Model-based time-varying clustering of multivariate longitudinal data with covariates and outliers

Antonello Maruotti[a,b,*], Antonio Punzo[c]

[a] *Dipartimanto di Scienze Economiche, Politiche e delle Lingue Moderne, LUMSA, Roma, Italy.*
[b] *Centre for Innovation and Leadership in Health Sciences, University of Southampton, Southampton, UK.*
[c] *Dipartimento di Economia e Impresa, Università di Catania, Catania, Italy.*

Abstract

A class of multivariate linear models under the longitudinal setting, in which unobserved heterogeneity may evolve over time, is introduced. A latent structure is considered to model heterogeneity, having a discrete support and following a first-order Markov chain. Heavy-tailed multivariate distributions are introduced to deal with outliers. Maximum likelihood estimation is performed to estimate parameters by using expectation-maximization and expectation-conditional-maximization algorithms. Notes on model identifiability and robustness are provided, along with all computational details needed to implement the proposal. Three applications on artificial and real data are illustrated. These focus on the potential effects of outliers on clustering and their identification.

Keywords: hidden Markov models, robust regression, multivariate contaminated Gaussian distribution, ECM algorithm

1. Introduction

Hidden Markov regression models (HMRMs) are the state of the art in the analysis of time-dependent data. Since the seminal works of Goldfeld and Quandt (1973) and Hamilton (1990), HMRMs have been used in a wide range of empirical applications such as ecology (Langrock and King, 2013 and Schliehe-Diecks et al., 2012), environmetrics (Martinez-Zarzoso and Maruotti, 2013, Ailliot et al., 2015, and Maruotti et al., 2016), medicine (Langrock et al., 2013 and Lagona et al., 2014), and more. The HMRM is the most suitable approach to deal with time-dependent data features as serial dependence, (time-varying) heterogeneity, and dependence of the response vector on several covariates. Its mathematical properties, based on a hidden Markov structure (see e.g. Zucchini and MacDonald, 2009), render the HMRM approach appealing from a theoretical perspective as well. Indeed, mean, variance, and autocorrelations are available, the likelihood is straightforward to compute and its computation is linear in the number of observations.

To remark the importance of HMRMs in longitudinal data analysis, a book has been recently published by Bartolucci et al. (2013), in which the HMRM approach is introduced and extensively discussed. Furthermore, few review papers discuss features of HMRMs under different settings (see, e.g., Maruotti, 2011, Visser, 2011). In the analysis of multivariate longitudinal data through HMRMs, the attention has been mainly focused on non-continuous (Lagona et al., 2015 and Bartolucci and Farcomeni, 2009) or mixed-support (Raffa and Dubin, 2015 and Bartolucci and Farcomeni, 2015) data, with few notable exceptions (see, e.g., Lee et al., 2014). We want to contribute to this literature by considering longitudinal response vectors of continuous type. Under the HMRM framework, the benchmark natural model to cope with continuous response vectors is represented by the HMRM based on state-specific Gaussian distribution for the error term, simply referred to as Gaussian HMRM herein. However, real longitudinal data are often "contaminated" by outliers that affect the estimation of the parameters

[*] Corresponding author: Antonello Maruotti – Dipartimanto di Scienze Economiche, Politiche e delle Lingue Moderne, Libera Università Maria Ss. Assunta. Via Pompeo Magno 22, 00192 Roma, Italy. Email: a.maruotti@lumsa.it,a.maruotti@soton.ac.uk,

Email addresses: a.maruotti@lumsa.it,a.maruotti@soton.ac.uk (Antonello Maruotti), antonio.punzo@unict.it (Antonio Punzo)

for the Gaussian HMRM.Accordingly, the detection of these outliers, and the development of robust methods of parameters estimation insensitive to their presence, is an important problem.

There is a wide literature on robust estimation of mixture regression models (see, e.g., Bai et al., 2016, Bai et al., 2012, and García-Escudero et al., 2010). Vermunt (2010) and Frühwirth-Schnatter (2011) provide excellent overviews of using finite mixture models in longitudinal research. To introduce robustness against outliers, Frühwirth-Schnatter and Kaufmann (2008) consider a multivariate t distribution and Juárez and Steel (2010) introduce a skew t distribution for the error term to capture conditional skewness. Wang (2013) and Wang et al. (2015) extend the aforementioned approaches in a mixed-effects regression framework. Between- and within-subject variations through subject-specific random effects and intra-subject errors can be modeled, as well as missingness and censoring can be easily dealt with in such a framework. There are not many papers dealing with robustness issues in HMRMs. In the univariate case, Maruotti (2014) considers a bi-square scale estimator, and Farcomeni (2012) introduces a quantile regression for longitudinal data based on latent Markov subject-specific parameters. Up to our knowledge, these are the only attempts to deal with outliers in HMRMs. Other approaches have been recently developed in a general time-dependent clustering framework for multivariate data (see e.g. Farcomeni and Greco, 2015 and Punzo and Maruotti, 2016), ignoring the functional relationships between the response vector and covariates.

We propose two robust generalizations of the Gaussian HMRM obtained by replacing the conditional multivariate Gaussian distribution with two well-known multivariate symmetric heavy-tailed distributions: the t and the contaminated Gaussian. As it will be better explained in the following, the proposed HMRMs offer practical alternatives needed for mild outliers robustness (cf. Ritter, 2015, pp. 79–80) where the multivariate Gaussian distribution, often used as the reference distribution for the typical observations, as in the Gaussian HMRM, lacks sufficient fit. Such mild outliers document mainly the difficulty of the specification problem. In their presence the statistician is recommended to choose a model flexible enough to accommodate all data points, including the outliers.

A maximum likelihood approach is pursued to achieve parameters estimation. Expectation-maximization (EM) and expectation-conditional-maximization (ECM) algorithms are discussed in depth to allow non-experts to fit our models. Suggestions on the initialization strategies are provided, along with parameters interpretation and procedures to identify outliers. Artificial and real data examples are provided to highlight the usefulness and the potentials of the proposed approach. These examples address several research questions and provide evidence of the adequateness of robust HMRMs under a longitudinal setting.

The remind of the paper is as follow. In Section 2, we introduce the methodology, reviewing the basics of the HMRMs, and providing details on the multivariate t and multivariate contaminated Gaussian distributions. An important issue as identifiability of the proposed HMRMs is addressed in Section 3. Maximum likelihood estimation is discussed in Section 4. To give the opportunity to readers to apply our proposal, we provide further computational details and model features in Appendix A. A wide discussion on robustness is proposed in Section 5, along with practical guidance to identify outliers. Empirical applications on artificial and real data are provided in Section 6. Section 7 summarizes the key aspects of the proposal along with future possible extensions.

2. Methodology

2.1. Main assumptions and notation

Let $\{\boldsymbol{Y}_{it}; i = 1,\dots,I, t = 1,\dots,T\}$ denote sequences of multivariate real-valued longitudinal observations of dimension $d_{\boldsymbol{Y}}$ recorded on I units and T times. Moreover, let $\{S_{it}; i = 1,\dots,I, t = 1,\dots,T\}$ be a first-order Markov chain defined on the state space $\{1,\dots,k,\dots,K\}$. A hidden Markov model (HMM) is a particular kind of dependent mixture. It is a stochastic process consisting of two parts: the underlying unobserved process $\{S_{it}\}$, fulfilling the Markov property, i.e.

$$\Pr\left(S_{it} = s_{it} \mid S_{i1} = s_{i1}, S_{i2} = s_{i2}, \dots, S_{i(t-1)} = s_{i(t-1)}\right) = \Pr\left(S_{it} = s_{it} \mid S_{i(t-1)} = s_{i(t-1)}\right),$$

and the state-dependent observation process $\{\boldsymbol{Y}_{it}\}$ for which the conditional independence property holds, i.e.

$$f\left(\boldsymbol{Y}_{it} = \boldsymbol{y}_{it} \mid \boldsymbol{Y}_{i1} = \boldsymbol{y}_{i1}, \dots, \boldsymbol{Y}_{iT} = \boldsymbol{y}_{iT}, S_{i1} = s_{i1}, \dots, S_{iT} = s_{iT}\right) = f\left(\boldsymbol{Y}_{it} = \boldsymbol{y}_{it} \mid S_{it} = s_{it}\right),$$

where $f(\cdot)$ is a generic probability density function.

The hidden Markov chain has K states with initial probabilities $\pi_{ik} = \Pr(S_{i1} = k)$, $k = 1, \dots, K$, and transition probabilities

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

In (1), k refers to the current state, whereas j refers to the one previously visited; this convention will be used throughout the paper. In the following, for simplicity of explanation, we will consider $\pi_{i,k|j} = \pi_{k|j}$ and $\pi_{ik} = \pi_k$, $i = 1, \dots, I$. Such an assumption can be easily relaxed to include covariates and/or unit-specific random effects as described in Maruotti and Rocci (2012). Thus, we collect the initial probabilities in the K-dimensional vector $\boldsymbol{\pi}$, whereas the time-homogeneous transition probabilities are collected in the $K \times K$ transition matrix $\boldsymbol{\Pi}$.

In many applied longitudinal studies, we also have sequences of (fixed) covariates $\{\boldsymbol{X}_{it}; i = 1, \dots, I, t = 1, \dots, T\}$, with each $\boldsymbol{X}_{it}$ being of dimension d_X, that we would like to use to explain the sequences of response random vectors $\{\boldsymbol{Y}_{it}; i = 1, \dots, I, t = 1, \dots, T\}$. In such a case, a valid alternative, in the HMM framework, is represented by hidden Markov regression models (HMRMs; see, e.g., Lee et al., 2014 and Maruotti, 2014). Here, in each latent state k, we are interested in modeling the conditional distribution

$$f\left(\boldsymbol{Y}_{it} = \boldsymbol{y}_{it} \mid \boldsymbol{X}_{it} = \boldsymbol{x}_{it}, S_{it} = k\right) \tag{2}$$

by assuming a functional, typically parametric, form for the expectation $E\left(\boldsymbol{Y}_{it} \mid \boldsymbol{X}_{it} = \boldsymbol{x}_{it}, S_{it} = k\right)$. In the following, for simplicity, we will consider the classical linear case

$$E\left(\boldsymbol{Y}_{it} \mid \boldsymbol{X}_{it} = \boldsymbol{x}_{it}, S_{it} = k; \boldsymbol{\beta}_k\right) = \boldsymbol{\mu}\left(\boldsymbol{x}_{it}; \boldsymbol{\beta}_k\right) = \boldsymbol{\beta}_k' \boldsymbol{x}_{it}^*, \tag{3}$$

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

2.2. Considered multivariate heavy-tailed distributions

The standard HMRM for multivariate continuous outcomes is based on the multivariate Gaussian distribution

$$f_N\left(\boldsymbol{Y}_{it} = \boldsymbol{y}_{it} \mid \boldsymbol{X}_{it} = \boldsymbol{x}_{it}, S_{it} = k; \boldsymbol{\mu}\left(\boldsymbol{x}_{it}; \boldsymbol{\beta}_k\right), \boldsymbol{\Sigma}_k\right) = (2\pi)^{-\frac{d_Y}{2}} \left|\boldsymbol{\Sigma}_k\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\delta\left(\boldsymbol{y}_{it}, \boldsymbol{\mu}\left(\boldsymbol{x}_{it}; \boldsymbol{\beta}_k\right); \boldsymbol{\Sigma}_k\right)\right\}, \tag{4}$$

where $\boldsymbol{\mu}\left(\boldsymbol{x}_{it}; \boldsymbol{\beta}_k\right)$ and $\boldsymbol{\Sigma}_k$ denote the mean and the covariance matrix, respectively, and where

$$\delta\left(\boldsymbol{y}_{it}, \boldsymbol{\mu}\left(\boldsymbol{x}_{it}; \boldsymbol{\beta}_k\right); \boldsymbol{\Sigma}_k\right) = \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}\left(\boldsymbol{x}_{it}; \boldsymbol{\beta}_k\right)\right)' \boldsymbol{\Sigma}_k^{-1} \left(\boldsymbol{y}_{it} - \boldsymbol{\mu}\left(\boldsymbol{x}_{it}; \boldsymbol{\beta}_k\right)\right)$$

denotes the squared Mahalanobis distance between $\boldsymbol{y}_{it}$ and $\boldsymbol{\mu}\left(\boldsymbol{x}_{it}; \boldsymbol{\beta}_k\right)$, with covariance matrix $\boldsymbol{\Sigma}_k$. In symbols, $\boldsymbol{Y}_{it} \mid \boldsymbol{X}_{it} = \boldsymbol{x}_{it}, S_{it} = k \sim N_{d_Y}\left(\boldsymbol{\mu}\left(\boldsymbol{x}_{it}; \boldsymbol{\beta}_k\right), \boldsymbol{\Sigma}_k\right)$.

As densities to be adopted for (2), we consider:

- the multivariate t distribution

$$f_t\left(\boldsymbol{Y}_{it} = \boldsymbol{y}_{it} \mid \boldsymbol{X}_{it} = \boldsymbol{x}_{it}, S_{it} = k; \boldsymbol{\mu}\left(\boldsymbol{x}_{it}; \boldsymbol{\beta}_k\right), \boldsymbol{\Sigma}_k, \nu_k\right) = \frac{\Gamma\left(\frac{\nu_k + d_Y}{2}\right) \left|\boldsymbol{\Sigma}_k\right|^{-\frac{1}{2}}}{\Gamma\left(\frac{\nu_k}{2}\right) (\pi \nu_k)^{\frac{d_Y}{2}} \left[1 + \frac{1}{\nu_k}\delta\left(\boldsymbol{y}_{it}, \boldsymbol{\mu}\left(\boldsymbol{x}_{it}; \boldsymbol{\beta}_k\right); \boldsymbol{\Sigma}_k\right)\right]^{\frac{\nu_k + d_Y}{2}}}, \tag{5}$$

  where $\boldsymbol{\mu}\left(\boldsymbol{x}_{it}; \boldsymbol{\beta}_k\right)$ and $\boldsymbol{\Sigma}_k$ denote the mean and the scale matrix, respectively, while ν_k denotes the degrees of freedom. In symbols, $\boldsymbol{Y}_{it} \mid \boldsymbol{X}_{it} = \boldsymbol{x}_{it}, S_{it} = k \sim t_{d_Y}\left(\boldsymbol{\mu}\left(\boldsymbol{x}_{it}; \boldsymbol{\beta}_k\right), \boldsymbol{\Sigma}_k, \nu_k\right)$. Note that (5) approaches (4) as $\nu_k \to \infty$;

- the multivariate contaminated Gaussian distribution

$$\begin{aligned} f_{CN}\left(\boldsymbol{Y}_{it} = \boldsymbol{y}_{it} \mid \boldsymbol{X}_{it} = \boldsymbol{x}_{it}, S_{it} = k; \boldsymbol{\mu}\left(\boldsymbol{x}_{it}; \boldsymbol{\beta}_k\right), \boldsymbol{\Sigma}_k, \alpha_k, \eta_k\right) = \\ \alpha_k f_N\left(\boldsymbol{Y}_{it} = \boldsymbol{y}_{it} \mid \boldsymbol{X}_{it} = \boldsymbol{x}_{it}, S_{it} = k; \boldsymbol{\mu}\left(\boldsymbol{x}_{it}; \boldsymbol{\beta}_k\right), \boldsymbol{\Sigma}_k\right) \\ + (1 - \alpha_k) f_N\left(\boldsymbol{Y}_{it} = \boldsymbol{y}_{it} \mid \boldsymbol{X}_{it} = \boldsymbol{x}_{it}, S_{it} = k; \boldsymbol{\mu}\left(\boldsymbol{x}_{it}; \boldsymbol{\beta}_k\right), \eta_k \boldsymbol{\Sigma}_k\right), \end{aligned} \tag{6}$$

  where $\boldsymbol{\mu}\left(\boldsymbol{x}_{it}; \boldsymbol{\beta}_k\right)$ is the mean, $\boldsymbol{\Sigma}_k$ is the scale matrix, $\alpha_k \in (0, 1)$ is the proportion of typical points in state k, and $\eta_k > 1$ is an inflation parameter accounting for the degree of outlierness in state k. In symbols, $\boldsymbol{Y}_{it} \mid \boldsymbol{X}_{it} = \boldsymbol{x}_{it}, S_{it} = k \sim CN_{d_Y}\left(\boldsymbol{\mu}\left(\boldsymbol{x}_{it}; \boldsymbol{\beta}_k\right), \boldsymbol{\Sigma}_k, \alpha_k, \eta_k\right)$. Note that (6) approaches (4) as $\alpha_k \to 1^-$ and $\eta_k \to 1^+$.

