Computational Statistics and Data Analysis 211 (2025) 108198

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Heavy-tailed matrix-variate hidden Markov models

Salvatore D. Tomarchio

Department of Economics and Business, University of Catania, Catania, Italy

ARTICLE INFO

Keywords:
Atypical data
Heavy-tails
Hidden Markov models
Matrix-variate

ABSTRACT

The matrix-variate framework for hidden Markov models (HMMs) is expanded with two families of models using matrix-variate t and contaminated normal distributions. These models improve the handling of tail behavior, clustering, and address challenges in identifying outlying matrices in matrix-variate data. Two Expectation-Conditional Maximization (ECM) algorithms are implemented in the R package **MatrixHMM** for parameter estimation. Simulations assess parameter recovery, robustness, anomaly detection, and show the advantages over alternative approaches. The models are applied to real-world data to analyze labor market dynamics across Italian provinces.

1. Introduction

Matrix-variate data have gained significant importance across various scientific fields, particularly in the context of model-based clustering. Applications span multiple domains, including economics and education (Tomarchio, 2024), medicine and crime (Viroli, 2011b; Melnykov and Zhu, 2019), image recognition (Gallaugher and McNicholas, 2018), and environment (Tomarchio and Gallaugher, 2024).

When a matrix of dimensions $P \times R$ is observed for each statistical unit, matrix-variate models provide a more comprehensive and flexible framework compared to traditional multivariate approaches applied to vectorized PR-dimensional data. As established in the literature (see, e.g., Anderlucci et al., 2014; Viroli, 2012; Gallaugher and McNicholas, 2018; Tomarchio et al., 2022b, 2024; Ma et al., 2023), directly utilizing matrix-variate structures enables simultaneous modeling of variability across the P rows and R columns using two distinct covariance (or scale) matrices. This dual representation preserves structural relationships within the data more effectively than vectorization, which can obscure dependencies and lead to information loss.

An additional advantage of matrix-variate modeling lies in its more parsimonious parameterization of covariance (or scale) matrices. Converting $P \times R$ matrices into PR-dimensional vectors requires the estimation of $(P^2R^2 + PR)/2$ parameters. In contrast, maintaining the matrix-variate structure reduces this number to $P(P+1)/2 + R(R+1)/2 - 1$, significantly lowering model complexity. Furthermore, the increased number of parameters induced by vectorization can affect model selection when employing information criteria (Sarkar et al., 2020; Tomarchio et al., 2021; Tomarchio and Punzo, 2025).

A common practice involves representing time along either the rows or columns of matrices (see, e.g., Viroli, 2011a,b; Melnykov and Zhu, 2019; Tomarchio et al., 2020; Tomarchio, 2024). From a model-based clustering perspective, this approach assumes time-constant clustering, meaning that sample units cannot transition between clusters over time, thereby disregarding the temporal evolution of the clustering structure. However, accounting for temporal dynamics can reveal valuable insights. Hidden Markov Models (HMMs) offer a robust framework for capturing temporal dependencies through hidden states that evolve dynamically (Zucchini et al., 2017).

E-mail address: daniele.tomarchio@unict.it.

https://doi.org/10.1016/j.csda.2025.108198
Received 30 August 2024; Received in revised form 10 April 2025; Accepted 25 April 2025

Available online 12 May 2025

The methodology of HMMs within the matrix-variate framework remains relatively narrow and has only recently begun to be explored, as shown by the contributions of Asilkalkan and Zhu (2021); Tomarchio et al. (2022b); Sarkar and Zhu (2022); Tomarchio et al. (2024); Gallaugher and Zhu (2024). This gap presents opportunities for further research and development, and this manuscript contributes to the literature by introducing heavy-tailed HMMs for the analysis of matrix-variate longitudinal data.

By starting with the concept of heavy-tailed HMMs, the commonly used matrix-variate normal (MVN) distribution, while mathematically elegant, may be overly restrictive for real-world applications involving atypical matrices. Failure to accommodate such characteristics can impact classification and parameter inference (Hossain and Naik, 1991). To mitigate these issues, two heavy-tailed elliptical generalizations of the MVN distribution are considered: the matrix-variate t (MVT) (Doğru et al., 2016) and the matrix-variate contaminated-normal (MVCN) (Tomarchio et al., 2022a) distributions. These distributions offer greater flexibility in modeling data with heavy tails and atypical matrices, thereby enhancing robustness in various scenarios.

The proposed approach also facilitates the detection of atypical matrices. As demonstrated in this manuscript, post-estimation procedures can be readily applied for this purpose, which is particularly valuable in matrix-variate data analysis, where visual anomaly detection poses significant challenges.

Shifting to the concept of matrix-variate longitudinal data, this format arises when a set of $P \times R$ matrices is observed for I units over T time frames, resulting in a four-way array of dimensions $P \times R \times I \times T$. This structure enables richer data representations than standard multivariate HMMs. For instance, while Sarkar et al. (2020) have utilized two-factor data within three-way cross-sectional frameworks, incorporating a temporal dimension requires an additional layer, naturally accommodated within the four-way structure, as shown in Tomarchio et al. (2022b, 2024).

To further increase model parsimony, this work employs the well-known eigen-decomposition of covariance (or scale) matrices, a classic approach in model-based clustering (Celeux and Govaert, 1995). This decomposition is applied to the scale matrices of the MVT and MVCN distributions, leading to two families comprising 98 parsimonious MV-HMMs.

From a computational perspective, model parameters are estimated using Expectation Conditional Maximization (ECM) algorithms (Meng and Rubin, 1993) combined with recursions commonly utilized in the HMM literature (Baum et al., 1970). These methodologies are implemented in the **MatrixHMM** package for the R software, recently released on the CRAN repository. An overview of the package is provided in this manuscript.

The proposed algorithms and models are initially evaluated through simulated studies. First, the ability to recover the parameters from data-generating models is assessed. Next, the robustness of parameter estimation is examined in the presence of atypical matrices, followed by an investigation into the effectiveness of detecting such matrices. Comparisons with multivariate HMMs and matrix-variate mixtures are also considered.

A real data application is also performed. Specifically, three key work-related variables (employment, unemployment, and inactivity rates) are jointly analyzed for the Italian provinces over the last six years of available data. These variables provide a comprehensive overview of labor market activity. The identification of hidden states, their characteristics, and the temporal transitions of provinces yield valuable insights into the structure and trends of the Italian job market.

The remainder of this paper is organized as follows. Section 2 introduces the two families of 98 parsimonious MV-HMMs, while Section 3 details the ECM algorithms. Computational and operational aspects, including initialization strategies, atypical matrix detection tools, and the functionalities of the **MatrixHMM** package, are discussed in Section 4. Section 5 presents the simulation studies and their results. Section 6 provides an in-depth analysis of the real data application. Finally, Section 7 concludes the study.

2. Heavy-tailed matrix-variate hidden Markov models

2.1. Model specifications

Let $\{\mathcal{X}_{it}; i = 1, \ldots, I, t = 1, \ldots, T\}$ be a sequence of matrix-variate longitudinal observations measured over I units and T time points, where each $\mathcal{X}_{it}$ has dimensions $P \times R$. Additionally, let $\{S_{it}; i = 1, \ldots, I, t = 1, \ldots, T\}$ be a family of I independent first-order Markov chains of length T operating within the state space $\{1, \ldots, k, \ldots, K\}$. In an HMM framework, the following conditional independence property holds:

$$\begin{aligned} &f\left(\mathcal{X}_{it} = \mathbf{X}_{it} | \mathcal{X}_{i1} = \mathbf{X}_{i1}, \ldots, \mathcal{X}_{it-1} = \mathbf{X}_{it-1}; S_{i1} = s_{i1} \ldots, S_{it} = s_{it}; \boldsymbol{\theta}\right) \\ &= f\left(\mathcal{X}_{it} = \mathbf{X}_{it} | S_{it} = s_{it}; \boldsymbol{\theta}\right), \end{aligned} \tag{1}$$

where $f(\cdot)$ is a generic probability density function (pdf) with parameter $\boldsymbol{\theta}$. Here, $f(\cdot)$ in (1) can take the functional form of either the MVT or MVCN distribution.

A generic $P \times R$ random matrix $\mathcal{X}_{it}$ follows an MVT distribution if its pdf is given by:

$$\frac{|\boldsymbol{\Sigma}_k|^{-\frac{R}{2}} |\boldsymbol{\Psi}_k|^{-\frac{P}{2}} \Gamma(\frac{PR+\nu_k}{2})}{(\pi\nu_k)^{\frac{PR}{2}} \Gamma(\frac{\nu_k}{2})} \left[1 + \frac{\delta_k(\mathbf{X}_{it}; \mathbf{M}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Psi}_k)}{\nu_k}\right]^{-\frac{PR+\nu_k}{2}}, \tag{2}$$

where $\mathbf{M}_k$ is the $P \times R$ mean matrix, $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\Psi}_k$ are the $P \times P$ and $R \times R$ scale matrices, respectively, $\nu_k > 0$ is the degrees of freedom parameter, and $\delta_k(\mathbf{X}_{it}; \mathbf{M}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Psi}_k) = \text{tr}\left[\boldsymbol{\Sigma}_k^{-1}(\mathbf{X}_{it} - \mathbf{M}_k)\boldsymbol{\Psi}_k^{-1}(\mathbf{X}_{it} - \mathbf{M}_k)'\right]$ is the squared Mahalanobis distance. In symbols, $\mathcal{X}_{it} | S_{it} = k \sim \mathbb{T}(\mathbf{M}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Psi}_k, \nu_k)$.

Table 1
Family, nomenclature, type, and number of free parameters in $\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K$ for the parsimonious models obtained via the eigen decomposition. $\boldsymbol{I}$ is the identity matrix.

Family	Model	Type	Number of free parameters in $\Sigma_1, \ldots, \Sigma_K$
Spherical	EII	$\lambda \boldsymbol{I}$	1
Spherical	VII	$\lambda_k \boldsymbol{I}$	K
Diagonal	EEI	$\lambda \boldsymbol{\Delta}$	P
Diagonal	VEI	$\lambda_k \boldsymbol{\Delta}$	$K + P - 1$
Diagonal	EVI	$\lambda \boldsymbol{\Delta}_k$	$K(P-1)+1$
Diagonal	VVI	$\lambda_k \boldsymbol{\Delta}_k$	KP
General	EEE	$\lambda \boldsymbol{\Gamma}\boldsymbol{\Delta}\boldsymbol{\Gamma}'$	$P(P+1)/2$
General	VEE	$\lambda_k \boldsymbol{\Gamma}\boldsymbol{\Delta}\boldsymbol{\Gamma}'$	$P(P+1)/2 + K - 1$
General	EVE	$\lambda \boldsymbol{\Gamma}\boldsymbol{\Delta}_k\boldsymbol{\Gamma}'$	$P(P-1)/2 + K(P-1) + 1$
General	VVE	$\lambda_k \boldsymbol{\Gamma}\boldsymbol{\Delta}_k\boldsymbol{\Gamma}'$	$P(P-1)/2 + KP$
General	EEV	$\lambda \boldsymbol{\Gamma}_k\boldsymbol{\Delta}\boldsymbol{\Gamma}_k'$	$KP(P-1)/2 + P$
General	VEV	$\lambda_k \boldsymbol{\Gamma}_k\boldsymbol{\Delta}\boldsymbol{\Gamma}_k'$	$KP(P-1)/2 + K + P - 1$
General	EVV	$\lambda \boldsymbol{\Gamma}_k\boldsymbol{\Delta}_k\boldsymbol{\Gamma}_k'$	$KP(P+1)/2 - K + 1$
General	VVV	$\lambda_k \boldsymbol{\Gamma}_k\boldsymbol{\Delta}_k\boldsymbol{\Gamma}_k'$	$KP(P+1)/2$

Similarly, $\mathcal{X}_{it}$ follows an MVCN distribution if its pdf is:

$$\alpha_k \phi(\mathbf{X}_{it}; \mathbf{M}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Psi}_k) + (1-\alpha_k)\phi(\mathbf{X}_{it}; \mathbf{M}_k, \eta_k\boldsymbol{\Sigma}_k, \boldsymbol{\Psi}_k), \tag{3}$$

where $\phi(\mathbf{X}_{it}; \mathbf{M}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Psi}_k)$ is the pdf of the MVN distribution, i.e.,

$$\frac{1}{(2\pi)^{\frac{PR}{2}} |\boldsymbol{\Sigma}_k|^{\frac{R}{2}} |\boldsymbol{\Psi}_k|^{\frac{P}{2}}} \exp\left[-\frac{\delta_k(\mathbf{X}_{it}; \mathbf{M}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Psi}_k)}{2}\right], \tag{4}$$

where $\mathbf{M}_k$ is the $P \times R$ mean matrix, and $\boldsymbol{\Sigma}_k$ and $\boldsymbol{\Psi}_k$ are the $P \times P$ and $R \times R$ covariance matrices, respectively. In (3), $\alpha_k \in (0,1)$ is the proportion of typical points, and $\eta_k > 1$ is an inflation parameter accounting for the degree of contamination in the data, and is a measure of how different the atypical matrices are from the bulk of the data. In symbols, $\mathcal{X}_{it}|S_{it} = k \sim \mathbb{CN}(\mathbf{M}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Psi}_k, \alpha_k, \eta_k)$.

