hat{\Sigma}_k^{(t-1)}$ represents the estimate at the previous iteration. This problem is convex and therefore any local minimum is guaranteed to be the global minimum. The authors propose a generalized gradient descent algorithm to seeks the matrix $\Sigma_k$ that minimizes the given objective function (8). Generalized gradient descent algorithms are optimization methods that minimize a cost function by improving upon traditional gradient descent. They tackle challenges like non-convexity and non-smoothness by incorporating advanced techniques or modifications.
In this case, the update equation of $\Sigma_k$ is given by:

$$\Sigma_k < -\mathcal{S}(\Sigma_k - t(\Sigma_{0k}^{-1} - \Sigma_k^{-1} S_k \Sigma_k^{-1}), \lambda t P) \tag{10}$$

where $\mathcal{S}$ denotes the soft-thresholding operator, commonly used in estimation or regularization contexts. It has the general form:

$$\mathcal{S}(x, \lambda) = \operatorname{sign}(x) \cdot \max(0, |x| - \lambda)$$

Where:

- x is the value undergoing soft-thresholding;
- $\lambda$ is the regularization or threshold parameter;
- sign(x) returns the sign of x;
- $\max(0, |x| - \lambda)$ is the "soft" part of thresholding;

In simple terms, the soft-thresholding operator sets values smaller than the threshold $\lambda$ to zero and reduces other values by an amount $\lambda$, while maintaining the original sign.

t represents a parameter that regulates the magnitude of the step taken during each iteration of the algorithm. The update equation iteratively modifies $\Sigma_k$ to approach the optimal solution by considering its current state $\Sigma_{0k}$, adjusting it based on the gradient information, and applying soft-thresholding to induce sparsity, as guided by the regularization term. In this case, the tuning parameter $\lambda$ is chosen by cross-validation.

The choice to adopt Covariance Graphical Lasso in the context of our methodology is motivated by considerations related to the nature of high-dimensional data and the specific goals of our analysis.
Lasso regularization is indeed particularly well-suited in this context because it introduces the aforementioned sparsity-inducing mechanism, promoting the presence of zeros in the coefficients associated with less relevant variables. In our case, the objective is to obtain robust and interpretable estimates of covariance parameters, allowing for an accurate representation of underlying patterns in high-dimensional data.

Our method integrates this technique with a specific focus on enhancing the model's performance in high-dimensional clustering scenarios. Central to our adaptation is the reformulation of the model's objective function (7) to account for the penalty term induced by the use of Covariance Graphical Lasso, which directly impacts the estimation of sparse covariance matrices. The revised objective function is expressed as follows:

$$L(\Psi; \mathbf{\Gamma}) = \sum_{i=1}^{N} \sum_{k=1}^{g} \beta_k f_W(\Gamma_i; \Sigma_k; \nu_k) - \sum_{k=1}^{g} (\lambda \dot{\|} \Sigma_k \|_1) \tag{11}$$

By subtracting the sum of the L1 norms of the covariance matrices, scaled by $\lambda$, we introduce a robust mechanism to prioritize sparsity in our model's estimation process, as well as control overfitting, facilitate variable selection, and provide flexibility in modeling complex relationships within high-dimensional data.

### 3.2.1 Estimating $\Sigma_k$ through Covariance Graphical Lasso

As mentioned earlier, the sparse estimation of $\Sigma_k$ is entrusted to the covariance graphical lasso. Specifically, for each $k$ in 1, ..., g, $\Sigma_k$ is estimated with the formula (8). In this case, P is the identity matrix. More precisely, in the context of this study:

- $S_k$ is the sample covariance matrix, calculated as the weighted sum of covariance matrices for the data points assigned to the specific cluster
- $n_k$ is the effective sample size used in the estimation, set to the product between the sum of probabilities assigned to the cluster $k$, and $\nu_k$, the degrees of freedom associated with the cluster

This process is performed iteratively by updating the degrees of freedom $\nu_k$ by numerically solving at each iteration $t$:

$$\sum_{i=1}^{N} z_{ik}^{(t-1)} \log \left( \frac{|\Gamma_i (\Sigma_k^{(t-1)})^{-1}|}{2} \right) = \sum_{i=1}^{N} z_{ik}^{(t-1)} \sum_{j=1}^{p} \psi \left( \frac{1}{2} (\nu_k^{(t)} - j + 1) \right) \tag{12}$$

and inserting the found root into the formula for $\Sigma_k$, all until convergence of $\Sigma_k$.
The convergence relies on the Frobenius norm of the difference $\Sigma_k^t - \Sigma_k^{t-1}$ for k=1, ..., g, where $t$ and $t-1$ represent two following iterations. It is required that the norm is less or equal to a given threshold $\epsilon_\Sigma$, with default value $10^{-6}$.
Given two square matrices S and R of dimension p×p, the Frobenius norm of their difference is defined as:

$$\|S - R\|_F = \sqrt{\sum_{i=1}^{p} \sum_{j=1}^{p} (S_{ij} - R_{ij})^2}$$

where $S_{ij}$ and $R_{ij}$ represent the elements in position (i,j) of matrix S and matrix R, respectively.
In the context of covariance matrices, the Frobenius norm can provide an overall measure of how different two matrices are in terms of all their elements. When this norm is small, it indicates that the two matrices are

similar, while a larger norm indicates a greater discrepancy between the matrices.
In general, the Frobenius norm is a convenient measure for evaluating the distance between matrices, and is often preferred when one wishes to consider all elements of the matrices in evaluating their similarity or discrepancy.

To provide a comprehensive overview of the methodology, we present the pseudo-algorithm for the entire process.

**Algorithm 1** Algorithm

This E-M algorithm clusters covariance matrices data using a mixture of Wishart distributions, iteratively optimizing cluster assignments and parameters for sparsity and likelihood maximization, concluding upon convergence with final cluster determinations and model parameters' estimates.

**Input**: $\mathbf{\Gamma}$, $g$ number of clusters, $p$ dimension of the $\Gamma$ matrices, matrix $Z$, $\nu$, shrinkage parameter $\lambda$
**repeat**
  E-step:
  $z_{ik}^{(t)} = \beta_k^{(t)} \cdot \frac{f_w(\Gamma_i, \Sigma_k^{(t)}, \nu_k^{(t)})}{\sum_{l=1}^{g} \beta_l^{(t)} \cdot f_w(\Gamma_i, \Sigma_l^{(t)}, \nu_l^{(t)})}$
  M-step:
  $\beta_k^{(t)} = \frac{1}{N} \sum_{i=1}^{N} z_{ik}^{(t-1)}$
  **repeat**
    $\nu_k^{(t)}$ by solving numerically
    $\sum_{i=1}^{N} z_{ik}^{(t-1)} \log\left(\frac{|\Gamma_i(\Sigma_k^{(t-1)})^{-1}|}{2}\right) = \sum_{i=1}^{N} z_{ik}^{(t-1)} \sum_{j=1}^{p} \psi\left(\frac{1}{2}(\nu_k^{(t)} - j + 1)\right)$
    $\Sigma_k^{(t)} = \underset{\Sigma_k}{\operatorname{argmin}} \ \log(\det(\Sigma_k^{(t-1)})) + \operatorname{trace}((\Sigma_k^{(t-1)})^{-1} S_k) + \lambda \|\Sigma_k^{(t-1)}\|_1$
  **until** convergence of $\Sigma_k$
**until** convergence
**Output**: $P(\Gamma_i) = \underset{k}{\operatorname{argmax}}\ z_{ik}^{(t)}$, $\mathbf{\Sigma}$, $\nu$, BIC, objective function

As mentioned before, P is a vector indicating to which cluster each observation is believed to belong to, created through the MAP rule.