These multivariate heavy-tailed conditional distributions give rise to two different HMRMs, respectively abbreviated as t-HMRM and CN-HMRM. The HMRM based on the multivariate Gaussian distribution in (4) will be abbreviated as N-HMRM hereafter. While N-HMRMs are well-consolidated in literature (see, e.g., Lee et al., 2014 for the multivariate case), both t-HMRMs and CN-HMRMs are new, up to our knowledge.

3. Identifiability

An important issue in dealing with the proposed HMRMs is to establish their identifiability. Identifiability is a necessary requirement, *inter alia*, for the usual asymptotic theory to hold for maximum likelihood (ML) estimation of the model parameters.

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 (see also Dannemann et al., 2014). In our case, the marginal mixtures are mixtures of regression models (MRMs) of the form

$$f\left(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\vartheta}\right)=\sum_{k=1}^{K}\pi_k f\left(\boldsymbol{y}|\boldsymbol{x};\boldsymbol{\vartheta}_k\right), \tag{7}$$

where the basic conditions

$$\pi_k > 0, \quad \forall\, k = 1,\dots,K,$$

and

$$\forall\, k,l \in \{1,\dots,K\} : k \neq l \Rightarrow \boldsymbol{\vartheta}_k \neq \boldsymbol{\vartheta}_l,$$

are fulfilled. These two conditions prevent overfitting (a potential problem for identifiability first noted by Crawford, 1994) and identifiability problems which occur due to empty components where $\boldsymbol{\vartheta}_k$ cannot be uniquely determined and due to components with equal component parameter vectors where different values for π_k are possible (see Frühwirth-Schnatter, 2006, Chapter 1.3, for details).

For a general enough class of MRMs, those based on component distributions belonging to the exponential family (the so-called mixtures of generalized linear models), Grün and Leisch (2008) give a sufficient condition for their identifiability (see Hennig, 2000 for some previous results). Roughly speaking, the fulfillment of this sufficient condition requires caution about three issues. Below, we give an intuitive explanation of these issues as well as the way they are handled in the sufficient condition by Grün and Leisch (2008).

Issue 1. Identifiability of the finite mixtures of distributions obtained by (7) conditioning on $\boldsymbol{x}$ is essential.

Issue 2. Identifiability issues may arise if there are only a limited number of different covariate points. Such issues might occur in applications because the covariates are often categorical. Caution about these issues is formulated by constraining K to be lower than the minimum number of $(d_X - 1)$-dimensional hyperplanes to cover the distinct covariates values.

Issue 3. The covariate matrix must be full column rank. Such a rank condition ensures that the regression coefficients can be uniquely determined given the linear predictor.

The sufficient condition in Grün and Leisch (2008) applies in our case. With respect to the first issue, and based on the distributions considered herein, sufficient conditions for identifiability of finite mixtures of multivariate t distributions are given in Holzmann et al. (2006), while sufficient conditions for identifiability of finite mixtures of multivariate contaminated Gaussian distributions are given in Punzo and McNicholas (2016).

We further investigate identifiability issues by using the parametric bootstrap with random initialization. The Hartigans' DIP test for unimodality (Hartigan and Hartigan, 1985) is applied to check if the regression parameter estimates follow a unimodal distribution. This is done for the state-specific regression coefficients, in which rejecting the null hypothesis of unimodality implies that identifiability issues are present.

4. Maximum likelihood estimation

In order to perform ML estimation of the parameters for the proposed HMRMs on the basis of the sample $\{(\boldsymbol{x}_{it}, \boldsymbol{y}_{it}); i = 1, \ldots, I, t = 1, \ldots, T\}$, the need arises of computing

$$\mathcal{L}(\boldsymbol{\vartheta}) = \prod_{i=1}^{I} \mathcal{L}_i(\boldsymbol{\vartheta}) = \prod_{i=1}^{I} \boldsymbol{\pi}' \boldsymbol{f}_{i1} \boldsymbol{\Pi} \boldsymbol{f}_{i2} \cdots \boldsymbol{\Pi} \boldsymbol{f}_{iT} \mathbf{1}_K, \tag{8}$$

where $\boldsymbol{\vartheta}$ corresponds to the set of all model parameters, $\mathbf{1}_K$ denotes a vector of K ones, and $\boldsymbol{f}_{it}$ denotes a $K \times K$ diagonal matrix having on the main diagonal the conditional densities $f(\boldsymbol{Y}_{it} = \boldsymbol{y}_{it} \mid \boldsymbol{X}_{it} = \boldsymbol{x}_{it}, S_{it} = k)$, $k = 1, \ldots, K$. Finding the value of the parameters $\boldsymbol{\vartheta}$ that maximizes the log-transformation of (8) under the constraints $\pi_k > 0$, $\pi_{k|j} > 0$, $\sum_{k=1}^{K} \pi_k = 1$, $\sum_{k=1}^{K} \pi_{k|j} = 1$, $k, j = 1, \ldots, K$, in addition to the constraints required by the parameters of the chosen multivariate heavy-tailed distributions, is not an easy problem since (8) is not available in an analytically convenient form. Efficient computation of (8) may be performed by exploiting a forward recursion described in the HMM literature (see, e.g., Zucchini and MacDonald, 2009).

In this relatively general framework, when the multivariate t distribution is involved in the formulation of the model, an expectation-maximization (EM) algorithm (Baum et al., 1970 and Dempster et al., 1977), which is a natural approach for ML estimation when data are incomplete, is described for fitting. When the multivariate contaminated Gaussian distribution is involved, an expectation-conditional maximization (ECM) algorithm (Meng and Rubin, 1993) is considered. Of course, ML estimation in this framework can be even carried out using a direct numerical maximization (Turner, 2008 and MacDonald, 2014) or a hybrid (Bulla and Berzel, 2008) algorithm. The EM/ECM algorithm tends to be preferred to its alternatives due to its robustness and ease of application in various scenarios, especially when the model parameters are constrained.

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 heavy-tailed distribution while the others are common to both the models. The common source – the classical one in the use of HMMs and the unique one when the multivariate Gaussian distribution is considered for HMRMs – 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 $\boldsymbol{z}_{it} = (z_{it1}, \ldots, z_{itk} \ldots, z_{itK})'$ and the unobserved states transition

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

respectively, with

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

Based on this source of incompleteness, we can write the complete-data log-likelihood in the following way

$$\ell_c(\boldsymbol{\vartheta}) = \ell_{c_1}(\boldsymbol{\pi}) + \ell_{c_2}(\boldsymbol{\Pi}) + \ell_{c_3}(\boldsymbol{\vartheta}_{\boldsymbol{Y}}), \tag{9}$$

where

$$\ell_{c_1}(\boldsymbol{\pi}) = \sum_{i=1}^{I}\sum_{k=1}^{K} z_{i1k}\log(\pi_k), \tag{10}$$

$$\ell_{c_2}(\boldsymbol{\Pi}) = \sum_{i=1}^{I}\sum_{t=2}^{T}\sum_{k=1}^{K}\sum_{j=1}^{K} zz_{itjk}\log\left(\pi_{k|j}\right), \tag{11}$$

$$\ell_{c_3}(\boldsymbol{\vartheta}_Y) = \sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{k=1}^{K} z_{itk}\ln\left[f\left(\boldsymbol{Y}_{it} = \boldsymbol{y}_{it} \mid \boldsymbol{X}_{it} = \boldsymbol{x}_{it}, S_{it} = k; \boldsymbol{\vartheta}_k\right)\right], \tag{12}$$

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

The other sources of incompleteness are distribution-dependent and yield to different specifications for ℓ_{c_3} in (12); see Appendix A.1 for details.

4.2. *EM and ECM algorithms*

As said above, an EM algorithm is used for fitting the t-HMRM and an ECM algorithm is considered for fitting the CN-HMRM. The latter iterates between an E-step and two CM-steps, until convergence. The only difference from the EM algorithm is that each M-step is replaced by simpler CM-steps.

The E-step, on the $(r+1)$th iteration of all 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 zz_{itjk} with their conditional expectations, namely, $z_{itk}^{(r)}$ and $zz_{itjk}^{(r)}$ (for computational details, see Appendix A.2). The rest of the E-step 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 three terms on the right-hand side of (9) 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 (10) and $\ell_{c_2}(\boldsymbol{\Pi})$ in (11) – with respect to $\boldsymbol{\pi}$ and $\boldsymbol{\Pi}$, respectively, subject to the constraints on these parameters, yields

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

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

4.2.1. *Multivariate t distribution.*

If $\boldsymbol{Y}_{it} \mid \boldsymbol{X}_{it} = \boldsymbol{x}_{it}, S_{it} = k \sim t_{d_Y}\left(\boldsymbol{\mu}\left(\boldsymbol{x}_{it};\boldsymbol{\beta}_k\right), \boldsymbol{\Sigma}_k, \nu_k\right)$, then the E-step, on the $(r+1)$th iteration, further requires the replacement of u_{itk} with

$$u_{itk}^{(r)} = \frac{\nu_k^{(r)} + d_Y}{\nu_k^{(r)} + \delta\left(\boldsymbol{x}_{it}, \boldsymbol{\mu}\left(\boldsymbol{y}_{it};\boldsymbol{\beta}_k^{(r)}\right); \boldsymbol{\Sigma}_k^{(r)}\right)}. \tag{13}$$

Thus, by substituting z_{itk} and u_{itk} in (A.1), with $z_{itk}^{(r)}$ and $u_{itk}^{(r)}$, respectively, we obtain $Q_3(\boldsymbol{\vartheta}_Y)$.

The M-step, on the same iteration, requires the calculation of $\boldsymbol{\vartheta}_Y^{(r+1)}$ as the value of $\boldsymbol{\vartheta}_Y$ that maximizes $Q_3(\boldsymbol{\vartheta}_Y)$.

Such a maximization yields

$$\boldsymbol{\beta}_k^{(r+1)} = \left[\sum_{i=1}^{I}\sum_{t=1}^{T} z_{itk}^{(r)} u_{itk}^{(r)} \boldsymbol{x}_{it}^{*} \boldsymbol{x}_{it}^{*\prime}\right]^{-1} \left[\sum_{i=1}^{I}\sum_{t=1}^{T} z_{itk}^{(r)} u_{itk}^{(r)} \boldsymbol{x}_{it}^{*} \boldsymbol{y}_{it}\right], \tag{14}$$

$$\boldsymbol{\Sigma}_k^{(r+1)} = \frac{\displaystyle\sum_{i=1}^{I}\sum_{t=1}^{T} z_{itk}^{(r)} u_{itk}^{(r)} \left[\boldsymbol{y}_{it} - \boldsymbol{\mu}\left(\boldsymbol{x}_{it};\boldsymbol{\beta}_k^{(r+1)}\right)\right]\left[\boldsymbol{y}_{it} - \boldsymbol{\mu}\left(\boldsymbol{x}_{it};\boldsymbol{\beta}_k^{(r+1)}\right)\right]'}{\displaystyle\sum_{i=1}^{I}\sum_{t=1}^{T} z_{itk}^{(r)}}. \tag{15}$$

A closed form solution is not analytically available for the update $\nu_k^{(r+1)}$ of ν_k. However, by differentiating $Q_3\left(\boldsymbol{\vartheta}_{\boldsymbol{Y}}\right)$ with respect to ν_k, we note that $\nu_k^{(r+1)}$ is a solution of the equation

$$-\psi\left(\frac{\nu_k}{2}\right) + \ln\left(\frac{\nu_k}{2}\right) + 1 + \frac{1}{\displaystyle\sum_{t=1}^{T} z_{itk}^{(r)} u_{itk}^{(r)}} \sum_{i=1}^{I}\sum_{t=1}^{T} z_{itk}^{(r)}\left[\ln\left(u_{itk}^{(r)}\right) - u_{itk}^{(r)}\right] + \psi\left(\frac{\nu_k^{(r)} + d_{\boldsymbol{Y}}}{2}\right) - \ln\left(\frac{\nu_k^{(r)} + d_{\boldsymbol{Y}}}{2}\right) = 0, \tag{16}$$

where $\psi(\cdot)$ is the Digamma function. Operationally, the `uniroot()` function in the **stats** package for R (R Core Team, 2013) is used to numerically find the root of (16) over the interval $\left(2, \nu_{\boldsymbol{Y}}^{*}\right)$, with $\nu_{\boldsymbol{Y}}^{*} > 2$. In the analyses of Section 6, we fix $\nu_{\boldsymbol{Y}}^{*} = 200$.

4.2.2. Multivariate contaminated Gaussian distribution.

If $\boldsymbol{Y}_{it} \mid \boldsymbol{X}_{it} = \boldsymbol{x}_{it}, S_{it} = k \sim CN_{d_Y}\left(\boldsymbol{\mu}\left(\boldsymbol{x}_{it};\boldsymbol{\beta}_k\right), \boldsymbol{\Sigma}_k, \alpha_k, \eta_k\right)$, then two CM-steps, instead of a single M-step, are considered. The two CM-steps arise from the partition $\boldsymbol{\vartheta}_{\boldsymbol{Y}} = \left(\boldsymbol{\vartheta}_{\boldsymbol{Y},1}, \boldsymbol{\vartheta}_{\boldsymbol{Y},2}\right)$, where $\boldsymbol{\vartheta}_{\boldsymbol{Y},1} = \left(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}\right)$ and $\boldsymbol{\vartheta}_{\boldsymbol{Y},2} = \boldsymbol{\eta}$.