The MVT distribution is more parsimonious than the MVCN, as it involves fewer parameters. However, the additional parameters introduced in the MVCN distribution offer practical interpretability, as they relate directly to contamination detection and data structure (Tomarchio et al., 2022a), and have closed-form expressions for parameter estimation. Having both distributions available is beneficial, as it provides multiple options for modeling real data.

In addition to the parameters of the MVT or MVCN distributions, the model includes parameters related to the Markov chain. Specifically, these consist of the initial probabilities $\pi_{ik} = \Pr\left(S_{i1} = k\right)$, $k = 1, \ldots, K$, and the transition probabilities

$$\pi_{ik|j} = \Pr\left(S_{it} = k | S_{it-1} = j\right), \quad t = 2, \ldots, T \text{ and } j, k = 1, \ldots, K,$$

where j refers to the previously visited state. The initial probabilities are collected in the K-dimensional vector $\boldsymbol{\pi}$, while the transition probabilities are stored in the $K \times K$ transition matrix $\boldsymbol{\Pi}$.

Heavy-tailed elliptical distributions, such as those in (2) and (3), offer the flexibility needed to robustly handle mildly atypical observations. In contrast, the MVN distribution in (4), commonly used as a reference distribution, does not fit as well. Supporting results within a matrix-variate model-based clustering context are discussed in Tomarchio et al. (2022a) and Tomarchio and Gallaugher (2024). Consequently, using the MVT and MVCN distributions within an HMM framework is expected to offer these advantages when modeling matrix-variate longitudinal data.

2.2. Parsimonious structures

To introduce parsimony in HMMs, the eigen-decomposition of the covariance/scale matrices is considered. Specifically, $\boldsymbol{\Sigma}_k$ is decomposed as:

$$\boldsymbol{\Sigma}_k = \lambda_k \boldsymbol{\Gamma}_k \boldsymbol{\Delta}_k \boldsymbol{\Gamma}_k', \tag{5}$$

where $\lambda_k = |\boldsymbol{\Sigma}_k|^{1/P}$ is a proportionality constant, $\boldsymbol{\Gamma}_k$ is a $P \times P$ orthogonal matrix of eigenvectors and $\boldsymbol{\Delta}_k$ is the $P \times P$ diagonal matrix with scaled eigenvalues (with $|\boldsymbol{\Delta}_k| = 1$) on the main diagonal. The constant λ_k can be constrained to be equal across the hidden states, while the matrices $\boldsymbol{\Gamma}_k$ and $\boldsymbol{\Delta}_k$ can be constrained to be equal across the hidden states or to be the identity matrix. As a result, the 14 parsimonious structures reported in Table 1 are obtained.

For $\boldsymbol{\Psi}_k$, the identifiability constraint $|\boldsymbol{\Psi}_k| = 1$ is imposed (Sarkar et al., 2020; Tomarchio et al., 2022b), eliminating λ_k in the eigen-decomposition. This reduces the number of parsimonious structures for $\boldsymbol{\Psi}_k$ from 14 to the 7 presented in Table 2.

Table 2
Family, nomenclature, type, and number of free parameters in $\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_K$ for the parsimonious models obtained via the eigen decomposition. $\boldsymbol{I}$ is the identity matrix.

Family	Model	Type	Number of free parameters in $\Psi_1, \ldots, \Psi_K$
Spherical	II	$\boldsymbol{I}$	0
Diagonal	EI	$\boldsymbol{\Delta}$	$R-1$
Diagonal	VI	$\boldsymbol{\Delta}_k$	$K(R-1)$
General	EE	$\boldsymbol{\Gamma\Delta\Gamma}'$	$R(R+1)/2-1$
General	VE	$\boldsymbol{\Gamma\Delta}_k\boldsymbol{\Gamma}'$	$R(R-1)/2+K(R-1)$
General	EV	$\boldsymbol{\Gamma}_k\boldsymbol{\Delta\Gamma}_k'$	$KR(R-1)/2+R-1$
General	VV	$\boldsymbol{\Gamma}_k\boldsymbol{\Delta}_k\boldsymbol{\Gamma}_k'$	$KR(R+1)/2-K$

Based on the parsimonious structures detailed in Table 1 and Table 2, $14 \times 7 = 98$ parsimonious HMMs are introduced for each distribution: MVT-HMMs and MVCN-HMMs. To ensure clear notation, each parsimonious HMM is uniquely labeled by hyphenating the acronyms representing the row and column covariance/scale matrices.

3. Parameter estimation via ECM algorithms

Let $\left\{\mathbf{X}_{it}; i = 1, \ldots, I, t = 1, \ldots, T\right\}$ be a sample of matrix-variate longitudinal observations. To estimate the parameters for the proposed models using maximum likelihood (ML), the following likelihood function must be calculated

$$L(\boldsymbol{\Theta}) = \prod_{i=1}^{I} \boldsymbol{\pi}' \mathbf{f}_{i1} \boldsymbol{\Pi} \mathbf{f}_{i2} \cdots \boldsymbol{\Pi} \mathbf{f}_{iT} \mathbf{1}_K, \tag{6}$$

where $\mathbf{f}_{it}$ represents a $K \times K$ matrix with the main diagonal containing the densities specified in (2) or in (3), $\mathbf{1}_K$ is a vector of K ones and $\boldsymbol{\Theta}$ encompasses all the parameters related to each model.

To efficiently maximize the logarithm of (6), the ECM algorithm is utilized. This algorithm is a variant of the traditional expectation-maximization (EM) algorithm (Dempster et al., 1977), differing in that the M-step is substituted with a series of simpler and computationally more convenient CM steps. Direct implementation of the EM algorithm is not feasible because, as is often the case with many matrix-variate distributions, there is no closed-form solution for the covariance/scale matrices. Specifically, the value of one matrix depends on the value of the other from the previous iteration.

In the ECM framework, the observed data is treated as incomplete. This incompleteness arises from various sources: some are specific to the matrix-variate distribution under consideration, while others are common across all models. The shared source of incompleteness stems from the absence of information regarding each unit's state membership and its evolution over time. Therefore, it is necessary to define the unobserved state membership $\mathbf{z}_{it} = \left(z_{it1}, \ldots, z_{itk}, \ldots, z_{itK}\right)'$ and the unobserved state transitions

$$\mathbf{zz}_{it} = \begin{bmatrix} zz_{it11} & \cdots & zz_{it1k} & \cdots & zz_{it1K} \\ \vdots & & \vdots & & \vdots \\ zz_{itj1} & \cdots & zz_{itjk} & \cdots & zz_{itjK} \\ \vdots & & \vdots & & \vdots \\ zz_{itK1} & \cdots & zz_{itKk} & \cdots & zz_{itKK} \end{bmatrix},$$

where

$$z_{itk} = \begin{cases} 1 & \text{if } S_{it} = k \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad zz_{itjk} = \begin{cases} 1 & \text{if } S_{it-1} = j \text{ and } S_{it} = k \\ 0 & \text{otherwise} \end{cases}.$$

Taking this source of incompleteness into account, a first-level complete-data log-likelihood can be formulated as follows

$$l(\boldsymbol{\Theta}) = \sum_{i=1}^{I}\sum_{k=1}^{K} z_{i1k} \log\left(\pi_k\right) + \sum_{i=1}^{I}\sum_{t=2}^{T}\sum_{k=1}^{K}\sum_{j=1}^{K} zz_{itjk} \log\left(\pi_{k|j}\right) + l_d(\boldsymbol{\theta}), \tag{7}$$

where $l_d(\boldsymbol{\theta})$ is distribution-dependent and $\boldsymbol{\theta}$ includes all its parameters; this will be addressed in more detail in the following sections. Notice that, in the algorithms below, parameters with a single dot indicate updates from the previous iteration, while parameters with two dots denote updates from the current iteration.

3.1. MVT-HMM

When the pdf in (2) is considered, then a further source of incompleteness arises from the fact that the MVT distribution can be written as a scale mixture of MVN distributions whose row (or column)-covariance matrix is scaled by the reciprocal of a gamma

distribution (Doğru et al., 2016). In practice, for each $\mathbf{X}_{it}$ in the state k, this source of incompleteness is denoted by $W_{itk} \sim$ Gamma $(\nu_k/2, \nu_k/2)$. This leads the writing of $l_d(\boldsymbol{\theta})$ in (7) in terms of a second-level complete-data log-likelihood as

$$l_d(\boldsymbol{\theta}) = l_{d1}(\boldsymbol{\Phi}) + l_{d2}(\boldsymbol{\nu}),$$

where

$$\begin{aligned} l_{d1}(\boldsymbol{\Phi}) &= \sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{k=1}^{K} z_{itk}\left[-\frac{PR}{2}\ln(2\pi) - \frac{R}{2}\ln|\boldsymbol{\Sigma}_k| - \frac{P}{2}\ln|\boldsymbol{\Psi}_k| + \frac{PR}{2}\ln(w_{itk})\right. \\ &\quad \left. -\frac{w_{itk}}{2}\delta_k(\mathbf{X}_{it};\mathbf{M}_k,\boldsymbol{\Sigma}_k,\boldsymbol{\Psi}_k)\right], \\ l_{d2}(\boldsymbol{\nu}) &= \sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{k=1}^{K} z_{itk}\left\{\frac{\nu_k}{2}\ln\left(\frac{\nu_k}{2}\right) + \left(\frac{\nu_k}{2}-1\right)\ln(w_{itk}) - \frac{\nu_k}{2}w_{itk}\right. \\ &\quad \left. -\ln\left[\Gamma\left(\frac{\nu_k}{2}\right)\right]\right\}, \end{aligned}$$

with $\boldsymbol{\Phi} = \{\mathbf{M}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Psi}_k; k = 1, \ldots, K\}$, $\boldsymbol{\nu} = \{\nu_k; k = 1, \ldots, K\}$, and $\boldsymbol{\theta} = \{\boldsymbol{\Phi}, \boldsymbol{\nu}\}$.

E-Step. The E-step involves calculating the conditional expectation of (7), given the data and the current estimates of $\dot{\boldsymbol{\Theta}}$. As part of this process, z_{itk} and zz_{itjk} must be replaced with their conditional expectations, denoted as $\ddot{z}_{itk}$ and $\ddot{zz}_{itjk}$, respectively. This can be efficiently achieved using a forward recursion approach (Baum et al., 1970; Welch, 2003). More precisely, let

$$\gamma_{itk} = \Pr\left(\mathcal{X}_{i1} = \mathbf{X}_{i1}, \ldots, \mathcal{X}_{it} = \mathbf{X}_{it}, S_{it} = k\right), \tag{8}$$

represent the forward probability, i.e., the probability of observing the partial sequence ending in state k at time t. Similarly, the corresponding backward probability is defined as

$$\beta_{itk} = \Pr\left(\mathcal{X}_{it+1} = \mathbf{X}_{it+1}, \ldots, \mathcal{X}_{iT} = \mathbf{X}_{iT} | S_{it} = k\right). \tag{9}$$

Then, the necessary updates can be calculated as follows

$$\ddot{z}_{itk} = \frac{\gamma_{itk}\beta_{itk}}{\sum_{h=1}^{K}\gamma_{ith}\beta_{ith}} \quad \text{and} \quad \ddot{zz}_{itjk} = \frac{\gamma_{i(t-1)j}f(\cdot)\beta_{itk}}{\sum_{h=1}^{K}\gamma_{iTh}}, \tag{10}$$

where $f(\cdot)$ is the pdf in (2). It is worth noting that the computation of (8) and (9) is susceptible to numerical overflow errors. To decrease the risk of such errors, the scaling procedure discussed in Zucchini et al. (2017) is generalized to the matrix-variate longitudinal setting and accordingly implemented in the `Eigen.HMM fit()` function, whose usage will be disclosed in Section 4.2.

In addition to (10), the following quantities must be computed

$$\begin{aligned} \ddot{w}_{itk} &= E(W_{itk}|\mathbf{X}_{it}, \mathbf{z}_{it}, \mathbf{zz}_{it}; \boldsymbol{\Theta}) = \frac{PR + \dot{\nu}_k}{\dot{\nu}_k + \dot{\delta}_k(\mathbf{X}_{it};\dot{\mathbf{M}}_k,\dot{\boldsymbol{\Sigma}}_k,\dot{\boldsymbol{\Psi}}_k)}, \\ \ddot{m}_{itk} &= E[\ln(W_{itk})|\mathbf{X}_{it}, \mathbf{z}_{it}, \mathbf{zz}_{it}; \boldsymbol{\Theta}] = \\ &= \varphi\left(\frac{PR + \dot{\nu}_k}{2}\right) - \ln\left\{\frac{1}{2}\left[\dot{\nu}_k + \dot{\delta}_k(\mathbf{X}_{it};\dot{\mathbf{M}}_k,\dot{\boldsymbol{\Sigma}}_k,\dot{\boldsymbol{\Psi}}_k)\right]\right\}, \end{aligned}$$

with $\varphi(\cdot)$ denoting the digamma function.

