



Multivariate hidden Markov regression models: random covariates and heavy-tailed distributions

Antonio Punzo¹ · Salvatore Ingrassia¹ · Antonello Maruotti^{2,3}

Received: 15 July 2018 / Revised: 3 August 2019 / Published online: 18 November 2019

© Springer-Verlag GmbH Germany, part of Springer Nature 2019

Abstract

Despite recent methodological advances in hidden Markov regression models and a rapid increase in their application in a wide range of empirical settings, complex clustering-based research questions that include the contribution of the covariates set to the classification and the presence of atypical observations are often addressed ignoring the possible effects of wrong model assumptions. Hidden Markov regression models with random covariates (HMRMRCs) have been recently proposed as an improvement over the classical fixed covariates approach, allowing the covariates to contribute to the underlying clustering structure. To make the approach more flexible, when all the considered random variables are continuous, HMRMRCs are here defined focusing on three multivariate elliptical distributions: the normal (reference distribution), the t , and the contaminated normal. The latter two, heavy-tailed generalizations of the normal distribution, are introduced to protect the reference model for the occurrence of mildly atypical points and also allow us their automatic detection. Identifiability conditions are provided, EM-based algorithms are outlined for parameter estimation, and various implementation and operational issues are discussed. Properties of the estimators of the regression coefficients, as well as of the hidden path parameters, are evaluated through Monte Carlo experiments with the aim of showing the consequences of wrong model assumptions on parameters estimates and inferred clustering. Artificial and real data analyses are provided to investigate models behavior in presence of heterogeneity and atypical observations.

Keywords Hidden Markov models · Multivariate outcome · Atypical observations · Clustering · Heavy-tailed distributions

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s00362-019-01146-3>) contains supplementary material, which is available to authorized users.

Extended author information available on the last page of the article

1 Introduction

In a regression setting, hidden Markov models (HMMs) are often defined through the inclusion of observed covariates to represent (fixed) effects shared by all units in the same hidden state (see, e.g., Maruotti 2014; Martinez-Zarzoso and Maruotti 2013). This yields to hidden Markov regression models with fixed covariates (HMRMFCs); see Sect. 2.1. However, the hidden structure often depends on the values taken by the covariates as well, and considering the covariate as fixed could lead to misleading inferential results. In such a situation, HMRMFCs fail as the assignment of the observations to the states ignores, and is independent of, the covariates distribution (*assignment independence*; Hennig 2000); see Sect. 2.2.

To overcome the problem, Punzo et al. (2018a) introduced hidden Markov regression models with random covariates (HMRMRCs). They fit the joint distribution of outcomes and covariates, in each hidden state of the HMM, as the product of the marginal distribution of the covariates and the conditional distribution of the outcomes given the covariates; this allows the random covariates to contribute to properly recover the latent structure in the data (*assignment dependence*; Hennig 2000), a crucial aspect of cluster analysis; see Sect. 2.2.

In this paper, by focusing on multivariate continuous outcomes and covariates, we improve HMRMRCs from a twofold point of view. Firstly, with respect to the formulation given in Punzo et al. (2018a), we relax the assumption of local (with respect to the states) independence about covariates, as well as outcomes given covariates, by considering the multivariate normal distribution, whose covariance matrix allows us to describe the pairwise dependence structure among variables. Secondly, by further considering two heavy-tailed elliptical generalizations of the normal distribution, namely the t and contaminated normal, we give to the resulting HMRMRCs the flexibility needed for achieving mildly atypical observations robustness, such that each state-specific joint density adheres to the commonly considered taxonomy of the observations in regression analysis as typical (bulk of the data), good leverage, bad leverage, and outlier (see, e.g., Rousseeuw and Leroy 2005, Chapter 1).

By combining the three considered multivariate distributions with respect to covariates, and outcomes given covariates, in Sect. 2.3 we introduce a novel family of nine different HMRMRCs. For each member of this family, in Sect. 3 we give sufficient conditions for the identifiability, with the proof postponed to Web Appendix A. We conveniently obtain estimates of the model parameters using maximum likelihood (ML). In Sect. 4 we outline the expectation maximization (EM) algorithm to deal with HMRMRCs based on multivariate normal and t distributions, and an *ad hoc* version of the expectation-conditional maximization (ECM) algorithm in the case of HMRMRCs based on the contaminated normal distribution. Regardless of the considered distribution, further general computational and operational details are illustrated in Sect. 5. In particular, in Sect. 5.2 we explain how robustness is automatically obtained with some of the models of our family and, for these models, we outline detection of mildly atypical observations too.

The rest of the paper is organized as follows. In Sect. 6, we test the proposed models by analyzing artificial and real data, focusing on the effect of ignoring the contribution of the covariates to clustering and of atypical observations on both parameters esti-

mation and obtained clustering partition. Indeed, different aspects of robustness are described and analyzed. We further illustrate the benefits of the proposed models in Online Appendix B by a large-scale simulation study. At last, in Sect. 7, we summarize the key aspects of the proposal along with future possible extensions.

2 Methodology

2.1 Preliminaries

A hidden Markov model (HMM) is a particular kind of (time-)dependent mixture. For each unit $i, i = 1, \dots, I$, an underlying unobserved process and a state-dependent process are defined (see Zucchini et al. 2016, Chapter 2). Formally, let $\{S_{it}; i = 1, \dots, I, t = 1, \dots, T\}$ be a first-order Markov chain defined on the discrete state space $\{1, \dots, k, \dots, K\}$. The process $\{S_{it}\}$, which represents the underlying unobserved process of the HMM, fulfills the Markov property

$$\Pr(S_{it} = s_{it} | S_{i1} = s_{i1}, \dots, S_{it-1} = s_{it-1}) = \Pr(S_{it} = s_{it} | S_{it-1} = s_{it-1}),$$

meaning that the state at any given time t depends on the previous states only through the state at time $t-1, t = 2, \dots, T$. The initial probabilities of the hidden Markov chain are denoted as $\pi_{ik} = \Pr(S_{i1} = k), k = 1, \dots, K$, while the transition probabilities as

$$\pi_{i,k|j} = \Pr(S_{it} = k | S_{it-1} = j), \quad t = 2, \dots, T \text{ and } j, k = 1, \dots, K. \quad (1)$$

In (1), k refers to the current state, whereas j refers to the one previously visited. In the following, for simplicity, we will consider a homogeneous HMM, that is $\pi_{i,k|j} = \pi_{k|j}$ and $\pi_{ik} = \pi_k, i = 1, \dots, I$. Moreover, the K -dimensional vector $\boldsymbol{\pi}$ collects the initial probabilities, whereas the $K \times K$ transition probability matrix $\boldsymbol{\Pi}$ collects the transition probabilities, which govern the state switching behavior of the chain.

Let $\{\mathbf{W}_{it}; i = 1, \dots, I, t = 1, \dots, T\}$ denote sequences of multivariate continuous longitudinal observations of dimension d_W recorded on I units and T times. The process $\{\mathbf{W}_{it}\}$ represents the state-dependent process of the HMM and fulfills the conditional (on the hidden states) independence property.

The random vector of interest \mathbf{W} is composed, in many applied longitudinal studies, by a d_Y -variate outcome vector \mathbf{Y} and by a random vector of covariates \mathbf{X} of dimension d_X , with $d_X + d_Y = d_W$; that is, $\mathbf{W} = (\mathbf{X}, \mathbf{Y})$. The interest is usually on modelling \mathbf{Y} on \mathbf{X} . In this literature, hidden Markov regression models (HMRMs) play a special role (see e.g., Zucchini et al. 2016; Bartolucci et al. 2014; Maruotti 2011). In this family the standard approach, often referred as hidden Markov regression models with fixed covariates (HMRMFCs; Maruotti and Punzo 2017), focuses on modeling

$$f(\mathbf{y}_{it} | \mathbf{X}_{it} = \mathbf{x}_{it}, S_{it} = k), \quad (2)$$

where $f(\cdot)$ is a generic multivariate probability density function, by assuming a functional form for the expectation $E(\mathbf{Y}_{it} | \mathbf{X}_{it} = \mathbf{x}_{it}, S_{it} = k)$. In the following, for simplicity, we will consider the classical linear case

Table 1 Observation labelling

Outlier/leverage	Yes	No
Yes	Bad leverage	Outlier
No	Good leverage	Typical

$$E(Y_{it}|X_{it} = \mathbf{x}_{it}, S_{it} = k; \boldsymbol{\beta}_k) = \boldsymbol{\mu}_Y(\mathbf{x}_{it}; \boldsymbol{\beta}_k) = \boldsymbol{\beta}'_k \mathbf{x}_{it}^*,$$

with $\boldsymbol{\beta}_k$ being a vector of local regression coefficients of dimension $[(1 + d_X) \times d_Y]$ and being $\mathbf{x}_{it}^* = (1, \mathbf{x}'_{it})'$ to account for the intercept(s).

2.2 Assignment independence versus assignment dependence

An implicit assumption of HMRMFCs, in a clustering perspective, is the so-called *assignment independence* (Hennig 2000): the assignment of the data points $(\mathbf{x}_{it}, \mathbf{y}_{it})$ to the hidden states is independent from the covariates distribution. To relax the *assignment independence* assumption, Punzo et al. (2018a) present hidden Markov regression models with random covariates (HMRMRCs), where the interest, in each hidden state k , is in modeling the joint distribution

$$f(\mathbf{x}_{it}, \mathbf{y}_{it} | S_{it} = k) = f(\mathbf{y}_{it} | X_{it} = \mathbf{x}_{it}, S_{it} = k) f(\mathbf{x}_{it} | S_{it} = k). \quad (3)$$

This formulation implies *assignment dependence*: the state-specific distributions for the covariates can also be distinct and they can affect the assignment of the data points to the hidden states.

2.3 Robustness in hidden Markov models with random covariates

As a further issue, real longitudinal regression data are often “contaminated” by mildly atypical observations (cf. Ritter 2015, pp. 79–80) and inference on model parameters, with particular interest to the regression coefficients, could be strongly affected (Hos-sain and Naik 1991). Accordingly, the detection of these atypical observations, and the development of robust estimation methods, are important problems. According to a common-to-regression taxonomy, observations can be classified as: typical (bulk of the data), good leverage, bad leverage, and outlier (see Table 1, Croux and Dehon 2003; Rousseeuw and Leroy 2005, Chapter 1).

Heavy-tailed elliptical distributions, such as the t and the contaminated normal, offer the flexibility needed for achieving mildly atypical observations robustness, whereas the normal distribution, often used as the reference distribution, lacks sufficient fit (see, e.g., Lachos et al. 2011; Niu et al. 2016). For the use of t and contaminated normal distributions in the HMM framework, see Bernardi et al. (2017) and Punzo and Maruotti (2016), respectively, while for alternative heavy-tailed distributions see Maruotti et al. (2019). By considering the t and contaminated normal distributions for outcomes and/or covariates, HMRMRCs are also flexible enough to fit longitudinal data presenting mildly atypical observations.

Thus, as densities to be adopted in (3), with reference to a generic random vector \mathbf{H} which, for our purposes, may be either $\mathbf{Y}|\mathbf{X} = \mathbf{x}$ or \mathbf{X} , we consider:

- the multivariate normal distribution

$$f_N(\mathbf{h}_{it}; \boldsymbol{\mu}_{\mathbf{H}|k}, \boldsymbol{\Sigma}_{\mathbf{H}|k}) = (2\pi)^{-\frac{d_{\mathbf{H}}}{2}} |\boldsymbol{\Sigma}_{\mathbf{H}|k}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \delta(\mathbf{h}_{it}, \boldsymbol{\mu}_{\mathbf{H}|k}; \boldsymbol{\Sigma}_{\mathbf{H}|k}) \right\}, \quad (4)$$

where $\boldsymbol{\mu}_{\mathbf{H}|k}$ and $\boldsymbol{\Sigma}_{\mathbf{H}|k}$ denote the mean and the covariance matrix, respectively, and where

$$\delta(\mathbf{h}_{it}, \boldsymbol{\mu}_{\mathbf{H}|k}; \boldsymbol{\Sigma}_{\mathbf{H}|k}) = (\mathbf{h}_{it} - \boldsymbol{\mu}_{\mathbf{H}|k})' \boldsymbol{\Sigma}_{\mathbf{H}|k}^{-1} (\mathbf{h}_{it} - \boldsymbol{\mu}_{\mathbf{H}|k})$$

denotes the squared Mahalanobis distance between \mathbf{h}_{it} and $\boldsymbol{\mu}_{\mathbf{H}|k}$, with covariance matrix $\boldsymbol{\Sigma}_{\mathbf{H}|k}$. In symbols, $\mathbf{H}_{it}|S_{it} = k \sim \mathcal{N}_{d_{\mathbf{H}}}(\boldsymbol{\mu}_{\mathbf{H}|k}, \boldsymbol{\Sigma}_{\mathbf{H}|k})$;

- the multivariate t distribution

$$\begin{aligned} f_t(\mathbf{h}_{it}; \boldsymbol{\mu}_{\mathbf{H}|k}, \boldsymbol{\Sigma}_{\mathbf{H}|k}, \nu_{\mathbf{H}|k}) \\ = \frac{\Gamma\left(\frac{\nu_{\mathbf{H}|k} + d_{\mathbf{H}}}{2}\right) |\boldsymbol{\Sigma}_{\mathbf{H}|k}|^{-\frac{1}{2}}}{\Gamma\left(\frac{\nu_{\mathbf{H}|k}}{2}\right) (\pi \nu_{\mathbf{H}|k})^{\frac{d_{\mathbf{H}}}{2}} \left[1 + \frac{1}{\nu_{\mathbf{H}|k}} \delta(\mathbf{h}_{it}, \boldsymbol{\mu}_{\mathbf{H}|k}; \boldsymbol{\Sigma}_{\mathbf{H}|k})\right]^{\frac{\nu_{\mathbf{H}|k} + d_{\mathbf{H}}}{2}}}, \end{aligned} \quad (5)$$

where $\boldsymbol{\mu}_{\mathbf{H}|k}$, $\boldsymbol{\Sigma}_{\mathbf{H}|k}$, and $\nu_{\mathbf{H}|k}$ denote the mean, the scale matrix, and the the degrees of freedom. In symbols, $\mathbf{H}_{it}|S_{it} = k \sim t_{d_{\mathbf{H}}}(\boldsymbol{\mu}_{\mathbf{H}|k}, \boldsymbol{\Sigma}_{\mathbf{H}|k}, \nu_{\mathbf{H}|k})$. Note that (5) approaches (4) as $\nu_{\mathbf{H}|k} \rightarrow \infty$;

- the multivariate contaminated normal distribution

$$\begin{aligned} f_{\text{CN}}(\mathbf{h}_{it}; \boldsymbol{\mu}_{\mathbf{H}|k}, \boldsymbol{\Sigma}_{\mathbf{H}|k}, \alpha_{\mathbf{H}|k}, \eta_{\mathbf{H}|k}) \\ = \alpha_{\mathbf{H}|k} f_N(\mathbf{h}_{it}; \boldsymbol{\mu}_{\mathbf{H}|k}, \boldsymbol{\Sigma}_{\mathbf{H}|k}) + (1 - \alpha_{\mathbf{H}|k}) f_N(\mathbf{h}_{it}; \boldsymbol{\mu}_{\mathbf{H}|k}, \eta_{\mathbf{H}|k} \boldsymbol{\Sigma}_{\mathbf{H}|k}), \end{aligned} \quad (6)$$

where $\boldsymbol{\mu}_{\mathbf{H}|k}$ is the mean, $\boldsymbol{\Sigma}_{\mathbf{H}|k}$ is the scale matrix, $\alpha_{\mathbf{H}|k} \in (0, 1)$ is the proportion of typical points in state k , and $\eta_{\mathbf{H}|k} > 1$ is an inflation parameter accounting for the degree of atypicality in state k . In symbols, $\mathbf{H}_{it}|S_{it} = k \sim \mathcal{CN}_{d_{\mathbf{H}}}(\boldsymbol{\mu}_{\mathbf{H}|k}, \boldsymbol{\Sigma}_{\mathbf{H}|k}, \alpha_{\mathbf{H}|k}, \eta_{\mathbf{H}|k})$. Note that (6) approaches (4) as $\alpha_{\mathbf{H}|k} \rightarrow 1^-$ and $\eta_{\mathbf{H}|k} \rightarrow 1^+$.