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

$$u_{itk}^{(r)} = \frac{\alpha_k^{(r)} f_N\left(\boldsymbol{Y}_{it} = \boldsymbol{y}_{it} \mid \boldsymbol{X}_{it} = \boldsymbol{x}_{it}, S_{it} = k; \boldsymbol{\mu}\left(\boldsymbol{x}_{it};\boldsymbol{\beta}_k^{(r)}\right), \boldsymbol{\Sigma}_k^{(r)}\right)}{f_{CN}\left(\boldsymbol{Y}_{it} = \boldsymbol{y}_{it} \mid \boldsymbol{X}_{it} = \boldsymbol{x}_{it}, S_{it} = k; \boldsymbol{\mu}\left(\boldsymbol{x}_{it};\boldsymbol{\beta}_k^{(r)}\right), \boldsymbol{\Sigma}_k^{(r)}, \alpha_k^{(r)}, \eta_k^{(r)}\right)}. \tag{17}$$

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

$$\alpha_k^{(r+1)} = \frac{1}{\displaystyle\sum_{i=1}^{I}\sum_{t=1}^{T} z_{itk}^{(r)}} \sum_{i=1}^{I}\sum_{t=1}^{T} z_{itk}^{(r)} u_{itk}^{(r)},$$

$$\boldsymbol{\beta}_k^{(r+1)} = \left[\sum_{i=1}^{I}\sum_{t=1}^{T} z_{itk}^{(r)}\left(u_{itk}^{(r)} + \frac{1 - u_{itk}^{(r)}}{\eta_k^{(r)}}\right) \boldsymbol{x}_{it}^{*} \boldsymbol{x}_{it}^{*\prime}\right]^{-1} \left[\sum_{i=1}^{I}\sum_{t=1}^{T} z_{itk}^{(r)}\left(u_{itk}^{(r)} + \frac{1 - u_{itk}^{(r)}}{\eta_k^{(r)}}\right) \boldsymbol{x}_{it}^{*} \boldsymbol{y}_{it}\right], \tag{18}$$

$$\boldsymbol{\Sigma}_k^{(r+1)} = \frac{\displaystyle\sum_{i=1}^{I}\sum_{t=1}^{T} z_{itk}^{(r)}\left(u_{itk}^{(r)} + \frac{1 - u_{itk}^{(r)}}{\eta_k^{(r)}}\right)\left[\boldsymbol{y}_{it} - \boldsymbol{\mu}\left(\boldsymbol{x}_{it};\boldsymbol{\beta}_k^{(r+1)}\right)\right]\left[\boldsymbol{y}_{it} - \boldsymbol{\mu}\left(\boldsymbol{x}_{it};\boldsymbol{\beta}_k^{(r+1)}\right)\right]'}{\displaystyle\sum_{i=1}^{I}\sum_{t=1}^{T} z_{itk}^{(r)}}. \tag{19}$$

The second CM-step, on the same iteration, requires the calculation of $\boldsymbol{\vartheta}_{\boldsymbol{Y},2}^{(r+1)}$ as the value of $\boldsymbol{\vartheta}_{\boldsymbol{Y},2}$ that maximizes $Q_3\left(\boldsymbol{\vartheta}_{\boldsymbol{Y},2} \mid \boldsymbol{\vartheta}_{\boldsymbol{Y},1} = \boldsymbol{\vartheta}_{\boldsymbol{Y},1}^{(r+1)}\right)$. In particular, for each $k = 1, \ldots, K$, we have to maximize

$$-\frac{d_{\boldsymbol{Y}}}{2}\sum_{i=1}^{I}\sum_{t=1}^{T} z_{itk}^{(r)}\left(1 - u_{itk}^{(r)}\right)\ln\left(\eta_k\right) - \frac{1}{2}\sum_{i=1}^{I}\sum_{t=1}^{T} z_{itk}^{(r)}\frac{1 - u_{itk}^{(r)}}{\eta_k}\delta\left(\boldsymbol{y}_{it}, \boldsymbol{\mu}\left(\boldsymbol{x}_{it};\boldsymbol{\beta}_k^{(r+1)}\right); \boldsymbol{\Sigma}_k^{(r+1)}\right), \tag{20}$$

with respect to η_k, under the constraint $\eta_k > 1$. Operationally, as a closed form solution is not analytically available, the `optimize()` function in the **stats** package for R is used to perform a numerical search of the maximum of (20) over the interval $\left(1, \eta_Y^*\right)$, with $\eta_Y^* > 1$. In the analyses of Section 6, we fix $\eta_Y^* = 10,000$.

5. Notes about robustness and outliers detection

The N-HMRM constitutes a benchmark model. However, as said in Section 1, real longitudinal regression data are often "contaminated" by outliers that affect the estimation of the model parameters with particular interest, in the regression context, to the regression coefficients. Accordingly, the detection of these outliers, and the development of robust methods of parameters estimation insensitive to their presence, is an important problem.

5.1. Robust parameters estimation

In Sections 5.1.1 and 5.1.2 we illustrate how the use of multivariate *t* and contaminated Gaussian distributions, respectively, generates robust methods of parameters estimation insensitive to the presence of outliers.

5.1.1. Robust parameters estimation via the t-HMRM

Based on (14), the regression coefficients $\boldsymbol{\beta}_k^{(r+1)}$ can be considered a weighted least squares estimate with weights depending on (13). These weights clearly decrease with increasing squared Mahalanobis distance (i.e., the squared standardized residuals) δ. Moreover, the degree of downweighting of outliers for $\boldsymbol{\beta}_k^{(r+1)}$ in (14) increases with decreasing $u_{itk}^{(r)}$. Therefore, the weights inside (14) reduce the effect of outliers in the estimation of $\boldsymbol{\beta}_k$, so providing a robust way to estimate $\boldsymbol{\beta}_k$, $k = 1, \ldots, K$. In addition, from (15), the larger squared residuals δ also have smaller effects on $\boldsymbol{\Sigma}_k$, $k = 1, \ldots, K$, due to the quantity in (13).

5.1.2. Robust parameters estimation via the CN-HMRM

Based on (18), the regression coefficients $\boldsymbol{\beta}_k^{(r+1)}$ can be considered a weighted least squares estimate with weights depending on

$$u_{itk}^{(r)} + \frac{1 - u_{itk}^{(r)}}{\eta_k^{(r)}}. \tag{21}$$

Now, consider the update (17) for $u_{itk}^{(r)}$ as a function of the squared Mahalanobis distance (i.e., the squared standardized residuals) δ; the updating function in (17) can be so written as

$$g(\delta; \alpha, \eta) = \frac{\alpha \exp\left\{-\frac{\delta}{2}\right\}}{\alpha \exp\left\{-\frac{\delta}{2}\right\} + \frac{(1-\alpha)}{\sqrt{\eta}} \exp\left\{-\frac{\delta}{2\eta}\right\}} = \frac{1}{1 + \frac{(1-\alpha)}{\alpha} \frac{1}{\sqrt{\eta}} \exp\left\{\frac{\delta}{2}\left(1 - \frac{1}{\eta}\right)\right\}}, \tag{22}$$

with $\delta \geq 0$. Due to the constraint $\eta > 1$, from the last expression of (22) it is straightforward to realize that $g(\delta; \alpha, \eta)$ is a decreasing function of δ. Based on (22), formula (21) can be written as

$$w(\delta; \alpha, \eta) = g(\delta; \alpha, \eta) + \frac{1 - g(\delta; \alpha, \eta)}{\eta} = \frac{1}{\eta}\left[1 + (\eta - 1) g(\delta; \alpha, \eta)\right]. \tag{23}$$

From the last expression of (23), it easy to realize that $w(\delta; \alpha, \eta)$ is an increasing function of $g(\delta; \alpha, \eta)$; this also means that $w(\delta; \alpha, \eta)$ is a decreasing function of δ. Therefore, the weights in (21) reduce the effect of outliers in the estimation of $\boldsymbol{\beta}_k$, so providing a robust way to estimate $\boldsymbol{\beta}_k$, $k = 1, \ldots, K$. In addition, from (19), the larger squared residuals δ also have smaller effects on $\boldsymbol{\Sigma}_k$, $k = 1, \ldots, K$, due to the weights in (21). See Little (1988) for a discussion on downweighting of outliers for the contaminated Gaussian distribution.

5.2. Outliers detection

t-HMRMs and CN-HMRMs can be also used as models for mild outliers detection according to the strategies illustrated in Sections 5.2.1 and 5.2.2, respectively.

5.2.1. *Outliers detection via t-HMRMs*

For t-HMRMs, an *a posteriori* procedure (i.e., a procedure taking place once the model is fitted) may be considered to detect mild outliers. Extending the idea illustrated by McLachlan and Peel (2000), p. 232, for mixtures of multivariate t distributions, each observation $(\boldsymbol{x}_{it}, \boldsymbol{y}_{it})$ can be treated as an outlier if

$$\sum_{k=1}^{K} \text{MAP}\left(\hat{z}_{itk}\right) \delta\left(\boldsymbol{y}_{it}, \boldsymbol{\mu}\left(\boldsymbol{x}_{it}; \hat{\boldsymbol{\beta}}_k\right); \hat{\boldsymbol{\Sigma}}_k\right) \tag{24}$$

is sufficiently large, where

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

denotes the maximum *a posteriori* probabilities (MAP) operator, being $\hat{z}_{itk}$, $\hat{\boldsymbol{\beta}}_k$, and $\hat{\boldsymbol{\Sigma}}_k$ the values of z_{itk}, $\boldsymbol{\beta}_k$, and $\boldsymbol{\Sigma}_k$, respectively, at convergence of the EM algorithm. To decide on how large the statistic (24) must be in order for $(\boldsymbol{x}_{it}, \boldsymbol{y}_{it})$ to be classified as outlier, we can compare it to the $(1-\alpha)$ 100th percentile of the χ^2 distribution with $d_{\boldsymbol{Y}}$ degrees of freedom, where the χ^2 distribution is used to approximate the distribution of $\delta\left(\boldsymbol{y}_{it}, \boldsymbol{\mu}\left(\boldsymbol{x}_{it}; \hat{\boldsymbol{\beta}}_k\right); \hat{\boldsymbol{\Sigma}}_k\right)$.

5.2.2. *Outliers detection via CN-HMRMs*

Once the CN-HMRM is fitted to the observed longitudinal data, by means of maximum *a posteriori* probabilities, each observation $(\boldsymbol{x}_{it}, \boldsymbol{y}_{it})$ can be first assigned to one of the K latent states and then classified as typical or outlier; thus, we have a model for simultaneous robust clustering and automatic detection of mild outliers in a longitudinal regression context. Note that, differently from the approach illustrated in Section 5.2.1, the approach to detect outliers based on the contaminated Gaussian distribution makes no additional distributional assumptions and is not based on subjective choices such as the percentile of the χ^2 distribution.

Automatic outliers detection. For CN-HMRMs, the classification of an observation $(\boldsymbol{x}_{it}, \boldsymbol{y}_{it})$ means:

Step 1. determine its state of membership;

Step 2. establish if it is either typical or outlier in that state.

Let $\hat{\boldsymbol{u}}_{it}$ and $\hat{\boldsymbol{z}}_{it}$ denote, respectively, the expected values of $\boldsymbol{u}_{it}$ and $\boldsymbol{z}_{it}$ arising from the ECM algorithm, i.e., $\hat{u}_{itk}$ and $\hat{z}_{itk}$ are the values of u_{itk} and z_{itk}, respectively, at convergence. We then consider $\hat{u}_{ith}$, where h is selected such that $\text{MAP}\left(\hat{z}_{ith}\right) = 1$. Although $(1-\hat{u}_{ith})$ provides the richest information about the probability that $(\boldsymbol{x}_{it}, \boldsymbol{y}_{it})$ is an outlier in state h, the user could be interested in obtaining a classification of this observation as either typical or outlier. In such a case, $(\boldsymbol{x}_{it}, \boldsymbol{y}_{it})$ is classified as outlier if $\hat{u}_{ith} < 0.5$. Thus, once the observation has been classified in one of the K states, the approach reveals richer information about the role of that observation in that state. Note also that, the resulting information can be used to possibly eliminate the outliers if such an outcome is desired (Berkane and Bentler, 1988); in such a case, the remaining data may then be treated as effectively being distributed according to a N-HMRM, and the clustering results can be reported as usual.

The state assignment procedure, as well as the outlying observations detection, resembles the adoption of a zero-one loss function according to whether the reconstructed partition is perfect or not (see McLachlan and Peel, 2000, p. 33 and McLachlan, 1992, p. 416). The assignments are taken to be the mode of the posterior probabilities $\hat{z}_{ith}$ and $\hat{u}_{ith}$, respectively.

Constraints for outliers detection. When the CN-HMRM is used for outliers detection, $(1-\alpha_k)$ represents the proportion of outliers in state k. As suggested by Punzo and McNicholas (2016, see also Punzo and McNicholas, 2014), for these parameters one could require that in the kth state, $k = 1, \dots, K$, the proportion of outliers is at least equal to a pre-determined value α^*. In this case, the `optimize()` function is also used for a numerical search of the maximum $\alpha_k^{(r+1)}$, over the interval $(\alpha^*, 1)$, of the function

$$\sum_{i=1}^{I} \sum_{t=1}^{T} z_{itk}^{(r)} \left[u_{itk}^{(r)} \ln \alpha_k + \left(1 - u_{itk}^{(r)}\right) \ln\left(1-\alpha_k\right)\right].$$

In the analyses of Section 6, we use this approach to update α_k and we take $\alpha^* = 0.5$. Note that, if desired, it is possible to fix α_k *a priori*.

6. Illustrative examples

The aim of this section is to provide a detailed analysis of longitudinal data to highlight the most important features of the proposed HMRMs and to show how to interpret model parameters. This section can also be considered as a guidance that allows the readers to run their own similar analyses being aware of the effects of outliers in the data.

The choice of the starting values for EM-based algorithms constitutes an important issue (see, e.g., Biernacki et al., 2003 and Bagnato and Punzo, 2013). The EM/ECM algorithms described in Section 4.2, for fitting our HMRMs, are initialized by the solution provided by the corresponding MRMs by further considering $(\mathbf{1}_K\mathbf{1}_K' + s\boldsymbol{I}_K)/(K+s)$ as the starting value for $\boldsymbol{\Pi}$-matrix, where $\boldsymbol{I}_K$ is a $K \times K$ diagonal matrix and s is a suitable constant; in the analyses herein, we fix $s = 9$ as in Bartolucci and Farcomeni (2009). In turn, MRMs are initialized according to the partition provided by the K-means method as implemented by the `kmeans()` function of the **stats** package for R.

Comparison between models is handled in terms of Bayesian information criterion (BIC; Schwarz, 1978)

$$\text{BIC} = 2\log\mathcal{L} - \#\text{par} \times \log(I)$$

and integrated complete likelihood (ICL; Biernacki et al., 2000). As concerns the latter, in practice (see, e.g., Ingrassia et al., 2014, Punzo, 2014, and Subedi et al., 2013, 2015) an approximate ICL is used, given by

$$\text{ICL} \approx \text{BIC} + \sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{k=1}^{K}\text{MAP}\left(\hat{z}_{itk}\right)\log\left(\hat{z}_{itk}\right).$$

The ICL essentially penalizes the BIC for estimated mean entropy. For the alternative use of likelihood-ratio tests for the selection of the best model, see Punzo et al. (2016).