### 3.3. A Note on Numerical Optimization of Degrees of Freedom $\nu_k$

Within the iterative optimization process, a crucial step involves numerically determining the degrees of freedom $\nu_k$ for a Wishart distribution. This parameter, representing the covariance matrix's degrees of freedom, is bounded by the dimensionality of the matrix ($p$). To solve equation (5), it is necessary to find an interval constructed in such a way that, at the two ends, the function has opposite signs. This ensures the existence of a root of the function in the given interval. To guarantee that $\nu_k$ is always equal to or greater than $p$, the dimension of the matrices, the interval must have $p$ as the lower bound. In situations where the function in (5) is negative at $\nu_k = p$, given the decreasing nature of this function as depicted in Figure 1, finding an interval satisfying this condition becomes impossible. To solve this issue, we decided to check whether function (5) is negative at $\nu_k = p$, and, if that is the case, we just approximate $\nu_k = p$. This handling of the edge case enhances the stability and efficiency of the algorithm, ensuring reliable convergence even under challenging conditions.

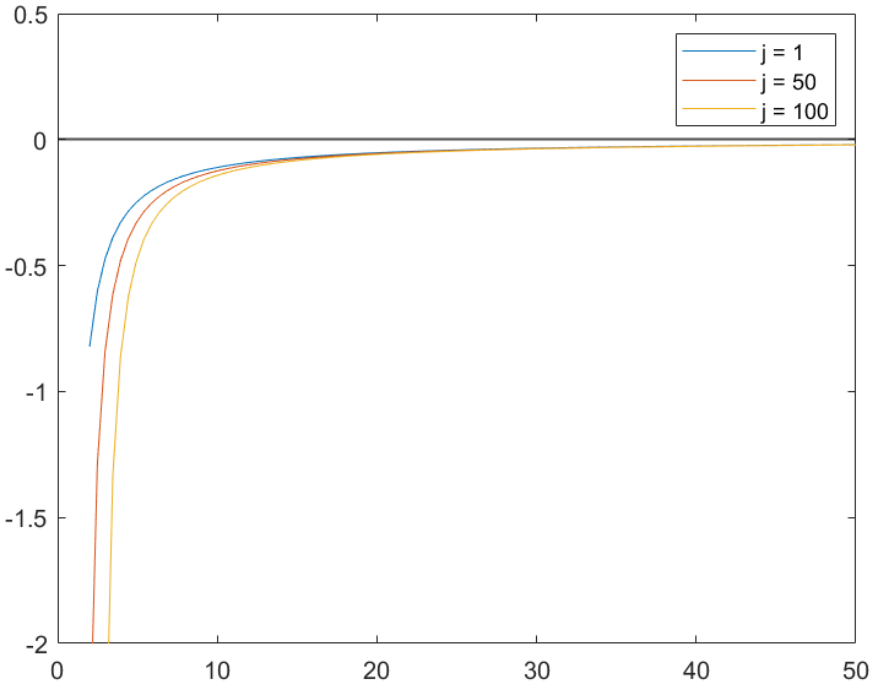


Figure 1: Plot of the derivative of the digamma ($\psi$) function in equation (5), since it is the only term that depends on $\nu_k$. As equation (5) comprises a finite sum of digamma terms, we plotted the derivative of a single term. The derivative of the sum will be the sum of the derivatives. Since the function is continuous and differentiable everywhere, a negative derivative implies that the function is decreasing.

## 3.4. Further Aspects

We address practical considerations regarding the algorithm designed to maximize equation (11), as discussed in the previous section.

### 3.4.1 Convergence

The E-M algorithm reaches convergence when the change in the penalized log-likelihood, calculated as the logarithm of the objective function in equation (11), between consecutive iterations falls below a predefined tolerance threshold. Specifically, convergence is achieved when the following condition holds:

$$\left|\log(L(\Gamma;\Psi^{(t)})) - \log(L(\Gamma;\Psi^{(t-1)}))\right| < \text{Tol}$$

Here, $\Psi^{(t)}$ and $\Psi^{(t-1)}$ denote the parameter estimates at the current and previous iterations, respectively. The algorithm stabilizes when the change in the penalized log-likelihood becomes sufficiently small, indicating that further iterations are not significantly improving the fit of the model. In our analysis, Tol is set to $10^{-6}$.

### 3.4.2 Model Selection

Exploring the significance of the parameter $\lambda$ in the covariance graphical lasso and its impact on the estimation process reveals its crucial role. The parameter $\lambda$ acts as a tuning parameter controlling the degree of sparsity induced in the estimated covariance matrices. A higher $\lambda$ leads to more aggressive shrinkage and, consequently, sparser covariance matrices.
Given a grid of different $\lambda$, model selection is performed using the Bayesian Information Criterion (BIC). The BIC is a criterion for model selection among a finite set of models, balancing goodness of fit and model complexity. It is defined as follows:

$$\text{BIC} = -2\log(\hat{L}(\Psi;\mathbf{\Gamma})) + d\log(N)$$

where $\hat{L}(\Psi; \mathbf{\Gamma})$ is the maximized likelihood of the model, $N$ is the sample size and $d$ is the number of parameters defined as:

$$d = g \cdot \left(2 + \frac{\text{NZ} \cdot (\text{NZ} + 1)}{2}\right) - 1 \tag{13}$$

where NZ is the number of non-zero entries of all the $\Sigma_k$ matrices, for k in 1, ..., g.
The model with the lowest BIC is considered the most suitable, striking a balance between goodness of fit and simplicity.

### 3.4.3 Implementation

The source **R** code containing all the routines and algorithms utilized for implementing the sparse covariance matrices estimation method can be accessed on the GitHub repository: https://github.com/elisab0/sparseWishartmix.git. The provided code contains the implementation of the core function utilizing the Expectation-Maximization algorithm, along with codes used to generate simulation results and conduct real dataset analysis. These details will be presented in the upcoming sections.

# 4. Simulation Study

## 4.1. Experimental Setup

In this section, we present a simulation study aimed at validating the effectiveness and efficiency of the proposed algorithm, accompanied by numerical results.
We generated a dataset consisting of $N = 300$ observations from a mixture of Wishart distributions with $g = 3$ components.

An essential aspect of our simulation involves the structure of $\Sigma_k$ for $k$ in $1, \ldots, g$, which are generated according to a sparse Erdős-Rényi graph model. An Erdős-Rényi graph, denoted as $G(n, \pi)$, is a random graph model comprising $n$ nodes, where each pair of nodes is connected independently with a fixed probability $\pi$. This model serves as the basis for introducing sparsity in $\Sigma_k$: when $\pi$ is small, the graph and consequently the covariance matrix exhibit corresponding sparsity, with fewer edges between nodes.
To explore various scenarios, we considered two distinct settings for the sparseness of the $\Sigma_k$:

- $\Sigma_k$ with the same levels of sparsity, where the probability $\pi$ is equal to 0.5 for all $\Sigma_k$.
- $\Sigma_k$ with different levels of sparsity, where the probabilities $\pi$ are set to 0.6, 0.4, and 0.2 for the corresponding $\Sigma_k$.

A visual example of the different settings is provided in Figures 2 and 3.