By combining the three considered multivariate distributions with respect to $\mathbf{Y}|\mathbf{x}$ and \mathbf{X} in (3), we introduce a family of nine different HMRMCs (see Table 2).

For the use of these distributions in HMRMFCs, see Maruotti and Punzo (2017).

Table 2 Proposed family of HMRMRCs

Model	$Y_{it} X_{it} = x_{it}, S_{it} = k$	$X_{it} S_{it} = k$
N-N-HMRMRC	$\mathcal{N}_{d_Y}(\mu_Y(x_{it}; \beta_k), \Sigma_{Y k})$	$\mathcal{N}_{d_X}(\mu_{X k}, \Sigma_{Y k})$
N- <i>t</i> -HMRMRC	$\mathcal{N}_{d_Y}(\mu_Y(x_{it}; \beta_k), \Sigma_{Y k})$	$t_{d_X}(\mu_{X k}, \Sigma_{X k}, \nu_{X k})$
N-CN-HMRMRC	$\mathcal{N}_{d_Y}(\mu_Y(x_{it}; \beta_k), \Sigma_{Y k})$	$\mathcal{CN}_{d_X}(\mu_{X k}, \Sigma_{X k}, \alpha_{X k}, \eta_{X k})$
<i>t</i> -N-HMRMRC	$t_{d_Y}(\mu_Y(x_{it}; \beta_k), \Sigma_{Y k}, \nu_{Y k})$	$\mathcal{N}_{d_X}(\mu_{X k}, \Sigma_{Y k})$
<i>t</i> - <i>t</i> -HMRMRC	$t_{d_Y}(\mu_Y(x_{it}; \beta_k), \Sigma_{Y k}, \nu_{Y k})$	$t_{d_X}(\mu_{X k}, \Sigma_{X k}, \nu_{X k})$
<i>t</i> -CN-HMRMRC	$t_{d_Y}(\mu_Y(x_{it}; \beta_k), \Sigma_{Y k}, \nu_{Y k})$	$\mathcal{CN}_{d_X}(\mu_{X k}, \Sigma_{X k}, \alpha_{X k}, \eta_{X k})$
CN-N-HMRMRC	$\mathcal{CN}_{d_Y}(\mu_Y(x_{it}; \beta_k), \Sigma_{Y k}, \alpha_{Y k}, \eta_{Y k})$	$\mathcal{N}_{d_X}(\mu_{X k}, \Sigma_{Y k})$
CN- <i>t</i> -HMRMRC	$\mathcal{CN}_{d_Y}(\mu_Y(x_{it}; \beta_k), \Sigma_{Y k}, \alpha_{Y k}, \eta_{Y k})$	$t_{d_X}(\mu_{X k}, \Sigma_{X k}, \nu_{X k})$
CN-CN-HMRMRC	$\mathcal{CN}_{d_Y}(\mu_Y(x_{it}; \beta_k), \Sigma_{Y k}, \alpha_{Y k}, \eta_{Y k})$	$\mathcal{CN}_{d_X}(\mu_{X k}, \Sigma_{X k}, \alpha_{X k}, \eta_{X k})$

3 Identifiability

An important issue in dealing with the proposed HMRMRCs is to establish their identifiability. Identifiability is a necessary requirement, *inter alia*, for the usual asymptotic theory to hold for ML estimation.

For HMMs, whose state-dependent distributions are assumed to belong to some parametric family, Leroux (1992) shows that identifiability up to label switching follows from identifiability of the marginal (finite) mixtures (cf. Dannemann et al. 2014, Section 2). In our case, the marginal mixtures are represented by mixtures of regression models with random covariates (MRMRCs).

Following Leroux (1992), Proposition 1 gives sufficient conditions for the identifiability of all the HMRMRCs in Table 2 by simply considering the corresponding MRMRCs.

Proposition 1 *Let*

$$f(x, y; \vartheta) = \sum_{k=1}^K \pi_k f(y|x; \vartheta_{Y|k}) f(x; \vartheta_{X|k})$$

and

$$f(x, y; \tilde{\vartheta}) = \sum_{s=1}^{\tilde{K}} \tilde{\pi}_s f(y|x; \tilde{\vartheta}_{Y|s}) f(x; \tilde{\vartheta}_{X|s})$$

be the densities of two MRMRCs based on the same component conditional/marginal densities. Regardless from the choice made about the density of X , consider the following cases.

N) $Y|x \sim \mathcal{N}_{d_Y}(\mu_Y(x; \beta_k), \Sigma_{Y|k})$. If $k \neq l$, with $k, l \in \{1, \dots, K\}$, implies

$$\|\beta_k - \beta_l\|_2^2 + \|\Sigma_{Y|k} - \Sigma_{Y|l}\|_2^2 \neq 0, \quad (7)$$

then the equality $f(\mathbf{x}, \mathbf{y}; \boldsymbol{\vartheta}) = f(\mathbf{x}, \mathbf{y}; \tilde{\boldsymbol{\vartheta}})$, for almost all $\mathbf{x} \in \mathbb{R}^{d_x}$, implies that $K = \tilde{K}$ and also implies that for each $k \in \{1, \dots, K\}$ there exists an $s \in \{1, \dots, K\}$ such that $\pi_k = \tilde{\pi}_s$, $\boldsymbol{\mu}_{X|k} = \tilde{\boldsymbol{\mu}}_{X|s}$, $\boldsymbol{\Sigma}_{X|k} = \tilde{\boldsymbol{\Sigma}}_{X|s}$, $\boldsymbol{\beta}_k = \tilde{\boldsymbol{\beta}}_s$, and $\boldsymbol{\Sigma}_{Y|k} = \tilde{\boldsymbol{\Sigma}}_{Y|s}$.

t) $Y|\mathbf{x} \sim t_{d_Y}(\boldsymbol{\mu}_Y(\mathbf{x}; \boldsymbol{\beta}_k), \boldsymbol{\Sigma}_{Y|k}, \nu_{Y|k})$. If $k \neq l$, with $k, l \in \{1, \dots, K\}$, implies

$$\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_l\|_2^2 + \|\boldsymbol{\Sigma}_{Y|k} - \boldsymbol{\Sigma}_{Y|l}\|_2^2 + \|\nu_{Y|k} - \nu_{Y|l}\|_2^2 \neq 0, \quad (8)$$

then the equality $f(\mathbf{x}, \mathbf{y}; \boldsymbol{\vartheta}) = f(\mathbf{x}, \mathbf{y}; \tilde{\boldsymbol{\vartheta}})$, for almost all $\mathbf{x} \in \mathbb{R}^{d_x}$, implies that $K = \tilde{K}$ and also implies that for each $k \in \{1, \dots, K\}$ there exists an $s \in \{1, \dots, K\}$ such that $\pi_k = \tilde{\pi}_s$, $\boldsymbol{\mu}_{X|k} = \tilde{\boldsymbol{\mu}}_{X|s}$, $\boldsymbol{\Sigma}_{X|k} = \tilde{\boldsymbol{\Sigma}}_{X|s}$, $\nu_{X|k} = \nu_{X|s}$, $\boldsymbol{\beta}_k = \tilde{\boldsymbol{\beta}}_s$, $\boldsymbol{\Sigma}_{Y|k} = \tilde{\boldsymbol{\Sigma}}_{Y|s}$, and $\nu_{Y|k} = \nu_{Y|s}$.

CN) $Y|\mathbf{x} \sim \mathcal{CN}_{d_Y}(\boldsymbol{\mu}_Y(\mathbf{x}; \boldsymbol{\beta}_k), \boldsymbol{\Sigma}_{Y|k}, \alpha_{Y|k}, \eta_{Y|k})$. If $k \neq l$, with $k, l \in \{1, \dots, K\}$, implies

$$\|\boldsymbol{\beta}_k - \boldsymbol{\beta}_l\|_2^2 + \|\boldsymbol{\Sigma}_{Y|k} - a\boldsymbol{\Sigma}_{Y|l}\|_2^2 \neq 0 \quad (9)$$

for all $a > 0$, then the equality $f(\mathbf{x}, \mathbf{y}; \boldsymbol{\vartheta}) = f(\mathbf{x}, \mathbf{y}; \tilde{\boldsymbol{\vartheta}})$, for almost all $\mathbf{x} \in \mathbb{R}^{d_x}$, implies that $K = \tilde{K}$ and also implies that for each $k \in \{1, \dots, K\}$ there exists an $s \in \{1, \dots, K\}$ such that $\pi_k = \tilde{\pi}_s$, $\alpha_{X|k} = \tilde{\alpha}_{X|s}$, $\boldsymbol{\mu}_{X|k} = \tilde{\boldsymbol{\mu}}_{X|s}$, $\boldsymbol{\Sigma}_{X|k} = \tilde{\boldsymbol{\Sigma}}_{X|s}$, $\eta_{X|k} = \tilde{\eta}_{X|s}$, $\alpha_{Y|k} = \tilde{\alpha}_{Y|s}$, $\boldsymbol{\beta}_k = \tilde{\boldsymbol{\beta}}_s$, $\boldsymbol{\Sigma}_{Y|k} = \tilde{\boldsymbol{\Sigma}}_{Y|s}$, and $\eta_{Y|k} = \tilde{\eta}_{Y|s}$.

Proof The proof is given in the Online Appendix A □

4 Maximum likelihood estimation

To perform ML estimation of the parameters for the proposed HMRMRCs, on the basis of the sample $\{(\mathbf{x}_{it}, \mathbf{y}_{it}); i = 1, \dots, I, t = 1, \dots, T\}$, we need to compute the likelihood function

$$\mathcal{L}(\boldsymbol{\vartheta}) = \prod_{i=1}^I \mathcal{L}_i(\boldsymbol{\vartheta}) = \prod_{i=1}^I \boldsymbol{\pi}' \mathbf{f}_{i1} \boldsymbol{\Pi} \mathbf{f}_{i2} \cdots \boldsymbol{\Pi} \mathbf{f}_{iT} \mathbf{1}_K, \quad (10)$$

where $\boldsymbol{\vartheta}$ corresponds to the set of all model parameters, $\mathbf{1}_K$ denotes a vector of K ones, and \mathbf{f}_{it} denotes a $K \times K$ diagonal matrix having on the main diagonal the joint densities in (3), $k = 1, \dots, K$. Finding the value of $\boldsymbol{\vartheta}$ that maximizes the log-transformation of (10) under the constraints $\pi_k > 0$, $\pi_{k|j} > 0$, $\sum_{k=1}^K \pi_k = 1$, $\sum_{k=1}^K \pi_{k|j} = 1$, $k, j = 1, \dots, K$, in addition to the constraints required by the parameters of the chosen multivariate distributions, is not an easy problem since (10) is not available in an analytically convenient form.

When the multivariate contaminated normal distribution is not involved in the formulation of the model, we describe an expectation-maximization (EM) algorithm (Baum et al. 1970) to fit the models. Whenever the contaminated normal distribution

is involved in model specification, we consider an expectation-conditional maximization (ECM) algorithm (Meng and Rubin 1993).

4.1 Sources of incompleteness and complete-data log-likelihood

To illustrate the algorithms, we need to specify the sources of incompleteness in our case: some of them depend on the considered multivariate distribution while the others are common to all the models. The common source – and the unique one when the N-N-HMRMRC is considered – arises from the fact that we do not know the state membership and its evolution over time; this source of incompleteness is introduced in the formulation of the model via the definition of the unobserved state membership $\mathbf{z}_{it} = (z_{it1}, \dots, z_{itk}, \dots, z_{itK})'$ and the unobserved states transition

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

respectively, with

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

Based on this source of incompleteness, we can start to write a sort of “first level” complete-data log-likelihood in the following way

$$\ell_c(\boldsymbol{\vartheta}) = \ell_{c_1}(\boldsymbol{\pi}) + \ell_{c_2}(\boldsymbol{\Pi}) + \ell_{c_3}(\boldsymbol{\vartheta}_Y) + \ell_{c_4}(\boldsymbol{\vartheta}_X), \quad (11)$$

where

$$\ell_{c_1}(\boldsymbol{\pi}) = \sum_{i=1}^I \sum_{k=1}^K z_{i1k} \log(\pi_k), \quad (12)$$

$$\ell_{c_2}(\boldsymbol{\Pi}) = \sum_{i=1}^I \sum_{t=2}^T \sum_{k=1}^K \sum_{j=1}^K zz_{itjk} \log(\pi_{k|j}), \quad (13)$$

$$\ell_{c_3}(\boldsymbol{\vartheta}_Y) = \sum_{i=1}^I \sum_{t=1}^T \sum_{k=1}^K z_{itk} \ln[f(y_{it}; \boldsymbol{\vartheta}_{Y|k})], \quad (14)$$

$$\ell_{c_4}(\boldsymbol{\vartheta}_X) = \sum_{i=1}^I \sum_{t=1}^T \sum_{k=1}^K z_{itk} \ln[f(x_{it}; \boldsymbol{\vartheta}_{X|k})], \quad (15)$$

with $\boldsymbol{\vartheta}_Y$ and $\boldsymbol{\vartheta}_X$ denoting the sets of all model parameters related to $Y|x$ and X , respectively.

The other sources of incompleteness are distribution-dependent and are described in the following.