6.1. Sensitivity study based on the blue crabs data

A first sensitivity study, based on an artificial longitudinal version of the very popular crab dataset of Campbell and Mahon (1974), is here described to compare how outliers affect the N-HMRM and how them are instead handled by the t-HMRM and the CN-HMRM. Attention is focused on the sample of $I = 100$ blue crabs of the genus *Leptograpsus*, subdivided in two groups of equal size ($\pi_1 = \pi_2 = 0.5$). For each specimen, we consider two measurements (in millimeters), namely the rear width (RW), considered here as the covariate, and the length along the midline of the carapace (CL), considered as response variable. The ML estimates of the regression coefficients in the two groups are

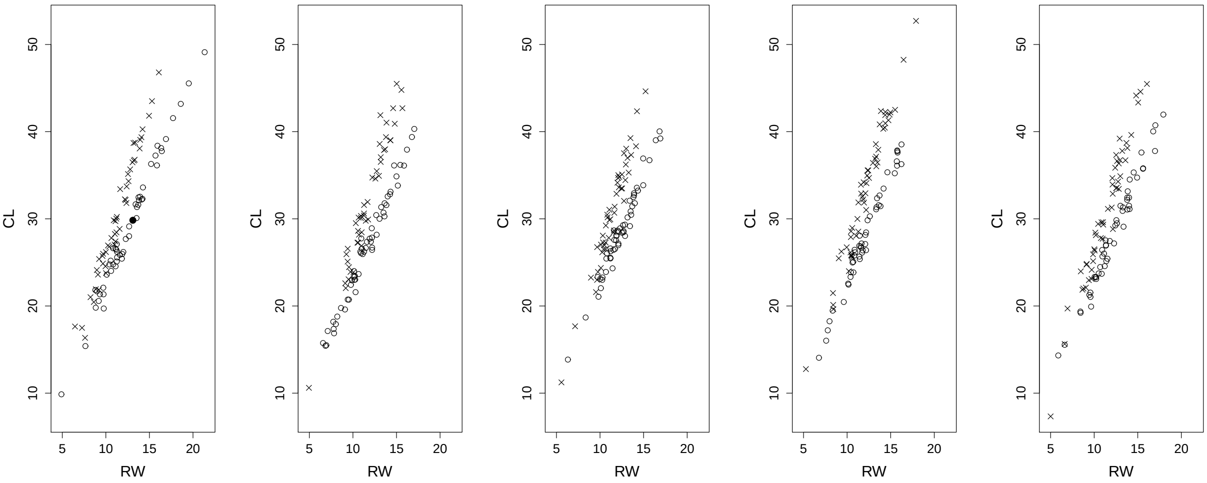
$$\boldsymbol{\beta}_1 = \begin{pmatrix}-7.612\\ 3.382\end{pmatrix}, \quad \boldsymbol{\beta}_2 = \begin{pmatrix}-0.988\\ 2.397\end{pmatrix}, \quad \Sigma_1 = 2.377, \quad \text{and} \quad \Sigma_2 = 0.866.$$

By assuming a normal distribution for the covariates in each group, the ML estimates of the mean and the standard deviation are 11.718 and 2.090 in group 1, and 12.138 and 2.414 in group 2 (see Greselin et al., 2011, Greselin and Punzo, 2013, and Bagnato et al., 2014, for details). Based on these estimates, and further introducing a transition probabilities matrix

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

we randomly generate a longitudinal version of this dataset on $T = 5$ times, based on the N-HMRM with a normally distributed covariate in each state. The scatterplots of the generated data, for each $t \in \{1,\ldots,5\}$, are displayed in Figure 1.

Ten "perturbed" datasets are created by substituting the original value of CL for the first point at time 1 (highlighted by a bullet in Figure 1(a)) with ten outliers shown in the first column of Table 1. We directly fit N-HMRMs,



(a) $t = 1$ (b) $t = 2$ (c) $t = 3$ (d) $t = 4$ (e) $t = 5$

Figure 1: Scatterplots of the artificial data (× and ○ denote group 1 and group 2, respectively; ● denotes the observation perturbed for the analysis of Section 6.1).

	Misallocation error			t-HMRM		CN-HMRM			
Value	N-HMRM	t-HMRM	CN-HMRM	$\hat{u}_{11k}$	$\hat{\nu}_k$	$\hat{u}_{11k}$	weight	$\hat{\eta}_k$	# of outliers detected
-40	0.080	0.034	0.030	0.00118	3.71948	0	0.00018	5477.57216	1
-35	0.080	0.034	0.030	0.00137	3.76054	0	0.00021	4701.41773	1
-30	0.080	0.034	0.030	0.00160	3.80636	0	0.00025	3984.46026	1
-25	0.076	0.034	0.030	0.00190	3.85802	0	0.00030	3326.72905	1
-20	0.078	0.034	0.030	0.00229	3.91691	0	0.00037	2728.19888	1
-15	0.076	0.034	0.030	0.00280	3.98499	0	0.00046	2188.94494	1
-10	0.072	0.034	0.030	0.00351	4.06504	0	0.00059	1708.91360	1
-5	0.068	0.034	0.030	0.00452	4.16118	0	0.00078	1288.25800	1
0	0.062	0.030	0.030	0.00379	4.28816	0	0.00108	926.98032	1
5	0.054	0.030	0.030	0.00566	4.46765	0	0.00160	625.21885	1

Table 1: Artificial blue crabs data: misallocation error for three HMRMs. Details from the fitted t-HMRM and CN-HMRM are also reported.

t-HMRMs, and CN-HMRMs, with $K = 2$. For each of the three competing techniques, Table 1 reports the proportion of misallocated observations (misallocations error) for each perturbed data set. The CN-HMRM is systematically the most robust to these perturbations, with the misallocations error remaining fixed at 0.03 regardless of the particular value perturbed. This is especially in contrast to the N-HMRM where the misallocations error changes (and does not necessarily decrease) as the extent of the perturbation increases.

As concerns the fitted t-HMRMs, as expected, by recalling that the original value of CL for the first point at time 1 was 29.836, the weight $\hat{u}_{11k}$ assigned to the outlier by the t-HMRM decreases as the value of this point further departs from its true value. A similar reasoning holds for the estimated degrees of freedom $\hat{\nu}_k$; this means that we need a conditional t distribution with heavier tails as the outlier departs from the bulk of its state of membership.

As concerns the fitted CN-HMRMs, we note that the probability to be a typical point for the outlier is practically null regardless of the particular value perturbed. As for the t-HMRM, the weight assigned to the outliers decreases as the value of this point further departs from its *true value*. Similarly, the estimated value of η_k (in the state containing the outlier at time 1) increases as the value of this point further departs from its true value; in these terms, this parameter can be also meant as a sort of "degree of outlierness", i.e as a measure of how vertically different outliers are from the regression model of their state of membership. Finally, as it can be seen by the last column of Table 1, the CN-HMRM always detects a single outlier, and it coincides with the perturbed point.

6.2. Sensitivity study based on pinus nigra data

A second study, based on an artificial longitudinal version of the Pinus nigra dataset analyzed in García-Escudero et al. (2010), is here described to evaluate the behavior of N-HMRMs, *t*-HMRMs, and CN-HMRMs in the presence of more than two states with outliers.

The complete data set is made of measurements of diameters (in millimeters) and heights (in meters) of 1089 trees in a cultivated forest of Pinus nigra located in the north of Palencia (Spain). Data were collected because searching for models to predict the "height" (trunk height without branches) in terms of the "diameter" (in the tree foot) is a classical problem in forestry and forest industry (see, e.g., Schreuder and Hafley, 1977). The interest is due to the fact that, from heights and diameters of the trees, we can obtain several parameters that determine the development and characteristics of a forest, such as the timber volume, indexes of competition between trees, growth of the forest mass, sustainability of the forest, etc.. Relations height-diameter tend to be linear (directly or after logarithmic transformations) and can vary depending on species, age, soil richness or others forest characteristics. The diameter of a tree is very easy to measure without using sophisticated instruments, but the height is not so easy to measure without cutting down the tree. Therefore, the role of "height" and "diameter" as response and covariate, respectively, is clearly justified.

Figure 2 shows the scatterplot of a subset of $I = 350$ observations from the original data. The scatterplot suggests the presence of three (almost parallel) linear groups. As well-explained by García-Escudero et al. (2010), the factor that caused the different location of the three groups is related with the fact that pines were sampled in three zones of varying quality, age, and density. That is, feet belonging to the lower cluster are located in an area of poor quality soil with low concentration of nutrients, which implies that the trees have no potential for root development and, therefore, the development is lower in height. The central tendency is formed by feet located in an area where no silvicultural treatments have been made, so there is a high density of trees, and the competition among trees is very large. Proof of this is the existence of the so-called foot-dominated trees, that is feet having small diameters, due to a small growth. The upper linear cluster represents an area of good quality soil, characterized by flat land, with soil rich in nutrients and where silvicultural treatments have enabled the trees to be free of competition and expose themselves to light, which creates a greater growth in diameter and height.

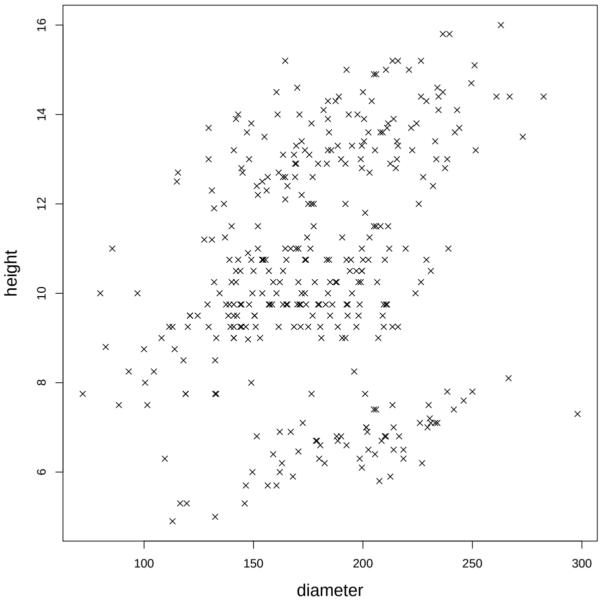


Figure 2: Pinus nigra data: Scatterplot.

Because the true group-membership is not available, an N-MRM with 3 mixture components and assuming a Gaussian distribution for the diameter in each group (see, e.g., Ingrassia et al., 2015, Punzo and Ingrassia, 2015, Ingrassia and Punzo, 2016), is estimated on the data in Figure 2. The obtained ML estimates for the mixture weights and the regression parameters are

$$\pi_1 = 0.353, \quad \pi_2 = 0.467, \quad \pi_3 = 0.180,$$

$$\boldsymbol{\beta}_1 = \begin{pmatrix} 10.594 \\ 0.015 \end{pmatrix}, \quad \boldsymbol{\beta}_2 = \begin{pmatrix} 7.633 \\ 0.014 \end{pmatrix}, \quad \boldsymbol{\beta}_3 = \begin{pmatrix} 3.936 \\ 0.014 \end{pmatrix},$$

$$\Sigma_1 = 0.766, \quad \Sigma_2 = 0.701, \quad \text{and} \quad \Sigma_3 = 0.226.$$

As concerns the parameters for the covariate, the estimated means are 191.098 in group 1, 164.483 in group 2, and 194.251 in group 3, while the standard deviations are 37.526 in group 1, 34.597 in group 2, and 37.456 in group 3. Based on these estimates, and further introducing a transition probabilities matrix

$$\boldsymbol{\Pi} = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{pmatrix},$$

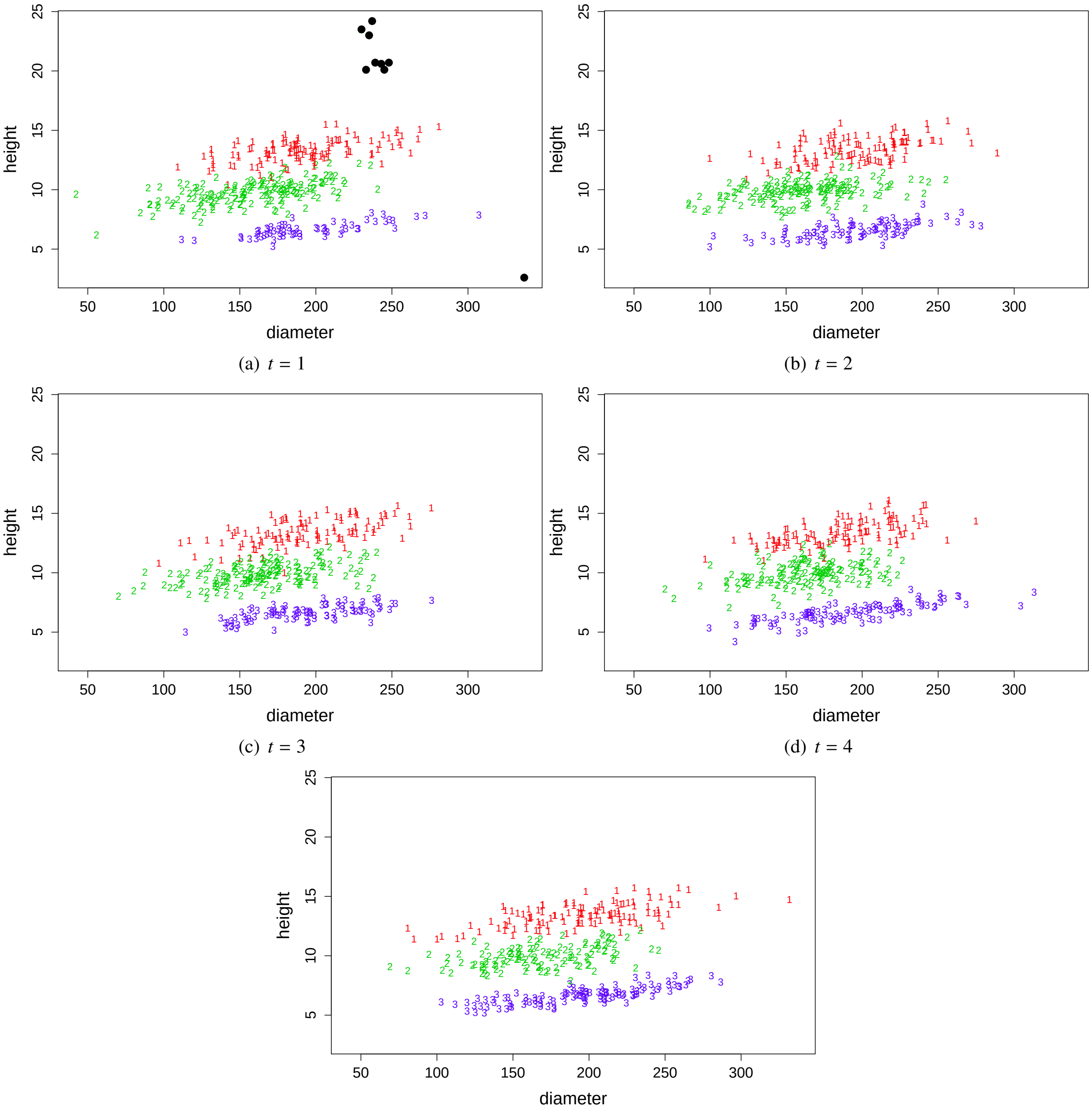
we randomly generate a longitudinal version of this dataset on $T = 5$ times, based on the N-HMRM with a normally distributed covariate in each state. Furthermore, nine observations at time 1 are randomly selected and substituted (perturbed) by nine observations extrapolated by the original dataset. The scatterplots of the generated data, for each $t \in \{1, \ldots, 5\}$, are displayed in Figure 3. The perturbed observations are denoted by black bullets in Figure 3(a); we can easily note how these observations are split into one isolated point on the bottom right corner and one "group" of eight outliers on the top.