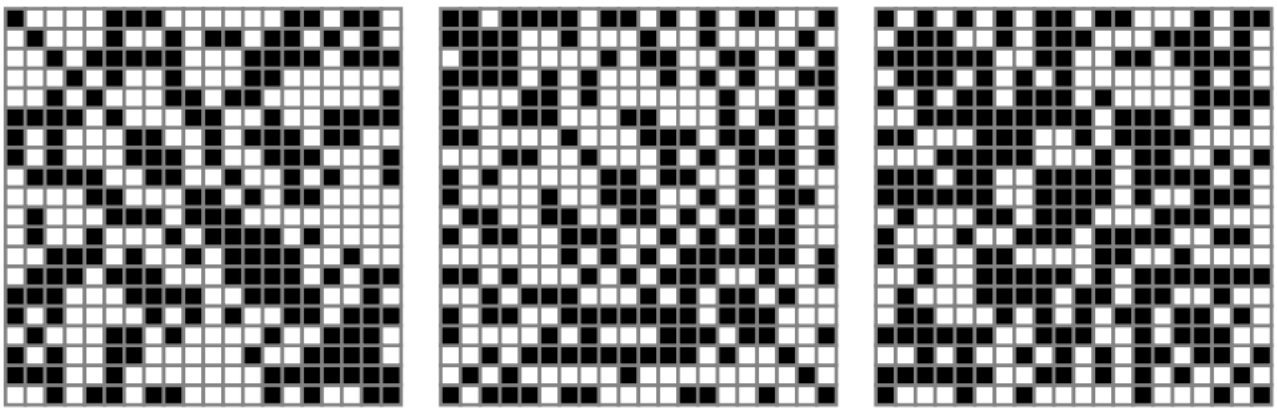

Figure 2: Example of simulation setting Same level of sparsity in $\Sigma_k$. Black squares denote values different from zero

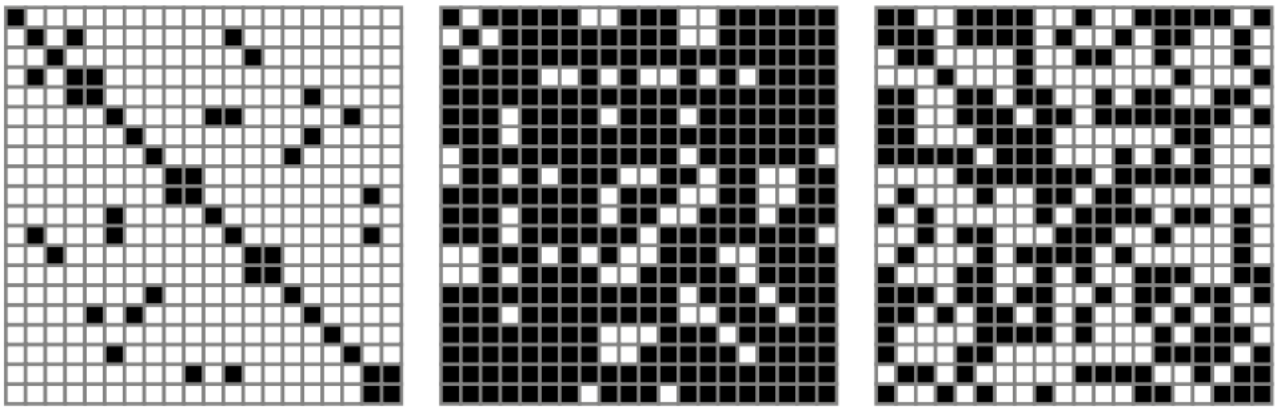

Figure 3: Example of simulation setting Different level of sparsity in $\Sigma_k$. Black squares denote values different from zero

For all of the scenarios, p is equal to 20, the degrees of freedom vector $\nu$ is set to $(20, 30, 40)$ and the mixture proportions $\beta_k$ are equal to $\frac{1}{3}$ for every k in $1, \ldots, g$. The matrix Z is initialized by randomly assign a mistaken cluster to some observations (approximately half of the total number of observations), in order to test the ability of the algorithm to recover the true partition.

The experiment is repeated B = 50 times.

## 4.2. Evaluation Metrics

First of all, we considered a grid of shrinkage values that has been built in such a way that the matrices $\Sigma_k$ are almost full for the lowest value of $\lambda$, while they are almost diagonal for the highest value.
The best model is chosen according to the Bayesian Information Criterion (BIC).

At the end of this process, for every simulation in $1, \ldots,$ B, four different evaluation metrics are computed: F1 Score, ARI, MSE for $\Sigma_k$ and MSE for $\nu_k$, for k in 1,...,g.

**F1 Score** is calculated as:

$$F1\_score = \frac{\text{tp}}{\text{tp} + 0.5 \cdot (\text{fp} + \text{fn})} \tag{14}$$

where:

- $tp$: The number of correctly identified edges (non-zero entries in the covariance matrix correctly identified as such).
- $fp$: The number of incorrectly identified edges (zero entries in the true covariance matrix identified as non-zero).
- $fn$: The number of missed edges (non-zero entries in the true covariance matrix that were incorrectly identified as zero).

It ranges from 0 to 1, with higher values indicating better performance.The F1-score is the harmonic mean of precision and recall.

- Precision measures the accuracy of positive predictions by calculating the ratio of true positive predictions to the total number of positive predictions. The formula for precision is given by: $Precision = \frac{tp}{tp+fp}$.
- Recall, also known as sensitivity, evaluates the model's ability to identify all relevant positive instances. It is calculated as the ratio of true positive predictions to the total number of actual positive instances. The formula for recall is: $Recall = \frac{tp}{tp+fn}$.

**Adjusted Rand Index (ARI)** assesses the similarity between two partitions, taking into account the possibility of random coincidence. It considers the number of pairs of items in the same cluster in both the true and estimated partitions, along with the total number of items in each cluster in both partitions. It ranges from $-1$ to 1, where an ARI close to 1 indicates good agreement between partitions and an ARI equal to 0 represents a

random partition.

**Mean Squared Error (MSE)** calculates the average squared discrepancy between estimated and true values. A lower MSE indicates a more accurate estimate of the parameter.
In our case, for each group $k$ in 1, ..., g, the Frobenius norm of the difference between the estimated and true covariance matrix is calculated as previously described in subsection (3.2.1). A standard MSE was calculated using the absolute value of the difference of the estimated degrees of freedom $\nu_k$ and the corresponding true values:

$$\text{MSE\_}\nu_k = \frac{1}{N}\sum_{i=1}^{N}(\nu_{k_i} - \overline{\nu}_k)^2 \tag{15}$$

where $\overline{\nu}_k$ represents the true value of the degrees of freedom of the k-th component.

## 4.3. Results

The results of this simulation experiment offer a detailed overview on the efficacy and efficiency of the proposed algorithm.

In the examination of covariance matrices $\Sigma_k$, sharing the same sparsity level, the mean F1-score surpasses 0.7 across all three clusters, underscoring the algorithm's robust proficiency in reconstructing the sparsity patterns inherent in the matrices (Figure 4). The F1-score, a composite measure of precision and recall, proves particularly informative in assessing the algorithm's precision in identifying non-zero entries, with higher scores nearing the desirable value of 1. Notably, the variability of these F1-score values, visualized through box plots, is minimal for each of the three clusters. It is important to highlight the presence of outliers in clusters 1 and 2, despite the generally high F1-scores. Moving to the corresponding Mean Squared Error (MSE), an indicator of the dissimilarity between estimated and true values, the remarkably low values across all clusters are notable, considering the matrices' high dimensions (Figure 5). Although the MSE box plots exhibit narrow distributions, indicating consistency in performance, the presence of outliers suggests occasional discrepancies in specific instances.
Turning attention to degrees of freedom $\nu_k$, the plot displays the distribution of estimated values for the three clusters with boxplots, and true values are indicated by horizontal dashed lines in corresponding colors. (Figure 6). The plot highlights a well-concentrated distribution of estimated $\nu_k$ around the true values for each cluster, demonstrating the algorithm's precision, with only a couple of noticeable outliers.