4.1.1 Multivariate t distribution for $Y|x$

If $Y_{it} | X_{it} = x_{it}, S_{it} = k \sim t_{d_Y}(\boldsymbol{\mu}_Y(x_{it}; \boldsymbol{\beta}_k), \boldsymbol{\Sigma}_{Y|k}, \nu_{Y|k})$, then a further source of incompleteness arises from the fact that a multivariate t random vector can be written as a multivariate normal vector whose covariance matrix is scaled by the reciprocal of a Gamma random variable. In practice, for each observation (x_{it}, y_{it}) in state k , this source of incompleteness is denoted by $U_{itk} \sim \text{Gamma}(\nu_{Y|k}/2, \nu_{Y|k}/2)$. This leads to write ℓ_{c_3} in (14), in order to define the “second level” complete-data log-likelihood for HMRMRCs related to this case, in the following way

$$\ell_{c_3}(\boldsymbol{\vartheta}_Y) = \ell_{c_{3a}}(\mathbf{v}_Y) + \ell_{c_{3b}}(\boldsymbol{\beta}, \boldsymbol{\Sigma}_Y), \quad (16)$$

where

$$\begin{aligned} \ell_{c_{3a}}(\mathbf{v}_Y) &= \sum_{i=1}^I \sum_{t=1}^T \sum_{k=1}^K z_{itk} \left\{ -\ln \left[\Gamma \left(\frac{\nu_{Y|k}}{2} \right) \right] + \frac{\nu_{Y|k}}{2} \ln \left(\frac{\nu_{Y|k}}{2} \right) \right. \\ &\quad \left. + \frac{\nu_{Y|k}}{2} [\ln(u_{itk}) - u_{itk}] - \ln(u_{itk}) \right\}, \\ \ell_{c_{3b}}(\boldsymbol{\beta}, \boldsymbol{\Sigma}_Y) &= -\frac{1}{2} \sum_{i=1}^I \sum_{t=1}^T \sum_{k=1}^K z_{itk} \left\{ d_Y \ln(2\pi) + \ln |\boldsymbol{\Sigma}_{Y|k}| \right. \\ &\quad \left. + u_{itk} \delta(y_{it}, \boldsymbol{\mu}_Y(x_{it}; \boldsymbol{\beta}_k); \boldsymbol{\Sigma}_{Y|k}) \right\}, \end{aligned}$$

$\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K)$, $\boldsymbol{\Sigma}_Y = (\boldsymbol{\Sigma}_{Y|1}, \dots, \boldsymbol{\Sigma}_{Y|K})$, $\mathbf{v}_Y = (\nu_{Y|1}, \dots, \nu_{Y|K})$, and $\boldsymbol{\vartheta}_Y = (\boldsymbol{\beta}, \boldsymbol{\Sigma}_Y, \mathbf{v}_Y)$.

4.1.2 Multivariate t distribution for X

Analogously, if $X_{it} | S_{it} = k \sim t_{d_X}(\boldsymbol{\mu}_{X|k}, \boldsymbol{\Sigma}_{X|k}, \nu_{X|k})$, the further source of incompleteness, for each observation (x_{it}, y_{it}) in state k , is denoted by $V_{itk} \sim \text{Gamma}(\nu_{X|k}/2, \nu_{X|k}/2)$. This leads to write ℓ_{c_4} in (15), in order to define the “second level” complete-data log-likelihood for HMRMRCs related to this case, in the following way

$$\ell_{c_4}(\boldsymbol{\vartheta}_X) = \ell_{c_{4a}}(\mathbf{v}_X) + \ell_{c_{4b}}(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X), \quad (17)$$

where

$$\ell_{c_{4a}}(\mathbf{v}_X) = \sum_{i=1}^I \sum_{t=1}^T \sum_{k=1}^K z_{itk} \left\{ -\ln \left[\Gamma \left(\frac{\nu_{X|k}}{2} \right) \right] + \frac{\nu_{X|k}}{2} \ln \left(\frac{\nu_{X|k}}{2} \right) \right\}$$

$$\begin{aligned}
& + \frac{v_{X|k}}{2} [\ln(v_{itk}) - v_{itk}] - \ln(v_{itk}) \Big\}, \\
\ell_{c_{4b}}(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X) = & -\frac{1}{2} \sum_{i=1}^I \sum_{t=1}^T \sum_{k=1}^K z_{itk} \left\{ d_X \ln(2\pi) + \ln|\boldsymbol{\Sigma}_{X|k}| \right. \\
& \left. + v_{itk} \delta(\mathbf{x}_{it}, \boldsymbol{\mu}_{X|k}; \boldsymbol{\Sigma}_{X|k}) \right\},
\end{aligned}$$

$\boldsymbol{\mu}_X = (\boldsymbol{\mu}_{X|1}, \dots, \boldsymbol{\mu}_{X|K})$, $\boldsymbol{\Sigma}_X = (\boldsymbol{\Sigma}_{X|1}, \dots, \boldsymbol{\Sigma}_{X|K})$, $\mathbf{v}_X = (v_{X|1}, \dots, v_{X|K})$, and $\boldsymbol{\vartheta}_X = (\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X, \mathbf{v}_X)$.

4.1.3 Multivariate contaminated normal distribution for $Y|x$

If $Y_{it} | X_{it} = \mathbf{x}_{it}, S_{it} = k \sim \mathcal{CN}_{d_Y}(\boldsymbol{\mu}_Y(\mathbf{x}_{it}; \boldsymbol{\beta}_k), \boldsymbol{\Sigma}_{Y|k}, \alpha_{Y|k}, \eta_{Y|k})$, a further source of incompleteness arises from the fact that for each observation $(\mathbf{x}_{it}, \mathbf{y}_{it})$ in state k we do not know if it is either good or bad “vertically”, i.e. with respect to $Y|x$. To denote this source of incompleteness, we use $\mathbf{u}_{it} = (u_{it1}, \dots, u_{itk}, \dots, u_{itK})'$, where $u_{itk} = 1$ if $(\mathbf{x}_{it}, \mathbf{y}_{it})$ in state k is a good vertical point and $u_{itk} = 0$ if it is a bad vertical point. Thus, in the “second level” complete-data log-likelihood for HMRMCs related to this case, we can specify (14) as

$$\begin{aligned}
\ell_{c_3}(\boldsymbol{\vartheta}_Y) &= \sum_{i=1}^I \sum_{t=1}^T \sum_{k=1}^K z_{itk} \ln \left\{ [\alpha_{Y|k} f_N(\mathbf{y}_{it}; \boldsymbol{\mu}_Y(\mathbf{x}_{it}; \boldsymbol{\beta}_k), \boldsymbol{\Sigma}_{Y|k})]^{u_{itk}} \right. \\
&\quad \times \left. [(1 - \alpha_{Y|k}) f_N(\mathbf{y}_{it}; \boldsymbol{\mu}_Y(\mathbf{x}_{it}; \boldsymbol{\beta}_k), \eta_{Y|k} \boldsymbol{\Sigma}_{Y|k})]^{1-u_{itk}} \right\} \\
&= \ell_{c_{3a}}(\boldsymbol{\alpha}_Y) + \ell_{c_{3b}}(\boldsymbol{\beta}, \boldsymbol{\Sigma}_Y, \boldsymbol{\eta}_Y),
\end{aligned}$$

where

$$\begin{aligned}
\ell_{c_{3a}}(\boldsymbol{\alpha}_Y) &= \sum_{i=1}^I \sum_{t=1}^T \sum_{k=1}^K z_{itk} [u_{itk} \ln \alpha_{Y|k} + (1 - u_{itk}) \ln (1 - \alpha_{Y|k})], \\
\ell_{c_{3b}}(\boldsymbol{\beta}, \boldsymbol{\Sigma}_Y, \boldsymbol{\eta}_Y) &= -\frac{1}{2} \sum_{i=1}^I \sum_{t=1}^T \sum_{k=1}^K \left\{ z_{itk} \ln |\boldsymbol{\Sigma}_{Y|k}| + d_Y z_{itk} (1 - u_{itk}) \ln \eta_{Y|k} \right. \\
&\quad \left. + z_{itk} \left(u_{itk} + \frac{1 - u_{itk}}{\eta_{Y|k}} \right) \delta(\mathbf{y}_{it}, \boldsymbol{\mu}_Y(\mathbf{x}_{it}; \boldsymbol{\beta}_k); \boldsymbol{\Sigma}_{Y|k}) \right\},
\end{aligned}$$

$\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K)$, $\boldsymbol{\Sigma}_Y = (\boldsymbol{\Sigma}_{Y|1}, \dots, \boldsymbol{\Sigma}_{Y|K})$, $\boldsymbol{\alpha}_Y = (\alpha_{Y|1}, \dots, \alpha_{Y|K})$, $\boldsymbol{\eta}_Y = (\eta_{Y|1}, \dots, \eta_{Y|K})$, and $\boldsymbol{\vartheta}_Y = (\boldsymbol{\beta}, \boldsymbol{\Sigma}_Y, \boldsymbol{\alpha}_Y, \boldsymbol{\eta}_Y)$.

4.1.4 Multivariate contaminated normal distribution for X

Analogously, if $X_{it} | S_{it} = k \sim \mathcal{CN}_{d_X}(\boldsymbol{\mu}_{X|k}, \boldsymbol{\Sigma}_{X|k}, \alpha_{X|k}, \eta_{X|k})$, a further source of incompleteness arises from the fact that for each observation $(\mathbf{x}_{it}, \mathbf{y}_{it})$ in state k we do

not know if it is either good or bad “horizontally”, i.e. with respect to \mathbf{X} . To denote this source of incompleteness, we use $\mathbf{v}_{it} = (v_{it1}, \dots, v_{itk}, \dots, v_{itK})'$, where $v_{itk} = 1$ if $(\mathbf{x}_{it}, \mathbf{y}_{it})$ in state k is a good horizontal point and $u_{itk} = 0$ if it is a bad horizontal point. Thus, in the “second level” complete-data log-likelihood for HMRMCs related to this case, we can specify (15) as

$$\begin{aligned}\ell_{c_4}(\boldsymbol{\vartheta}_X) &= \sum_{i=1}^I \sum_{t=1}^T \sum_{k=1}^K z_{itk} \ln \left\{ [\alpha_{X|k} f_N(\mathbf{x}_{it}; \boldsymbol{\mu}_{X|k}, \boldsymbol{\Sigma}_{X|k})]^{v_{itk}} \right. \\ &\quad \times \left. [(1 - \alpha_{X|k}) f_N(\mathbf{x}_{it}; \boldsymbol{\mu}_{X|k}, \eta_{X|k} \boldsymbol{\Sigma}_{X|k})]^{1-v_{itk}} \right\} \\ &= \ell_{c_{4a}}(\boldsymbol{\alpha}_X) + \ell_{c_{4b}}(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X, \boldsymbol{\eta}_X),\end{aligned}$$

where

$$\begin{aligned}\ell_{c_{4a}}(\boldsymbol{\alpha}_X) &= \sum_{i=1}^I \sum_{t=1}^T \sum_{k=1}^K z_{itk} [v_{itk} \ln \alpha_{X|k} + (1 - v_{itk}) \ln (1 - \alpha_{X|k})], \\ \ell_{c_{4b}}(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X, \boldsymbol{\eta}_X) &= -\frac{1}{2} \sum_{i=1}^I \sum_{t=1}^T \sum_{k=1}^K \left\{ z_{itk} \ln |\boldsymbol{\Sigma}_{X|k}| + d_X z_{itk} (1 - v_{itk}) \ln \eta_{X|k} \right. \\ &\quad \left. + z_{itk} \left(v_{itk} + \frac{1 - v_{itk}}{\eta_{X|k}} \right) \delta(\mathbf{x}_{it}, \boldsymbol{\mu}_{X|k}; \boldsymbol{\Sigma}_{X|k}) \right\},\end{aligned}$$

$\boldsymbol{\mu}_X = (\boldsymbol{\mu}_{X|1}, \dots, \boldsymbol{\mu}_{X|K})$, $\boldsymbol{\Sigma}_X = (\boldsymbol{\Sigma}_{X|1}, \dots, \boldsymbol{\Sigma}_{X|K})$, $\boldsymbol{\alpha}_X = (\alpha_{X|1}, \dots, \alpha_{X|K})$, $\boldsymbol{\eta}_X = (\eta_{X|1}, \dots, \eta_{X|K})$, and $\boldsymbol{\vartheta}_X = (\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X, \boldsymbol{\alpha}_X, \boldsymbol{\eta}_X)$.

4.2 EM and ECM algorithms

As said above, when a multivariate contaminated normal distribution is considered in the formulation of the model, an ECM algorithm is considered for fitting. The ECM algorithm iterates between an E-step and a convenient number of CM-steps, until convergence. The only difference from the EM algorithm is that each M-step is replaced by CM-steps (see McLachlan and Krishnan 2007, for details about these algorithms).

The E-step, on the $(r + 1)$ th iteration of these algorithms, requires the calculation of $Q(\boldsymbol{\vartheta})$, the current conditional expectation of $\ell_c(\boldsymbol{\vartheta})$ given the observed data and the current estimates $\boldsymbol{\vartheta}^{(r)}$ of the parameters. As a part of this calculation, regardless from the considered distribution, we replace z_{itk} and z_{itjk} with their conditional expectations, namely, $z_{itk}^{(r)}$ and $z_{itjk}^{(r)}$ (for computational details see, for example, Punzo and Maruotti 2016 and Maruotti and Punzo 2017). The rest of the E-step, if needed, depends on the adopted distribution and it will be detailed in the following. M and CM steps require the maximization of $Q(\boldsymbol{\vartheta})$ with respect to $\boldsymbol{\vartheta}$. As the four terms on the right-hand side of (11) have zero cross-derivatives, they can be maximized separately. In particular, the maximization of $Q_1(\boldsymbol{\pi})$ and $Q_2(\boldsymbol{\Pi})$ – expected counterparts

of $\ell_{c_1}(\boldsymbol{\pi})$ in (12) and $\ell_{c_2}(\boldsymbol{\Pi})$ in (13) – with respect to $\boldsymbol{\pi}$ and $\boldsymbol{\Pi}$, respectively, subject to the constraints on these parameters, yields

$$\pi_k^{(r+1)} = \frac{1}{I} \sum_{i=1}^I z_{i1k}^{(r)} \quad \text{and} \quad \pi_{k|j}^{(r+1)} = \frac{\sum_{i=1}^I \sum_{t=2}^T z_{itjk}^{(r)}}{\sum_{i=1}^I \sum_{t=2}^T \sum_{k=1}^K z_{itjk}^{(r)}},$$

regardless of the considered distribution and regardless of the type of maximization, direct (M-step) or conditional (CM-step). The updates of the remaining parameters $\boldsymbol{\vartheta}_X$ and $\boldsymbol{\vartheta}_Y$ depend on the considered distribution as well as on the type of maximization; these updates are detailed in the following.