After the E-Step, two CM steps follow, each corresponding to the partition of $\boldsymbol{\Theta}$ into $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$. Specifically, $\boldsymbol{\Theta}_1 = \left\{\boldsymbol{\Pi}, \pi_k, \mathbf{M}_k, \boldsymbol{\Sigma}_k, \nu_k; k = 1, \ldots, K\right\}$ and $\boldsymbol{\Theta}_2 = \left\{\boldsymbol{\Psi}_k; k = 1, \ldots, K\right\}$.

CM-Step 1. At the first CM-step, the expectation of (7) is maximized with respect to $\boldsymbol{\Theta}_1$, fixing $\boldsymbol{\Theta}_2$ at $\dot{\boldsymbol{\Theta}}_2$. Thus, we have the following updates

$$\ddot{\pi}_k = \frac{\sum_{i=1}^{I}\ddot{z}_{i1k}}{I}, \quad \ddot{\pi}_{k|j} = \frac{\sum_{i=1}^{I}\sum_{t=2}^{T}\ddot{zz}_{itjk}}{\sum_{i=1}^{I}\sum_{t=2}^{T}\sum_{k=1}^{K}\ddot{zz}_{itjk}}, \tag{11}$$

$$\ddot{\mathbf{M}}_k = \frac{\sum_{i=1}^{I}\sum_{t=1}^{T}\ddot{z}_{itk}\ddot{w}_{itk}\mathbf{X}_{it}}{\sum_{i=1}^{I}\sum_{t=1}^{T}\ddot{z}_{itk}\ddot{w}_{itk}}. \tag{12}$$

The update of $\boldsymbol{\Sigma}_k$, depends on the specific parsimonious structure considered. To avoid making this section excessively long, the results of Tomarchio et al. (2022b) are referenced for details on each parsimonious structure and update, as they are similar to those required here. The key difference lies in the definition of the update related to the row scatter matrix of the kth state, which is now defined as

$$\ddot{\mathbf{U}}_k = \sum_{i=1}^{I}\sum_{t=1}^{T} \ddot{z}_{itk}\ddot{w}_{itk}\left(\mathbf{X}_{it} - \ddot{\mathbf{M}}_k\right)\dot{\boldsymbol{\Psi}}_k^{-1}\left(\mathbf{X}_{it} - \ddot{\mathbf{M}}_k\right)'.$$

Thus, for illustrative purposes, only the update for the VVV parsimonious structure is provided, which is

$$\ddot{\boldsymbol{\Sigma}}_k = \frac{\ddot{\mathbf{U}}_k}{R\sum_{i=1}^{I}\sum_{t=1}^{T}\ddot{z}_{itk}}. \tag{13}$$

Lastly, a closed-form solution is not analytically available for updating ν_k. Thus, it must be numerically estimated by solving the following equation

$$\ln\left(\frac{\nu_k}{2}\right) + 1 - \varphi\left(\frac{\nu_k}{2}\right) + \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{z}_{itk}(\ddot{m}_{itk} - \ddot{w}_{itk})}{\sum_{i=1}^{N}\sum_{t=1}^{T}\ddot{z}_{itk}} = 0.$$

Computationally, the numerical search for the solution to the above equation is performed using the `optimize()` function from the **stats** package in R.

CM-Step 2. In the second CM-step, the expectation of (7) is maximized with respect to $\boldsymbol{\Theta}_2$, keeping $\boldsymbol{\Theta}_1$ fixed at $\ddot{\boldsymbol{\Theta}}_1$. The update for $\boldsymbol{\Psi}_k$, varies depending on which of the seven parsimonious structures is applied. Detailed information on each parsimonious structure and its corresponding update can be found in Tomarchio et al. (2022b), due to similarities with those required for the proposed models. The primary difference lies in the definition of the update related to the column scatter matrix of the kth state, now defined as

$$\ddot{\mathbf{V}}_k = \sum_{i=1}^{I}\sum_{t=1}^{T} \ddot{z}_{itk}\ddot{w}_{itk}\left(\mathbf{X}_{it} - \ddot{\mathbf{M}}_k\right)'\ddot{\boldsymbol{\Sigma}}_k^{-1}\left(\mathbf{X}_{it} - \ddot{\mathbf{M}}_k\right).$$

For the sake of simplicity, only the update for the VV case is presented, i.e.,

$$\ddot{\boldsymbol{\Psi}}_k = \frac{\ddot{\mathbf{V}}_k}{|\ddot{\mathbf{V}}_k|^{\frac{1}{R}}}.$$

3.2. MVCN-HMM

When the pdf in (3) is used, another source of incompleteness emerges since the MVCN distribution can be expressed as a scale mixture of MVN distributions whose row (or column)-covariance matrix is scaled by a Bernoulli distribution (Gupta et al., 2013; Tomarchio et al., 2022a). In practice, for each $\mathbf{X}_{it}$ in the state k, this incompleteness is represented by $V_{itk} \sim$ Bernoulli (α_k). Consequently, $l_d(\boldsymbol{\theta})$ in (7) can be formulated in terms of a second-level complete-data log-likelihood as

$$l_d(\boldsymbol{\theta}) = l_{d1}(\boldsymbol{\Omega}) + l_{d2}(\boldsymbol{\alpha}),$$

where

$$\begin{aligned} l_{d1}(\boldsymbol{\Omega}) = \sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{k=1}^{K} z_{itk}\Big[&-\frac{R}{2}\ln|\boldsymbol{\Sigma}_k| - \frac{P}{2}\ln|\boldsymbol{\Psi}_k| - \frac{PR}{2}(1-v_{itk})\ln(\eta_k) \\ &-\frac{1}{2}(v_{itk} + \frac{1-v_{itk}}{\eta_k})\delta_k\left(\mathbf{X}_{it};\mathbf{M}_k,\boldsymbol{\Sigma}_k,\boldsymbol{\Psi}_k\right)\Big], \\ l_{d2}(\boldsymbol{\alpha}) = \sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{k=1}^{K} z_{itk}\left[v_{itk}\ln(\alpha_k) + (1-v_{itk})\ln(1-\alpha_k)\right], \end{aligned}$$

with $\boldsymbol{\Omega} = \{\mathbf{M}_k, \boldsymbol{\Sigma}_k, \boldsymbol{\Psi}_k, \eta_k; k = 1,\ldots,K\}$, $\boldsymbol{\alpha} = \{\alpha_k; k = 1,\ldots,K\}$, and $\boldsymbol{\theta} = \{\boldsymbol{\Omega}, \boldsymbol{\alpha}\}$.

E-Step. The E-step involves calculating the conditional expectation of (7), given the data and the current estimates of $\dot{\boldsymbol{\Theta}}$. As part of this process, the same updates reported in (10) must be considered, where now $f(\cdot)$ is the pdf in (3). Furthermore, the following calculations need to be performed

$$\ddot{v}_{itk} = E(V_{itk}|\mathbf{X}_{it}, \mathbf{z}_{it}, \mathbf{z}\mathbf{z}_{it}; \boldsymbol{\Theta}) = \frac{\dot{\alpha}_k\phi(\mathbf{X}_{it};\dot{\mathbf{M}}_k,\dot{\boldsymbol{\Sigma}}_k,\dot{\boldsymbol{\Psi}}_k)}{\xi(\mathbf{X}_{it};\dot{\mathbf{M}}_k,\dot{\boldsymbol{\Sigma}}_k,\dot{\boldsymbol{\Psi}}_k,\dot{\eta}_k)}, \tag{14}$$

with $\xi(\cdot)$ denoting the pdf in (3).

As before, after the E-Step, two CM steps follow, due to the partition of $\boldsymbol{\Theta}$ into $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$. Specifically, $\boldsymbol{\Theta}_1 = \left\{\boldsymbol{\Pi}, \pi_k, \mathbf{M}_k, \boldsymbol{\Sigma}_k, \alpha_k; k = 1,\ldots,K\right\}$ and $\boldsymbol{\Theta}_2 = \left\{\boldsymbol{\Psi}_k, \eta_k; k = 1,\ldots,K\right\}$.

CM-Step 1. At the first CM-step, the expectation of (7) is maximized with respect to $\boldsymbol{\Theta}_1$, fixing $\boldsymbol{\Theta}_2$ at $\dot{\boldsymbol{\Theta}}_2$. The same updates reported in (11) are calculated. Then, we have

$$\ddot{\mathbf{M}}_k = \frac{\sum_{i=1}^{I}\sum_{t=1}^{T} \ddot{z}_{itk}\left(\ddot{v}_{itk} + \frac{1-\ddot{v}_{itk}}{\dot{\eta}_k}\right)\mathbf{X}_{it}}{\sum_{i=1}^{I}\sum_{t=1}^{T} \ddot{z}_{itk}\left(\ddot{v}_{itk} + \frac{1-\ddot{v}_{itk}}{\dot{\eta}_k}\right)},$$
$$\ddot{\alpha}_k = \max\left\{\alpha_k^{\min}, \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{z}_{itk}\ddot{v}_{itk}}{\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{z}_{itk}}\right\}, \tag{15}$$

where $\alpha_k^{\min} = 0.50$. This follows the common assumption in robust statistics that at least half of the observations are typical (Punzo and McNicholas, 2016; Tomarchio and Punzo, 2020; Tomarchio et al., 2025).

As before, the update of $\boldsymbol{\Sigma}_k$, depends on the specific parsimonious structure considered. We continue to refer to Tomarchio et al. (2022b) for details on each parsimonious structure and its corresponding update, as they closely resemble those required here. The primary distinction is in the definition of the update associated with the row scatter matrix of the kth state, which is now given by

$$\ddot{\mathbf{U}}_k = \sum_{i=1}^{I}\sum_{t=1}^{T} \ddot{z}_{itk}\left(\ddot{v}_{itk} + \frac{1-\ddot{v}_{itk}}{\dot{\eta}_k}\right)\left(\mathbf{X}_{it} - \ddot{\mathbf{M}}_k\right)\dot{\boldsymbol{\Psi}}_k^{-1}\left(\mathbf{X}_{it} - \ddot{\mathbf{M}}_k\right)'.$$

As before, for illustrative purposes, only the update for the VVV parsimonious structure is provided, which is

$$\ddot{\boldsymbol{\Sigma}}_k = \frac{\ddot{\mathbf{U}}_k}{R\sum_{i=1}^{I}\sum_{t=1}^{T} \ddot{z}_{itk}}.$$

CM-Step 2. In the second CM-step, the expectation of (7) is maximized with respect to $\boldsymbol{\Theta}_2$, keeping $\boldsymbol{\Theta}_1$ fixed at $\ddot{\boldsymbol{\Theta}}_1$. The update for $\boldsymbol{\Psi}_k$, varies depending on which of the seven parsimonious structures is applied. As previously, detailed information on each parsimonious structure and its corresponding update can be found in Tomarchio et al. (2022b), as they closely resemble those required for the proposed models. The main distinction lies in the definition of the update for the column scatter matrix of the kth state, which is now given by

$$\ddot{\mathbf{V}}_k = \sum_{i=1}^{I}\sum_{t=1}^{T} \ddot{z}_{itk}\left(\ddot{v}_{itk} + \frac{1-\ddot{v}_{itk}}{\dot{\eta}_k}\right)\left(\mathbf{X}_{it} - \ddot{\mathbf{M}}_k\right)'\ddot{\boldsymbol{\Sigma}}_k^{-1}\left(\mathbf{X}_{it} - \ddot{\mathbf{M}}_k\right).$$

For the sake of simplicity, only the update for the VV case is presented, i.e.,

$$\ddot{\boldsymbol{\Psi}}_k = \frac{\ddot{\mathbf{V}}_k}{|\ddot{\mathbf{V}}_k|^{\frac{1}{R}}}.$$

Finally, we also have to calculate

$$\ddot{\eta}_k = \max\left\{\eta_k^{\min}, \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{z}_{itk}(1-\ddot{v}_{itk})\delta_k\left(\mathbf{X}_{it}; \ddot{\mathbf{M}}_k, \ddot{\boldsymbol{\Sigma}}_k, \ddot{\boldsymbol{\Psi}}_k\right)}{PR\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{z}_{ik}(1-\ddot{v}_{itk})}\right\},$$

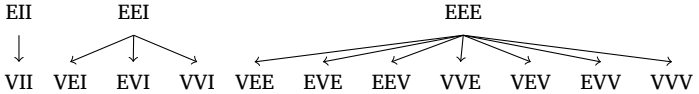
where $\eta_k^{\min}$ is the minimum value for $\ddot{\eta}_k$; herein, $\eta_k^{\min} = 1.0001$ is set in the fashion of Tomarchio et al. (2022a).