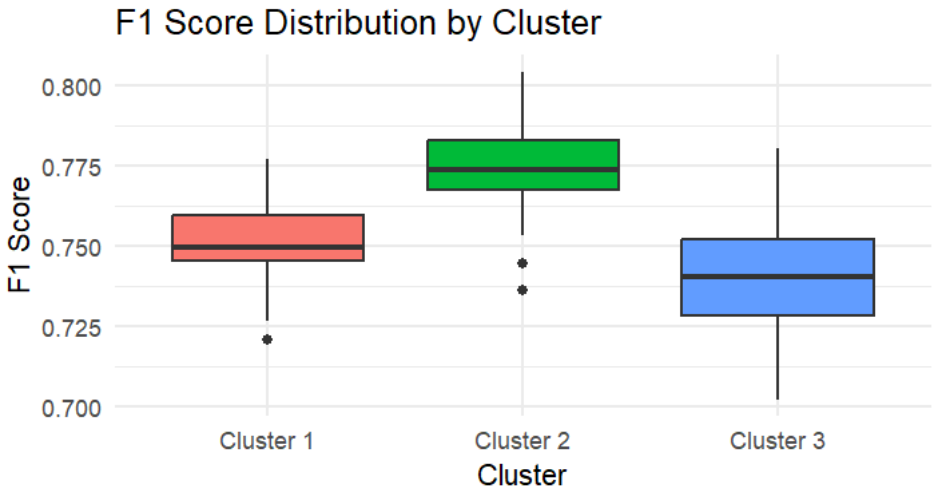


Figure 4: F1-scores for Clusters 1, 2, and 3 in the same sparsity case.

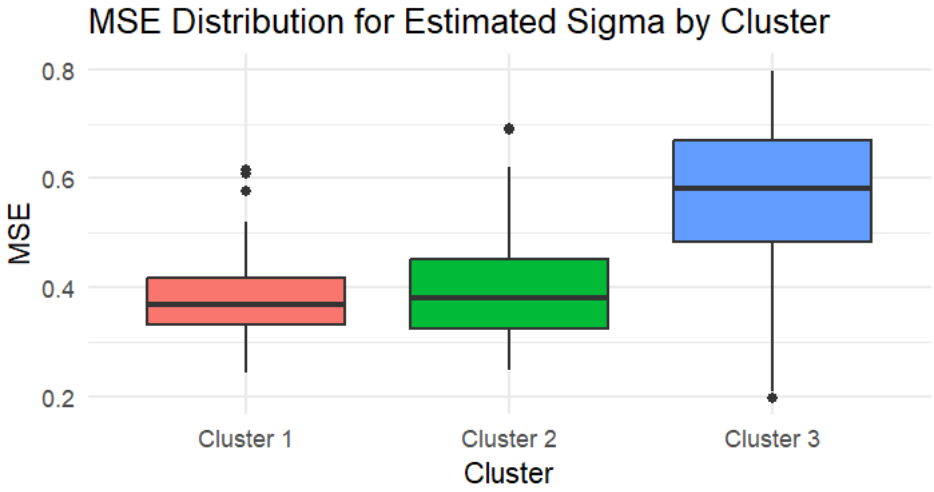


Figure 5: MSE for $\Sigma_1$, $\Sigma_2$ and $\Sigma_3$ in the same sparsity case.

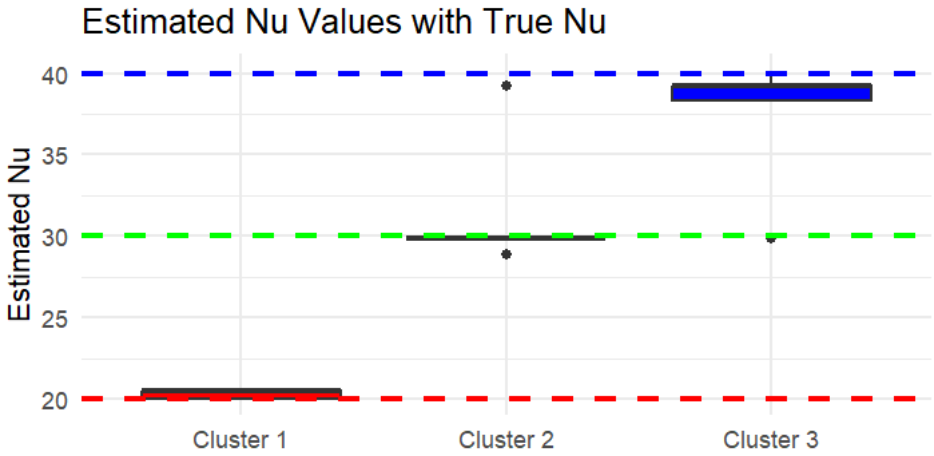


Figure 6: Estimated $\nu_1$, $\nu_2$, and $\nu_3$ compared with true values in the same sparsity case.

In the scenario of different sparsity levels, the F1-score is lower for sparser covariance matrices, which aligns with expectations, as a higher sparsity level might lead to an increased number of false negatives, which lowers recall. Overall, across all three clusters, the F1-score remains satisfactory, albeit slightly lower than in the first scenario (Figure 7).
Despite these variations in F1-scores, the Mean Squared Error for covariance matrices consistently maintains low values, below 1 (Figure 8). However, a few outliers persist, and the box-plots indicate a higher degree of variability compared to the first scenario. This implies that, while the overall accuracy remains robust, there are instances where the algorithm encounters disparities in performance.
Shifting focus to the degrees of freedom $\nu_k$, the plot in Figure 9 illustrates a close alignment between the estimated $\nu_k$ and the true values across all clusters, indicating an overall effective estimation process with minimal outliers.


Figure 7: F1-scores for Clusters 1, 2, and 3 in the different sparsity case.

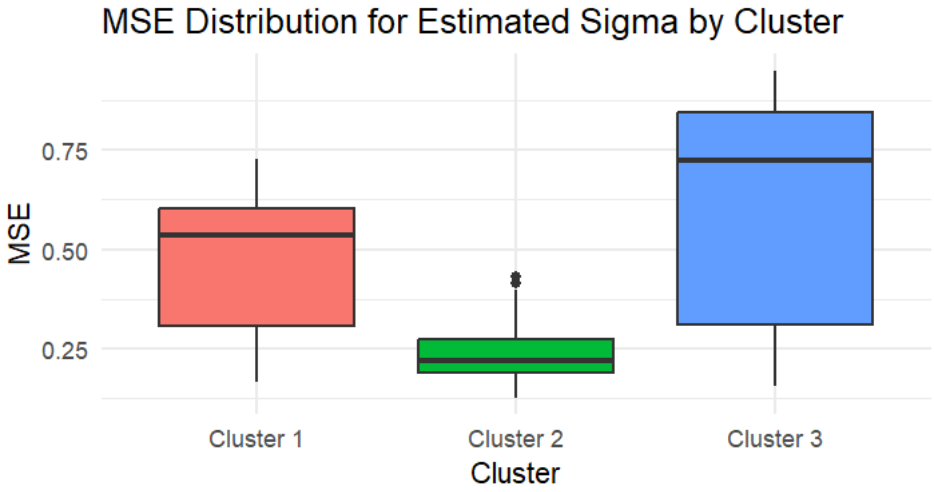


Figure 8: MSE for $\Sigma_1$, $\Sigma_2$ and $\Sigma_3$ in the different sparsity case.