4.2.1 Multivariate normal distribution for $Y \mid x$

If $Y_{it} \mid X_{it} = x_{it}, S_{it} = k \sim \mathcal{N}_{d_Y}(\boldsymbol{\mu}_Y(x_{it}; \boldsymbol{\beta}_k), \boldsymbol{\Sigma}_{Y|k})$, then no further calculations are needed for the E-step on the $(r+1)$ th iteration. Hence, on the same iteration, we can skip to the M-step; it requires the calculation of $\boldsymbol{\vartheta}_Y^{(r+1)}$ as the value of $\boldsymbol{\vartheta}_Y$ that maximizes

$$\begin{aligned} Q_3(\boldsymbol{\vartheta}_Y) = & -\frac{1}{2} \sum_{i=1}^I \sum_{t=1}^T \sum_{k=1}^K z_{itk}^{(r)} \left\{ d_Y \ln(2\pi) + \ln |\boldsymbol{\Sigma}_{Y|k}| \right. \\ & \left. + \delta(y_{it}, \boldsymbol{\mu}_Y(x_{it}; \boldsymbol{\beta}_k); \boldsymbol{\Sigma}_{Y|k}) \right\}. \end{aligned}$$

After some algebra, such a maximization yields

$$\begin{aligned} \boldsymbol{\beta}_k^{(r+1)} &= \left[\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)} \mathbf{x}_{it}^* \mathbf{x}_{it}^{*'} \right]^{-1} \left[\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)} \mathbf{x}_{it}^* y_{it}' \right], \\ \boldsymbol{\Sigma}_{Y|k}^{(r+1)} &= \frac{\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)} \left[y_{it} - \boldsymbol{\mu}_Y(x_{it}; \boldsymbol{\beta}_k^{(r+1)}) \right] \left[y_{it} - \boldsymbol{\mu}_Y(x_{it}; \boldsymbol{\beta}_k^{(r+1)}) \right]'}{\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)}}. \end{aligned}$$

4.2.2 Multivariate t distribution for $Y \mid x$

Let us assume $Y_{it} \mid X_{it} = x_{it}, S_{it} = k \sim t_{d_Y}(\boldsymbol{\mu}_Y(x_{it}; \boldsymbol{\beta}_k), \boldsymbol{\Sigma}_{Y|k}, \nu_{Y|k})$, then the E-step, on the $(r+1)$ th iteration, further requires the replacement of u_{itk} with

$$u_{itk}^{(r)} = \frac{v_{Y|k}^{(r)} + d_Y}{v_{Y|k}^{(r)} + \delta \left(y_{it}, \mu_Y \left(x_{it}; \beta_k^{(r)} \right); \Sigma_{Y|k}^{(r)} \right)}. \quad (18)$$

Thus, by substituting z_{itk} and u_{itk} in (16), with z_{itk} and u_{itk} , respectively, we obtain $Q_3(\vartheta_Y)$.

The M-step, on the same iteration, requires the calculation of $\vartheta_Y^{(r+1)}$ as the value of ϑ_Y that maximizes $Q_3(\vartheta_Y)$. Such a maximization yields

$$\begin{aligned} \beta_k^{(r+1)} &= \left[\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)} u_{itk}^{(r)} x_{it}^* x_{it}' \right]^{-1} \left[\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)} u_{itk}^{(r)} x_{it}^* y_{it}' \right], \\ \Sigma_{Y|k}^{(r+1)} &= \frac{\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)} u_{itk}^{(r)} \left[y_{it} - \mu_Y \left(x_{it}; \beta_k^{(r+1)} \right) \right] \left[y_{it} - \mu_Y \left(x_{it}; \beta_k^{(r+1)} \right) \right]'}{\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)}}. \end{aligned}$$

A closed form solution is not analytically available for the update $v_{Y|k}^{(r+1)}$ of $v_{Y|k}$. However, by differentiating $Q_3(\vartheta_Y)$ with respect to $v_{Y|k}$, we note that $v_{Y|k}^{(r+1)}$ is a solution of the equation

$$\begin{aligned} & -\psi \left(\frac{v_{Y|k}}{2} \right) + \ln \left(\frac{v_{Y|k}}{2} \right) + 1 + \frac{1}{\sum_{t=1}^T z_{itk}^{(r)} u_{itk}^{(r)}} \sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)} \left[\ln \left(u_{itk}^{(r)} \right) - u_{itk}^{(r)} \right] \\ & + \psi \left(\frac{v_{Y|k}^{(r)} + d_Y}{2} \right) - \ln \left(\frac{v_{Y|k}^{(r)} + d_Y}{2} \right) = 0, \end{aligned} \quad (19)$$

where $\psi(\cdot)$ is the Digamma function. Operationally, the `uniroot()` function in the **stats** package for R (R Core Team 2018) can be used to numerically find the root of (19) over the interval $(2, v_Y^*)$, with $v_Y^* > 2$. In all the numerical examples, we fix $v_Y^* = 200$ to facilitate faster convergence.

4.2.3 Multivariate contaminated normal distribution for $Y \mid x$

If $Y_{it} \mid X_{it} = x_{it}, S_{it} = k \sim \mathcal{CN}_{d_Y}(\mu_Y(x_{it}; \beta_k), \Sigma_{Y|k}, \alpha_{Y|k}, \eta_{Y|k})$, then two CM-steps, instead of a single M-step, are considered. The two CM-steps arise from the partition $\vartheta_Y = (\vartheta_{Y,1}, \vartheta_{Y,2})$, where $\vartheta_{Y,1} = (\beta, \Sigma_Y, \alpha_Y)$ and $\vartheta_{Y,2} = \eta_Y$.

The E-step, on the $(r+1)$ th iteration, further requires the replacement of u_{itk} with

$$u_{itk}^{(r)} = \frac{\alpha_{Y|k}^{(r)} f_N \left(y_{it}; \mu_Y \left(x_{it}; \beta_k^{(r)} \right), \Sigma_{Y|k}^{(r)} \right)}{f_{CN} \left(y_{it}; \mu_Y \left(x_{it}; \beta_k^{(r)} \right), \Sigma_{Y|k}^{(r)}, \alpha_{Y|k}^{(r)}, \eta_{Y|k}^{(r)} \right)}. \quad (20)$$

The first CM-step, on the same iteration, requires the calculation of $\vartheta_{Y,1}^{(r+1)}$ as the value of $\vartheta_{Y,1}$ that maximizes $Q_3(\vartheta_{Y,1} \mid \vartheta_{Y,2} = \vartheta_{Y,2}^{(r)})$. In particular, after some

algebra, we obtain

$$\alpha_{Y|k}^{(r+1)} = \max \left\{ \alpha_Y^*, \frac{1}{\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)}} \sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)} u_{itk}^{(r)} \right\}, \quad (21)$$

$$\beta_k^{(r+1)} = \left[\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)} \left(u_{itk}^{(r)} + \frac{1 - u_{itk}^{(r)}}{\eta_{Y|k}^{(r)}} \right) \mathbf{x}_{it}^* \mathbf{x}_{it}' \right]^{-1} \left[\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)} \left(u_{itk}^{(r)} + \frac{1 - u_{itk}^{(r)}}{\eta_{Y|k}^{(r)}} \right) \mathbf{x}_{it}^* \mathbf{y}_{it}' \right], \quad (22)$$

$$\Sigma_{Y|k}^{(r+1)} = \frac{\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)} \left(u_{itk}^{(r)} + \frac{1 - u_{itk}^{(r)}}{\eta_{Y|k}^{(r)}} \right) [y_{it} - \mu_Y(\mathbf{x}_{it}; \beta_k^{(r+1)})] [y_{it} - \mu_Y(\mathbf{x}_{it}; \beta_k^{(r+1)})]'}{\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)}}. \quad (23)$$

Formula (21) takes into account that the proportion of good “vertical” data in state k , $k = 1, \dots, K$, may be required to be at least equal to a pre-determined value α_Y^* . In the analyses of Sect. 6, as habit in robust statistics, we assume $\alpha_Y^* = 0.5$; see also Punzo and McNicholas (2016, 2017), Mazza and Punzo (2017) and Punzo et al. (2018b).

The second CM-step, on the same iteration, requires the calculation of $\vartheta_{Y,2}^{(r+1)}$ as the value of $\vartheta_{Y,2}$ that maximizes $Q_3(\vartheta_{Y,2} \mid \vartheta_{Y,1} = \vartheta_{Y,1}^{(r+1)})$. For each $k = 1, \dots, K$, this yields

$$\eta_{Y|k}^{(r+1)} = \max \left\{ \eta_{\min}, \frac{b_{Y|k}^{(r+1)}}{d_Y a_{Y|k}^{(r+1)}} \right\}, \quad (24)$$

where

$$a_{Y|k}^{(r+1)} = \sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)} (1 - u_{itk}^{(r)})$$

and

$$b_{Y|k}^{(r+1)} = \sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)} (1 - u_{itk}^{(r)}) \delta(y_{it}, \mu_Y(\mathbf{x}_{it}; \beta_k^{(r+1)}); \Sigma_{Y|k}^{(r+1)}),$$

with $\eta_{\min} > 1$ being a minimum value for $\eta_{Y|k}^{(r+1)}$; in the analyses herein, we use $\eta_{\min} = 1.001$.

4.2.4 Multivariate normal distribution for X

If $X_{it} \mid S_{it} = k \sim \mathcal{N}_{d_X}(\boldsymbol{\mu}_{X|k}, \boldsymbol{\Sigma}_{X|k})$, then no further calculations are needed for the E-step on the $(r + 1)$ th iteration. Hence, on the same iteration, we can skip to the M-step; it requires the calculation of $\boldsymbol{\vartheta}_X^{(r+1)}$ as the value of $\boldsymbol{\vartheta}_X$ that maximizes

$$Q_4(\boldsymbol{\vartheta}_X) = -\frac{1}{2} \sum_{i=1}^I \sum_{t=1}^T \sum_{k=1}^K z_{itk}^{(r)} \left\{ d_X \ln(2\pi) + \ln |\boldsymbol{\Sigma}_{X|k}| + \delta(\mathbf{x}_{it}, \boldsymbol{\mu}_{X|k}; \boldsymbol{\Sigma}_{X|k}) \right\}.$$

After some algebra, such a maximization yields

$$\begin{aligned} \boldsymbol{\mu}_{X|k}^{(r+1)} &= \frac{\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)} \mathbf{x}_{it}}{\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)}}, \\ \boldsymbol{\Sigma}_{X|k}^{(r+1)} &= \frac{\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)} (\mathbf{x}_{it} - \boldsymbol{\mu}_{X|k}) (\mathbf{x}_{it} - \boldsymbol{\mu}_{X|k})'}{\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)}}. \end{aligned}$$

4.2.5 Multivariate t distribution for X

If $X_{it} \mid S_{it} = k \sim t_{d_X}(\boldsymbol{\mu}_{X|k}, \boldsymbol{\Sigma}_{X|k}, v_{X|k})$, then the E-step, on the $(r + 1)$ th iteration, further requires the replacement of v_{itk} with

$$v_{itk}^{(r)} = \frac{v_{X|k}^{(r)} + d_X}{v_{X|k}^{(r)} + \delta(\mathbf{x}_{it}, \boldsymbol{\mu}_{X|k}; \boldsymbol{\Sigma}_{X|k})}. \quad (25)$$

Thus, by substituting z_{itk} and v_{itk} in (17), with z_{itk} and v_{itk} , respectively, we obtain $Q_4(\boldsymbol{\vartheta}_X)$.

The M-step, on the same iteration, requires the calculation of $\boldsymbol{\vartheta}_X^{(r+1)}$ as the value of $\boldsymbol{\vartheta}_X$ that maximizes $Q_4(\boldsymbol{\vartheta}_X)$. After some algebra, we obtain

$$\begin{aligned} \boldsymbol{\mu}_{X|k}^{(r+1)} &= \frac{\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)} v_{itk}^{(r)} \mathbf{x}_{it}}{\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)} v_{itk}^{(r)}}, \\ \boldsymbol{\Sigma}_{X|k}^{(r+1)} &= \frac{\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)} v_{itk}^{(r)} (\mathbf{x}_{it} - \boldsymbol{\mu}_{X|k}) (\mathbf{x}_{it} - \boldsymbol{\mu}_{X|k})'}{\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)}}. \end{aligned}$$

Unfortunately, a closed form solution is not analytically available for the update $v_{X|k}^{(r+1)}$ of $v_{X|k}$. However, by differentiating $Q_4(\boldsymbol{\vartheta}_X)$ with respect to $v_{X|k}$, we note that $v_{X|k}^{(r+1)}$ is a solution of the equation

$$\begin{aligned}
& -\psi\left(\frac{v_{X|k}}{2}\right) + \ln\left(\frac{v_{X|k}}{2}\right) + 1 + \frac{1}{\sum_{t=1}^T z_{itk}^{(r)} v_{itk}^{(r)}} \sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)} \left[\ln\left(v_{itk}^{(r)}\right) - v_{itk}^{(r)} \right] + \\
& + \psi\left(\frac{v_{X|k} + d_X}{2}\right) - \ln\left(\frac{v_{X|k} + d_X}{2}\right) = 0,
\end{aligned} \quad (26)$$

where $\psi(\cdot)$ is the Digamma function. Operationally, the `uniroot()` function in the **stats** package for R is used to numerically find the root of (26) over the interval $(2, v_X^*)$, with $v_X^* > 2$. We fix $v_X^* = 200$ to facilitate faster convergence.

4.2.6 Multivariate contaminated normal distribution for X

If $X_{it} \mid S_{it} = k \sim \mathcal{CN}_{d_X}(\mu_{X|k}, \Sigma_{X|k}, \alpha_{X|k}, \eta_{X|k})$, then two CM-steps are considered. The two CM-steps arise from the partition $\vartheta_X = (\vartheta_{X,1}, \vartheta_{X,2})$, where $\vartheta_{X,1} = (\mu_X, \Sigma_X, \alpha_X)$ and $\vartheta_{X,2} = \eta_X$.

The E-step, on the $(r+1)$ th iteration, further requires the replacement of v_{itk} with

$$v_{itk}^{(r)} = \frac{\alpha_{X|k}^{(r)} f_N(X_{it}; \mu_{X|k}^{(r)}, \Sigma_{X|k}^{(r)})}{f_{CN}(X_{it}; \mu_{X|k}^{(r)}, \Sigma_{X|k}^{(r)}, \alpha_{X|k}^{(r)}, \eta_{X|k}^{(r)})}. \quad (27)$$

The first CM-step, on the same iteration, requires the calculation of $\vartheta_{X,1}^{(r+1)}$ as the value of $\vartheta_{X,1}$ that maximizes $Q_4(\vartheta_{X,1} \mid \vartheta_{X,2} = \vartheta_{X,2}^{(r)})$. In particular, after some algebra, we obtain

$$\begin{aligned}
\alpha_{X|k}^{(r+1)} &= \left\{ \alpha_X^*, \frac{1}{\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)}} \sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)} v_{itk}^{(r)} \right\}, \\
\mu_{X|k}^{(r+1)} &= \frac{\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)} \left(v_{itk}^{(r)} + \frac{1 - v_{itk}^{(r)}}{\eta_{X|k}^{(r)}} \right) y_{it}}{\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)} \left(v_{itk}^{(r)} + \frac{1 - v_{itk}^{(r)}}{\eta_{X|k}^{(r)}} \right)}, \\
\Sigma_{X|k}^{(r+1)} &= \frac{\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)} \left(v_{itk}^{(r)} + \frac{1 - v_{itk}^{(r)}}{\eta_{X|k}^{(r)}} \right) (x_{it} - \mu_{X|k}) (x_{it} - \mu_{X|k})'}{\sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)}}. \quad (28)
\end{aligned}$$

The second CM-step, on the same iteration, requires the calculation of $\boldsymbol{\vartheta}_{X,2}^{(r+1)}$ as the value of $\boldsymbol{\vartheta}_{X,2}$ that maximizes $Q_4\left(\boldsymbol{\vartheta}_{X,2} \mid \boldsymbol{\vartheta}_{X,1} = \boldsymbol{\vartheta}_{X,1}^{(r+1)}\right)$. In particular, for each $k = 1, \dots, K$, we have

$$\eta_{X|k}^{(r+1)} = \max \left\{ \eta_{\min}, \frac{b_{X|k}^{(r+1)}}{d_X a_{X|k}^{(r+1)}} \right\}, \quad (30)$$

where

$$a_{X|k}^{(r+1)} = \sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)} \left(1 - v_{itk}^{(r)}\right)$$

and

$$b_{X|k}^{(r+1)} = \sum_{i=1}^I \sum_{t=1}^T z_{itk}^{(r)} \left(1 - v_{itk}^{(r)}\right) \delta\left(\mathbf{x}_{it}, \boldsymbol{\mu}_{X|k}^{(r+1)}; \boldsymbol{\Sigma}_{X|k}^{(r+1)}\right),$$

with η_{\min} defined as in (24).