On the artificial data in Figure 3 we fit the N-HMRM, the t-HMRM, and the CN-HMRM with $K = 3$ states. The scatterplots of the data, with labeling and regression lines from the fitted models, are shown in Figure 4 for the N-HMRM, Figure 5 for the t-HMRM, and Figure 6 for the CN-HMRM. While the regression line on the top from the t-HMRM and the CN-HMRM is not substantially affected by the group of outliers, the regression line on the top for the N-HMRM is dragged towards these points. Thus, the natural trend of the top state is biased by the presence of the group of outliers at time 1. As concerns the robust methods, the t-HMRM and the CN-HMRM provide similar results with the states being characterized by essentially parallel straight lines. Detected outliers from these models are also shown, via bullets colored based on their state of membership, in Figure 5 for the t-HMRM and in Figure 6 for the CN-HMRM. Note that, for the t-HMRM, we consider the 95th percentile of the χ^2-distribution (cf. Section 5.2.1). In detail, over $350 \cdot 5 = 1,750$ observations, there are 122 detected outliers by the t-HMRM and 9 by the CN-HMRM. Although both the models correctly identify the perturbed values as outliers, the detection rule for the t-HMRM, with the selected percentile, has an high false positive rate (proportion of typical points incorrectly classified as outliers) of 0.065. On the contrary, the detection rule for the CN-HMRM provides an optimal result: false positive rate equal to zero and true positive rate equal to one, being the true positive rate the proportion of outliers that are correctly identified as outliers. A further difference among these robust methods is in terms of state of membership of some of the detected true outliers: while all of them belong to the first state for the t-HMRM, they are split between the first and the second state for the CN-HMRM. This has implications for the estimated parameters. The estimated degrees of freedom for the t-HMRM are 2.973 (state 1), 17.470 (state 2), and 16.275 (state 3); the lowest value in correspondence of the first state highlights the need for heavier tails to accommodate for the presence of outliers in that state. The estimated proportions of typical points for the CN-HMRM are 0.987 (state 1), 0.987 (state 2), 0.999 (state 3), while the degrees of outlierness are 70.830 (state 1), 92.002 (state 2), and 4.020 (state 3); the outliers in the first two states have an impact on the lower values of $\hat{\alpha}_1$ and $\hat{\alpha}_2$ and on the higher values of $\hat{\eta}_1$ and $\hat{\eta}_2$. Always for this model, the perturbed values have an estimated posterior probability $\hat{u}_{itk}$ to be typical, in the state they are assigned, which ranges from $3.372 \cdot 10^{-51}$ to $2.092 \cdot 10^{-08}$. The fact that these probabilities are all very close to zero, also underlines that these points will be severely downweighted in the estimation of the regression parameters (cf. Section 5.1.2).

Summarizing, the CN-HMRM is the best model among the fitted ones. This consideration is also corroborated in terms of BIC and ICL (cf. Table 2).

6.3. Real data

The HMRMs so far proposed will be illustrated on a subset of the data from a Mayo Clinic trial on patients with primary biliarycirrhosis (PBC) conducted in 1974-1984 (Dickson et al., 1989). Only $I = 105$ subjects with



(a) $t = 1$

(b) $t = 2$

(c) $t = 3$

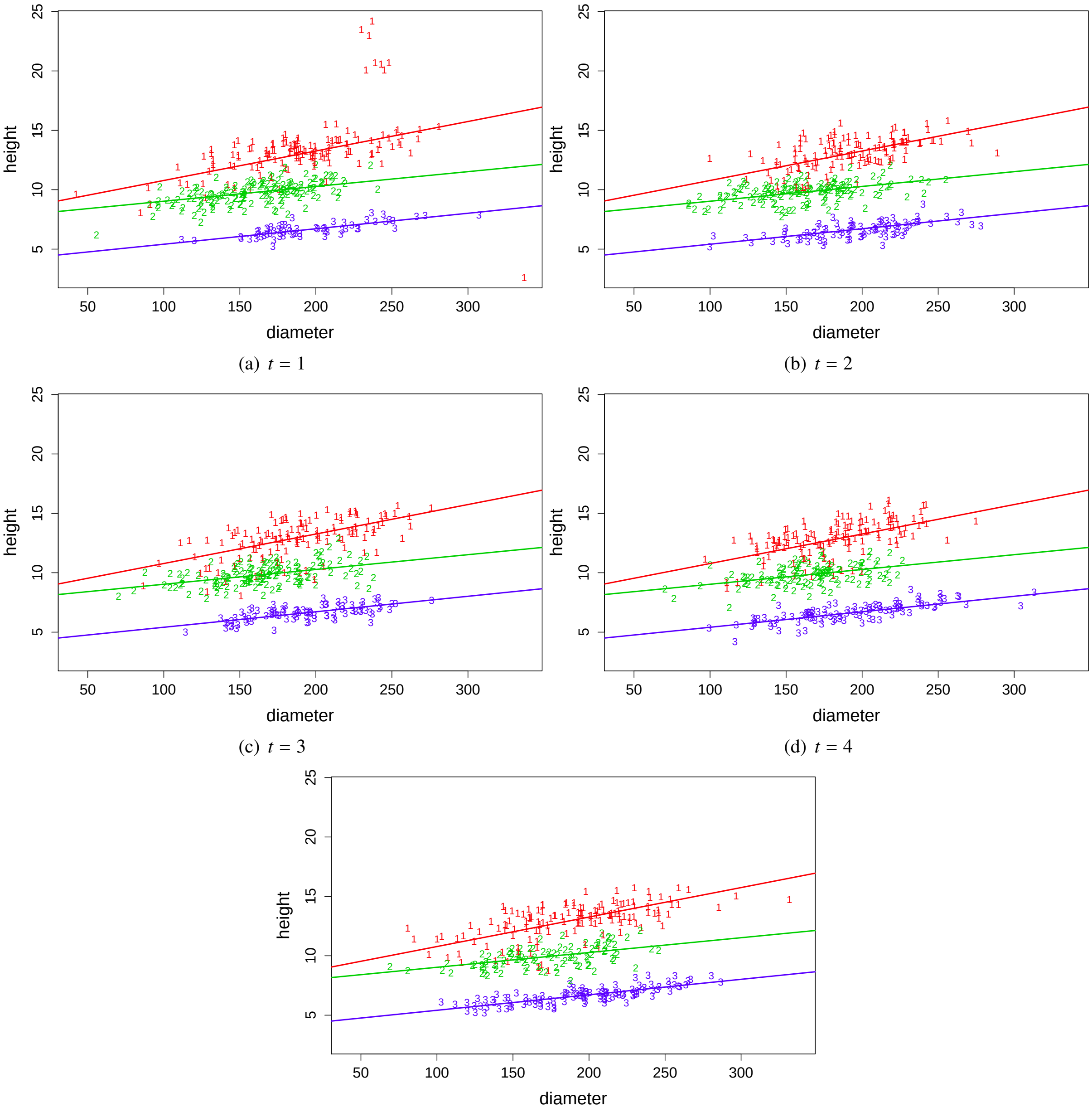
(d) $t = 4$

(e) $t = 5$

Figure 3: Pinus nigra data: Scatterplots of the artificial data (latent states are diversified by numbers and colors; • denotes the perturbed observations).

	N-HMRM	t-HMRM	CN-HMRM
BIC	-6755.003	-6484.416	**-6467.950**
ICL	-6851.075	-6529.902	**-6502.548**

Table 2: BIC and ICL values for the fitted HMRMs with $K = 3$ states. The best models, according to each criterion (i.e., for each row), is in bold.



(a) $t = 1$

(b) $t = 2$

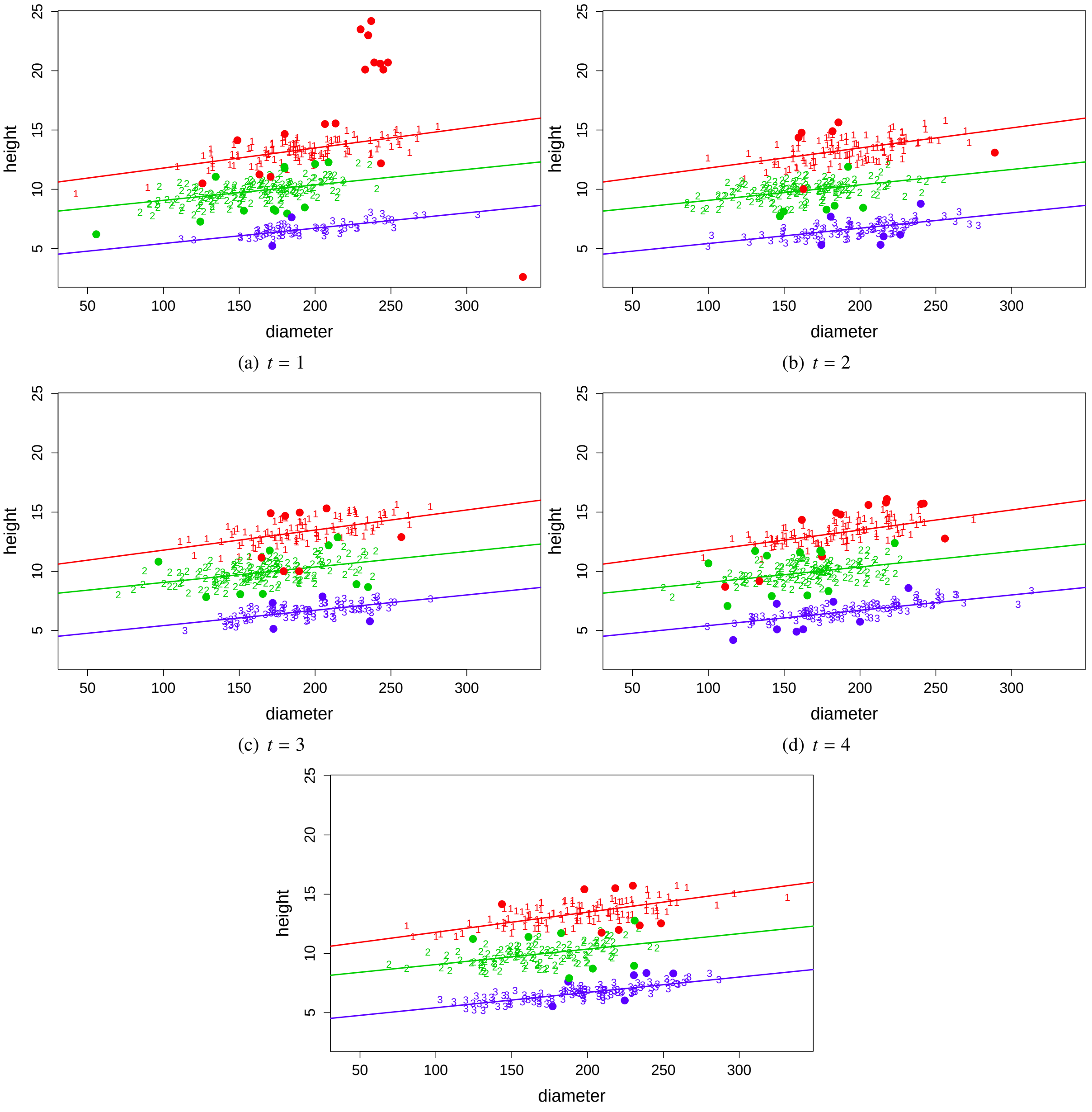
(c) $t = 3$

(d) $t = 4$

(e) $t = 5$

Figure 4: Pinus nigra data: Scatterplots, regression lines, and classification from the fitted N-HMRM with $K = 3$ states.

$T = 5$ longitudinal measurements are considered to avoid misleading inference due to the unbalancedness of the data. Seven response variables are recorded, corresponding to the natural logarithm of: serum bilirubin (mg/dl; *lbili*), serum albumin (mg/dl; *lalbumin*), alkaline phosphatase (U/liter; *lalk.phos*), serum cholesterol (mg/dl; *lchol*), serum glutamic-oxaloacetic transaminase (U/ml; *lsgot*), platelet count (*lplatelet*), and standardized blood clotting time (*lprotime*). Along with response variables, several covariates values at the baseline are recorded. In the following, we will focus on age, as it is available as a time-varying covariate, and gender only, which may help in the understanding of the evolution of subjects health status. Summary statistics over time are reported in Table 3. From a clinical perspective, the evolution over time of laboratory values reflects the evolution of subject health status, which might be more important for reasonable classification than e.g. simply the last known measurement.



(a) $t = 1$

(b) $t = 2$

(c) $t = 3$

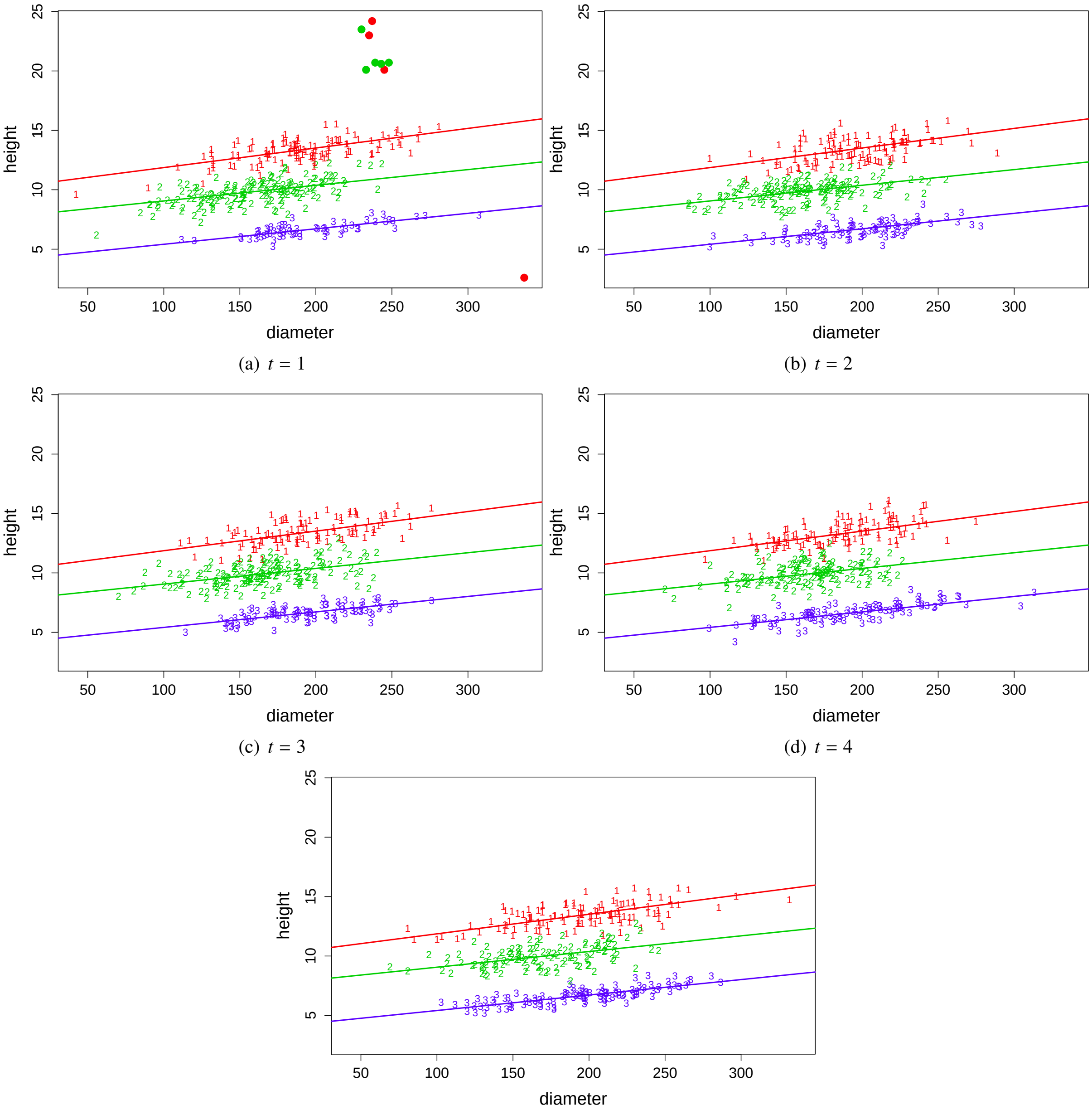
(d) $t = 4$

(e) $t = 5$

Figure 5: Pinus nigra data: Scatterplots, regression lines, and classification from the fitted t-HMRM with $K = 3$ states. Detected outliers are denoted by bullets colored based on the state of membership.