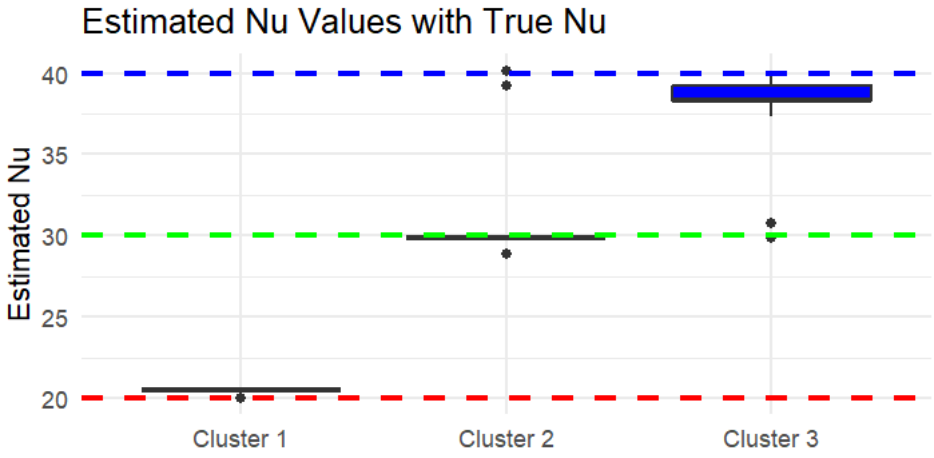


Figure 9: Estimated $\nu_1$, $\nu_2$, and $\nu_3$ compared with true values in the different sparsity case.

Notably, in both cases, the Adjusted Rand Index (ARI) is exactly 1, indicating that the partition has been

perfectly recovered in all the repetitions of the simulated settings. This metric highlights the algorithm's accurate clustering of observations, emphasizing its effectiveness in recovering the true underlying structure. It's important to note that achieving an ARI of 1 in clustering simulations is common, as it is difficult to simulate data that truly challenges the algorithm's ability to recognize the partition.
Overall, these results underscore the algorithm's strong performance in reconstructing sparse covariance matrices, even in a scenario with varying sparsity levels.

In line with the methodology previously described, simulations were executed across various shrinkage values, and only the outcome with the minimum Bayesian Information Criterion (BIC) was retained for computing the evaluation metrics. Notably, the results predominantly gravitate towards two specific shrinkage values. These preferred values are situated in the middle-to-high range of the chosen shrinkage grid, indicating that the use of a stronger lasso penalty led to more parsimonious and generalizable models.

# 5. Cluster Correlation Matrices from fMRI Signals

This section presents the dataset to which the proposed method was applied.
The dataset used in this study is derived from a pilot investigation within the Enhanced Nathan Kline Institute-Rockland Sample (NKI-RS) project ([5]).
It includes multimodal imaging data and subject-specific covariates for $n = 24$ subjects. This pilot study offers scan–rescan imaging data for a subset of subjects, allowing validation and inference on subject-specific variability in brain functions and structures. The dataset comprises the following main sources of information:

1. **Structural Networks:** Maps the brain's structure using Diffusion Tensor Imaging (DTI) to show how white matter fibers connect different brain areas. Information is quantified in symmetric adjacency matrices. This provides a structural foundation for understanding how different parts of the brain are physically connected.
2. **Dynamic Functional Activity:** Measures the dynamic activity of each brain region as multivariate time-series data for 70 regions, observed over 404 time points, capturing the variability and dynamics of brain function during resting-state functional MRI (R-fMRI) sessions.
3. **Functional Networks:** Represents synchronization in brain activity for each pair of brain regions, obtained from the correlation in dynamic functional activity. These networks are expressed through a four-dimensional array W, with elements indicating the pairwise correlation between brain regions' activities, where values range from $-1$ to 1.

Additional detailed information about brain regions and patients' characteristics are also available.
In this section, we will focus on analyzing results on Functional Networks matrices using our implemented method, based also on information given by Structural Networks data.

## 5.1. Exploratory Analysis

The exploratory analysis of the data begins with the examination of structural network matrices derived from DTI scans. Due to the availability of second scans for a limited number of patients, analysis focused on the first scan only. Additionally, matrices were excluded for certain subjects where data was incomplete or missing, resulting in a final dataset of 20 patients. The average matrix was then computed element-wise from the available data, forming the basis for subsequent analyses. Moreover, a heatmap of the average matrix revealed, consistent with the reference paper ([5]), that many brain regions exhibited nearly zero white fiber connections with others (Figure 10). Considering that the regions located in the left hemisphere are found in the first half of the matrix, and those pertaining to the right hemisphere are in the second half, it is evident that there are more connections among regions within the same hemisphere of the brain (either left or right) than between regions located in different hemispheres.

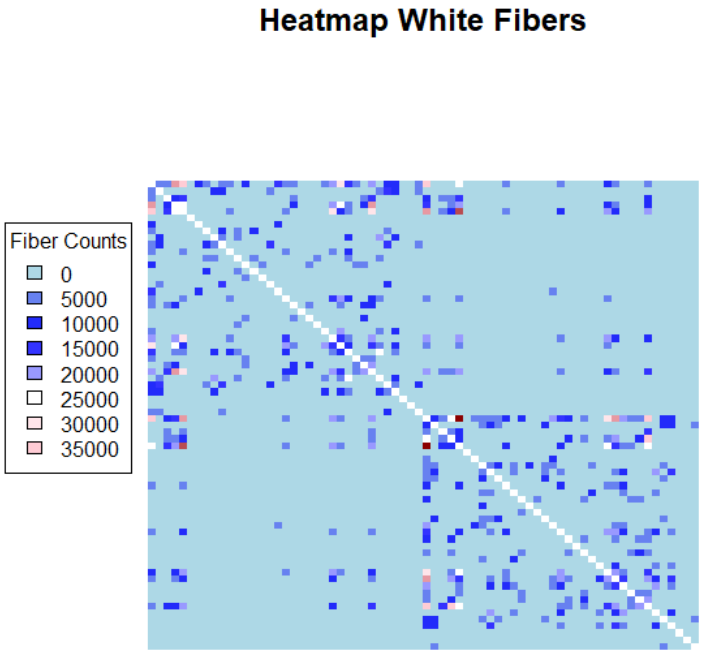


Figure 10: Heatmap of the Average Structural Networks Matrices

To capture this sparsity, a binary matrix was generated. Each element equaled one if the corresponding element in the average matrix was greater than or equal to a given threshold (specifically, the 30th percentile); otherwise, it was set to zero. The resulting binary matrix effectively highlighted brain regions with significant white fiber connectivity. The heatmap in Figure 11 illustrates this step. We can observe a central cross-like pattern formed by zeros, highlighting areas of the brain that are essentially disconnected from other regions: specifically, *lh-supramarginal* (left hemisphere, parietal lobe) and *lh-frontalpole* (left hemisphere, frontal lobe).