5 Operational aspects

5.1 Initialization strategy

The EM/ECM algorithms described in Sect. 4 are initialized by the solution provided by the corresponding MRMRCs by further considering $(\mathbf{1}_K \mathbf{1}'_K + s \mathbf{I}_K) / (K + s)$ as the starting value for the $\boldsymbol{\Pi}$ -matrix, where $\mathbf{1}_K$ is a column vector of K ones, \mathbf{I}_K denotes the $K \times K$ identity matrix, and s is a suitable constant; in our analyses, we fix $s = 9$ as in Bartolucci and Farcomeni (2009). In turn, MRMRCs are initialized according to the partition provided by the K -means method as implemented by the `kmeans()` function of the `stats` package.

5.2 Robustness and detection of atypical observations

HMRMRCs with heavy-tailed models for $f(\mathbf{y}_{it} | \mathbf{x}_{it}, S_{it} = k)$ and $f(\mathbf{x}_{it} | S_{it} = k)$ in (3), adhere to the taxonomy of atypical observations given in Table 1. HMRMFCs with a heavy-tailed model for $f(\mathbf{y}_{it} | \mathbf{x}_{it}, S_{it} = k)$ in (2), only adhere to the classification of each observation as typical or outlier. These “adherences” are useful from a twofold point of view, as we describe below.

Firstly, we obtain robust estimates of $\boldsymbol{\beta}_k$ and $\boldsymbol{\Sigma}_{Y|k}$ when $f(\mathbf{y}_{it} | \mathbf{x}_{it}, S_{it} = k)$ is heavy-tailed, and robust estimates of $\boldsymbol{\mu}_{X|k}$ and $\boldsymbol{\Sigma}_{X|k}$ when $f(\mathbf{x}_{it} | S_{it} = k)$ is heavy-tailed, $k = 1, \dots, K$. Robustness is attained because the estimates of $\boldsymbol{\beta}_k$, $\boldsymbol{\Sigma}_{Y|k}$, $\boldsymbol{\mu}_{X|k}$, and $\boldsymbol{\Sigma}_{X|k}$ are weighted means where the weights are a function of the squared Mahalanobis distances δ . This is the underlying idea of M -estimation (Maronna 1976),

where a decreasing weighting function $w(\delta) : (0, \infty) \rightarrow (0, \infty)$ is used to down-weight the observations with large δ values. For the t distribution, the weights arise from the E-step of the EM algorithm and are given in (18) and (25). For the contaminated normal distribution, the weights are

$$u_{itk}^{(r)} + \frac{1 - u_{itk}^{(r)}}{\eta_{Y|k}^{(r)}} \quad (31)$$

for β_k and $\Sigma_{Y|k}$, as shown in (22)–(23), and

$$v_{itk}^{(r)} + \frac{1 - v_{itk}^{(r)}}{\eta_{X|k}^{(r)}} \quad (32)$$

for $\mu_{X|k}$ and $\Sigma_{X|k}$, as shown in (28)–(29).

Secondly, we can consider an a posteriori procedure (i.e. a procedure taking place once the model is fitted) to automatically detect mildly atypical observations. Below, we describe such a procedure for the t - t -HMRMRCs and CN-CN-HMRMRCs. Regardless of the considered model, each observation $(\mathbf{x}_{it}, \mathbf{y}_{it})$ is first assigned to one of the K states through the maximum a posteriori probabilities (MAP) operator

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

where \hat{z}_{itk} is the expected value of Z_{itk} at convergence of the EM/ECM algorithm, and then classified in one of the four categories of Table 1. For the t - t -HMRMRCs, extending the idea illustrated by McLachlan and Peel (2000, p. 232) for mixtures of multivariate t distributions, we define

$$\sum_{k=1}^K \text{MAP}(\hat{z}_{itk}) \delta(\mathbf{y}_{it}, \mu_Y(\mathbf{x}_{it}, \hat{\beta}_k); \hat{\Sigma}_{Y|k}) \quad (33)$$

and

$$\sum_{k=1}^K \text{MAP}(\hat{z}_{itk}) \delta(\mathbf{x}_{it}, \hat{\mu}_{X|k}; \hat{\Sigma}_{X|k}), \quad (34)$$

and we categorize $(\mathbf{x}_{it}, \mathbf{y}_{it})$ as bad leverage if (33) and (34) are both sufficiently large, as outlier or good leverage if only (33) or (34), respectively, are sufficiently large, and as typical otherwise. In (33) and (34), the hat denotes the values of the parameters at convergence of the EM algorithm. To decide on how large the statistics (33) and (34) must be in order to be defined as “sufficiently large”, always extending the idea of McLachlan and Peel (2000, p. 232), we can compare them to the $(1 - \alpha)$ 100th percentile of the χ^2 distribution with d_Y and d_X degrees of freedom, where the χ^2 distribution is used to approximate the distribution of the squared Mahalanobis distances in (33) and (34), respectively. For the CN-CN-HMRMRCs, let \hat{u}_{itk} and \hat{v}_{itk} be the

Table 3 CN-CN-HMRMRCs: rule for classifying a generic observation (x_{it}, y_{it}) in one of the four categories of Table 1

$\hat{u}_{itk} \wedge \hat{v}_{itk}$	$[0, 0.5)$	$[0.5, 1]$
$[0, 0.5)$	Bad leverage	Outlier
$[0.5, 1]$	Good leverage	Typical

expected values of U_{itk} and V_{itk} , respectively, at convergence of the ECM algorithm. Thus, we can straightforwardly apply the rule in Table 3 to classify (x_{it}, y_{it}) in one of the 4 categories of Table 1. Summarizing, once the observation has been classified in one of the K states, t - t -HMRMRCs and CN-CN-HMRMRCs reveal richer information about the role of that observation in that state. Moreover, the resulting information can be used to possibly eliminate some of the atypical observations (such as outliers and bad leverage points).

6 Data analysis

6.1 Sensitivity study based on artificial data

A sensitivity study, based on an artificial longitudinal version of a real data set, is here described to compare how a single atypical observation affects N-N-HMRMRCs and how it is instead handled by t - t -HMRMRCs and CN-CN-HMRMRCs. Insights about these robust methods are also given.

The Students data set—analyzed by Ingrassia et al. (2014) and available in the **flexCWM** package (Mazza et al. 2018) for R — is a suitable benchmark data set for this purpose. The data come from a survey of $I = 270$ students attending a statistics course at the Department of Economics and Business of the University of Catania in the academic year 2011/2012. The following analysis only concerns, for illustrative purposes, the variables HEIGHT (height of the respondent, measured in centimeters) and HEIGHT.F (height of respondent's father, measured in centimeters). Therefore, the role of HEIGHT and HEIGHT.F as outcome and covariate, respectively, is clearly justified. The students are subdivided in two groups of size 119 and 151 (corresponding to $\pi_1 = 0.441$ and $\pi_2 = 0.559$). The ML estimates of the parameters $\beta_1, \beta_2, \Sigma_{Y|1}, \Sigma_{Y|2}, \mu_{X|1}, \mu_{X|2}, \Sigma_{X|1}$, and $\Sigma_{X|2}$ are

$$\begin{aligned} \hat{\beta}_1 &= \begin{pmatrix} 61.243 \\ 0.667 \end{pmatrix}, \quad \hat{\beta}_2 = \begin{pmatrix} 54.894 \\ 0.608 \end{pmatrix}, \quad \hat{\Sigma}_{Y|1} = 11.312, \quad \hat{\Sigma}_{Y|2} = 15.337, \\ \hat{\mu}_{X|1} &= 174.143, \quad \hat{\mu}_{X|2} = 175.609, \quad \hat{\Sigma}_{X|1} = 36.089, \quad \text{and} \quad \hat{\Sigma}_{X|2} = 33.602. \end{aligned}$$

Based on these estimates, and further introducing a transition probabilities matrix

$$\Pi = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix},$$

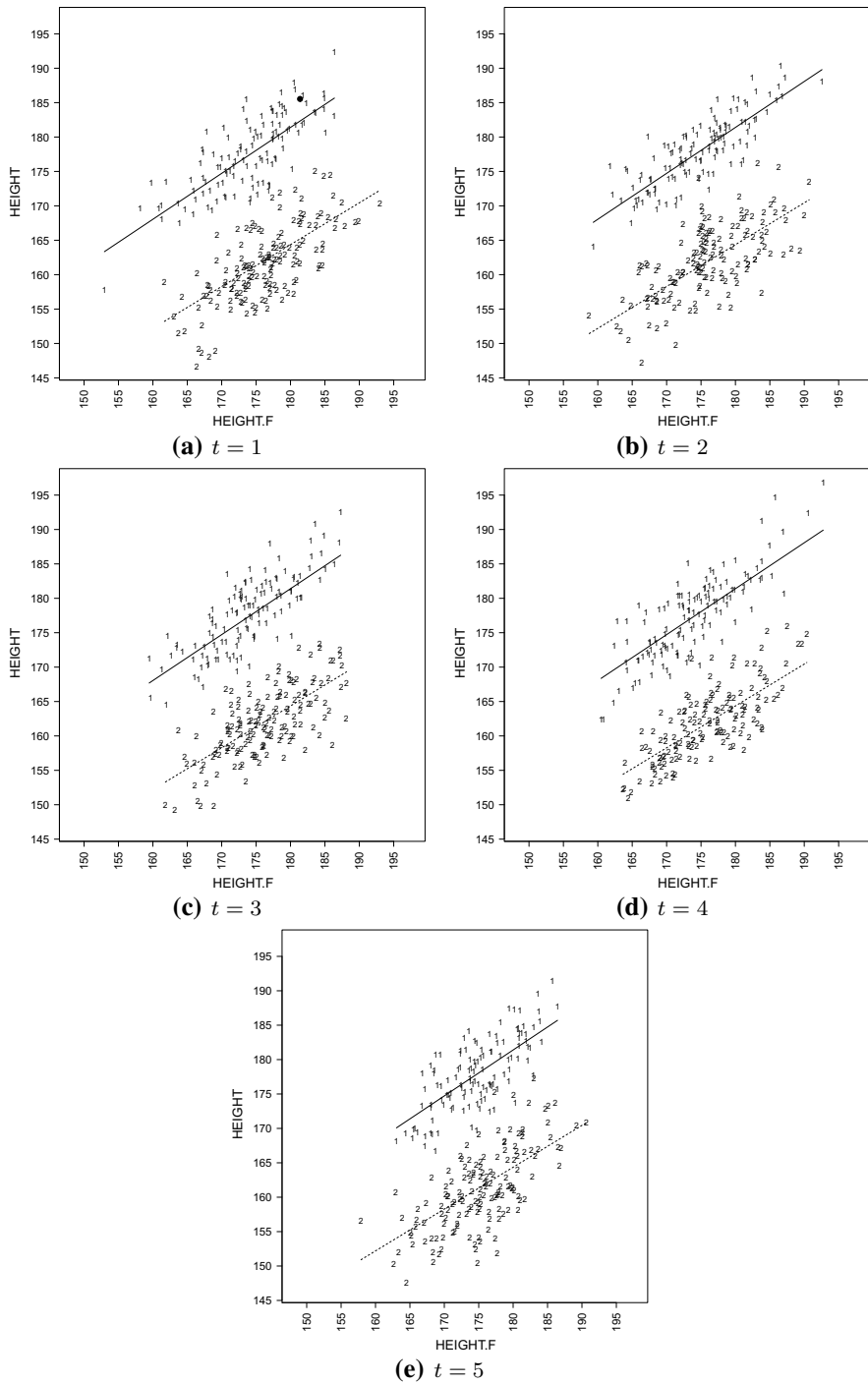
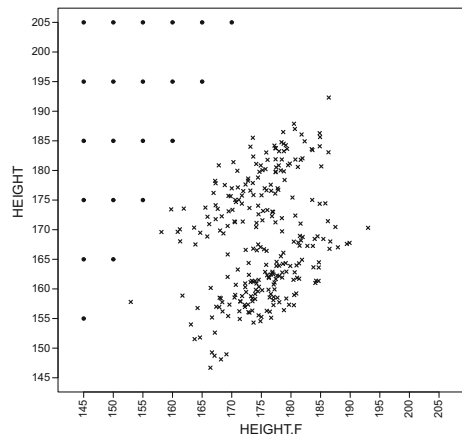


Fig. 1 Scatterplots of the artificial data (1 and 2 denote group 1 and group 2, respectively; (filled circle) denotes the observation perturbed for the analysis of Section 6.1)

Fig. 2 Artificial longitudinal Students data: scatter plot, at time 1, with each (filled circle) denoting the perturbed observations



we randomly generate, from a N-N-HMRMRCs, a longitudinal version of this data set on $T = 5$ times; the data set is available at <http://www.economia.unict.it/punzo/Data.htm>. The scatterplots of the generated data, for each $t \in \{1, \dots, 5\}$, are displayed in Fig. 1. By substituting one randomly selected point at time 1 (highlighted by a bullet in Fig. 1a) with 21 atypical points, 21 “perturbed” data sets are created. These atypical points are all displayed together, as bullets, in Fig. 2. They represent different types of (local) atypical observations in accordance to Table 1.