4. Computational and operational aspects

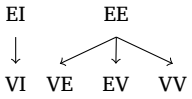
4.1. Initialization strategy

Selecting appropriate initial values is a crucial step in HMM estimation, as poor initialization can lead to convergence to local maxima and inaccurate results. In this paper, a short-EM initialization strategy is considered (Biernacki et al., 2003), which involves running the algorithm for a few numbers of iterations (n_1) from various random starting points (S) and selecting the parameter set yielding the highest log-likelihood. While this approach improves the variability of initial values, it can be computationally demanding, particularly when fitting a large set of models, as is the case with the parsimonious structures of the covariance/scale matrices.

To mitigate this issue, the strategy proposed in Tomarchio et al. (2024) is implemented, which applies the short-EM initialization only to a subset of parsimonious models. Specifically, the following scheme is adopted to initialize $\boldsymbol{\Sigma}_k$:



Similarly, the following scheme is employed to initialize $\boldsymbol{\Psi}_k$:



In other terms, initial values are provided exclusively for the six simplest parsimonious structures within each of the families detailed in Table 1 and Table 2: EII, EEI, EEE, EI, EE, and II (the identity matrix). Combining these structures for the two covariance/scale matrices results in nine possible initializations. Thus, the short-EM initialization is run only for this subset, and the resulting values are used to initialize the ECM algorithm of the remaining models following the hierarchical structure above. For example, to fit the EVI-VE MVT-HMM, the short-EM initialization for the EEI-EE MVT-HMM is used.

Note that, throughout the manuscript, $S = 50$ and $n_1 = 1$ are considered.

4.2. Detection of atypical matrices

A key feature of the proposed models is their ability to identify mildly atypical observations, which is particularly relevant given the challenges in visualizing matrix-variate data. The detection methods differ between MVT-HMMs and MVCN-HMMs. Specifically, for MVT-HMMs, the approach of Peel and McLachlan (2000) (see, also Greselin and Ingrassia, 2010; Punzo et al., 2021; Tomarchio and Gallaugher, 2024) is extended in the matrix-variate HMM framework. In contrast, MVCN-HMMs use a procedure enabled by the distinctive characteristics of the MVCN distribution (Tomarchio et al., 2022a).

Regardless of the considered model, each observation $\mathbf{X}_{it}$ is first assigned to one of the K states at each time point using the maximum *a posteriori* probabilities (MAP) operator

$$\text{MAP}(\widehat{z}_{itk}) = \begin{cases} 1 & \text{if } \max_h \left\{\widehat{z}_{ith}\right\} \text{occurs in group } h = k, \\ 0 & \text{if otherwise,} \end{cases}$$

where $\widehat{z}_{itk}$ is the value obtained from (10) at convergence of the ECM algorithm. Then, for the MVT-HMMs, by calculating

$$\sum_{k=1}^{K} \text{MAP}(\widehat{z}_{itk})\, \delta_k \left(\mathbf{X}_{it}; \widehat{\mathbf{M}}_k, \widehat{\boldsymbol{\Sigma}}_k, \widehat{\boldsymbol{\Psi}}_k\right), \tag{16}$$

each $\mathbf{X}_{it}$ is labeled as atypical if (16) is sufficiently large, and typical otherwise. In (16), the hat denotes the estimated values of the parameters at the convergence of the algorithm.

To determine how large the statistic (16) must be to be considered sufficiently large, the obtained values are compared to selected percentiles ϵ of the chi-squared distribution with pr degrees of freedom. The chi-squared distribution is used to approximate the distribution of the squared Mahalanobis distances in (16). Herein, $\epsilon \in (0.95, 0.99, 0.999)$ is considered (Peel and McLachlan, 2000; Punzo and McNicholas, 2017; Tomarchio and Gallaugher, 2024).

Regarding MVCN-HMMs, let $\widehat{\upsilon}_{itk}$ be the value obtained from (14) at the convergence of the ECM algorithm. The commonly adopted decision rule when contaminated distributions are used (see, e.g., Punzo and McNicholas, 2016; Tomarchio and Punzo, 2020; Punzo et al., 2021), is to consider a point typical if $\widehat{\upsilon}_{itk} > 0.5$ and atypical otherwise.

To summarize, after classifying the observation into one of the K states, MVT-HMMs and MVCN-HMMs offer more detailed insights about the role of that observation within that state.

4.3. An overview of the MatrixHMM package

The ECM algorithms and strategies discussed in this paper have been implemented in the recently released R package **MatrixHMM**, available on the CRAN repository. This section provides a brief overview of its key functions.

The function `Eigen.HMM_init()` handles the initialization as described in Section 4.1, and its output must be passed to the `Eigen.HMM_fit()` function using the `init.par` argument. The function `Eigen.HMM_fit()` implements the ECM algorithms of Section 3 to fit MVT-HMMs and MVN-HMMs. Additionally, it supports fitting MVN-HMMs, which are based on the MVN distribution.

Once all models within a family have been fitted, the function `extract.bestM()` identifies the best-fitting model according to the Bayesian information criterion (BIC; Schwarz, 1978). By default, setting the `top` argument to 1 extracts the highest-ranked model, but the function can return additional models based on their BIC ranking.

The tools for detecting atypical matrices, discussed in Section 4.2, are implemented in the functions `atp.MVT()` and `atp.MVCN()` for MVT-HMM and MVCN-HMM, respectively. The `atp.MVT()` function requires the data, estimated parameters, state memberships, and a threshold value ϵ. In contrast, the `atp.MVCN()` function needs the data, estimated state memberships, and $\widehat{\upsilon}_{itk}$, which is provided through the `pgood` argument.

Additionally, the package includes the `r.HMM()` function, designed to simulate matrix-variate longitudinal observations for MVN-HMMs, MVT-HMMs, and MVCN-HMMs. Two simulated datasets (`simData` and `simData2`), complete the actual version of the package.

Table 3
Average MAEs of the parameter estimates for the MVT-HMM over the three values of T.

	$T=7$		$T=15$		$T=30$	
	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2
$\mathbf{M}_k$	0.0432	0.1046	0.0296	0.0748	0.0204	0.0537
$\mathbf{\Sigma}_k$	0.0315	0.1035	0.0206	0.0845	0.0136	0.0570
$\mathbf{\Psi}_k$	0.0285	0.0365	0.0188	0.0268	0.0131	0.0194
ν_k	0.3766	0.7437	0.2338	0.5284	0.1580	0.4258
$\pi_{1,2}$	0.0369		0.0393		0.0376	
$\mathbf{\Pi}$	0.0182		0.0122		0.0087	

Table 4
Average MAEs of the parameter estimates for the MVCN-HMM over the three values of T.

	$T=7$		$T=15$		$T=30$	
	Group 1	Group 2	Group 1	Group 2	Group 1	Group 2
$\mathbf{M}_k$	0.0435	0.1105	0.0287	0.0839	0.0193	0.0594
$\mathbf{\Sigma}_k$	0.0206	0.0934	0.0149	0.0637	0.0100	0.0400
$\mathbf{\Psi}_k$	0.0271	0.0390	0.0177	0.0275	0.0121	0.0189
α_k	0.0143	0.0279	0.0105	0.0232	0.0062	0.0140
η_k	1.6857	0.6008	1.0259	0.4223	0.5003	0.2010
$\pi_{1,2}$	0.0404		0.0397		0.0360	
$\mathbf{\Pi}$	0.0231		0.0168		0.0098	

5. Simulated data analyses

This section covers multiple aspects of the proposed models. Specifically, in Section 5.1, the algorithms are assessed on their ability to recover the parameters of the data-generating processes. Section 5.2 showcases the robustness of the models in estimating parameters when the data includes atypical matrices. Section 5.3 investigates the effectiveness of identifying these atypical matrices. In Section 5.4, the clustering of the proposed models is compared to those of matrix-variate mixture models. Lastly, Section 5.5 illustrates some drawbacks of vectorizing and using multivariate HMMs when dealing with matrix-variate data.

5.1. Parameter recovery assessment

In this section, the following values are set: $P=2$, $R=5$, $I=100$, $T\in\{7,15,30\}$, and $K=2$. For illustrative purposes, data are simulated from models with a VEE-VE parsimonious structure for the covariance/scale matrices. The parameters used for simulating the data are accordingly reported in the supplementary material.

Here, data are simulated using both MVT-HMM and MVCN-HMM. In detail, for each value of T and each model, 100 datasets are generated. Each dataset is then fitted by the corresponding data-generating model.

To evaluate the accuracy of parameter estimates, the mean absolute errors (MAEs) are calculated. Given the dimensions and the large number of parameters that need to be reported, an averaging approach is implemented (see, e.g., Farcomeni and Punzo, 2020 and Tomarchio et al., 2024). Specifically, for a generic parameter $\mathbf{\Phi}_k$ of dimensions $M\times N$, its MAE across A datasets is determined as follows

$$\text{MAE}(\widehat{\mathbf{\Phi}}_k)=\frac{1}{MN}\sum_{MN}\left[\frac{1}{A}\sum_{a=1}^{A}|\widehat{\mathbf{\Phi}}_{ka}-\mathbf{\Phi}_k|\right],\quad k=1,\dots,K, \tag{17}$$

where $\widehat{\mathbf{\Phi}}_{ka}$ refers to the estimate of $\mathbf{\Phi}_k$ for the ath dataset. In other terms, an average among the MAEs of the elements of each estimated parameter is calculated, thus shrinking the results into a single number. Note that, for the parameters π_k, ν_k, α_k, and η_k, which are not matrices, the MAE is calculated using the conventional formula, i.e., that indicated inside the square brackets in (17). Furthermore, given that $K=2$, the weights π_k are perfectly complementary, resulting in identical MAE for both components; thus the notation $\pi_{1,2}$ will be used.

Results are reported in Table 3 and Table 4 for MVT-HMM and MVCN-HMM, respectively.

Overall, across both models, the MAEs consistently decrease as T increases. This pattern is evident for all parameters, highlighting that both models enhance their estimation accuracy as more data becomes available. More in detail, both models exhibit low MAEs for all the parameters, except ν_k and η_k, i.e., those strictly related to tail heaviness. As noted in Tomarchio et al. (2020); Punzo and Bagnato (2021), and Gallaugher et al. (2022), parameters that govern tail weight are generally the most challenging to estimate. Additionally, the estimation of ν_k is further complicated by the lack of closed-form solutions and reliance on numerical methods. It is also important to recognize that parameters with larger magnitudes tend to have higher MAEs. Nevertheless, even for these parameters, there is a clear improvement in the estimation accuracy as the amount of available data grows.

In summary, the results demonstrate good consistency and accuracy in capturing the parameters of the underlying models.

Table 5
Average MAEs of the parameter estimates in Scenario S_1 when $T = 7$.

l	q	MVN-HMM		MVT-HMM		MVCN-HMM	
		$\mathbf{M}_1$	$\mathbf{M}_2$	$\mathbf{M}_1$	$\mathbf{M}_2$	$\mathbf{M}_1$	$\mathbf{M}_2$
	1%	0.0416	0.1063	0.0415	0.0981	0.0414	0.0967
1	4%	0.0420	0.2011	0.0421	0.1019	0.0419	0.0981
	8%	0.0428	0.3870	0.0437	0.1190	0.0428	0.1056
	1%	0.0385	0.1196	0.0385	0.0994	0.0383	0.0956
2	4%	0.0411	0.2046	0.0412	0.1040	0.0411	0.0982
	8%	0.0426	0.3508	0.0425	0.1083	0.0426	0.1012
	1%	0.0432	0.1408	0.0424	0.1021	0.0424	0.1003
3	4%	0.0418	0.2293	0.0417	0.1053	0.0418	0.0992
	8%	0.0439	0.3229	0.0438	0.1058	0.0439	0.0990
		$\boldsymbol{\pi}_{1,2}$		$\boldsymbol{\pi}_{1,2}$		$\boldsymbol{\pi}_{1,2}$	
	1%	0.0386		0.0387		0.0385	
1	4%	0.0464		0.0465		0.0424	
	8%	0.0536		0.0460		0.0417	
	1%	0.0453		0.0452		0.0449	
2	4%	0.0447		0.0447		0.0421	
	8%	0.0533		0.0532		0.0428	
	1%	0.0420		0.0421		0.0419	
3	4%	0.0445		0.0448		0.0442	
	8%	0.0561		0.0560		0.0457	
		$\boldsymbol{\Pi}$		$\boldsymbol{\Pi}$		$\boldsymbol{\Pi}$	
	1%	0.0193		0.0192		0.0193	
1	4%	0.0332		0.0287		0.0215	
	8%	0.0628		0.0415		0.0227	
	1%	0.0215		0.0216		0.0210	
2	4%	0.0358		0.0358		0.0260	
	8%	0.0601		0.0597		0.0322	
	1%	0.0191		0.0189		0.0188	
3	4%	0.0342		0.0339		0.0269	
	8%	0.0615		0.0614		0.0396	

5.2. Robust estimates evaluation

By using the same parsimonious structure and parameters as in Section 5.1, data are now simulated via the MVN-HMM. Atypical values are then introduced into the generated datasets in two distinct ways, creating two scenarios (S_1 and S_2). In particular, for each scenario, a percentage q of observations is randomly selected, and

- in S_1, for each selected observation, one of the P rows is chosen at random, and the values in that row are replaced with random numbers drawn from a uniform distribution over the interval $[l \cdot a, l \cdot b]$, where a and b are the minimum and maximum values observed in the simulated dataset, respectively, and l is a multiplicative factor.
- in S_2, for each selected observation, all its values are replaced with random numbers generated from a uniform distribution over the interval $[l \cdot a, l \cdot b]$.