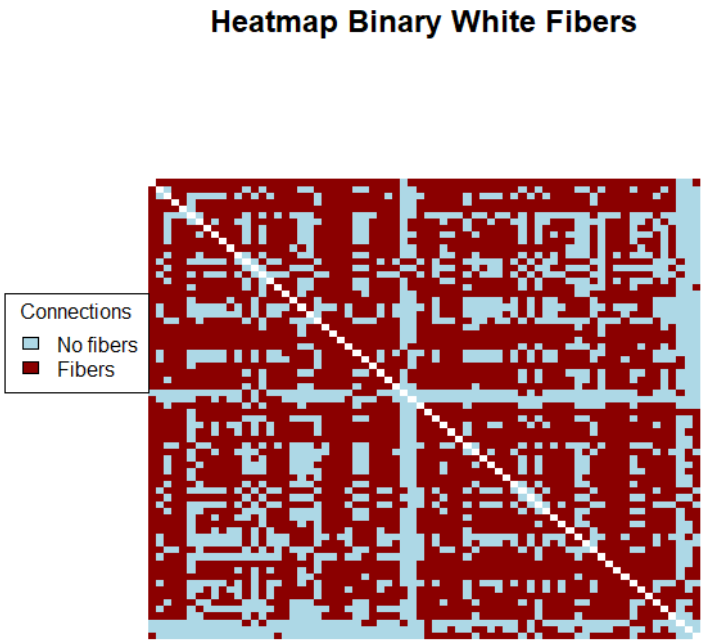


Figure 11: Heatmap of the Binary Structural Networks Matrix

Subsequently, the analysis extended to functional networks' matrices. After addressing missing data, the average matrix was computed. As suggested by the literature ([5]), based on heatmap values (Figure 12), five distinct ranges were created to categorize the average matrix values. This resulted in a matrix that encapsulates these ranges rather than specific values. It helps underlying the most significant connections and the relevant correlations among different brain regions. A corresponding heatmap was generated to facilitate graphical analysis of the results (Figure 13). Both figures exhibit the same cross-shaped pattern previously highlighted in Figure 11.

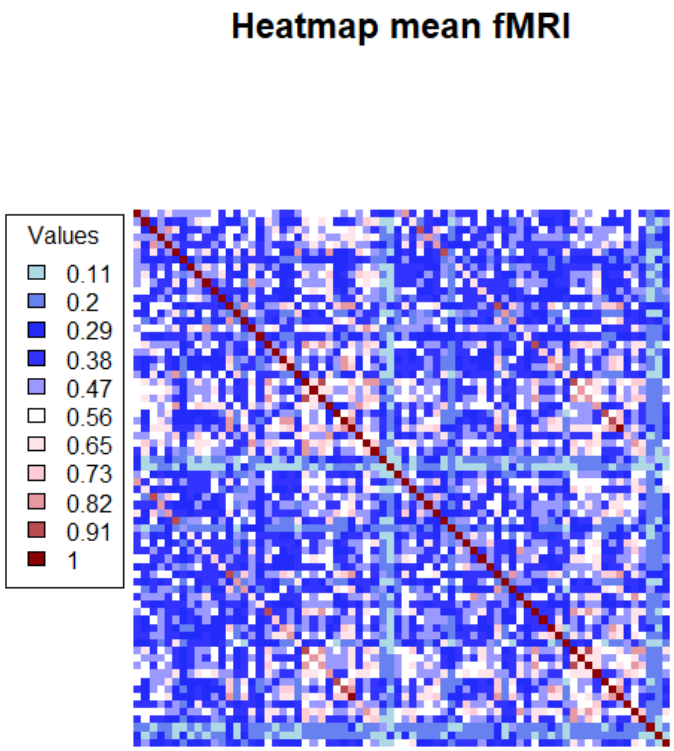


Figure 12: Heatmap of the Average Functional Networks Matrix

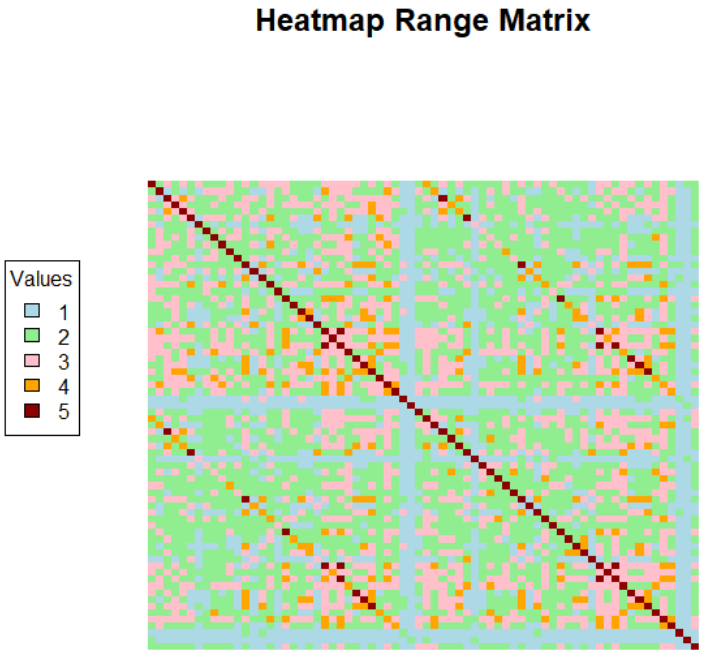


Figure 13: Heatmap of the Range Functional Networks Matrix

These analyses were crucial for obtaining a clear and comprehensive overview of the data. Understanding the patterns and relationships within the structural and functional connectivity matrices allowed for informed expectations regarding outcomes.

## 5.2. Data Pre-process and Initialization

The algorithm takes in input the 22 function networks correlation matrices, since matrices for subject 1 and subject 21 were missing.
It is worth noticing that our method was applied to complete fMRI matrices ($70 \times 70$ dimensions) to ensure that no information was lost through dimensionality reduction. This decision was taken trying to preserve the integrity of the data and obtaining meaningful results from the subsequent clustering algorithm.
To initialize the binary matrix Z, we employed Fuzzy C-Means algorithm ([1]), as suggested in [16], where the number of clusters ($g$) was determined to be equal to 2 through BIC. The Fuzzy C-Means (FCM) algorithm is a clustering method that assigns a degree of membership to each data point for each cluster, allowing for a fuzzy or probabilistic assignment. This algorithm is particularly useful when dealing with complex and overlapping structures in the data. In this study, the degrees of freedom $\nu$ were set to $(80, 80, 80)$. The algorithm was run on a grid of different shrinkage values, which was chosen to be, by trial and error, from $\lambda = 0.05$ to $\lambda = 18$. The best $\lambda$ was the one resulting in the smallest BIC.

## 5.3. Analysis of the Results

In this section, we present results obtained using both the shrinkage value chosen via BIC and a value for $\lambda$ that lead to a pronounced sparsity in the resulting estimates. In details, $\lambda_{BIC} = 0.85$ and $\lambda_{sparse} = 15$. We are going to analyze similarity and differences that arise from the two scenarios. The results are interesting from both a clustering perspective and the estimation of covariance matrices $\Sigma_k$.

We begin by analyzing the case of $\lambda = 15$. Regarding the identified clusters, notable differences in subject characteristics emerge. In Cluster 1, subjects tend to be generally younger, and none of them have a current diagnosis. In contrast, Cluster 2 consists of subjects of slightly older age, with four individuals having a current diagnosis and six with a lifetime diagnosis. The age difference may contribute to the higher prevalence of

diagnosed patients in Cluster 2. The two cluster are numerically balanced, in particular Cluster 1 is composed of 10 patients, while Cluster 2 by 12.

Notably, Cluster 1 includes only one subject with a lifetime diagnosis, which stands out from the diagnoses observed in Cluster 2. Specifically, the patient in the first cluster is diagnosed with an eating disorder, while all the lifetime diagnoses in Cluster 2 relate to alcohol or drugs abuse and/or depression. These findings suggest that both age and specific diagnostic patterns contribute to the distinct profiles observed within the identified clusters.