For each perturbed data set we directly fit N-N-HMRMRCs, t - t -HMRMRCs, and CN-CN-HMRMRCs, with $K = 2$. For each of the three competing models, Table 4 reports the number of misallocated observations for each perturbed data set. It can be seen that, as expected, clusterings from t - t -HMRMRCs and CN-CN-HMRMRCs are more robust than clustering from N-N-HMRMRCs to the perturbations close to the atypical point of coordinates (145, 205); after all, these are the points mainly suspected to be bad leverage or outliers. However, CN-CN-HMRMRCs are systematically the most robust to these perturbations, with the number of misallocated observations remaining fixed at 6 regardless of the particular value perturbed; this is especially in contrast to N-N-HMRMRCs where the number of misclassifications changes as the extent of the perturbation increases. Table 5 shows how the perturbed values are labelled by CN-CN-HMRMRCs, based on the rule outlined in Table 3. Interestingly, all the non-perturbed points are classified as typical by CN-CN-HMRMRCs. Moreover, while the regression lines from t - t -HMRMRCs and CN-CN-HMRMRCs are not substantially different from those displayed in Fig. 1(a), because of their robustness properties described in Sect. 5.2, there are some scenarios where one of the regression lines from the estimated N-N-HMRMRCs is dragged towards the atypical point. This happens for the atypical points on the top-left corner of Fig. 2.

To complete the analysis for t - t -HMRMRCs, Table 6 reports the values at convergence of the weights v_{i1k} and u_{i1k} , given in (25) and (18), respectively, for the perturbed value only, where k is chosen such that $\text{MAP}(\hat{z}_{i1k}) = 1$. As expected, v_{i1k} decreases as the value of HEIGHT.F, for the perturbed point, further departs from the bulk of the values of HEIGHT.F in that latent state, almost regardless from the value of

HEIGHT; this can be easily noted by looking at Table 6(a) column-by-column. Vertically speaking, u_{i1k} decreases as the atypical point further departs from the regression line of the state the perturbed point is assigned. Analogous considerations can be made about the estimated degrees of freedom $\hat{v}_{\text{HEIGHT.F}|k}$ and $\hat{v}_{\text{HEIGHT}|k}$, in the MAP-state of membership of the perturbed point, reported in Table 7.

To complete the analysis for CN-CN-HMRMRCs, Table 8 reports the values at convergence of the posterior probabilities v_{i1k} and u_{i1k} , given in (27) and (20), respectively, for the perturbed value only, where k is chosen such that $\text{MAP}(\hat{z}_{i1k}) = 1$. As expected, the probability v_{i1k} to be a typical horizontal point decreases as the value of HEIGHT.F, for the perturbed point, further departs from the bulk of the values of HEIGHT.F, regardless from the value of HEIGHT; this can be easily noted by looking at Table 8(a) column-by-column. Vertically speaking, the probability u_{i1k} decreases as the perturbed point further departs from the regression line of the state the point is assigned, as we can note in Table 8(b). Advantageously, these probabilities are of easier and immediate interpretation with respect to the weights in Table 6(b) for t - t -HMRMRCs. Through the functions given in (32) and (31), these probabilities are also related to the down-weighting of the atypical point in the estimation phase and this is an important aspect for the robust estimation of the parameters (cf. Sect. 5.2); the values of (32) and (31) are reported in Table 9 and they can be compared with the weights from t - t -HMRMRCs reported in Table 6. Analogous considerations can be made about the estimated degrees of atypicality $\hat{\eta}_{\text{HEIGHT.F}|k}$ and $\hat{\eta}_{\text{HEIGHT}|k}$ in the MAP-state of membership of the perturbed point, reported in Table 10. As an example, the values of $\hat{\eta}_{\text{HEIGHT}|k}$ in Table 10(b) increase as the perturbed point departs from the estimated regression line of that state; this is the reason why this parameter can be also meant as a sort of “degree of (vertical) atypicality” of the (vertical) atypical point(s), i.e. as a measure of how vertically different atypical observations are from the estimated regression line of that state.

6.2 Heart rate activity

The data here analysed come from a multicenter clinical study on heart conditions in an elderly population conducted in Italy. In the following we focus on heart rate and blood pressure response under exercise testing as main outcomes. Exercise testing is a cardiovascular stimulation test which is performed monitoring the electrocardiogram and the blood pressure. The exercise test used in this study measures the steps used to turn around 360 degree without help. It is simple, objective, inexpensive, reproducible and required no advanced equipment, and it is used to estimate prognosis, to determine the functional capacity, to assess the probability and extent of heart diseases, as such an exercise testing is an indicator of cardiovascular health and is a good predictor of future diseases. Measured SBP, DBP and heart rate are useful tools for risk stratification and are of interest when planning and evaluating medical treatment, surgery or rehabilitation. However, in practice, it can be rather difficult to define the figures for SBP and DBP during exercise testing, especially the latter. We determine these measurements through automatic equipment that offer many advantages over manual

Table 4 Artificial longitudinal students data: number of misallocations over $I \times T = 270 \times 5 = 1350$ observations

HEIGHT	HEIGHT.F					
	145	150	155	160	165	170
(a) N-N-HMRMRCs						
205	8	8	8	8	7	7
195	8	7	7	7	6	
185	7	7	6	6		
175	6	6	6			
165	6	6				
155	6					
(b) <i>t-t</i> -HMRMRCs						
205	7	7	7	6	6	6
195	6	6	6	6	6	
185	6	6	6	6		
175	6	6	6			
165	6	6				
155	6					
(c) CN-CN-HMRMRCs						
205	6	6	6	6	6	6
195	6	6	6	6	6	
185	6	6	6	6		
175	6	6	6			
165	6	6				
155	6					

Table 5 Artificial longitudinal students data: labelling of the perturbed value by CN-CN-HMRMRCs

HEIGHT	HEIGHT.F					
	145	150	155	160	165	170
205	Bad leverage	Outlier	Outlier	Outlier	Outlier	Outlier
195	Bad leverage	Outlier	Outlier	Outlier	Outlier	
185	Bad leverage	Outlier	Outlier	Outlier		
175	Bad leverage	Outlier	Typical			
165	Good leverage	Typical				
155	Good leverage					

techniques, having verified its utility and clinical validity. Also the weight has been included as a covariate.

A total of $I = 370$ individuals were visited every 3 months over a year (i.e., $T = 4$) and several information recorded. More individuals were recruited in the study, but they were excluded if they: had history of cardiovascular diseases; had electrocardiographic evidence of coronary heart disease or cardiac arrhythmia; were hypertensive, as defined by currently using any anti-hypertensive medication; had incomplete exer-

Table 6 Artificial longitudinal Students data: weights from t - t -HMRMCs for the perturbed value in its MAP-state of membership

HEIGHT	HEIGHT.F					
	145	150	155	160	165	170
(a) \widehat{v}_{i1k}						
205	0.517	0.750	0.902	0.959	0.988	1.005
195	0.517	0.750	0.902	0.959	0.988	
185	0.517	0.749	0.901	0.959		
175	0.516	0.748	0.901			
165	0.515	0.748				
155	0.515					
(b) \widehat{u}_{i1k}						
205	0.039	0.047	0.057	0.072	0.092	0.121
195	0.069	0.088	0.115	0.157	0.226	
185	0.150	0.214	0.325	0.541		
175	0.516	0.834	0.946			
165	0.979	0.997				
155	1.004					

Table 7 Artificial longitudinal students data: estimated degrees of freedom for the state containing the perturbed value

HEIGHT	HEIGHT.F					
	145	150	155	160	165	170
(a) $\widehat{v}_{\text{HEIGHT.F} k}$						
205	24.170	44.909	82.698	105.700	109.789	106.222
195	24.149	44.836	82.570	105.540	109.583	
185	24.116	44.714	82.333	105.210		
175	24.049	44.467	81.992			
165	23.982	44.346				
155	23.983					
(b) $\widehat{v}_{\text{HEIGHT} k}$						
205	8.389	8.697	9.079	9.565	10.207	11.103
195	9.494	10.104	10.948	12.203	14.295	
185	12.021	13.957	17.780	29.268		
175	27.631	80.295	156.958			
165	189.932	192.696				
155	186.998					

cise measurement data. Based on their exercise test results, blood pressure and heart rate response to exercise were evaluated. Clinical characteristics of the study sample are shown in Table 11.

Another noteworthy aspect is that the clinical outcomes considered in our setting measure different aspects of the same phenomenon, i.e. heart activity. Thus, a strong

Table 8 Artificial longitudinal students data: probabilities to be a good point, from CN-CN-HMRMCs, for the perturbed value in its MAP-state of membership

HEIGHT	HEIGHT.F					
	145	150	155	160	165	170
(a) v_{ik}						
205	0.010	0.985	1.000	1.000	1.000	1.000
195	0.010	0.985	1.000	1.000	1.000	
185	0.010	0.985	1.000	1.000		
175	0.010	0.985	1.000			
165	0.010	0.985				
155	0.010					
(b) u_{ik}						
205	0.000	0.000	0.000	0.000	0.000	0.000
195	0.000	0.000	0.000	0.000	0.000	
185	0.000	0.000	0.000	0.003		
175	0.002	0.153	1.000			
165	1.000	1.000				
155	1.000					

Table 9 Artificial longitudinal students data: weights from CN-CN-HMRMCs for the perturbed value in its MAP-state of membership

HEIGHT	HEIGHT.F					
	145	150	155	160	165	170
(a) Weights with respect to X (based on v_{ik})						
205	0.154	0.987	1.000	1.000	1.000	1.000
195	0.154	0.987	1.000	1.000	1.000	
185	0.154	0.987	1.000	1.000		
175	0.156	0.987	1.000			
165	0.154	0.987				
155	0.154					
(b) Weights with respect to Y (based on u_{ik})						
205	0.007	0.008	0.010	0.013	0.016	0.021
195	0.012	0.016	0.020	0.027	0.037	
185	0.026	0.035	0.050	0.083		
175	0.075	0.283	1.000			
165	1.000	1.000				
155	1.000					

dependence would be expected between outcomes. Therefore, the association between different outcomes constitutes a crucial aspect of the analysis as well, and should not be neglected or treated as nuisance. Moreover, a further type of dependence can be investigate between the outcomes and the available covariates. Figure 3 shows pairwise scatter plots of the variables considered: the green line results from a simple linear regression, the solid red line from a nonparametric-regression line via a generalized additive model, and the red dashed lines correspond to 90% confidence intervals of the

Table 10 Artificial longitudinal students data: estimated degrees of atypicality for the state containing the perturbed value

HEIGHT	HEIGHT.F					
	145	150	155	160	165	170
(a) $\hat{\eta}_{\text{HEIGHT.F} k}$						
205	6.880	8.515	1.899	1.357	1.300	1.307
195	6.871	8.516	1.896	1.346	1.277	
185	6.857	8.516	1.825	1.105		
175	6.802	8.513	1.901			
165	6.880	8.515				
155	6.877					
(b) $\hat{\eta}_{\text{HEIGHT} k}$						
205	143.612	119.727	97.870	78.756	62.104	47.675
195	82.079	64.805	50.051	37.469	27.142	
185	39.392	28.701	19.933	12.611		
175	13.692	6.516	1.196			
165	1.005	1.117				
155	1.006					

Table 11 Heart activity: clinical characteristics. Values are reported as mean \pm standard deviation

Variable	First visit	Second visit	Third visit	Fourth visit
SBP	150.96 \pm 23.30	138.60 \pm 21.26	135.44 \pm 16.32	132.41 \pm 18.65
DBP	85.38 \pm 13.74	80.74 \pm 12.16	85.29 \pm 12.92	83.27 \pm 11.22
Heart rate	75.68 \pm 12.08	73.55 \pm 9.53	72.37 \pm 8.00	72.11 \pm 9.40

smoothing line. Finally, the main diagonal contains marginal densities estimated by a nonparametric kernel smoother. From the visual perspective, it is apparent that patterns of non-linear correlation can be detected between outcomes, and that the correlation structure is rather heterogeneous since each variable (outcome or covariate) seems to be related to others in different ways.

For sake of brevity, we show and discuss results for the N-N-, t - t - and CN-CN-HMRMRCs only. We fitted several HMRMRCs, with one to six hidden states, to SBP, DBP and heart rate, as outcomes ($d_Y = 3$) and weight and steps as covariates ($d_X = 2$). We compare the fitted models on the basis of Bayesian information criterion (BIC) and integrated complete likelihood (ICL); see Table 12.

The four-state model is chosen by both the BIC the ICL, and the CN-CN-HMRMRC is preferred. Parameters estimates are given in Table 13. The estimation procedure outlined in Sect. 4 does not produce standard errors of the estimates. Visser et al. (2000) investigate the reliable estimation of confidence bands in the context of HMMs and recommend bootstrap-based techniques. We followed their proposal and implemented a parametric bootstrap approach to get the standard errors of estimated coefficients.

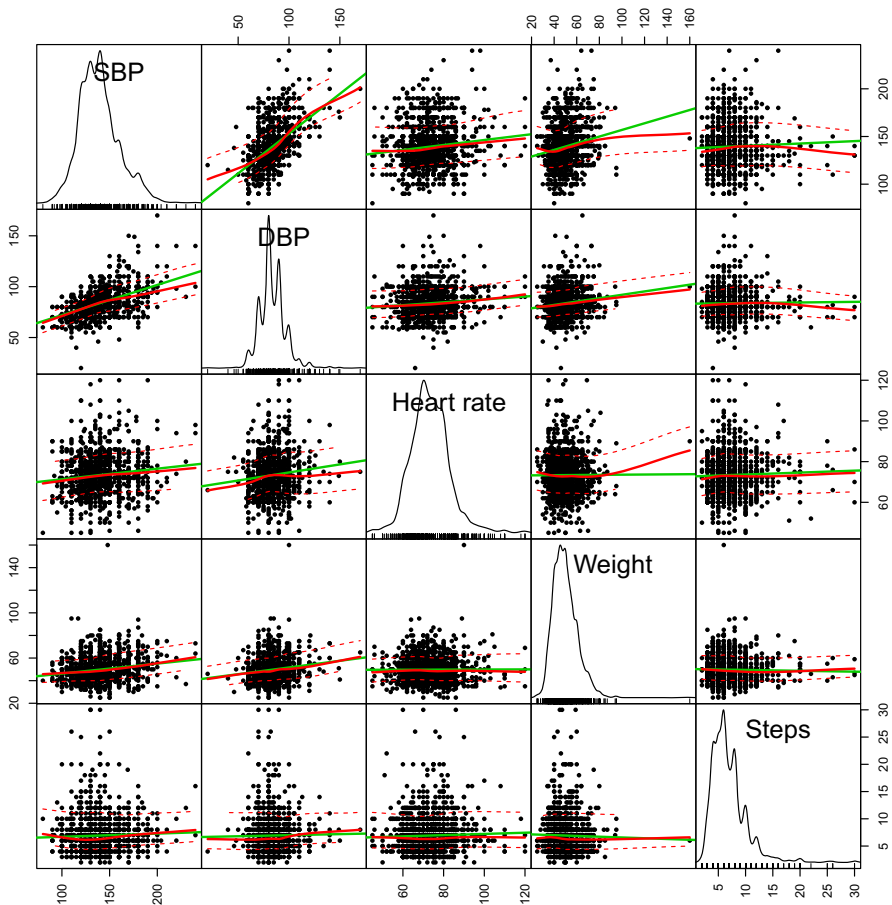


Fig. 3 Heart activity: marginal densities and scatter plots of clinical measurements. The green line results from linear regression, and the red line from a generalized additive model. The dashed red lines determine a 90% confidence interval of the smoother. The main diagonal shows marginal densities resulting from a kernel density estimator

Results are of different kind. Here, we mainly refer to the inferred clustering structure, i.e., the hidden states, and the detection of outlying observations. We can identify state-specific: regression models, covariance/correlation structures between outcomes and between covariates, mean values of the covariates. Moreover, we can describe the evolution over time of the obtained clusters. The clusters, obtained via a *local maximum a posteriori* procedure, are displayed in Fig. 4. Clusters are, in general, separated enough, though some overlapping can be observed with State 1 and the others; histograms of the posterior state probabilities are used to visually assess this aspect (see Fig. 5).