Thus, in Scenario S_1 only one row of the selected observations is modified, while in Scenario S_2, the entire observation is replaced. The following values are set: $q \in \{1\%, 4\%, 8\%\}$ and $l \in \{1, 2, 3\}$. In other terms, q controls the amount of contamination, whereas l can be used to tune the abnormality level of these points.

For each combination of S, q, l and T (with the values of T specified in Section 5.1), 100 datasets are generated, and the MVN-HMM, MVT-HMM, and MVCN-HMM are then fitted. Thus, $2 \times 3^3 = 54$ total configurations are created, 5400 datasets are simulated, and 16200 models are fitted.

Table 5 and Table 6 summarize the results for the two scenarios when $T = 7$. Results for $T = 15$ and $T = 30$ exhibit similar trends and are provided in Tables 1 to 4 of supplementary material to avoid lengthening the manuscript. For comparability's sake, and following a similar approach to Punzo and Maruotti (2016), MAEs are calculated only for the parameters with consistent interpretations across models: $\mathbf{M}_k$, π_k, and $\boldsymbol{\Pi}$, $k = 1, \ldots, K$.

The results reveal that both q and l have a substantial impact on the MAEs of the parameter estimates. Increasing q leads to a noticeable increase in MAEs across all models, with the effect being more pronounced for higher abnormal levels ($l = 3$) compared to lower levels ($l = 1$), particularly in the estimates for $\mathbf{M}_1$ and $\mathbf{M}_2$.

Across the two scenarios, MAEs are consistently higher in S_2, reflecting the increased difficulty due to a more challenging contamination setting. While the differences between models are less pronounced in S_1, MVT-HMM and MVCN-HMM still tend to perform better than MVN-HMM, particularly in the estimation of $\mathbf{M}_2$. Indeed, while MVN-HMM does well at lower contamination levels (e.g.,

Table 6
Average MAEs of the parameter estimates in Scenario S_2 when $T = 7$.

l	q	MVN-HMM		MVT-HMM		MVCN-HMM	
		$\mathbf{M}_1$	$\mathbf{M}_2$	$\mathbf{M}_1$	$\mathbf{M}_2$	$\mathbf{M}_1$	$\mathbf{M}_2$
	1%	0.0409	0.1091	0.0408	0.0986	0.0407	0.0974
1	4%	0.0422	0.1944	0.0422	0.1078	0.0422	0.1001
	8%	0.0429	0.3002	0.0431	0.1115	0.0430	0.1017
	1%	0.0419	0.1313	0.0418	0.0985	0.0418	0.0943
2	4%	0.0391	0.2257	0.0393	0.1062	0.0391	0.0989
	8%	0.0439	0.3144	0.0425	0.1126	0.0425	0.1041
	1%	0.0397	0.1700	0.0399	0.1033	0.0397	0.0998
3	4%	0.0425	0.3072	0.0428	0.1053	0.0425	0.0991
	8%	0.2044	0.7411	0.0420	0.1092	0.0528	0.2550
		$\pi_{1,2}$		$\pi_{1,2}$		$\pi_{1,2}$	
	1%	0.0388		0.0389		0.0389	
1	4%	0.0497		0.0497		0.0493	
	8%	0.0547		0.0545		0.0497	
	1%	0.0359		0.0359		0.0357	
2	4%	0.0428		0.0428		0.0425	
	8%	0.0444		0.0444		0.0401	
	1%	0.0389		0.0386		0.0389	
3	4%	0.0436		0.0430		0.0430	
	8%	0.0961		0.0483		0.0622	
		$\mathbf{\Pi}$		$\mathbf{\Pi}$		$\mathbf{\Pi}$	
	1%	0.0207		0.0207		0.0204	
1	4%	0.0372		0.0371		0.0326	
	8%	0.0630		0.0623		0.0515	
	1%	0.0226		0.0226		0.0223	
2	4%	0.0371		0.0371		0.0340	
	8%	0.0621		0.0621		0.0538	
	1%	0.0201		0.0200		0.0202	
3	4%	0.0368		0.0364		0.0343	
	8%	0.0970		0.0632		0.0692	

Table 7
Average TPR and FPR values in Scenario S_1 when $T = 7$.

l	q	MVT-HMM						MVCN-HMM	
		$\epsilon = 0.95$		$\epsilon = 0.99$		$\epsilon = 0.999$			
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
	1%	0.9857	0.0585	0.9729	0.0133	0.9557	0.0014	0.8557	0.0002
1	4%	0.9843	0.0563	0.9671	0.0122	0.9261	0.0014	0.9257	0.0009
	8%	0.9800	0.0521	0.9586	0.0102	0.9173	0.0012	0.9402	0.0010
	1%	1.0000	0.0614	1.0000	0.0137	0.9986	0.0016	0.9943	0.0005
2	4%	0.9993	0.0550	0.9993	0.0116	0.9989	0.0013	0.9950	0.0005
	8%	0.9991	0.0466	0.9984	0.0091	0.9966	0.0010	0.9954	0.0001
	1%	1.0000	0.0621	1.0000	0.0131	1.0000	0.0015	0.9957	0.0004
3	4%	0.9996	0.0558	0.9996	0.0114	0.9996	0.0013	0.9975	0.0001
	8%	1.0000	0.0456	1.0000	0.0097	1.0000	0.0012	0.9993	0.0002

$q = 1\%$), it struggles as the contamination levels rise. In S_2, the robustness of MVT-HMM and MVCN-HMM becomes more evident, as they achieve lower MAEs than MVN-HMM even under severe contamination.

5.3. Atypical matrices detection

To evaluate the performance of the MVT-HMM and MVCN-HMM in detecting atypical matrices, the datasets analyzed in Section 5.2 are used. The true positive rate (TPR), which measures the proportion of correctly identified atypical matrices, and the false positive rate (FPR), which indicates the proportion of typical matrices incorrectly classified as atypical, are calculated. The average values of these metrics over the 100 simulated datasets for each configuration, are presented in Table 7 and Table 8 for the two scenarios when $T = 7$. As before, results for $T = 15$ and $T = 30$ are provided in the supplementary material (Tables 5 to 8).

When comparing the two scenarios, both models demonstrate an easier task in detecting atypical matrices in Scenario S_2, where all values in the matrix are atypical. In particular, both models show superior results in terms of TPR in Scenario S_2 compared to Scenario S_1. On the other hand, the FPR does not exhibit substantial or consistent differences between the two scenarios.

Table 8
Average TPR and FPR values in Scenario S_2 when $T = 7$.

l	q	MVT-HMM						MVCN-HMM	
		$\epsilon = 0.95$		$\epsilon = 0.99$		$\epsilon = 0.999$			
		TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
1	1%	1.0000	0.0585	1.0000	0.0126	0.9957	0.0014	0.9786	0.0023
	4%	1.0000	0.0525	0.9979	0.0106	0.9918	0.0011	0.9925	0.0008
	8%	0.9993	0.0445	0.9962	0.0095	0.9859	0.0009	0.9943	0.0009
2	1%	1.0000	0.0615	1.0000	0.0137	1.0000	0.0016	1.0000	0.0004
	4%	1.0000	0.0555	1.0000	0.0119	1.0000	0.0013	1.0000	0.0002
	8%	1.0000	0.0467	1.0000	0.0088	1.0000	0.0011	1.0000	0.0002
3	1%	1.0000	0.0648	1.0000	0.0148	1.0000	0.0018	1.0000	0.0003
	4%	1.0000	0.0572	1.0000	0.0121	1.0000	0.0011	1.0000	0.0032
	8%	1.0000	0.0461	1.0000	0.0088	1.0000	0.0009	0.9600	0.0217

When examining the effect of increasing T, it becomes evident that TPR tends to improve as T grows, with some instances showing a notable rise in detection capability. Simultaneously, FPR decreases as T increases, suggesting that both models are better at minimizing false positives when larger T values are used. This trend holds for both MVT-HMM and MVCN-HMM across different configurations q and l.

Regarding the influence of q, an increase in q generally results in a slight reduction in TPR for MVT-HMM. In contrast, MVCN-HMM often exhibits an improvement in TPR as q increases, with some exceptions in Scenario S_2 when $q = 8\%$ and $l = 3$, which represents the most contaminated configuration. For FPR, there is a slight downward trend with increasing q, particularly for MVCN-HMM, indicating improved performance in avoiding false positives.

In relation to l, TPR increases for both models as l grows, indicating that higher abnormality levels enhance detection capability. Conversely, FPR remains relatively stable as l increases, showing minimal sensitivity to this parameter for both models.

Considering the effect of ϵ for the MVT-HMM, there is a general trade-off: the TPR decreases as ϵ grows, while FPR improves. This suggests that the model becomes more conservative in predicting positives, reducing false positives at the expense of missing some true positives. This trend is consistently observed across all combinations of q and l.

When comparing the two models, the primary differences in TPR occur in Scenario S_1, where the MVT-HMM provides higher TPR values. However, in terms of FPR, the MVCN-HMM consistently outperforms the MVT-HMM, showing significantly lower FPR values across both scenarios. This indicates that MVCN-HMM is better at avoiding false alarms, making it a more reliable option in this regard.

In summary, the MVCN-HMM stands out as a robust model across both scenarios due to its consistently low FPR and relatively high TPR, making it effective at identifying atypical points while minimizing false alarms. On the other hand, the performance of the MVT-HMM is dependent on the selected ϵ. Based on the results, $\epsilon = 0.99$ emerges as the optimal choice, balancing a high TPR with a low FPR, making it an effective option for detecting atypical matrices while minimizing the risk of misclassifying typical matrices as atypical.

5.4. A comparison with matrix-variate mixtures

In this study, we compare the performance of our HMMs with that of matrix-variate mixture models from a clustering perspective. While HMMs naturally account for temporal dependencies within their structure, matrix-variate mixture models rely on a per-time clustering approach without explicitly modeling temporal dynamics.

To evaluate clustering accuracy, we use the Adjusted Rand Index (ARI; Hubert and Arabie, 1985), which measures the agreement between the true classification and the one predicted by the model. Note that, in the fashion of Punzo and McNicholas (2016) and Tomarchio et al. (2025), the ARI is calculated only by considering the true typical observations, i.e., the atypical points are not taken into consideration.

For the HMMs, we apply a maximum *a posteriori* estimation approach (Punzo et al., 2018; Tomarchio et al., 2024), where each unit is classified to one of the hidden states at each time point. In contrast, for each time point, MVT mixtures (MVT-M) and MVCN mixtures (MVCN-M) are fitted to the extracted three-way data having dimensions $P \times R \times I$. Thus, an ARI is computed for each time point $t \in \{1, \ldots, T\}$ for both approaches. However, for HMMs, the ARIs are obtained simultaneously in a single model fitting, whereas MVT-M and MVCN-M require separate fittings for each t.

We rely on the same datasets analyzed in Sections 5.2 and 5.3. The ARI values computed for each dataset are averaged to obtain a single value. The average values of these ARIs over the 100 simulated datasets for each configuration are presented in Table 9 and Table 10 for the two scenarios when $T = 7$. As before, results for $T = 15$ and $T = 30$ are provided in the supplementary material (Tables 9 to 12).

From the analysis of the results, we note that in Scenario S_1, HMM-based models consistently achieved perfect ARI values across all configurations. In contrast, mixture models showed slight degradation in performance as q and l grew, with MVT-M particularly showing declines in ARI as contamination levels increased.

For Scenario S_2, all models faced greater challenges, particularly as the proportion of atypical observations increased. Both MVT-HMM and MVCN-HMM displayed greater stability with increasing values of q and l, while mixture models, notably MVT-M, exhibited

Table 9
Average ARI values in Scenario S_1 when $T = 7$.

l	q	MVT-HMM	MVCN-HMM	MVT-M	MVCN-M
1	1%	1.0000	1.0000	1.0000	1.0000
	4%	1.0000	1.0000	1.0000	1.0000
	8%	1.0000	1.0000	0.9999	1.0000
2	1%	1.0000	1.0000	1.0000	1.0000
	4%	1.0000	1.0000	1.0000	1.0000
	8%	1.0000	1.0000	1.0000	1.0000
3	1%	1.0000	1.0000	1.0000	1.0000
	4%	1.0000	1.0000	0.9972	1.0000
	8%	1.0000	1.0000	0.9744	0.9971

Table 10
Average ARI values in Scenario S_2 when $T = 7$.

l	q	MVT-HMM	MVCN-HMM	MVT-M	MVCN-M
1	1%	0.9753	0.9734	0.9758	0.9753
	4%	0.9070	0.8942	0.8984	0.8970
	8%	0.8262	0.7990	0.8087	0.8006
2	1%	0.9737	0.9728	0.9751	0.9750
	4%	0.9037	0.8936	0.8981	0.8974
	8%	0.8132	0.7927	0.7699	0.7918
3	1%	0.9761	0.9742	0.9740	0.9761
	4%	0.9023	0.8941	0.8040	0.8879
	8%	0.7771	0.7961	0.5185	0.7505

Table 11
Summary of parsimonious structure detection, selected number of components K, and the corresponding number of parameters (# npar) for the MVT-HMM and MT-HMM models.