The covariance matrices $\hat{\Sigma}_k$ exhibit a slightly different level of sparsity, with a more pronounced number of zeros in the case of Cluster 1. This result suggests that a higher degree of correlation among brain regions may indicate the potential presence of a mental disorder. The increased sparsity in the covariance matrix of Cluster 1 implies a more independent structure among brain regions within this cluster. Such variations in the covariance structure may provide valuable insights into the underlying neurological characteristics associated with mental health conditions, highlighting the potential relevance of inter-regional correlations in identifying patterns related to mental disorders. Despite their different levels of sparsity, both matrices reveal similar patterns. As can be seen in Figure 14, there are regions within each matrix where correlations with other regions seem to be minimal. In particular, distinct rows of zeros are evident, giving rise to a discernible cross-shaped pattern, a characteristic already noticed before in the exploratory analysis (Figures 11, 12 and 13). This commonality in the identified patterns across the two clusters underscores the potential significance of these specific inter-regional correlations in elucidating neurological characteristics related to mental health conditions.


Figure 14: On the left, the estimate of $\Sigma_1$, on the right, the estimate of $\Sigma_2$, both obtained with $\lambda = 15$. Black squares denote values different from zero.

Regarding the case where $\lambda = 0.85$, the clustering output exhibits minimal changes; only one patient has been reassigned from Cluster 2 to Cluster 1. Notably, this particular patient carries a lifetime diagnosis of a disorder, complicating the identification of discernible patterns within these clusters. The majority of patients with a life-time diagnosis still reside in Cluster 2. Cluster 1 maintains its lower average age, since the relocated patient is very young. In this scenario, the clusters are perfectly balanced with respect to the number of samples, each containing 11 individuals.
In terms of the estimated covariance matrices, they obviously manifest low sparsity (Figure 15). However, it is noteworthy that $\Sigma_1$ still displays a higher number of zero elements, exceeding twice the count found in $\Sigma_2$.

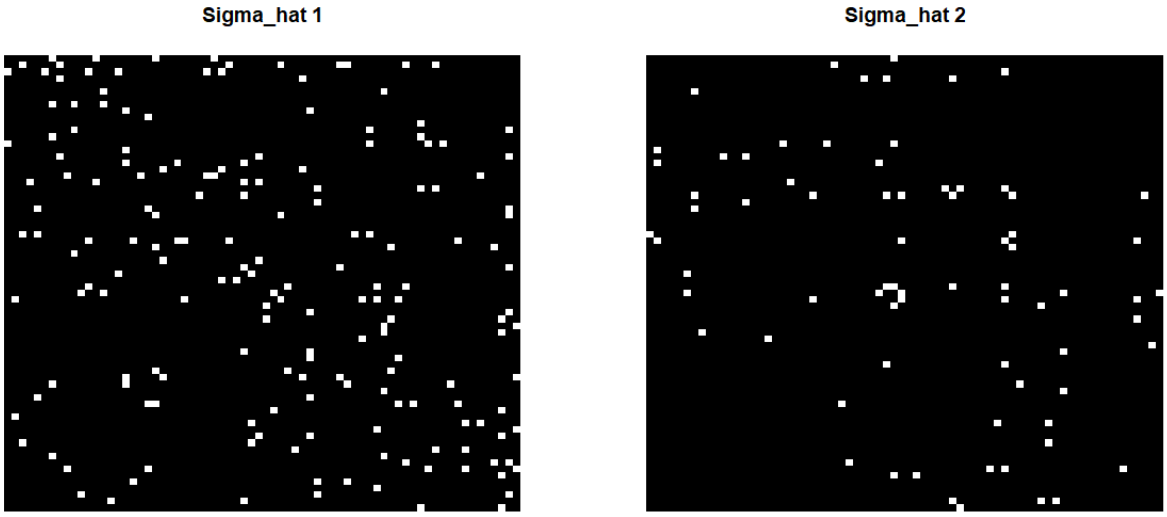


Figure 15: On the left, the estimate of $\Sigma_1$, on the right, the estimate of $\Sigma_2$, both obtained with $\lambda = 0.85$. Black squares denote values different from zero.

The study shows promising outcomes in terms of patient stratification, underscoring its potential impact in the field. However, given the small sample size of the dataset, it is necessary to conduct further analyses to validate these results thoroughly. Such additional investigations are crucial for confirming the reliability and robustness of the findings. It is crucial to underscore that the drawn conclusions should undergo thorough validation through further statistical analyses and, whenever feasible, be juxtaposed with existing scientific literature.

# 6. Discussion

Motivated by functional magnetic resonance imaging (fMRI) data, this work addresses limitations in model-based clustering within high-dimensional contexts. The challenges posed by high-dimensional datasets, where variables outnumber observations, necessitate advanced statistical modeling techniques. A crucial aspect in overcoming these challenges is the incorporation of sparsity, emphasizing the relevance of a subset of variables, with penalized likelihood estimation serving as a key tool in inducing sparsity.
This thesis focuses on enhancing model-based clustering for high-dimensional data by integrating sparsity techniques, particularly in estimating covariance matrices from a mixture of Wishart distributions. Drawing on the foundational work of Hidot and Saint-Jean, our approach utilizes the Expectation-Maximization algorithm, integrated with the Covariance Graphical Lasso, to induce sparsity effectively. This combination not only addresses the challenges inherent in high-dimensional data, but also enables the identification of distinct subgroups, enriching our understanding of the intricate covariance structures that exist among variables.
Rigorous testing on simulated data showcases the methodology's efficiency, successfully identifying meaningful patterns and subgroups within complex structures. When applied to correlation matrices derived from fMRI data, the method reveals two distinct clusters associated with age and mental health status.
The proposed methodology not only demonstrates its capability to uncover meaningful patterns even in scenarios with a limited number of observations, but also yields promising results, showing its potential effectiveness. However, it is crucial to acknowledge the impact of the reduced sample size on the reliability of the results.
Future research directions could focus on refining the methodology by incorporating group-specific shrinkage parameters ($\lambda$) to enhance precision in parameter estimation. This could be especially beneficial in constrained sample sizes. A relevant work in this context has been carried out by Casa, Cappozzo and Fop in [7]. The paper describes a methodological advancement in the context of finite Gaussian mixture models applied to clustering multivariate continuous data. One notable limitation addressed in this study is the assumption of similar levels of sparsity across classes, neglecting potential variations in the degrees of association between variables across groups. The proposed solution involves deriving group-wise penalty factors. Extending this methodology to covariance matrices derived from a mixture of Wishart distributions, as explored in our study, holds the potential for significant advancements. The incorporation of group-wise shrinkage factors could enhance the precision of parameter estimation in the presence of complex covariance structures. This extension aligns with

the broader aim of adapting statistical methodologies to the challenges posed by high-dimensional data. The proposed advancements could offer valuable insights into the intricate covariance patterns among brain regions, contributing to a deeper understanding of brain connectivity and potentially leading to more accurate diagnoses of neurological disorders.

Another promising direction for enhancing our model-based clustering method involves integrating an extended version of the covariance graphical lasso algorithm designed for multi-class settings (see [9], [14], [19] and [21]). This adaptation allows for the estimation of sparse covariance matrices that reflect the unique association patterns of observations arising from $g$ distinct sub-populations, each with potentially different covariance structures. Such an approach aligns with the complexity of datasets like those in fMRI studies. However, employing these modified graphical lasso techniques introduces potential drawbacks that merit careful consideration. While these methods are skilled at identifying commonalities and shared sparsity patterns, they inherently promote some degree of similarity among the groups. This can be advantageous for creating parsimonious and interpretable models that elucidate the relationships among variables within and across classes. Yet, this emphasis on structural similarities could impede the primary goal of clustering by making distinct groups appear more similar than they actually are. Therefore, while the application of these advanced graphical lasso techniques offers a sophisticated means to accommodate the structures of observations arising from different groups, it is crucial to balance the benefits against the potential risk of diminishing the clarity of cluster distinctions.