State 1 is characterized by high correlation of the heart rate with SBP and DBP. The effect of covariates, mainly the number of steps, is significant and strongly affects the main outcomes. State 2 clusters *not-at-risk* patients, having normal blood pressures,

Table 12 Heart activity: model selection

<i>K</i>	N-N-HMRMRC		<i>t-t</i> -HMRMRC		CN-CN-HMRMRC	
	BIC	ICL	BIC	ICL	BIC	ICL
1	− 54,467.29	− 54,467.29	− 53,706.23	− 53,706.23	− 53,740.47	− 53740.47
2	− 53,621.96	− 53,779.59	− 53,360.65	− 53,561.74	− 53,385.35	− 53602.09
3	− 53,435.38	− 53,710.39	− 53,099.25	− 53,313.42	− 53,176.83	− 53,470.42
4	− 53,147.97	− 53,376.88	− 53,076.62	− 53,394.25	− 53,049.33	− 53,315.24
5	− 53,258.58	− 53,545.64	− 53,086.40	− 53,395.55	− 53,068.57	− 53,352.58
6	− 53,242.84	− 53,559.70	− 53,191.85	− 53,539.62	− 53,138.26	− 53,442.50

Bold style highlights the best value according to each criterion

Table 13 Heart activity: Estimated parameters for the four-state HMRMRCs, with standard errors in *italic*

$\hat{\pi}$	$\hat{\eta}$			
	$\begin{pmatrix} 0.418 \\ 0.056 \\ 0.118 \\ 0.039 \\ 0.094 \\ 0.029 \\ 0.369 \\ (0.043) \end{pmatrix}$	$\begin{pmatrix} 0.401 & 0.598 & 0.001 & 0.000 \\ 0.090 & 0.072 & 0.001 & 0.000 \\ 0.012 & 0.924 & 0.063 & 0.000 \\ 0.020 & 0.034 & 0.012 & 0.000 \\ 0.000 & 0.033 & 0.827 & 0.140 \\ 0.004 & 0.027 & 0.059 & 0.040 \\ 0.000 & 0.001 & 0.502 & 0.497 \\ 0.005 & 0.001 & 0.054 & 0.054 \end{pmatrix}$		
$\hat{\beta}_1$	$\begin{pmatrix} 151.445 & 65.730 & 81.262 \\ 11.500 & 8.529 & 4.492 \\ -0.032 & 0.181 & 0.011 \\ 0.078 & 0.081 & 0.135 \\ 0.308 & 1.368 & -0.759 \\ 0.138 & 0.476 & 0.367 \end{pmatrix}$	$\begin{pmatrix} 131.771 & 79.449 & 76.329 \\ 6.030 & 4.304 & 2.390 \\ -0.042 & 0.052 & -0.128 \\ 0.107 & 0.061 & 0.078 \\ -0.119 & -0.063 & 0.084 \\ 0.331 & 0.188 & 0.123 \end{pmatrix}$	$\begin{pmatrix} 134.016 & 85.425 & 79.369 \\ 11.442 & 9.047 & 2.913 \\ -0.043 & -0.063 & -0.147 \\ 0.163 & 0.134 & 0.066 \\ 0.160 & 0.292 & 0.132 \\ 0.383 & 0.110 & 0.123 \end{pmatrix}$	$\begin{pmatrix} 169.274 & 81.506 & 77.602 \\ 13.385 & 3.691 & 7.530 \\ -0.171 & 0.089 & -0.081 \\ 0.163 & 0.077 & 0.047 \\ -0.851 & -0.191 & 0.242 \\ 0.138 & 0.188 & 0.063 \end{pmatrix}$
$\hat{\Sigma}^Y 1$	$\begin{pmatrix} 384.076 & 123.260 & 3.289 \\ 18.510 & 16.367 & 1.278 \\ 123.260 & 143.771 & 30.808 \\ 16.367 & 13.979 & 6.630 \\ 3.289 & 30.808 & 101.00 \\ 1.278 & 6.630 & 17.927 \end{pmatrix}$	$\begin{pmatrix} 147.973 & 61.405 & 6.599 \\ 13.797 & 10.136 & 2.334 \\ 61.405 & 89.277 & 5.418 \\ 10.136 & 7.920 & 1.895 \\ 6.599 & 5.418 & 47.297 \\ 2.334 & 1.895 & 9.269 \end{pmatrix}$	$\begin{pmatrix} 174.506 & 58.495 & 0.896 \\ 18.772 & 6.321 & 0.234 \\ 58.495 & 70.111 & 3.376 \\ 6.321 & 6.136 & 0.837 \\ 0.896 & 3.376 & 48.448 \\ 0.234 & 0.837 & 6.269 \end{pmatrix}$	$\begin{pmatrix} 424.039 & 106.380 & -15.234 \\ 21.955 & 3.553 & 3.547 \\ 106.380 & 137.674 & -1.013 \\ 3.553 & 7.968 & 0.479 \\ -15.234 & -1.013 & 64.338 \\ 3.553 & 0.479 & 7.769 \end{pmatrix}$
$\hat{\mu}^X 1$	$\begin{pmatrix} 44.086 \\ 2.633 \\ 7.085 \\ (0.200) \end{pmatrix}$	$\begin{pmatrix} 42.457 \\ 0.485 \\ 6.866 \\ (0.336) \end{pmatrix}$	$\begin{pmatrix} 52.949 \\ 1.664 \\ 5.335 \\ (0.194) \end{pmatrix}$	$\begin{pmatrix} 61.503 \\ 1.573 \\ 7.753 \\ (0.320) \end{pmatrix}$

Table 13 continued

$\hat{\Sigma}_{X 1}$	$\begin{pmatrix} 48.066 & -1.545 \\ 5.736 & 0.291 \\ -1.545 & 5.937 \\ 0.291 & 0.808 \end{pmatrix}$	$\hat{\Sigma}_{X 2}$	$\begin{pmatrix} 25.369 & 1.987 \\ 3.210 & 0.206 \\ 1.987 & 6.806 \\ 0.206 & 0.523 \end{pmatrix}$	$\hat{\Sigma}_{X 3}$	$\begin{pmatrix} 43.481 & -1.232 \\ 4.914 & 0.183 \\ -1.232 & 2.419 \\ 0.183 & 0.541 \end{pmatrix}$	$\hat{\Sigma}_{X 4}$	$\begin{pmatrix} 76.140 & -7.613 \\ 7.619 & 1.045 \\ -7.613 & 8.027 \\ 1.045 & 1.858 \end{pmatrix}$
$\hat{\alpha}_{X 1}$	$\begin{pmatrix} 0.995 \\ 0.020 \end{pmatrix}$	$\hat{\alpha}_{X 2}$	$\begin{pmatrix} 0.909 \\ 0.034 \end{pmatrix}$	$\hat{\alpha}_{X 3}$	$\begin{pmatrix} 1.000 \\ 0.007 \end{pmatrix}$	$\hat{\alpha}_{X 4}$	$\begin{pmatrix} 0.915 \\ 0.029 \end{pmatrix}$
$\hat{\alpha}_{Y 1}$	$\begin{pmatrix} 0.972 \\ 0.024 \end{pmatrix}$	$\hat{\alpha}_{Y 2}$	$\begin{pmatrix} 0.763 \\ 0.065 \end{pmatrix}$	$\hat{\alpha}_{Y 3}$	$\begin{pmatrix} 0.845 \\ 0.047 \end{pmatrix}$	$\hat{\alpha}_{Y 4}$	$\begin{pmatrix} 0.860 \\ 0.040 \end{pmatrix}$
$\hat{\eta}_{X 1}$	$\begin{pmatrix} 130.563 \\ 65.564 \end{pmatrix}$	$\hat{\eta}_{X 2}$	$\begin{pmatrix} 9.373 \\ 19.911 \end{pmatrix}$	$\hat{\eta}_{X 3}$	$\begin{pmatrix} 1.000 \\ 0.040 \end{pmatrix}$	$\hat{\eta}_{X 4}$	$\begin{pmatrix} 6.769 \\ 2.968 \end{pmatrix}$
$\hat{\eta}_{Y 1}$	$\begin{pmatrix} 12.690 \\ 9.310 \end{pmatrix}$	$\hat{\eta}_{Y 2}$	$\begin{pmatrix} 2.903 \\ 0.878 \end{pmatrix}$	$\hat{\eta}_{Y 3}$	$\begin{pmatrix} 4.507 \\ 0.692 \end{pmatrix}$	$\hat{\eta}_{Y 4}$	$\begin{pmatrix} 4.514 \\ 1.125 \end{pmatrix}$

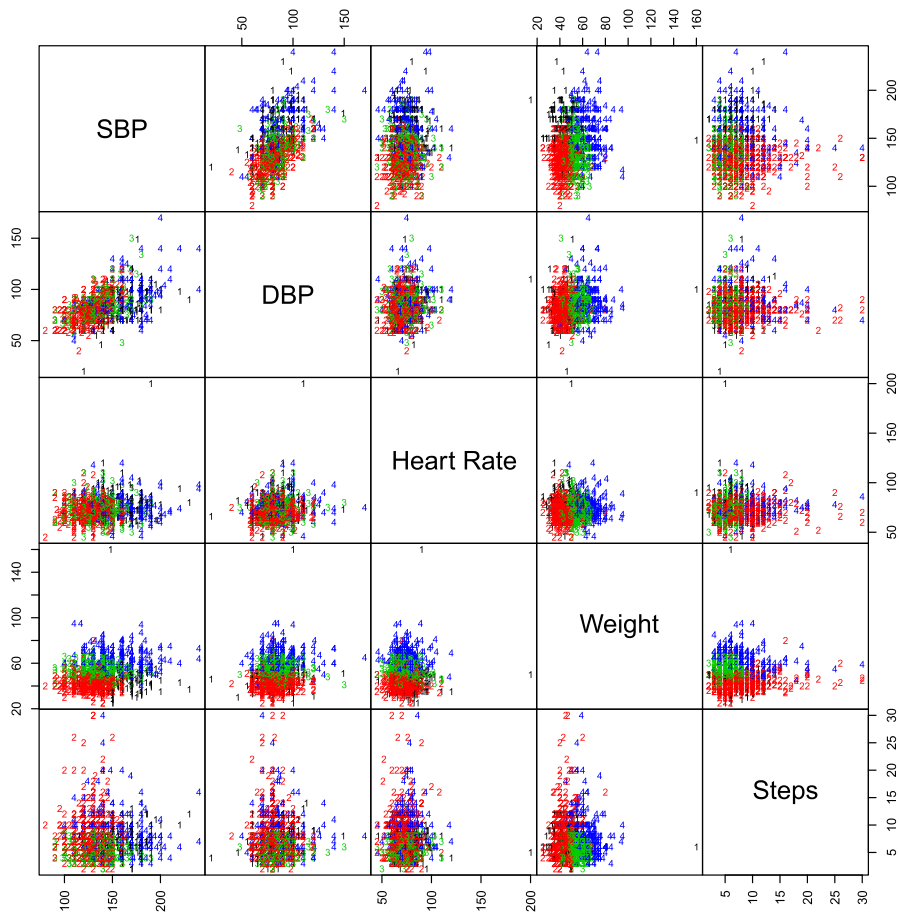


Fig. 4 Heart activity: matrix of scatter plots with inferred classification.

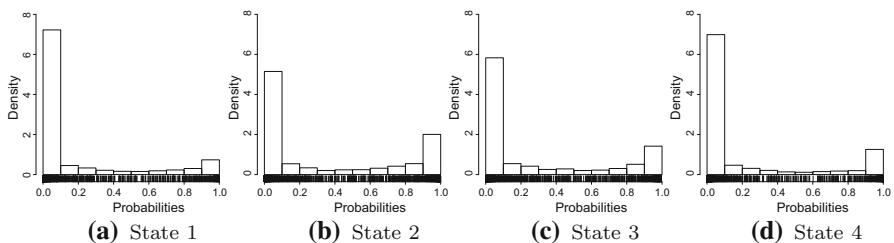


Fig. 5 Heart activity: histograms of the posterior probabilities

low weights and who need just few steps to turn around 360 degree without help. The effect of covariates is not-significant; hence, these patients can be considered as well-being. State 3 does not differ too much from State 2 in terms of heart activity, but in the relationships between the dependent variables and the covariates; slightly higher

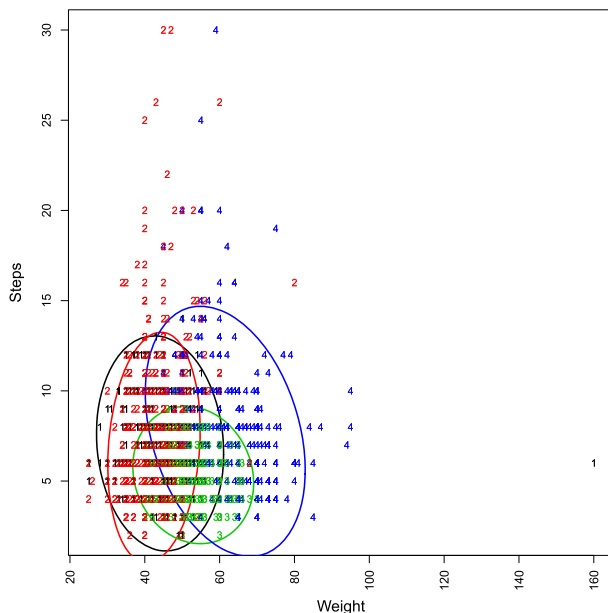


Fig. 6 Heart activity: scatter plot of the covariates only with inferred classification and bivariate contaminated normal density contours at the 95% level for each state

weights are also observed for patients clustered in this state and different correlations between covariates are estimated. For some patients, we observed SBP values rising during exercise, as an indication that the heart is working harder to pump more blood with each contraction to keep muscles supplied with oxygen. Small changes in the DBP are also observed in the same patients. This is not surprising as DBP changes very little, if at all, during exercise; that's because the blood vessels in working muscles widen, decreasing peripheral resistance. However, even such small changes may be an indication of potential risks. These characteristics are observed in patients clustered in State 4, the *at-risk* state, in which an exaggerated blood pressure response to exercise is observed and this indicates a greater risk of hypertension. Moreover, a higher weight is estimated on average ($\mu_{\text{Weight}|4} = 61.503$), indicating also an issue in the life-style conditions.