HMM	# Detected	# $K = 3$ Selected	# npar ($K = 3$)
VVV-VV MVT	100	100	242
EVE MT	50	45	697
VVE MT	50	44	699

substantial performance drops as q and l rose, with ARI values falling to as low as 0.4130 for $q = 8\%$ and $l = 3$ when $T = 30$. These results highlight the significant advantage of leveraging temporal dependencies with HMM-based models, particularly in scenarios involving substantial atypical points.

5.5. A comparison with multivariate HMMs

As introduced in Section 1, vectorizing matrix-variate data for subsequent use with multivariate models can introduce several disadvantages. Here, we demonstrate how this process increases the number of parameters and discuss its impact on model selection in the context of HMMs. To illustrate these effects, we simulate 100 datasets from an MVT-HMM using a VVV-VV parsimonious structure with $P = 4$, $R = 8$, $I = 100$, $T = 5$, and $K = 3$. The parameters used for generating the data are detailed in the supplementary material.

Each simulated dataset is first fitted using the MVT-HMM across all parsimonious structures. Next, the data are vectorized, and the multivariate t HMM (MT-HMM) parsimonious family is applied. Both families are fitted for $K \in \{1, \ldots, 4\}$, with model selection guided by the BIC. The results are summarized in Table 11.

As we can see, the number of parameters increases substantially due to the vectorization process, leading to significant over-parameterization in the multivariate models. While the BIC consistently identifies the true parsimonious structure and K for the matrix-variate model, for the multivariate HMM, the EVE and VVE parsimonious structures are equally detected, and in several cases, the wrong K is selected. These results highlight how vectorization can affect the results, and emphasize the importance of using matrix-variate models when dealing with such structured data.

Table 12
Parsimonious structure (Pars), number of states (K) and value of the information criterion (BIC) for the best among each competing model according to the BIC.

HMM	Pars	K	BIC
MVN	VVV-EE	9	-8322.55
MVT	EEV-EE	9	-8687.02
MVCN	EEV-VE	9	-8605.09

6. Real data analysis

6.1. Data presentation

Labor market dynamics across $I = 107$ Italian provinces are herein analyzed, focusing on $P = 3$ pivotal indicators: unemployment, employment, and inactivity rates. The unemployment rate quantifies the proportion of the labor force that is unemployed and actively seeking work. It serves as a primary indicator of economic health, reflecting both job availability and market efficiency in matching workers with opportunities. In contrast, the employment rate measures the fraction of the working-age population that is employed, offering a broader perspective on labor market engagement. This rate captures the extent to which the workforce is utilized, making it a vital measure of economic vitality.

The inactivity rate represents the percentage of the working-age population that is neither employed nor actively seeking employment. This group includes individuals engaged in activities such as education, retirement, or other non-labor market pursuits. The inactivity rate, alongside the unemployment and employment rates, provides a comprehensive view of labor market participation, helping to elucidate underlying socio-economic conditions (see, e.g., Carcillo and Grubb, 2006; Lauzadyte, 2007; Nieuwenhuis, 2022).

These variables are further disaggregated into $R = 4$ distinct age classes (15-24, 25-34, 35-49, 50-74), providing a refined understanding of the labor market across different demographic segments. This disaggregation is crucial, as labor market behaviors can vary significantly by age, influencing policy implications and socio-economic interpretations (see, e.g., Milner et al., 2014; Albæk, 2015).

The dataset utilized in this study is provided by the Italian National Institute of Statistics (ISTAT), the principal agency responsible for generating official statistics in Italy. The analysis covers the six years from 2018 to 2023, chosen to align with recent regulatory frameworks established by the European Parliament and Council (European Parliament and Council, 2019), thereby ensuring the inclusion of the most current data available. Consequently, the data is organized in a structure of a $3 \times 4 \times 107 \times 6$ array. Note that, to avoid the so-called boundary bias problem (Tomarchio and Punzo, 2020), data are mapped to the whole real line using the logit transformation.

By leveraging the HMM framework, it is possible to capture and analyze clustering dynamics over time. Specifically, HMMs enable the identification of latent states that represent different economic regimes or clusters, which can vary over time. In our context, these states may correspond to distinct labor market conditions within the Italian provinces, reflecting periods of high employment, rising unemployment, or increased inactivity rates. For instance, some provinces may transition between different latent states more frequently, indicating volatile labor market conditions, while others may remain in a stable state, suggesting more resilient economic structures. Additionally, our model incorporates heavy-tailed distributions, allowing for robust handling of atypicalities and extreme variations in the data—common occurrences in economic metrics. This feature is crucial for accurately capturing the full spectrum of labor market conditions, including atypical events.

6.2. Results

Data are fitted by the MVN-HMM, MVT-HMM, and MVCN-HMM for $K \in \{1, \dots, 10\}$ and the corresponding results are reported in Table 12.

In terms of the parsimonious structure of the covariance/scale matrices, there are notable differences among the models. While the detected MVN-HMM have an unconstrained structure for the row covariance matrices, the MVT-HMM and MVCN-HMM have greater constraints. The key distinction between MVT-HMM and MVCN-HMM lies in the variability allowed in the column scale matrices, with MVCN-HMM selecting a more flexible configuration.

Regarding the number of latent states, all models agree on detecting $K = 9$. However, the heavy-tailed models provide a significantly better fit than the MVN-HMM, as evidenced by the BIC values. This improvement is attributed to the additional flexibility these models offer in capturing tail behavior, leading to a more accurate representation of the data. Among them, the MVT-HMM demonstrates the best fit. Consequently, the subsequent analysis will concentrate exclusively on this model.

To gain a deeper understanding of the nature of the detected states, the estimated mean matrices are analyzed. These matrices are visualized using parallel coordinate plots in Fig. 1.

Distinct patterns in labor market dynamics appear across the states. Starting from the inactivity rates, the data generally reveal a high prevalence of inactivity among the youngest age group (15-24) across most states, with a notable decrease as age increases. Specifically, States 1 through 4 exhibit relatively high inactivity rates in the younger age group, which, however, decrease significantly

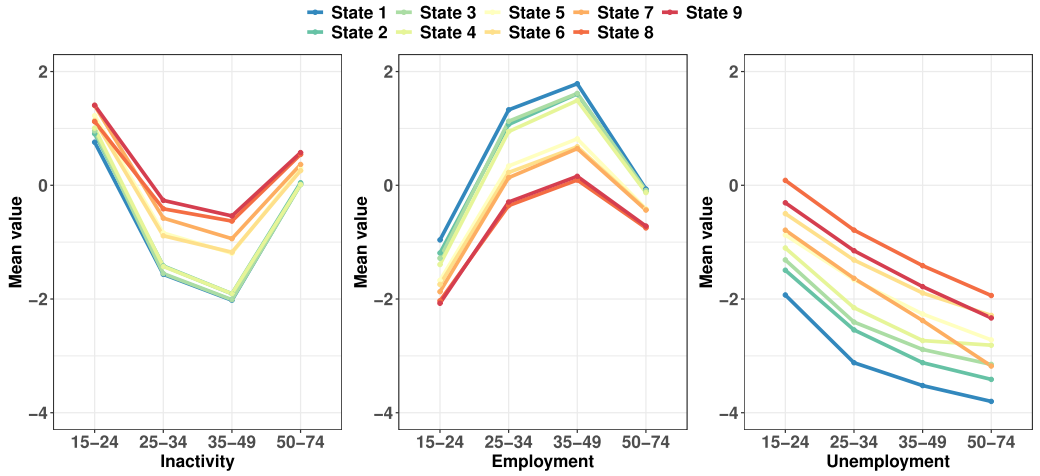


Fig. 1. Parallel coordinate plots of the mean matrices estimated by the EEV-EE MVT-HMM, shown across the three variables and distinguished by the estimated state.

in older age brackets. This trend indicates a more pronounced issue with youth inactivity in these states. Conversely, States 5 through 9 show even higher inactivity rates for the youngest age group, but with a more gradual decline across older age groups.

In terms of employment rates, the data suggest a general trend where employment levels improve with age. States 1 through 4 have relatively high values for the 25-49 age group, indicating stable employment conditions in mid-life, though the rates taper off for older individuals. In contrast, States 5 through 9 exhibit lower employment rates across all age groups, particularly for the youngest age group, with a less pronounced increase with age. This suggests that employment opportunities are more limited in these states, especially for younger individuals.

Regarding unemployment rates, the data indicate a decrease with age. States 1 through 4 show minimal unemployment, in contrast to States 5 through 9, which face more significant challenges as indicated by their relatively higher unemployment rates.

Overall, the data highlight heterogeneity across states, with younger populations experiencing higher rates of inactivity and unemployment, particularly in States 5-9. The disparity between younger and older age groups points to potential issues such as barriers to labor market entry or insufficient job opportunities for younger individuals.

The estimated scale matrices are now discussed. Starting with the row scale matrices, which characterize the relationships among the three labor market indicators, they are shown below

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.026 & -0.025 & 0.012 \\ -0.025 & 0.026 & -0.023 \\ 0.012 & -0.023 & 0.074 \end{bmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.032 & -0.029 & 0.014 \\ -0.029 & 0.031 & -0.026 \\ 0.014 & -0.026 & 0.063 \end{bmatrix},$$

$$\boldsymbol{\Sigma}_3 = \begin{bmatrix} 0.035 & -0.031 & 0.013 \\ -0.031 & 0.032 & -0.027 \\ 0.013 & -0.027 & 0.059 \end{bmatrix}, \quad \boldsymbol{\Sigma}_4 = \begin{bmatrix} 0.033 & -0.028 & 0.012 \\ -0.028 & 0.031 & -0.028 \\ 0.012 & -0.028 & 0.062 \end{bmatrix},$$

$$\boldsymbol{\Sigma}_5 = \begin{bmatrix} 0.033 & -0.030 & 0.012 \\ -0.030 & 0.034 & -0.028 \\ 0.012 & -0.028 & 0.059 \end{bmatrix}, \quad \boldsymbol{\Sigma}_6 = \begin{bmatrix} 0.035 & -0.032 & 0.010 \\ -0.032 & 0.037 & -0.029 \\ 0.010 & -0.029 & 0.053 \end{bmatrix},$$

$$\boldsymbol{\Sigma}_7 = \begin{bmatrix} 0.036 & -0.033 & 0.013 \\ -0.033 & 0.035 & -0.026 \\ 0.013 & -0.026 & 0.055 \end{bmatrix}, \quad \boldsymbol{\Sigma}_8 = \begin{bmatrix} 0.026 & -0.022 & 0.003 \\ -0.022 & 0.033 & -0.033 \\ 0.003 & -0.033 & 0.067 \end{bmatrix},$$

$$\boldsymbol{\Sigma}_9 = \begin{bmatrix} 0.028 & -0.026 & 0.008 \\ -0.026 & 0.033 & -0.030 \\ 0.008 & -0.030 & 0.065 \end{bmatrix}.$$

Across all states, we observe negative off-diagonal scale elements between employment (second row) and the other two indicators: inactivity (first row) and unemployment (third row). This pattern reflects an inverse association between these measures, consistent with the typical dynamics observed in labor markets. Notably, the magnitude of these scale elements varies across states, indicating different degrees of heterogeneity among provinces within each latent state.

Concerning the column scale matrices, which capture the dependency structure across the four age classes and are identical across all states, they are:

$$\boldsymbol{\Psi}_{1,\ldots,9} = \begin{bmatrix} 1.684 & 0.314 & 0.212 & 0.035 \\ 0.314 & 1.243 & 0.346 & 0.072 \\ 0.212 & 0.346 & 0.888 & 0.136 \\ 0.035 & 0.072 & 0.136 & 0.662 \end{bmatrix}.$$

The entries reveal a smooth pattern of association, with stronger connections between adjacent age classes and weaker ones between more distant cohorts. This reflects the gradual transitions in labor market behavior across the life cycle. The diagonal entries indicate that the youngest age group (15-24) displays the highest scale values, suggesting greater dispersion in labor market participation across provinces. This is economically plausible, as younger individuals typically face more fragmented and uncertain pathways in entering the labor market. Conversely, older cohorts (particularly 50-74) show lower dispersion, in line with more stable labor market statuses, such as permanent employment or retirement.