Our approach exploits jointly the whole history of longitudinal measurements of all considered markers, clustering subjects into different health states time-by-time capturing their health evolution over time.

We first discuss the possibilities of characterization of the hidden states in more detail, i.e. health status that were found by our analysis. Then, we look at the most likely hidden state sequence for clustering subjects into those states. A major issue in medical applications is to find the *proper* number of states corresponding to certain health conditions. Accordingly, the number of states should be inferred from the data. Table 4 shows the values of the penalized likelihood criteria for the models with $K = 1, \ldots, 5$ states. It shows that the multi-state models fit the data clearly better than a model with just a single state, i.e. heterogeneity arises in the data, for all the



(a) $t = 1$

(b) $t = 2$

(c) $t = 3$

(d) $t = 4$

(e) $t = 5$

Figure 6: Pinus nigra data: Scatterplots, regression lines, and classification from the fitted CN-HMRM with $K = 3$ states. Detected outliers are denoted by bullets colored based on the state of membership.

considered distributions. Among all the fitted models, the 3-state model with t conditional distributions is selected, with estimated state-specific degrees of freedom equal to 18.62, 4.77, and 27.73 for state 1, state 2, and state 3, respectively. In the following, we comment on estimated parameters of this *preferred* model.

The first state is characterized by a remarkably lower baseline bilirubin, albumin, cholesterol, and alkaline phosphatase levels compared to other states. The second state collects subjects with the highest alkaline phosphatase and standardized blood clotting time, and the third one those with the highest baseline values for the remaining variables. As shown in Table 5, the two covariates (age and gender) may or may not affect the observed levels, conditional to the health status. The algorithms described in Section 4.2 do not produce standard errors of the es-

Variables	First occasion Mean (Std dev)	Second occasion Mean (Std dev)	Third occasion Mean (Std dev)	Fourth occasion Mean (Std dev)	Fifth occasion Mean (Std dev)
lbili	0.08 (0.69)	0.14 (0.85)	0.29 (1.00)	0.45 (1.07)	0.54 (1.10)
lalbumin	1.29 (0.09)	1.24 (0.14)	1.22 (0.11)	1.21 (0.12)	1.17 (0.14)
lalk.phos	7.07 (0.58)	6.85 (0.59)	6.82 (0.62)	6.77 (0.57)	6.75 (0.59)
lchol	5.70 (0.37)	5.67 (0.32)	5.67 (0.36)	5.64 (0.32)	5.59 (0.38)
lsgot	4.61 (0.44)	4.52 (0.54)	4.47 (0.54)	4.47 (0.57)	4.49 (0.60)
lplatelet	5.52 (0.38)	5.37 (0.38)	5.31 (0.45)	5.28 (0.48)	5.27 (0.48)
lprotime	2.34 (0.06)	2.34 (0.06)	2.36 (0.13)	2.39 (0.07)	2.42 (0.07)
age	49.26 (9.64)	51.64 (9.61)	53.00 (9.61)	54.17 (9.64)	55.30 (9.63)

Table 3: PBC data: summary statistics over time

	N-HMRM		*t*-HMRM		CN-HMRM	
K	BIC	ICL	BIC	ICL	BIC	ICL
1	-1520.3349	-1520.3349	-1104.3876	-1104.3876	-1164.3811	-1164.3811
2	-746.7234	-758.7897	-564.0300	-576.3818	-572.4755	-585.9363
3	-684.7411	-701.3146	**-501.7825**	**-516.1390**	-522.9460	-543.6695
4	**-572.0132**	**-588.6623**	-514.0938	-535.8752	**-517.6907**	**-536.2827**
5	-631.6910	-647.5355	-628.0385	-642.2293	-537.4091	-558.8781

Table 4: Model selection. Bold numbers highlight the best value for each column.

timates, because approximations based on the observed information matrix often require a very large sample size. A parametric bootstrap approach has been, thus, implemented to get uncertainty measures of the estimates. The standard errors reported in Table 5 are based on 500 bootstrap samples. From the clinical point of view, the first state exhibits more favorable values and is also the most probable at the first occasion, i.e. $\boldsymbol{\pi} = (0.70, 0.13, 0.17)$, with (bootstrapped) standard errors equal to 0.03, 0.04, and 0.05 respectively; hence, it clusters patients with a better prognosis compared to other states. A clinical implication is clearly derived for other states as well. By looking at the transition probabilities matrix

$$\mathbf{\Pi} = \begin{bmatrix} \pi_{1|1} & \pi_{2|1} & \pi_{3|1} \\ \\ \pi_{1|2} & \pi_{2|2} & \pi_{3|2} \\ \\ \pi_{1|3} & \pi_{2|3} & \pi_{3|3} \\ \\ \end{bmatrix} = \begin{bmatrix} 0.92 & 0.06 & 0.02 \\ (0.11) & (0.04) & (0.03) \\ 0.00 & 0.97 & 0.03 \\ (0.15) & (0.20) & (0.09) \\ 0.00 & 0.10 & 0.90 \\ (0.04) & (0.05) & (0.05) \end{bmatrix},$$

where in parentheses we report (bootstrapped) standard errors, states 2 and 3 very rarely ($\pi_{1|j} \approx 0, j = 2, 3$) communicate with state 1. This implies that once the health conditions get worsen, it is unlikely (almost impossible) to improve them again. The only chance is to move between states 2 and 3, i.e. between different *bad* conditions.

We argued that from a clinical perspective, the first state corresponds to patients with a better prognosis compared to the others. It is possible to confirm this conclusion since the information concerning the residual progression free survival time, defined as time till death due to liver complications or till liver transplantation, is available in the form of the classical right-censored data. We calculated Kaplan-Meier estimates of the survival probabilities based on data from patients classified in each state at the last observation. These are plotted as solid lines on Figure 7. Indeed, the survival prognosis of state 1 is much better than that of states 2 and 3 with the estimated 10-year survival probability in state 1 of 0.98 compared to 0.63 in state 2 and to 0.50 in state 3. This illustration confirms that the hidden states have a physical meaning and reflect the progression free survival status.

Marginally, State 1 is visited 310 occasions, State 2 and State 3 collects 119 and 96 occasions, respectively. It is possible to quantify the uncertainty surrounding the obtained classification by looking at the posterior probabilities.

		State 1			State 2			State 3	
Variable	Intercept	Age	Female	Intercept	Age	Female	Intercept	Age	Female
lbili	-0.089 (0.117)	0.005 (0.002)	-0.479 (0.065)	0.082 (0.262)	0.020 (0.004)	-0.619 (0.097)	2.023 (0.218)	0.003 (0.004)	-0.765 (0.130)
lalbumin	1.297 (0.025)	0.000 (0.001)	-0.007 (0.014)	1.391 (0.056)	-0.003 (0.001)	-0.081 (0.021)	1.440 (0.025)	-0.005 (0.001)	0.028 (0.015)
lalk.phos	7.288 (0.172)	-0.009 (0.003)	-0.124 (0.096)	7.887 (0.166)	-0.019 (0.002)	-0.027 (0.061)	7.632 (0.107)	-0.004 (0.002)	-0.177 (0.063)
lchol	5.675 (0.080)	0.001 (0.001)	-0.112 (0.045)	5.584 (0.079)	-0.003 (0.001)	0.121 (0.029)	6.040 (0.121)	0.006 (0.002)	-0.447 (0.072)
lsgot	5.119 (0.126)	-0.011 (0.002)	-0.298 (0.070)	5.046 (0.141)	-0.006 (0.002)	-0.086 (0.052)	5.809 (0.115)	-0.009 (0.002)	-0.359 (0.068)
lplatelet	5.522 (0.077)	-0.002 (0.001)	0.116 (0.043)	4.992 (0.169)	-0.007 (0.002)	0.229 (0.062)	5.699 (0.073)	-0.001 (0.001)	-0.227 (0.043)
lprotime	2.356 (0.015)	0.000 (0.000)	-0.021 (0.008)	2.393 (0.042)	0.001 (0.001)	-0.025 (0.016)	2.263 (0.016)	0.002 (0.000)	-0.002 (0.010)

Table 5: PBC data: state-specific regression coefficients estimates with standard errors in brackets.

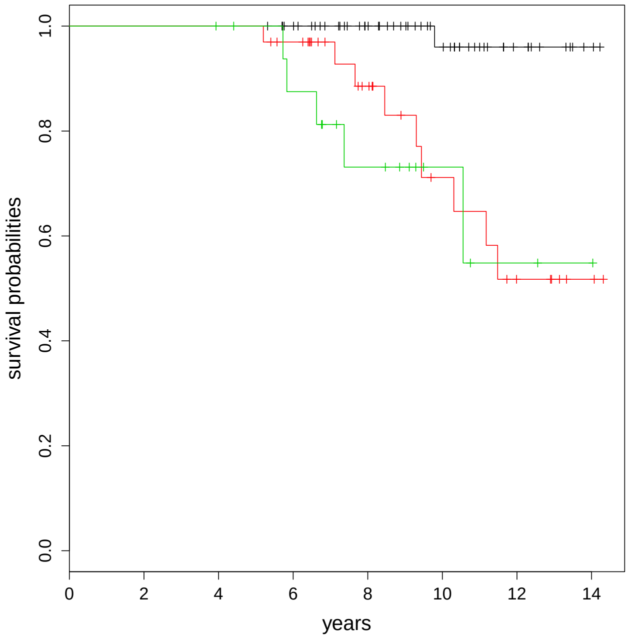


Figure 7: PBC data. Kaplan-Meier estimates of survival probability in each state (k = 1: black, k = 2: red; k = 3 green) created using the clustering procedure. Everybody classified using the estimated state according to the Viterbi algorithm at the fifth occasion.

To get a *good* and interpretable partition, we should have a lot of observations with posteriors close to zero or one. A peak at the posterior probability close to 1 indicates that a state is well separated from the others, while a mass in the middle of the unit interval indicates overlapping states. Accordingly, we want the distribution to be concentrated near the vertices (0,0), (0,1), and (1,0). In our simple example the components are well separated (see Figure 8). Thus, it is possible to classify a subject in one of the health status states with enough certainty at all the observed times.

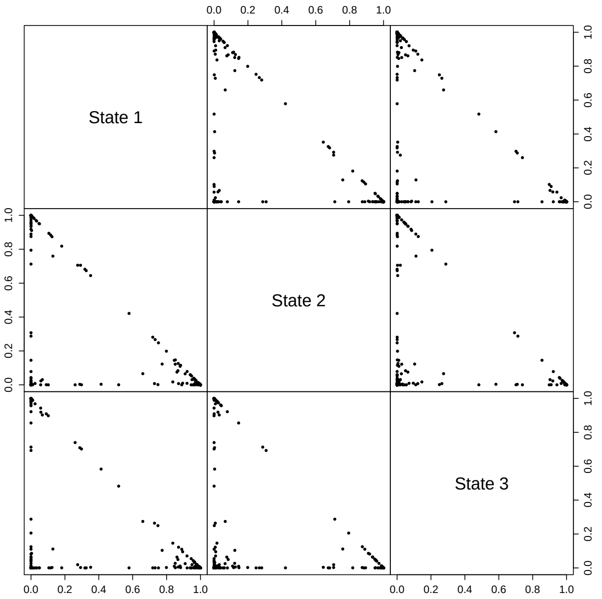


Figure 8: Uncertainty surrounding the classification. Bivariate scatterplots for the posterior state probabilities, which are reported on the axis

To complete the analysis, we look at outliers detection following the procedure described in Section 5.2. We recall that outliers are downweighted in the estimation steps (cf. Section 5.1.1), and this is an important aspect for the robust estimation of the parameters. By fixing $\alpha = 0.001$, we obtain a partition between typical observations and outliers. Figure 9 provides evidence of the ability of the t-HMRM to identify outliers. Indeed, the clear outlier is recognized, as well as other *atypical* observations that are *far* from the bulk of the data.

At last, the DIP-test is applied to check if the state-specific regression parameter estimates follow a unimodal distribution. The critical value is 0.119. In Table 6 we report the test statistic values for all regression parameters. According to these results, we can be confident on the identification of state-specific regression parameters.

	State 1			State 2			State 3		
Variable	Intercept	Age	Female	Intercept	Age	Female	Intercept	Age	Female
lbili	0.027	0.024	0.024	0.022	0.042	0.017	0.036	0.029	0.028
lalbumin	0.029	0.028	0.027	0.031	0.035	0.029	0.037	0.031	0.035
lalk.phos	0.038	0.021	0.026	0.022	0.022	0.022	0.037	0.034	0.021
lchol	0.019	0.021	0.026	0.048	0.031	0.031	0.027	0.036	0.039
lsgot	0.026	0.022	0.023	0.027	0.028	0.029	0.029	0.027	0.028
lplatelet	0.024	0.026	0.026	0.024	0.028	0.026	0.032	0.022	0.052
lprotime	0.027	0.029	0.018	0.027	0.022	0.029	0.022	0.025	0.021

Table 6: PBC data: DIP test statistic values for state-specific regression coefficients estimates.