As a major feature of the HMRMRC approach, the inferred clustering depends on the covariates as well (assignment dependence). From Fig. 4, the role played by the Weight variable in defining the cluster is rather evident as for different weight's levels we have different clusters. To capture differences due to the marginal (over time) bivariate contaminated normal distributions of the covariates in each hidden state, we provide their graphical representation, via isodensities at the 95% level, in Fig. 6.

It is clear that not all clusters are affected by differences in the covariates values, but at the same time the Weight variable contributes in better classifying observations in State 2 and 4.

By looking at the evolution of the clustering structure over time (see Table 13), we have that, according to the estimated transition probability matrix, State 2 and

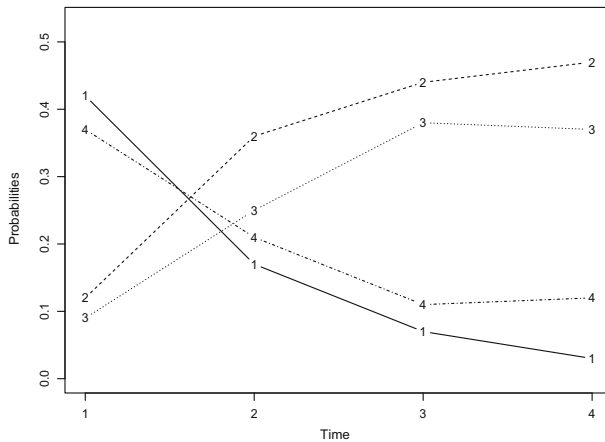


Fig. 7 Heart activity: estimated average probability of each hidden state at every time occasion

State 3, the *not-at-risk* states, are very persistent. The other two states are transient, i.e., visited and left at the following time. This means that patients can have difficulties to turnaround 360 degrees without help at one of the occasions, but not consecutively, and similarly being *at-risk* is also not a permanent condition. The latter implies that monitoring over time is necessary to capture such a condition. In Fig. 7, we report the proportion of patients clustered in each state at each time. The proportion of patients in State 2 and State 3 increases over time, i.e. patients' conditions in our sample improve over time.

The specific feature of the CN-CN-HMRMCs model is the identification of outlying observations, as defined in Table 1. Only 5 observations are identified as bad leverage, 41 as good leverage and 127 as outliers (see Fig. 8).

As discussed in this Section, there are several data features that can be modelled and investigated via HMRMCs. Here, we have focused on the hidden structured mainly, as our primary purpose is clustering patients over time. Of course, more details can be also provided with respect to other aspects of the analysis.

7 Discussion

Hidden Markov regression models with random covariates (HMRMCs) have been recently proposed by Punzo et al. (2018a) to improve the performance of hidden Markov regression models with fixed covariates (HMRMFCs) for model-based clustering purposes. In this paper, by focusing on outcomes and covariates of a continuous type, three multivariate elliptical distributions, the normal, the t , and the contaminated normal, have been considered to define particular cases of HMRMCs. The normal distribution has been considered as a reference distribution. The t and the contaminated normal distributions, heavy-tailed generalizations of the normal distribution, have been taken into account to protect the reference distribution for the occurrence of (mildly) atypical points and to allow for their automatic detection. The Monte Carlo

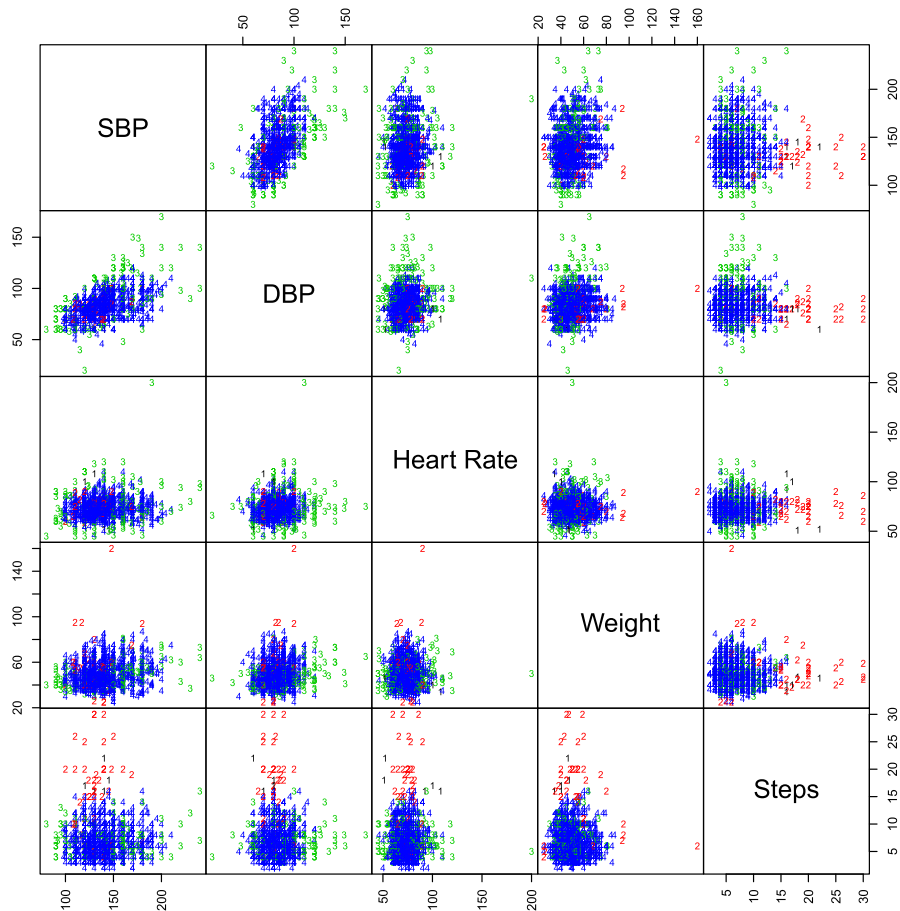


Fig. 8 Heart activity: matrix of scatter plots with inferred atypical points definition. Black (1): Bad leverage. Red (2): Good leverage. Green (3): Outliers. Blue (4): Typical

simulation studies of Online Appendix B, and the illustrative examples of Sect. 6, have been provided to find out more about the introduced family of HMRMRCs, to appreciate their advantages with respect to HMRMFCs, and to evaluate the effects of neglecting heavy-tailed distributions.

However, the proposed models have some limitations that we will try to overcome in future works following the ideas outlined below.

- A first limitation is that our models are restricted to continuous data. To overcome this issue, in the fashion of Mazza et al. (2018), we could allow for outcomes and/or covariates to be of mixed-type by: using the local independence assumption, considering convenient parametric models for the distribution of the covariates, and supporting modeling for the conditioned outcomes by means of the most common distributions of the exponential family via a generalized linear model.




- A second limitation is related to the number of parameters which increases quadratically with the number of outcomes and covariates, and this is due to each of the K scale matrices $\Sigma_{Y|k}$ and $\Sigma_{X|k}$, respectively. As a consequence, the models can be easily over-parameterized and the estimated parameters become of potentially difficult interpretation. A classical way to overcome the issue, for models characterized by scale matrices, consists in adding parsimony by considering convenient decompositions on the local scale matrices and putting constraints on the terms of the adopted decomposition. Classical examples in this direction, in the model-based clustering literature, are: the eigen-decomposition (see Punzo and Ingrassia 2015 and Dang et al. 2017 for mixtures of regression models with random covariates), the factor decomposition (see Maruotti et al. 2017 in the HMM framework, and Subedi et al. 2013, 2015 for mixtures of regression models with random covariates), and the variance/correlation decomposition (Biernacki and Lourme 2014).
- Working on the eigen-decomposed local scale matrices $\Sigma_{X|k}$ and $\Sigma_{Y|k}$, $k = 1, \dots, K$, in the fashion of Ingrassia and Rocci (2007), suitable constraints on their eigenvalues during the run of the EM/ECM algorithm could attenuate possible problems on the likelihood function such as unboundedness and spurious local maxima.

References

- Bartolucci F, Farcomeni A (2009) A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. *J Am Stat Assoc* 104:816–831
- Bartolucci F, Farcomeni A, Pennoni F (2014) Latent Markov models: a review of a general framework for the analysis of longitudinal data with covariates. *Test* 23(3):433–465
- Baum LE, 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(1):164–171
- Bernardi M, Maruotti A, Petrella L (2017) Multiple risk measures for multivariate dynamic heavy-tailed models. *J Empir Financ* 43:1–32
- Biernacki C, Lourme A (2014) Stable and visualizable Gaussian parsimonious clustering models. *Stat Comput* 24(6):953–969
- Croux C, Dehon C (2003) Estimators of the multiple correlation coefficient: local robustness and confidence intervals. *Stat Pap* 44(3):315–334
- Dang UJ, Punzo A, McNicholas PD, Ingrassia S, Browne RP (2017) Multivariate response and parsimony for Gaussian cluster-weighted models. *J Classif* 34(1):4–34
- Dannemann J, Holzmann H, Leister A (2014) Semiparametric hidden Markov models: identifiability and estimation. *Wiley Interdiscip Rev Comput Stat* 6(6):418–425
- Hennig C (2000) Identifiability of models for clusterwise linear regression. *J Classif* 17(2):273–296
- Hossain A, Naik DN (1991) A comparative study on detection of influential observations in linear regression. *Stat Pap* 32(1):55–69
- Ingrassia S, Rocci R (2007) Constrained monotone EM algorithms for finite mixture of multivariate Gaussians. *Comput Stat Data Anal* 51(11):5339–5351
- Ingrassia S, Minotti SC, Punzo A (2014) Model-based clustering via linear cluster-weighted models. *Comput Stat Data Anal* 71:159–182
- Lachos VH, Angolini T, Abanto-Valle CA (2011) On estimation and local influence analysis for measurement errors models under heavy-tailed distributions. *Stat Pap* 52(3):567–590
- Leroux BG (1992) Maximum-likelihood estimation for hidden Markov models. *Stoch Process Their Appl* 40(1):127–143
- Maronna RA (1976) Robust M -estimators of multivariate location and scatter. *Ann Stat* 4(1):51–67

- Martinez-Zarzoso I, Maruotti A (2013) The environmental kuznets curve: functional form, time-varying heterogeneity and outliers in a panel setting. *Environmetrics* 24(7):461–475
- Maruotti A (2011) Mixed hidden Markov models for longitudinal data: An overview. *Int Stat Rev* 79(3):427–454
- Maruotti A (2014) Robust fitting of hidden Markov regression models under a longitudinal setting. *J Stat Comput Simul* 84(8):1728–1747
- Maruotti A, Punzo A (2017) Model-based time-varying clustering of multivariate longitudinal data with covariates and outliers. *Comput Stat Data Anal* 113:475–496
- Maruotti A, Bulla J, Lagona F, Picone M, Martella F (2017) Dynamic mixtures of factor analyzers to characterize multivariate air pollutant exposures. *Ann Appl Stat* 11(3):1617–1648
- Maruotti A, Punzo A, Bagnato L (2019) Hidden Markov and semi-Markov models with multivariate leptokurtic-normal components for robust modeling of daily returns series. *J Financ Econom* 17(1):91–117
- Mazza A, Punzo A (2017) Mixtures of multivariate contaminated normal regression models. *Stat Pap*. <https://doi.org/10.1007/s00362-017-0964-y>
- Mazza A, Punzo A, Ingrassia S (2018) flexCWM: a flexible framework for cluster-weighted models. *J Stat Softw* 86(2):1–30
- McLachlan G, Krishnan T (2007) The EM algorithm and extensions, Wiley Series in Probability and Statistics, vol 382, 2nd edn. Wiley, New York
- McLachlan GJ, Peel D (2000) Finite mixture models. Wiley, New York
- Meng XL, Rubin DB (1993) Maximum likelihood estimation via the ECM algorithm: a general framework. *Biometrika* 80(2):267–278
- Niu X, Li P, Zhang P (2016) Testing homogeneity in a scale mixture of normal distributions. *Stat Pap* 57(2):499–516
- Punzo A, Ingrassia S (2015) Parsimonious generalized linear Gaussian cluster-weighted models. In: Morlini I, Minerva T, Vichi M (eds) *Advances in statistical models for data analysis. Studies in classification, data analysis and knowledge organization*. Springer, Switzerland, pp 201–209
- Punzo A, Maruotti A (2016) Clustering multivariate longitudinal observations: the contaminated Gaussian hidden Markov model. *J Comput Graph Stat* 25(4):1097–1116
- Punzo A, McNicholas PD (2016) Parsimonious mixtures of multivariate contaminated normal distributions. *Biom J* 58(6):1506–1537
- Punzo A, McNicholas PD (2017) Robust clustering in regression analysis via the contaminated Gaussian cluster-weighted model. *J Classif* 34(2):249–293
- Punzo A, Ingrassia S, Maruotti A (2018a) Multivariate generalized hidden Markov regression models with random covariates: physical exercise in an elderly population. *Stat Med* 37(19):2797–2808
- Punzo A, Mazza A, McNicholas PD (2018b) ContaminatedMixt: An R package for fitting parsimonious mixtures of multivariate contaminated normal distributions. *J Stat Softw* 85(10):1–25
- R Core Team (2018) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org/>
- Ritter G (2015) Robust cluster analysis and variable selection, Chapman & Hall/CRC monographs on statistics & applied probability, vol 137. CRC Press, Boca Raton
- Rousseeuw PJ, Leroy AM (2005) Robust regression and outlier detection. Wiley Series in probability and statistics. Wiley, Hoboken
- Subedi S, Punzo A, Ingrassia S, McNicholas PD (2013) Clustering and classification via cluster-weighted factor analyzers. *Adv Data Anal Classif* 7(1):5–40
- Subedi S, Punzo A, Ingrassia S, McNicholas PD (2015) Cluster-weighted *t*-factor analyzers for robust model-based clustering and dimension reduction. *Stat Methods Appl* 24(4):623–649
- Visser I, Raijmakers MEJ, Molenaar PCM (2000) Confidence intervals for hidden markov model parameters. *Br J Math Stat Psychol* 53(2):317–327
- Zucchini W, MacDonald IL, Langrock R (2016) Hidden Markov models for time series: an introduction using R, monographs on statistics & applied probability, vol 150, 2nd edn. CRC Press, Boca Raton

Affiliations

Antonio Punzo¹  · **Salvatore Ingrassia**¹  · **Antonello Maruotti**^{2,3} 

✉ Antonello Maruotti
a.maruotti@lumsa.it; antonello.maruotti@uib.no

Antonio Punzo
antonio.punzo@unict.it

Salvatore Ingrassia
s.ingrassia@unict.it

¹ Department of Economics and Business, University of Catania, Catania, Italy

² Department of Economics, Political Sciences and Modern Languages, LUMSA, Rome, Italy

³ Department of Mathematics, University of Bergen, Bergen, Norway