That said, transitions between the states are now investigated, given the valuable insights that can be gained. Specifically, the estimated transition probability matrix is

$$\boldsymbol{\Pi} = \begin{bmatrix} 1.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.17 & 0.58 & 0.22 & 0.03 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.03 & 0.30 & 0.46 & 0.21 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.04 & 0.36 & 0.60 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.04 & 0.00 & 0.02 & 0.71 & 0.06 & 0.17 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.45 & 0.55 & 0.00 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.23 & 0.00 & 0.77 & 0.00 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.75 & 0.25 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 0.07 & 0.05 & 0.02 & 0.87 \end{bmatrix}.$$

State 1 appears to be an absorbing state, with a probability of 1.00 of remaining in the same state and zero probability of transitioning to any other state. This could represent a labor market condition characterized by stability or persistence in certain economic characteristics. In particular, this state can be considered the more efficient from the point of view of the labor market, as seen in Fig. 1.

For the other states, transitions primarily occur between adjacent states, with minimal transitions among distant states. States 2, 3, 4, and 6 exhibit high degrees of mobility, indicating dynamic labor market conditions. Notably, three out of four of these states have relatively high labor market conditions (refer to Fig. 1.)

In contrast, States 5, 7, 8, and 9 demonstrate higher stability with fewer transitions. These states generally exhibit weaker labor market conditions, as illustrated in Fig. 1. In particular, State 9, considered the least favorable, has the second highest stability, underscoring its persistent labor market challenges. In summary, the transition probability matrix highlights significant heterogeneity and varying stability across the states.

To gain a graphical understanding of the detected states and their transitions over time, Fig. 2 presents the Italian provinces map for each year, with provinces colored according to their state memberships. From a geographical viewpoint, the northern provinces mainly belong to States 1 through 3. Given the characteristics of these states, it indicates that the labor market in the north is favorable and stable.

Central Italy shows a mix of States 2, 3, and 4. This suggests that labor market conditions in this area are generally favorable but feature high levels of mobility. The southern provinces and islands predominantly fall into States 5 through 9. These provinces have shown considerable persistence in these states, reflecting deep-rooted and severe labor market challenges.

From a temporal viewpoint, the initial period (2018–2019) shows a clear north-south divide, suggesting entrenched economic disparities. During 2020–2021, there are indications of transitions within the central regions, with certain provinces shifting between States 2, 3, and 4. This reflects moderate economic changes, likely influenced by broader national and global economic conditions, such as the COVID-19 pandemic. The later years (2022–2023) show a persistence of the established patterns, with the north maintaining its stability in more favorable states and the south remaining in less favorable states.

Finally, the provinces identified as atypical by the model are displayed separately for each year in Fig. 3. The provinces shown in white are those not classified as atypical, while the remaining provinces are colored according to their state membership.

The distribution of atypical provinces often varies across the country. The northern areas, generally associated with stable and favorable labor market conditions, exhibit relatively few atypical provinces, highlighting their consistent economic stability. In contrast, the central and southern parts of the country show a higher concentration of atypical provinces, particularly in 2022 and 2023, suggesting localized deviations from the broader economic trends within their states. The fluctuation in the number of atypical provinces over the years - from a low of 8 in 2018 to a peak of 19 in 2022 - indicates varying levels of instability, potentially driven by external factors affecting these areas.

6.2.1. *Results from alternative approaches*

As a first alternative approach, and in line with the discussion in Section 1, the 3×4 matrices can be rearranged into a 12-dimensional vector. This transformation results in a $12 \times 107 \times 6$ array, which can then be modeled using a multivariate HMM. To see the implications of such a choice, we fit the multivariate version of our proposed model to the data for $K \in \{1, \ldots, 10\}$.

The best-fitting model in this case is the EEV MT-HMM with $K = 6$. Notably, this model has 521 parameters–more than twice the number of parameters in the best matrix-variate HMM (EEV-EE MVT-HMM)–despite having fewer states. This inflation in the

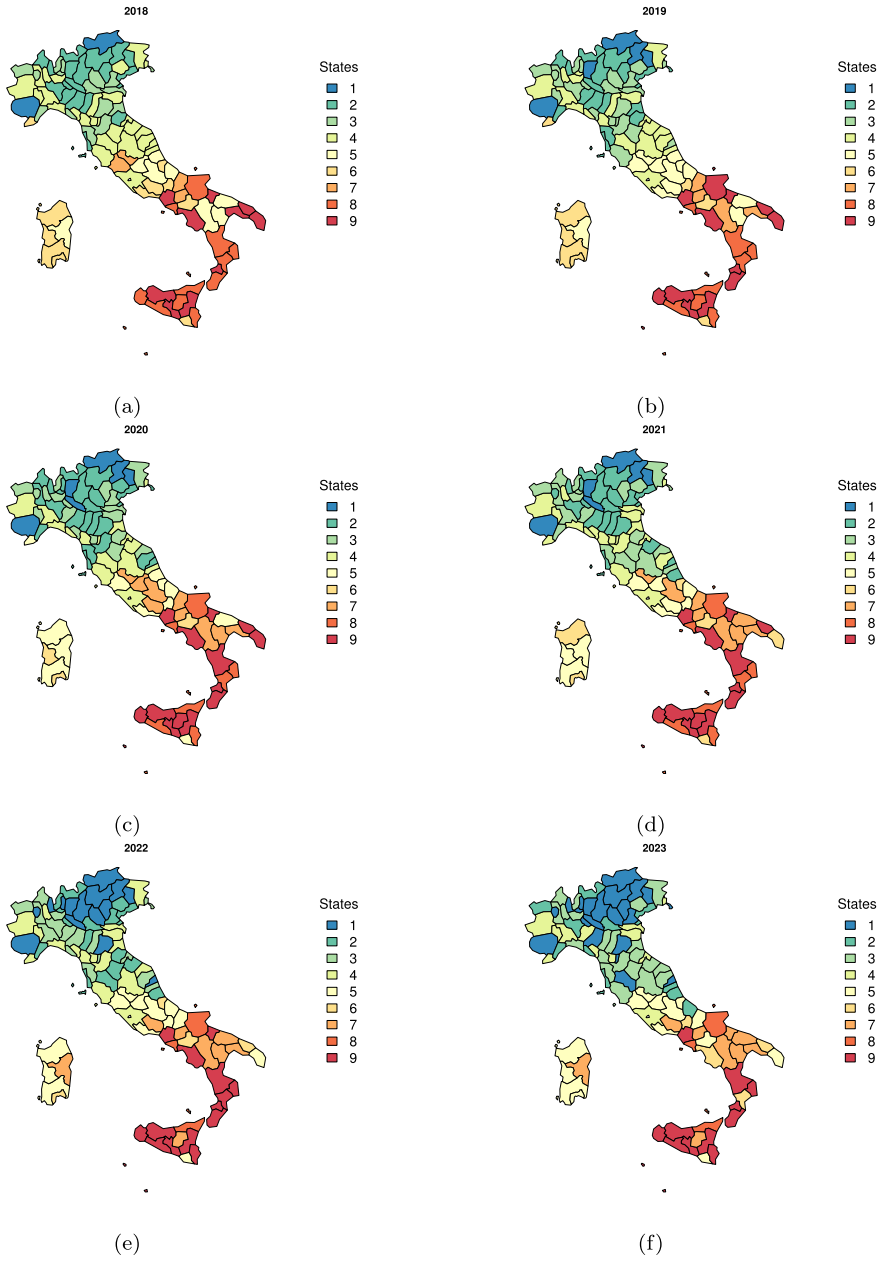


Fig. 2. Italian provinces colored according to the estimated state memberships by the EEV-EE MVT-HMM over the six years. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

number of parameters arises from the vectorization process, which dramatically increases model complexity, as also demonstrated in the simulation study of Section 5.

Furthermore, the clustering outcome differs significantly from that of the EEV-EE MVT-HMM. Comparing the classifications of the two models year by year using the ARI, the agreement ranges from 0.33 to 0.52, indicating a relatively weak correspondence. This discrepancy is partially driven by the differing numbers of states identified by the two models. For illustration, we visualize the clustering results for 2022 in Fig. 4. As seen, the clustering solution is a blurred version of that in Fig. 2, with many neighboring provinces grouped together. This confirms that the vectorization process leads to a loss of structural detail.

As a second alternative approach, we also considered a finite mixture model version of our proposed models. In particular, for each temporal point, a $3 \times 4 \times 107$ array is obtained, resulting in six separate datasets. Then, on each dataset, we fit mixture models for $K \in \{1, \dots, 10\}$.

In this case, the best-fitting models are consistently MVT-based mixtures. However, both the parsimonious structure and number of clusters vary across years, with K ranging from 6 to 10. Since this approach assumes no temporal dependency, clusters change

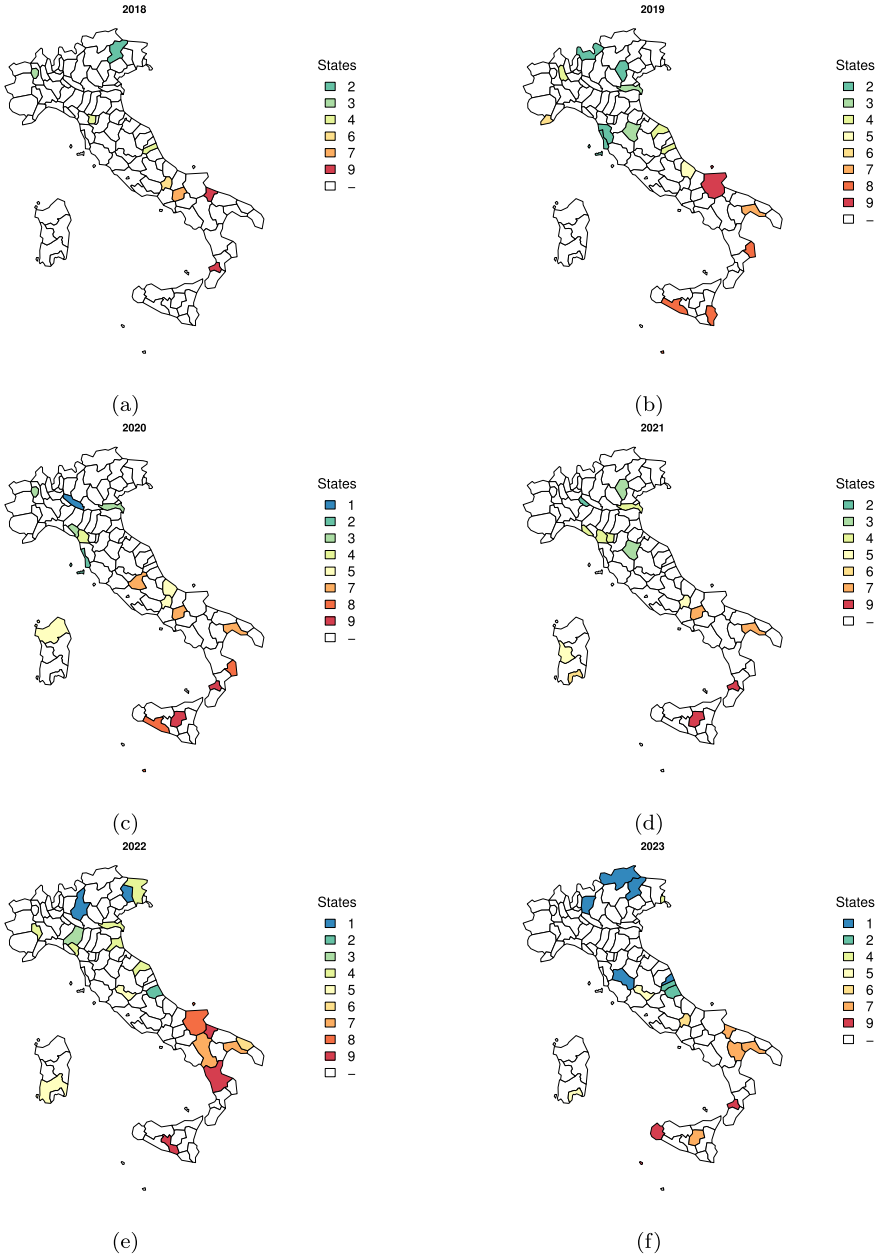


Fig. 3. Italian provinces labeled as atypical and colored according to the estimated state memberships by the EEV-EE MVT-HMM over the six years. Provinces in white are not identified as atypical in that year.

arbitrarily at each time point, making it difficult to interpret temporal patterns or track clusters over time. Conversely, HMMs provide transition probabilities between clusters, allowing us to model temporal dynamics and extract meaningful patterns, as discussed in Section 6.2.

Furthermore, the clustering outcomes from finite mixtures differ significantly from those of the EEV-EE MVT-HMM. Comparing their classifications year by year using ARI, the agreement ranges from 0.39 to 0.58, indicating a relatively limited correspondence. This further highlights the impact of explicitly modeling temporal dependencies in clustering.