7. Discussion

In this work we introduce two HMRMs, focusing on multivariate symmetric heavy-tailed distributions. In this setting, we consider the multivariate t and the multivariate contaminated Gaussian HMRMs to robustify HMRMs based on the multivariate Gaussian distribution. We provide all computational details needed to implement the proposed approaches and discuss parameters interpretation, in order to give guidance for the application of robust

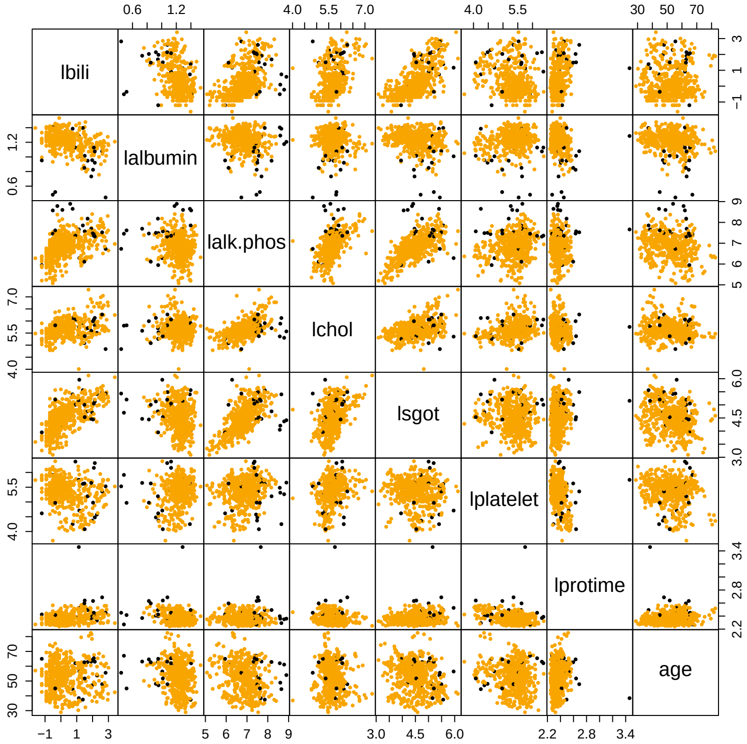


Figure 9: PBC data. Outliers (in black) and typical observations (in orange) according to the procedure outlined in Section 5.2 with $\alpha = 0.001$.

HMRMs. These models allow to account for several longitudinal data features (as time-varying heterogeneity) and, accordingly, can be used to model different types of real-world multivariate continuous data. Finite mixture of regression models, widely used in empirical applications to deal with heterogeneous populations, can be seen as a special case of our proposal. Furthermore, even simple linear regression models with pooled observations can be obtained in our framework by simply imposing $K = 1$.

Of course, dealing with multivariate longitudinal data can be challenging for high-dimensional response vectors. Indeed, the number of parameters would be extremely high and results difficult to be interpreted. In these cases, we would suggest to simultaneously perform regression, clustering, and dimensionality reduction by imposing constraints upon decomposed state-specific covariance matrices. Furthermore, our approach implies conditional elliptically contoured distributions for each state which, under specific empirical settings, could be rather restrictive. This is justified by the fact that non-elliptical distributions can be approximated quite well by a mixture of several basic elliptical distributions like the Gaussian one (McLachlan and Peel, 2000, p. 1 and Titterington et al., 1985, p. 24). While this can be very helpful for modeling purposes, it can be misleading when dealing with clustering/classification applications since one state may be represented by more than one mixture component just because it has, in fact, a non-elliptical distribution. A first possible route to continue to use our approach also in the presence of conditional non-elliptical distributions for each state, consists in considering transformations so as to make the components as elliptical as possible (Schork and Schork, 1988 and Zhu and Melnykov, 2016). Although such a treatment is very convenient to use, the achievement of joint ellipticity is rarely satisfied and the transformed variables become more difficult to be interpreted. Instead of applying transformations, we could extend our proposal by considering conditional skew distributions for each state, and this should be in line with the growing interest in proposing mixture models where the component distributions are skewed. Examples of existing approaches in this direction are: mixtures of skew-normal distributions (Lin, 2009 and Pyne et al., 2009), mixtures of shifted asymmetric Laplace distributions (Franczak et al., 2014), mixtures of multivariate skew-*t* distributions (see, e.g., Lin, 2010, and Lee and McLachlan, 2014), mixtures of multivariate *t* distributions with the Box-Cox transformation (Lo and Gottardo, 2012), mixtures of multivariate normal inverse Gaussian distributions (Karlis and Santourian, 2009), and mixtures of generalized hyperbolic distributions (Browne and McNicholas, 2015). For a recent enough survey about non-elliptical distributions in mixture modelling, see Lee and McLachlan (2013).

In our proposal, we assume a parametric form for the (conditional) expected value of the response vector. This assumption can be easily relaxed in the proposed framework by introducing nonparametric functions of the covariates in the linear predictor. Covariates can be also included to model all other distributions-specific parameters, as scale, shape and covariance with minor efforts. The role of covariates can be further investigated by allowing for random covariates. However, as widely documented in the literature on univariate HMRMs, including random effects in this framework can be cumbersome and computationally prohibitive to deal with, when the number of random covariates increases.

Appendix A. Further computational aspects

Appendix A.1. Complete data log-likelihood.

Appendix A.1.1. Multivariate t distribution.

If $\boldsymbol{Y}_{it} \mid \boldsymbol{X}_{it} = \boldsymbol{x}_{it}, S_{it} = k \sim t_{d_Y}\left(\boldsymbol{\mu}\left(\boldsymbol{x}_{it};\boldsymbol{\beta}_k\right), \boldsymbol{\Sigma}_k, \nu_k\right)$, 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 convenient Gamma random variable. In practice, for each observation $(\boldsymbol{x}_{it}, \boldsymbol{y}_{it})$ in state k, this source of incompleteness is denoted by $U_{itk} \sim \text{Gamma}\left(\nu_k/2, \nu_k/2\right)$. This leads to write ℓ_{c_3} in (12) as follows

$$\ell_{c_3}\left(\boldsymbol{\vartheta}_{\boldsymbol{Y}}\right) = \ell_{c_{3a}}\left(\boldsymbol{\nu}\right) + \ell_{c_{3b}}\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right), \tag{A.1}$$

where

$$\ell_{c_{3a}}\left(\boldsymbol{\nu}\right) = \sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{k=1}^{K} z_{itk}\left\{-\ln\left[\Gamma\left(\frac{\nu_k}{2}\right)\right] + \frac{\nu_k}{2}\ln\left(\frac{\nu_k}{2}\right) + \frac{\nu_k}{2}\left[\ln\left(u_{itk}\right) - u_{itk}\right] - \ln\left(u_{itk}\right)\right\},$$

$$\ell_{c_{3b}}\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}\right) = -\frac{1}{2}\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{k=1}^{K} z_{itk}\left\{d_{\boldsymbol{Y}}\ln\left(2\pi\right) + \ln\left|\boldsymbol{\Sigma}_k\right| + u_{itk}\delta\left(\boldsymbol{y}_{it}, \boldsymbol{\mu}\left(\boldsymbol{x}_{it};\boldsymbol{\beta}_k\right);\boldsymbol{\Sigma}_k\right)\right\},$$

with $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$, $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$, $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_K)$, and $\boldsymbol{\vartheta}_{\boldsymbol{Y}} = (\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\nu})$.

Appendix A.1.2. Multivariate contaminated Gaussian distribution.

If $\boldsymbol{Y}_{it} \mid \boldsymbol{X}_{it} = \boldsymbol{x}_{it}, S_{it} = k \sim CN_{d_Y}\left(\boldsymbol{\mu}\left(\boldsymbol{x}_{it};\boldsymbol{\beta}_k\right), \boldsymbol{\Sigma}_k, \alpha_k, \eta_k\right)$, a further source of incompleteness arises from the fact that for each observation $(\boldsymbol{x}_{it}, \boldsymbol{y}_{it})$ in state k we do not know if it is either typical or outlier. To denote this source of incompleteness, we use $\boldsymbol{u}_{it} = (u_{it1}, \ldots, u_{itk}, \ldots, u_{itK})'$, where $u_{itk} = 1$ if $(\boldsymbol{x}_{it}, \boldsymbol{y}_{it})$ in state k is a typical vertical point and $u_{itk} = 0$ if it is a bad vertical point. Thus, in the complete-data log-likelihood for the CN-HMRM, we can specify (12) as

$$\begin{aligned}\ell_{c_3}\left(\boldsymbol{\vartheta}_{\boldsymbol{Y}}\right) &= \sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{k=1}^{K} z_{itk}\ln\left\{\left[\alpha_k\phi\left(\boldsymbol{y}_{it};\boldsymbol{\mu}\left(\boldsymbol{x}_{it};\boldsymbol{\beta}_k\right), \boldsymbol{\Sigma}_k\right)\right]^{u_{itk}}\left[\left(1-\alpha_k\right)\phi\left(\boldsymbol{y}_{it};\boldsymbol{\mu}\left(\boldsymbol{x}_{it};\boldsymbol{\beta}_k\right), \eta_k\boldsymbol{\Sigma}_k\right)\right]^{1-u_{itk}}\right\}\\ &= \ell_{c_{3a}}\left(\boldsymbol{\alpha}\right) + \ell_{c_{3b}}\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\eta}\right),\end{aligned} \tag{A.2}$$

where

$$\ell_{c_{3a}}\left(\boldsymbol{\alpha}\right) = \sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{k=1}^{K} z_{itk}\left[u_{itk}\ln\alpha_k + \left(1-u_{itk}\right)\ln\left(1-\alpha_k\right)\right],$$

$$\ell_{c_{3b}}\left(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\eta}\right) = -\frac{1}{2}\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{k=1}^{K}\left\{z_{itk}\ln\left|\boldsymbol{\Sigma}_k\right| + d_{\boldsymbol{Y}}z_{itk}\left(1-u_{itk}\right)\ln\eta_k + z_{itk}\left(u_{itk} + \frac{1-u_{itk}}{\eta_k}\right)\delta\left(\boldsymbol{y}_{it}, \boldsymbol{\mu}\left(\boldsymbol{x}_{it};\boldsymbol{\beta}_k\right);\boldsymbol{\Sigma}_k\right)\right\},$$

with $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$, $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K)$, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)$, and $\boldsymbol{\vartheta}_{\boldsymbol{Y}} = (\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \boldsymbol{\eta})$.

Appendix A.2. Computational details

The quantities $z_{itk}^{(r)}$ and $zz_{itjk}^{(r)}$ can be computed recursively (Baum et al., 1970). Let us define the forward probability

$$\gamma_{itk} = \Pr\left(\boldsymbol{Y}_{i1} = \boldsymbol{y}_{i1}, \ldots, \boldsymbol{Y}_{it} = \boldsymbol{y}_{it}, S_{it} = k \mid \boldsymbol{X}_{i1} = \boldsymbol{x}_{i1}, \ldots, \boldsymbol{X}_{it} = \boldsymbol{x}_{it}\right),$$

which represents the probability of seeing the partial sequence ending up in state k at time t, and the corresponding backward probability

$$\tau_{itk} = \Pr\left(\boldsymbol{Y}_{i(t+1)} = \boldsymbol{y}_{i(t+1)}, \ldots, \boldsymbol{Y}_{iT} = \boldsymbol{y}_{iT} \mid \boldsymbol{X}_{i1} = \boldsymbol{x}_{i1}, \ldots, \boldsymbol{X}_{it} = \boldsymbol{x}_{it}, S_{it} = k\right).$$

The forward recursion is given by

$$\gamma_{i1k} = \pi_k f\left(\boldsymbol{Y}_{i1} = \boldsymbol{y}_{i1} \mid \boldsymbol{X}_{i1} = \boldsymbol{x}_{i1}, S_{i1} = k\right)$$

and for $t = 2, \ldots, T$ we compute

$$\gamma_{i(t+1)k} = \sum_{j=1}^{K} \gamma_{itj}\pi_{k|j} f\left(\boldsymbol{Y}_{i(t+1)} = \boldsymbol{y}_{i(t+1)} \mid \boldsymbol{X}_{i(t+1)} = \boldsymbol{x}_{i(t+1)}, S_{i(t+1)} = k\right).$$

As a by-product, the likelihood function (8) is given by

$$\mathcal{L}(\boldsymbol{\vartheta}) = \prod_{i=1}^{I} \sum_{k=1}^{K} \gamma_{iTk}.$$

Similarly, it is possible to implement the following backward recursion

$$\tau_{iTk} = 1,$$

and for $t = T-1, \ldots, 1$ we have

$$\tau_{itj} = \sum_{k=1}^{K} \pi_{k|j} f(\boldsymbol{y}_{i(t+1)} \mid \boldsymbol{x}_{i(t+1)}, S_{i(t+1)} = k)\tau_{i(t+1)k}.$$

The expected values of the quantities involved in the E-step can be computed as follows

$$z_{itk}^{(r)} = \frac{\gamma_{itk}\tau_{itk}}{\sum_{h=1}^{K} \gamma_{ith}\tau_{ith}}$$

and

$$zz_{itjk}^{(r)} = \frac{\gamma_{i(t-1)j}\pi_{k|j} f\left(\boldsymbol{y}_{it} \mid S_{it} = k\right)\tau_{itk}}{\sum_{k=1}^{K} \gamma_{iTk}}.$$

Appendix A.3. Path prediction

A major issue of interest in a HMM framework is the prediction of the most likely hidden states (under the fitted model) to have given rise to the observations sequence. Local and global decoding can be investigated to solve the problem. A global decoding procedure maximizes the posterior probability $\Pr(S_{i1} = s_{i1}, \ldots, S_{iT} = s_{iT} \mid \boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{iT}, \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{iT})$ with respect to $(s_{i1}, \ldots, s_{iT})$, i.e. identifies the most likely sequence of hidden states. To avoid inconsistent sequences and to account for the joint probability of the entire latent sequence, the Viterbi algorithm (Viterbi, 1967) is often employed. Let $\rho_{it}(s_{it}) = \max_{s_{i1},\ldots,s_{it}} \Pr(s_{i1}, \ldots, s_{it}, \boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{it} \mid \boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{it})$, the algorithm performs the following steps

1. For $k \in \{1, 2, \ldots, K\}$, compute $\rho_{i1}(k) = \pi_k f\left(\boldsymbol{Y}_{i1} = \boldsymbol{y}_{i1} \mid \boldsymbol{X}_{i1} = \boldsymbol{x}_{i1}, S_{i1} = k\right)$.
2. For $t = 2, \ldots, T$, $k = 1, \ldots, K$, calculate $\rho_{it}(k) = f\left(\boldsymbol{y}_{it} \mid \boldsymbol{x}_{it}, S_{it} = k\right) \max_j \left[\rho_{i(t-1)}(j)\pi_{k|j}\right]$.
3. Find the optimal $s_{iT} = \text{argmax}_k\, \rho_{iT}(k)$.
4. For $t = T-1, T-2, \ldots, 1$, determine s_{it} by $s_{it} = \text{argmax}_j\, \rho_{it}(j)\pi_{s_{i(t+1)}|j}$.

In other words, the algorithm performs a forward recursion to compute the above quantities, and then it finds the most likely latent sequence with a backward recursion. All of the above quantities are computed on the basis of the ML parameter estimates.

Appendix A.4. Forecasting

We now turn to the forecast distribution of an HMRM. The conditional forecast distribution of the response vector $\boldsymbol{Y}$ for unit i at time $T + t^*$, say $\boldsymbol{Y}_{i(T+t^*)}$, given both the information $\boldsymbol{y}_i$ up to time T and a set of covariates $\boldsymbol{x}_{i(T+t^*)}$, is a finite mixture of state-specific distributions; see e.g. Zucchini and MacDonald (2009, Section 5.2). Formally,

$$f\left(\boldsymbol{Y}_{i(T+t^*)} = \boldsymbol{y}_{i(T+t^*)} \mid \boldsymbol{X}_{i(T+t^*)} = \boldsymbol{x}_{i(T+t^*)}, \boldsymbol{y}_i; \hat{\boldsymbol{\vartheta}}_k\right) = \sum_{k=1}^{K} \tilde{\pi}_k^{(t^*)} f\left(\boldsymbol{Y}_{i(T+t^*)} = \boldsymbol{y}_{i(T+t^*)} \mid \boldsymbol{X}_{i(T+t^*)} = \boldsymbol{x}_{i(T+t^*)}, S_{i(T+t^*)} = k; \hat{\boldsymbol{\vartheta}}_k\right),$$

with mixing weights

$$\tilde{\pi}_k^{(t^*)} = \sum_{j=1}^{K} \pi_{k|j}^{t^*} \Pr(S_{iT} = j \mid \boldsymbol{y}_i),$$

where $\pi_{k|j}^{t^*}$ is the (j, k)-th entry of the transition probability matrix $\boldsymbol{\Pi}$ to the power t^*.