Additionally, exploring alternative sparsity-inducing techniques and their impact on clustering outcomes would contribute to a more comprehensive understanding of our method's strengths and limitations.

In conclusion, the thesis presents an innovative statistical methodology applied to neuroscience research, offering insights into brain connectivity. While demonstrating promising results, the study emphasizes the importance of considering limitations due to small sample sizes, urging a cautious interpretation of findings. Future work should concentrate on enhancing precision in parameter estimation and exploring alternative sparsity techniques to further refine and extend the methodology's applicability.

# References


[1] C. Bezdek. Pattern recognition with fuzzy objective function algorithms. *Plenum Press*, 1981.

[2] J. Bien and R.J. Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 98:807–820, 2011.

[3] P. Boileau, N.S. Hejazi, M. Van deer Lan, and S. Dudoit. Cross-validated loss-based covariance matrix estimator selection in high dimensions. *Journal of Computational and Graphical Statistics*, (32:2):601–612, 2023.

[4] A. Cabassi, A. Casa, M. Fontana, M. Russo, and A. Farcomeni. Three testing perspectives on connectome data. In Springer, editor, *Studies in Neural Data Science*, volume 257, pages 37–56, 2017.

[5] A. Canale, D. Durante, L. Paci, and B. Scarpa. Multimodal imaging data in neuroscience. In Springer, editor, *Studies in Neural Data Science-StartUp Research*, volume 257, 2017.

[6] A. Cappozzo, F. Ferraccioli, M. Stefanucci, and P. Secchi. An object oriented approach to multimodal imaging data in neuroscience. In Springer, editor, *Studies in Neural Data Science*, volume 257, pages 57–73, 2017.

[7] A. Casa, A. Cappozzo, and M. Fop. Group-wise shrinkage estimation in penalized model-based clustering. *Journal of Classifcation*, 2022.

[8] D. Crespo-Roces, I. Méndez-Jiménez, S. Salcedo-Sanz, and M. Cárdenas-Montes. Generalized probability distribution mixture model for clustering. In *International Conference on Hybrid Artificial Intelligence Systems*, pages 251–263, 2018.

[9] P. Danaher, P. Wang, and D.M. Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 76:373, 2014.

[10] L.I. Dryden, A. Koloydenko, and D. Zhou. Non-euclidean statistics for covariance matrices, with application to diffusion tensor imaging. *The Annals of Applied Statistics*, 3(3):1102–1123, 2009.

[11] C. Fraley and A.E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–632, 2002.

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

[13] M. Gallaugher, A. Punzo, and P. McNicholas. Finite mixtures of skewed matrix variate distributions. *Elsevier*, 80:83–93, 2018.

[14] J. Guo, E. Levina, G. Michailidis, and J. Zhu. Joint estimation of multiple graphical models. *Biometrika*, 98:1–15, 2011.

[15] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. Chapman and Hall/CRC, 2015.

[16] S. Hidot and C. Saint-Jean. An expectation–maximization algorithm for the wishart mixture model: Application to movement clustering. *Elsevier*, pages 2318–2324, 2009.

[17] V.H. Lachos Dávila, C.R. Barbosa Cabral, and C. Borelli Zeller. *Finite Mixture of Skewed Distributions*. Springer, New York, 2018.

[18] T.I. Lin. Maximum likelihood estimation for multivariate skew normal mixture models. *Elsevier*, 100:257–265, 2009.

[19] Y. Lyu, L. Xue, F. Zhang, H. Koch, L. Saba, K. Kechris, and Q. Li. Condition-adaptive fused graphical lasso (cfgl): An adaptive procedure for inferring condition-specifc gene co-expression network. *PLoS computational Biology*, 14, 2018.

[20] G.J. McLachlan and D. Peel. Robust cluster analysis via mixtures of multivariate t-distributions. *Springer*, 1451:658–666, 2005.

[21] K. Mohan, P. London, M. Fazel, D. Witten, and S. Lee. Node-based learning of multiple gaussian graphical models. *Journal of Machine Learning Research*, 15:445–488, 2014.

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

[23] W. Pan and X. Shen. Penalized model-based clustering with cluster-specifc diagonal covariance matrices and grouped variables. *Electronic Journal of Statistics*, 2, 2008.

[24] I. Vrbik and P.D. McNicholas. Parsimonious skew mixture models for model-based clustering and classification. *Elsevier*, 71:196–210, 2014.

[25] H. Wang. Coordinate descent algorithm for covariance graphical lasso. *Science+Business Media New York*, 24:521–529, 2014.

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

## Abstract in lingua italiana

La Risonanza Magnetica Funzionale (fMRI) ha trasformato radicalmente il nostro modo di esaminare il cervello umano in funzione, permettendoci di catturare con precisione i complessi pattern spaziotemporali dell'attività cerebrale attraverso dati dettagliati e di elevata dimensionalità. Una delle principali sfide nell'analisi di questi dati fMRI consiste nell'identificare le strutture sottostanti all'interno di questi ricchi dataset. Questo lavoro di tesi affronta tale sfida proponendo un approccio innovativo per classificare i pazienti secondo le strutture di covarianza comuni presenti nei loro segnali fMRI. Di fronte alla problematica dell'elevata dimensionalità dei dati fMRI, presentiamo una nuova metodologia basata sull'impiego di tecniche di stima penalizzata, in particolare attraverso l'utilizzo di un modello misto di distribuzioni Wishart sparse. Questo approccio si propone di superare il problema della grande dimensionalità favorendo la sparsità nella stima delle matrici di covarianza.
Il cuore di questa tesi è lo sviluppo di un metodo di analisi basato su modelli che, grazie all'uso di stime penalizzate, affina l'analisi statistica concentrando l'attenzione sulle matrici di covarianza sparse. Questo progresso rende possibile ottenere una comprensione più profonda dei complessi schemi di dati, migliorando notevolmente sia l'interpretazione che l'efficienza computazionale dei modelli.
Applicando il nostro modello di clustering ai dati fMRI, siamo stati in grado di identificare gruppi distinti di individui basati sui loro schemi di attività cerebrale. L'analisi ha messo in luce correlazioni significative tra i cluster identificati e le specificità dei soggetti coinvolti.
Questa ricerca punta a colmare la distanza tra le avanzate metodologie statistiche e la loro applicazione pratica nel campo delle neuroscienze. Mediante l'integrazione del clustering basato su modelli con la stima di matrici di covarianza sparse, offriamo uno strumento potente per analizzare la complessa rete di connessioni cerebrali, aprendo nuove strade per una migliore comprensione e diagnosi dei disturbi neurologici.



**Parole chiave:** Clustering basato sul modello, Verosimiglianza penalizzata, Matrici di covarianza sparse, Modelli di mistura di Wishart, Algoritmo E-M.



## Acknowledgements

Nell'esprimere la mia riconoscenza, tengo particolarmente a ringraziare i Professori Andrea Cappozzo e Alessandro Casa per il loro inestimabile sostegno, disponibilità, guida e comprensione forniti nel corso dello sviluppo di questo lavoro. Il loro prezioso orientamento e incoraggiamento sono stati determinanti nell'influenzare positivamente i risultati di questa ricerca.