7. Conclusions

In this manuscript, two families of parsimonious matrix-variate hidden Markov models (HMMs) have been introduced. Each family employs either the matrix-variate t (MVT) or matrix-variate contaminated normal (MVCN) distributions, aiming to enhance estimation

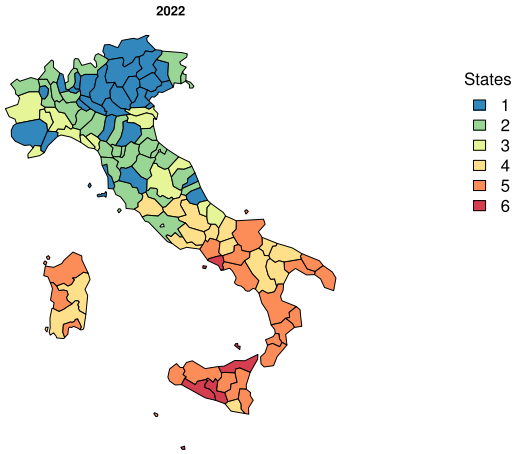


Fig. 4. Italian provinces colored according to the estimated state memberships by the EEV MT-HMM for 2022.

robustness compared to traditional methods based on normality assumptions. Two ECM algorithms for parameter estimation have been presented and implemented in the R package **MatrixHMM**, whose functionalities have been detailed.

Simulated analyses demonstrated the algorithms' capability to accurately recover the parameters of the data-generating models. The robustness of the estimates in the presence of atypical matrices has been discussed, with the proposed models showing superior performance compared to the normal-based HMM in terms of absolute squared errors. Additionally, the proposed models effectively detected atypical matrices, under varying configuration settings, providing promising results. Finally, two further simulation studies illustrated the advantages of using the proposed models against multivariate HMMs and matrix-variate mixtures.

The analysis of Italian labor market dynamics reveals persistent disparities, particularly between the northern and southern provinces. The northern provinces consistently exhibit more favorable and stable labor market conditions, falling mainly into States 1 through 3, characterized by higher employment and lower unemployment rates. In contrast, the central and southern provinces are predominantly categorized into States 5 through 9, reflecting higher unemployment and inactivity rates, particularly among younger populations.

Over the years 2018 to 2023, these disparities have remained pronounced, with the north maintaining stability while the south and parts of the central region continue to face significant labor market challenges. The analysis of atypical provinces further highlights these trends, with the north showing fewer deviations from expected economic patterns, while the south and central regions, especially in recent years, display higher frequencies of atypical provinces, indicating localized economic instability. These findings underscore the need for targeted policies to address labor market imbalances and support vulnerable areas in adapting to economic changes.

Concerning further developments of the present work, several avenues could be explored. While our study focuses on HMMs for matrix-variate data, it is worth noting that alternative Markovian models, such as Pairwise Markov Models (PMMs; Derrode and Pieczynski, 2004, 2013, 2016; Gorynin et al., 2018; Joumad et al., 2024) and Triplet Markov Models (TMMs; Pieczynski et al., 2003; Abbes et al., 2019; Chen and Jiang, 2020; Habbouchi et al., 2022; Gangloff et al., 2023), could provide further improvements. Given that these models have never been explored in a matrix-variate context, an interesting future direction could be their introduction and extension to heavy-tailed distributions such as the MVT and MVCN families considered in this study. These distributions have demonstrated their robustness in handling atypical observations, and incorporating them into more advanced Markovian frameworks may further improve clustering accuracy and interpretability.

Acknowledgements

Salvatore D. Tomarchio acknowledges the "PIAno di inCEntivi per la RIcerca di Ateneo 2020/2022", Linea di intervento 3 - STARTING GRANT 2020, University of Catania.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2025.108198.

References

Abbes, A.B., Farah, M., Farah, I.R., Barra, V., 2019. A non-stationary NDVI time series modelling using triplet Markov chain. Int. J. Inf. Dec. Sci. 11, 163–179.

Albæk, K., 2015. Youth Unemployment and Inactivity: A Comparison of School-to-Work Transitions and Labour Market Outcomes in Four Nordic Countries. Nordic Council of Ministers.

Anderlucci, L., Montanari, A., Viroli, C., et al., 2014. A matrix-variate regression model with canonical states: an application to elderly Danish twins. Statistica 74, 367–381.

Asilkalkan, A., Zhu, X., 2021. Matrix-variate time series modelling with hidden Markov models. Stat 10, e409.

Baum, L.E., Petrie, T., Soules, G., Weiss, N., 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. Ann. Math. Stat. 41, 164–171.
Biernacki, C., Celeux, G., Govaert, G., 2003. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. Comput. Stat. Data Anal. 41, 561–575.
Carcillo, S., Grubb, D., 2006. From Inactivity to Work: the Role of Active Labour Market Policies. Technical Report No. 36. Organisation for Economic Co-Operation and Development.
Celeux, G., Govaert, G., 1995. Gaussian parsimonious clustering models. Pattern Recognit. 28, 781–793.
Chen, S., Jiang, X., 2020. Modeling repayment behavior of consumer loan in portfolio across business cycle: a triplet Markov model approach. Complexity 2020, 5458941.
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, Methodol. 39, 1–22.
Derrode, S., Pieczynski, W., 2004. Signal and image segmentation using pairwise Markov chains. IEEE Trans. Signal Process. 52, 2477–2489.
Derrode, S., Pieczynski, W., 2013. Unsupervised data classification using pairwise Markov chains with automatic copulas selection. Comput. Stat. Data Anal. 63, 81–98.
Derrode, S., Pieczynski, W., 2016. Unsupervised classification using hidden Markov chain with unknown noise copulas and margins. Signal Process. 128, 8–17.
Doğru, F.Z., Bulut, Y.M., Arslan, O., 2016. Finite mixtures of matrix variate t distributions. Gazi Univ. J. Sci. 29, 335–341.
European Parliament and Council, 2019. Regulation (EU) 2019/1700. Off. J. Eur. Union, 1–32. http://data.europa.eu/eli/reg/2019/1700/oj.
Farcomeni, A., Punzo, A., 2020. Robust model-based clustering with mild and Gross outliers. Test 29, 989–1007.
Gallaugher, M.P., McNicholas, P.D., 2018. Finite mixtures of skewed matrix variate distributions. Pattern Recognit. 80, 83–93.
Gallaugher, M.P., Tomarchio, S.D., McNicholas, P.D., Punzo, A., 2022. Multivariate cluster weighted models using skewed distributions. Adv. Data Anal. Classif. 16, 93–124.
Gallaugher, M.P., Zhu, X., 2024. Modeling matrix variate time series via hidden Markov models with skewed emissions. Stat. Anal. Data Min. ASA Data Sci. J. 17, e11666.
Gangloff, H., Morales, K., Petetin, Y., 2023. Deep parameterizations of pairwise and triplet Markov models for unsupervised classification of sequential data. Comput. Stat. Data Anal. 180, 107663.
Gorynin, I., Gangloff, H., Monfrini, E., Pieczynski, W., 2018. Assessing the segmentation performance of pairwise and triplet Markov models. Signal Process. 145, 183–192.
Greselin, F., Ingrassia, S., 2010. Constrained monotone EM algorithms for mixtures of multivariate t distributions. Stat. Comput. 20, 9–22.
Gupta, A.K., Varga, T., Bodnar, T., et al., 2013. Elliptically Contoured Models in Statistics and Portfolio Theory. Springer.
Habbouchi, A., Boudaren, M.E.Y., Senouci, M.R., Aïssani, A., 2022. Markovian segmentation of non-stationary data corrupted by non-stationary noise. In: International Conference on Computing Systems and Applications. Springer, pp. 27–37.
Hossain, A., Naik, D., 1991. A comparative study on detection of influential observations in linear regression. Stat. Pap. 32, 55–69.
Hubert, L., Arabie, P., 1985. Comparing partitions. J. Classif. 2, 193–218.
Joumad, A., El Moutaouakkil, A., Nasroallah, A., Boutkhoum, O., Safran, M., Alfarhood, S., Ashraf, I., 2024. Unsupervised segmentation of images using bi-dimensional pairwise Markov chains model. AIMS Math. 9, 31057–31086.
Lauzadyte, A., 2007. Unemployment, employment and inactivity in Denmark: an analysis of event history data. University of Aarhus Economics Working Paper.
Ma, X., Zhao, J., Wang, Y., Shang, C., Jiang, F., 2023. Robust factored principal component analysis for matrix-valued outlier accommodation and detection. Comput. Stat. Data Anal. 179, 107657.
Melnykov, V., Zhu, X., 2019. Studying crime trends in the USA over the years 2000–2012. Adv. Data Anal. Classif. 13, 325–341.
Meng, X.L., Rubin, D.B., 1993. Maximum likelihood estimation via the ECM algorithm: a general framework. Biometrika 80, 267–278.
Milner, A., Morrell, S., LaMontagne, A.D., 2014. Economically inactive, unemployed and employed suicides in Australia by age and sex over a 10-year period: what was the impact of the 2007 economic recession? Int. J. Epidemiol. 43, 1500–1507.
Nieuwenhuis, R., 2022. No activation without reconciliation? The interplay between ALMP and ECEC in relation to women's employment, unemployment and inactivity in 30 OECD countries, 1985–2018. Soc. Policy Adm. 56, 808–826.
Peel, D., McLachlan, G.J., 2000. Robust mixture modelling using the t distribution. Stat. Comput. 10, 339–348.
Pieczynski, W., Hulard, C., Veit, T., 2003. Triplet Markov chains in hidden signal restoration. In: Image and Signal Processing for Remote Sensing VIII. SPIE, pp. 58–68.
Punzo, A., Bagnato, L., 2021. The multivariate tail-inflated normal distribution and its application in finance. J. Stat. Comput. Simul. 91, 1–36.
Punzo, A., Ingrassia, S., Maruotti, A., 2018. Multivariate generalized hidden Markov regression models with random covariates: physical exercise in an elderly population. Stat. Med. 37, 2797–2808.
Punzo, A., Ingrassia, S., Maruotti, A., 2021. Multivariate hidden Markov regression models: random covariates and heavy-tailed distributions. Stat. Pap. 62, 1519–1555.
Punzo, A., Maruotti, A., 2016. Clustering multivariate longitudinal observations: the contaminated gaussian hidden Markov model. J. Comput. Graph. Stat. 25, 1097–1098.
Punzo, A., McNicholas, P.D., 2016. Parsimonious mixtures of multivariate contaminated normal distributions. Biom. J. 58, 1506–1537.
Punzo, A., McNicholas, P.D., 2017. Robust clustering in regression analysis via the contaminated Gaussian cluster-weighted model. J. Classif. 34, 249–293.
Sarkar, S., Zhu, X., 2022. Finite mixture model of hidden Markov regression with covariate dependence. Stat 11, e469.
Sarkar, S., Zhu, X., Melnykov, V., Ingrassia, S., 2020. On parsimonious models for modeling matrix data. Comput. Stat. Data Anal. 142, 106822.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat., 461–464.
Tomarchio, S.D., 2024. Matrix-variate normal mean-variance Birnbaum–Saunders distributions and related mixture models. Comput. Stat. 39, 405–432.
Tomarchio, S.D., Gallaugher, M.P., 2024. Mixtures of regressions using matrix-variate heavy-tailed distributions. Adv. Data Anal. Classif., 1–28.
Tomarchio, S.D., Gallaugher, M.P., Punzo, A., McNicholas, P.D., 2022a. Mixtures of matrix-variate contaminated normal distributions. J. Comput. Graph. Stat. 31, 413–421.
Tomarchio, S.D., McNicholas, P.D., Punzo, A., 2021. Matrix normal cluster-weighted models. J. Classif. 38, 556–575.
Tomarchio, S.D., Punzo, A., 2020. Dichotomous unimodal compound models: application to the distribution of insurance losses. J. Appl. Stat. 47, 2328–2353.
Tomarchio, S.D., Punzo, A., 2025. On the number of components for matrix-variate mixtures: a comparison among information criteria. Int. Stat. Rev., 1–24.
Tomarchio, S.D., Punzo, A., Bagnato, L., 2020. Two new matrix-variate distributions with application in model-based clustering. Comput. Stat. Data Anal. 152, 107050.
Tomarchio, S.D., Punzo, A., Ferreira, J.T., Bekker, A., 2025. A new look at the Dirichlet distribution: robustness, clustering, and both together. J. Classif. 42, 31–53.
Tomarchio, S.D., Punzo, A., Maruotti, A., 2022b. Parsimonious hidden Markov models for matrix-variate longitudinal data. Stat. Comput. 32, 53.
Tomarchio, S.D., Punzo, A., Maruotti, A., 2024. Matrix-variate hidden Markov regression models: fixed and random covariates. J. Classif. 41, 429–454.
Viroli, C., 2011a. Finite mixtures of matrix normal distributions for classifying three-way data. Stat. Comput. 21, 511–522.
Viroli, C., 2011b. Model based clustering for three-way data structures. Bayesian Anal. 6, 573–602.
Viroli, C., 2012. On matrix-variate regression analysis. J. Multivar. Anal. 111, 296–309.
Welch, L.R., 2003. Hidden Markov models and the Baum-Welch algorithm. IEEE Inf. Theory Soc. Newsl. 53, 10–13.
Zucchini, W., MacDonald, I.L., Langrock, R., 2017. Hidden Markov Models for Time Series: An Introduction Using R. CRC Press.