Acknowledgments

This paper presents independent research part funded by the National Institute for Health Research Health Services & Delivery Research programme (grant number 13/114/17). The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health.

Antonio Punzo acknowledges the financial support from the grant (RBFR12SHVV) ”Finite mixture and latent variable models for causal inference and analysis of socio–economic data” (FIRB2012-Futuro in Ricerca) funded by the Italian Government.

References

Ailliot, P., Bessac, J., Monbet, V., Pene, F., 2015. Non-homogeneous hidden Markov-switching models for wind time series. Journal of Statistical Planning and Inference 160, 75 – 88.

Bagnato, L., Greselin, F., Punzo, A., 2014. On the spectral decomposition in normal discriminant analysis. Communications in Statistics - Simulation and Computation 43 (6), 1471–1489.

Bagnato, L., Punzo, A., 2013. Finite mixtures of unimodal beta and gamma densities and the k-bumps algorithm. Computational Statistics 28 (4), 1571–1597.

Bai, X., Chen, K., Yao, W., 2016. Mixture of linear mixed models using multivariate t distribution. Journal of Statistical Computation and Simulation 86 (4), 771–787.

Bai, X., Yao, W., Boyer, J. E., 2012. Robust fitting of mixture regression models. Computational Statistics & Data Analysis 56 (7), 2347–2359.

Bartolucci, F., Farcomeni, A., 2009. A multivariate extension of the dynamic logit model for longitudinal data based on a latent Markov heterogeneity structure. Journal of the American Statistical Association 104 (486), 816–831.

Bartolucci, F., Farcomeni, A., 2015. A discrete time event-history approach to informative drop-out in mixed latent Markov models with covariates. Biometrics 71 (1), 80–89.

Bartolucci, F., Farcomeni, A., Pennoni, F., 2013. Latent Markov models for longitudinal data. CRC Press.

Baum, L. E., Petrie, T., Soules, G., Weiss, N., 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. The Annals of Mathematical Statistics 41 (1), 164–171.

Berkane, M., Bentler, P. M., 1988. Estimation of contamination parameters and identification of outliers in multivariate data. Sociological Methods & Research 17 (1), 55–64.

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

Biernacki, C., Celeux, G., Govaert, G., 2003. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. Computational Statistics & Data Analysis 41 (3-4), 561–575.

Browne, R. P., McNicholas, P. D., 2015. A mixture of generalized hyperbolic distributions. Canadian Journal of Statistics 43 (2), 176–198.

Bulla, J., Berzel, A., 2008. Computational issues in parameter estimation for stationary hidden Markov models. Computational Statistics 23 (1), 1–18.

Campbell, N. A., Mahon, R. J., 1974. A multivariate study of variation in two species of rock crab of genus Leptograpsus. Australian Journal of Zoology 22 (3), 417–425.

Crawford, S. L., 1994. An application of the laplace method to finite mixture distributions. Journal of the American Statistical Association 89 (425), 259–267.

Dannemann, J., Holzmann, H., Leister, A., 2014. Semiparametric hidden Markov models: identifiability and estimation. Wiley Interdisciplinary Reviews: Computational Statistics 6 (6), 418–425.

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

Dickson, E. R., Grambsch, P. M., Fleming, T. R., Fisher, L. D., Langworthy, A., 1989. Prognosis inprimary biliary-cirrhosis: Model for decision-making. Hepatology 10, 1–7.

Farcomeni, A., 2012. Quantile regression for longitudinal data based on latent Markov subject-specific parameters. Statistics and Computing 22 (1), 141–152.

Farcomeni, A., Greco, L., 2015. S-estimation of hidden Markov models. Computational Statistics 30 (1), 57–80.

Franczak, B. C., Browne, R. P., McNicholas, P. D., 2014. Mixtures of shifted asymmetriclaplace distributions. IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (6), 1149–1157.

Frühwirth-Schnatter, S., 2006. Finite Mixture and Markov Switching Models. Springer, New York.

Frühwirth-Schnatter, S., 2011. Panel data analysis: a survey on model-based clustering of time series. Advances in Data Analysis and Classification 5 (4), 251–280.

Frühwirth-Schnatter, S., Kaufmann, S., 2008. Model-based clustering of multiple time series. Journal of Business & Economic Statistics 26 (1), 78–89.

García-Escudero, L. A., Gordaliza, A., Matrán, C., Mayo-Iscar, A., 2010. A review of robust clustering methods. Advances in Data Analysis and Classification 4 (2), 89–109.

García-Escudero, L. A., Gordaliza, A., Mayo-Iscar, A., San Martín, R., 2010. Robust clusterwise linear regression through trimming. Computational Statistics & Data Analysis 54 (12), 3057–3069.

Goldfeld, S. M., Quandt, R. E., 1973. A Markov model for switching regressions. Journal of Econometrics 1 (1), 3–15.

Greselin, F., Ingrassia, S., Punzo, A., 2011. Assessing the pattern of covariance matrices via an augmentation multiple testing procedure. Statistical Methods & Applications 20 (2), 141–170.

Greselin, F., Punzo, A., 2013. Closed likelihood ratio testing procedures to assess similarity of covariance matrices. The American Statistician 67 (3), 117–128.

Grün, B., Leisch, F., 2008. Recent Advances in Linear Models and Related Areas: Essays in Honour of Helge Toutenburg. Physica-Verlag HD, Heidelberg, Ch. Finite Mixtures of Generalized Linear Regression Models, pp. 205–230.

Hamilton, J. D., 1990. Analysis of time series subject to changes in regime. Journal of Econometrics 45 (1–2), 39–70.

Hartigan, J. A., Hartigan, P. M., 03 1985. The dip test of unimodality. Ann. Statist. 13 (1), 70–84.

Hennig, C., 2000. Identifiablity of models for clusterwise linear regression. Journal of Classification 17 (2), 273–296.

Holzmann, H., Munk, A., Gneiting, T., 2006. Identifiability of finite mixtures of elliptical distributions. Scandinavian Journal of Statistics 33 (4), 753–763.

Ingrassia, S., Minotti, S. C., Punzo, A., 2014. Model-based clustering via linear cluster-weighted models. Computational Statistics and Data Analysis 71, 159–182.

Ingrassia, S., Punzo, A., 2016. Decision boundaries for mixtures of regressions. Journal of the Korean Statistical Society. To appear, DOI: 10.1016/j.jkss.2015.11.005.

Ingrassia, S., Punzo, A., Vittadini, G., Minotti, S. C., 2015. The generalized linear mixed cluster-weighted model. Journal of Classification 32 (1), 85–113.

Juárez, M. A., Steel, M. F. J., 2010. Model-based clustering of non-gaussian panel data based on skew-*t* distributions. Journal of Business & Economic Statistics 28 (1), 52–66.

Karlis, D., Santourian, A., 2009. Model-based clustering with non-elliptically contoured distributions. Statistics and Computing 19 (1), 73–83.

Lagona, F., Jdanov, D., Shkolnikova, M., 2014. Latent time-varying factors in longitudinal analysis: a linear mixed hidden Markov model for heart rates. Statistics in Medicine 33 (23), 4116–4134.

Lagona, F., Maruotti, A., Padovano, F., 2015. Multilevel multivariate modelling of legislative count data, with a hidden Markov chain. Journal of the Royal Statistical Society - Series A 178, 705–723.

Langrock, R., King, R., 2013. Maximum likelihood estimation of mark-recapture-recovery models in the presence of continuous covariates. Annals of Applied Statistics 7 (3), 1709–1732.

Langrock, R., Swihart, B. J., Caffo, B. S., Punjabi, N. M., Crainiceanu, C. M., 2013. Combining hidden Markov models for comparing the dynamics of multiple sleep electroencephalograms. Statistics in Medicine 32 (19), 3342–3356.

Lee, S. X., McLachlan, G. J., 2013. Model-based clustering and classification with non-normal mixture distributions. Statistical Methods & Applications 22 (4), 427–454.

Lee, S. X., McLachlan, G. J., 2014. Finite mixtures of multivariate skew *t*-distributions: some recent and new results. Statistics and Computing 24 (2), 181–202.

Lee, Y., Ghosh, D., Hardison, R. C., Zhang, Y., 2014. Mrhmms: Multivariate regression hidden Markov models and the variants. Bioinformatics 30 (13), 1755–1756.

Leroux, B. G., 1992. Maximum-likelihood estimation for hidden Markov models. Stochastic Processes and Their Applications 40 (1), 127–143.

Lin, T. I., 2009. Maximum likelihood estimation for multivariate skew normal mixture models. Journal of Multivariate Analysis 100 (2), 257–265.

Lin, T. I., 2010. Robust mixture modeling using multivariate skew *t* distributions. Statistics and Computing 20 (3), 343–356.

Little, R. J. A., 1988. Robust estimation of the mean and covariance matrix from data with missing values. Applied Statistics 37 (1), 23–38.

Lo, K., Gottardo, R., 2012. Flexible mixture modeling via the multivariate *t* distribution with the box-cox transformation: an alternative to the skew-*t* distribution. Statistics and Computing 22 (1), 33–52.

MacDonald, I. L., 2014. Numerical maximisation of likelihood: A neglected alternative to EM? International Statistical Review 82 (2), 296–308.

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. International Statistical Review 79 (3), 427–454.

Maruotti, A., 2014. Robust fitting of hidden Markov regression models under a longitudinal setting. Journal of Statistical Computation and Simulation 84 (8), 1728–1747.

Maruotti, A., Punzo, A., Mastrantonio, G., Lagona, F., 2016. A time-dependent extension of the projected normal regression model for longitudinal circular data based on a hidden Markov heterogeneity structure. Stochastic Environmental Research and Risk Assessment, 1–16. To appear, doi: 10.1007/s00477-015-1183-5.

Maruotti, A., Rocci, R., 2012. A mixed non-homogeneous hidden Markov model for categorical data, with application to alcohol consumption. Statistics in Medicine 31 (9), 871–886.

McLachlan, G. J., 1992. Discriminant analysis and statistical pattern recognition. 2nd printing. John Wiley & Sons, Hoboken, New Jersey.

McLachlan, G. J., Peel, D., 2000. Finite Mixture Models. John Wiley & Sons, New York.

Meng, X.-L., Rubin, D. B., 1993. Maximum likelihood estimation via the ECM algorithm: A general framework. Biometrika 80 (2), 267–278.

Punzo, A., 2014. Flexible mixture modeling with the polynomial Gaussian cluster-weighted model. Statistical Modelling 14 (3), 257–291.

Punzo, A., Browne, R. P., McNicholas, P. D., 2016. Hypothesis testing for mixture model selection. Journal of Statistical Computation and Simulation. To appear, doi: 10.1080/00949655.2015.1131282.

Punzo, A., Ingrassia, S., 2015. Clustering bivariate mixed-type data via the cluster-weighted model. Computational Statistics. To appear, doi: 10.1007/s00180-015-0600-z.

Punzo, A., Maruotti, A., 2016. Clustering multivariate longitudinal observations: The contaminated Gaussian hidden Markov model. Journal of Computational and Graphical Statistics. To appear, doi: 10.1080/10618600.2015.1089776.

Punzo, A., McNicholas, P. D., 2014. Robust clustering in regression analysis via the contaminated Gaussian cluster-weighted model. arXiv.org e-print 1409.6019, available at: `http://arxiv.org/abs/1409.6019`.

Punzo, A., McNicholas, P. D., 2016. Parsimonious mixtures of multivariate contaminated normal distributions. Biometrical Journal. To appear.

Pyne, S., Hu, X., Wang, K., Rossin, E., Lin, T. I., Maier, L. M., Baecher-Allan, C., McLachlan, G. J., Tamayo, P., Hafler, D. A., De Jager, P. L., Mesirov, J. P., 2009. Automated high-dimensional flow cytometric data analysis. Proceedings of the National Academy of Sciences 106 (21), 8519–8524.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org/`

Raffa, J. D., Dubin, J. A., 2015. Multivariate longitudinal data analysis with mixed effects hidden Markov models. Biometrics 71 (3), 821–831.

Ritter, G., 2015. Robust Cluster Analysis and Variable Selection. Vol. 137 of Chapman & Hall/CRC Monographs on Statistics & Applied Probability. CRC Press.

Schliehe-Diecks, S., Kappeler, P., Langrock, R., 2012. On the application of mixed hidden Markov models to multiplebehavioural time series. Interface Focus 2, 180–189.

Schork, N. J., Schork, M. A., 1988. Skewness and mixtures of normal distributions. Communications in Statistics-Theory and Methods 17 (11), 3951–3969.

Schreuder, H. T., Hafley, W. L., 1977. A useful bivariate distribution for describing stand structure of tree heights and diameters. Biometrics 33 (3), 471–478.

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

Subedi, S., Punzo, A., Ingrassia, S., McNicholas, P. D., 2013. Clustering and classification via cluster-weighted factor analyzers. Advances in Data Analysis and Classification 7 (1), 5–40.

Subedi, S., Punzo, A., Ingrassia, S., McNicholas, P. D., 2015. Cluster-weighted *t*-factor analyzers for robust model-based clustering and dimension reduction. Statistical Methods & Applications 24 (4), 623–649.

Titterington, D. M., Smith, A. F. M., Makov, U. E., 1985. Statistical Analysis of Finite Mixture Distributions. John Wiley & Sons, New York.

Turner, R., 2008. Direct maximization of the likelihood of a hidden Markov model. Computational Statistics & Data Analysis 52 (9), 4147–4160.

Vermunt, J. K., 2010. Longitudinal Research with Latent Variables. Springer, Berlin, Heidelberg, Ch. Longitudinal Research Using Mixture Models, pp. 119–152.

Visser, I., 2011. Seven things to remember about hidden Markov models: A tutorial on Markovian models for time series. Journal of Mathematical Psychology 55 (6), 403–415.

Viterbi, A. J., 1967. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. IEEE Transactions on Information Theory 13 (2), 260–269.

Wang, W.-L., 2013. Multivariate t linear mixed models for irregularly observed multiple repeated measures with missing outcomes. Biometrical Journal 55 (4), 554–571.

Wang, W.-L., Lin, T.-I., Lachos, V. H., 2015. Extending multivariate-*t* linear mixed models for multiple longitudinal data with censored responses and heavy tails. Statistical Methods in Medical Research. To appear, doi: 10.1177/0962280215620229.

Zhu, X., Melnykov, V., 2016. Manly transformation in finite mixture modeling. Computational Statistics & Data Analysis.

Zucchini, W., MacDonald, I. L., 2009. Hidden Markov models for time series: An introduction using R. Chapman & Hall, Boca Raton, FL.