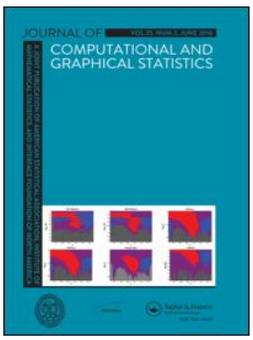


Taylor & Francis
Taylor & Francis Group

Journal of Computational and Graphical Statistics

ISSN: 1061-8600 (Print) 1537-2715 (Online) Journal homepage: https://amstat.tandfonline.com/loi/ucgs20

Clustering Multivariate Longitudinal Observations: The Contaminated Gaussian Hidden Markov Model

Antonio Punzo & Antonello Maruotti

To cite this article: Antonio Punzo & Antonello Maruotti (2016) Clustering Multivariate Longitudinal Observations: The Contaminated Gaussian Hidden Markov Model, Journal of Computational and Graphical Statistics, 25:4, 1097-1098, DOI: 10.1080/10618600.2015.1089776

To link to this article: https://doi.org/10.1080/10618600.2015.1089776

Supplementary materials for this article are available online.
Please go to www.tandfonline.com/r/JCGS

Clustering Multivariate Longitudinal Observations: The Contaminated Gaussian Hidden Markov Model

Antonio PUNZO and Antonello MARUOTTI

The Gaussian hidden Markov model (HMM) is widely considered for the analysis of heterogenous continuous multivariate longitudinal data. To robustify this approach with respect to possible elliptical heavy-tailed departures from normality, due to the presence of outliers, spurious points, or noise (collectively referred to as *bad points* herein), the contaminated Gaussian HMM is here introduced. The contaminated Gaussian distribution represents an elliptical generalization of the Gaussian distribution and allows for automatic detection of bad points in the same natural way as observations are typically assigned to the latent states in the HMM context. Once the model is fitted, each observation has a posterior probability of belonging to a particular state and, inside each state, of being a bad point or not. In addition to the parameters of the classical Gaussian HMM, for each state we have two more parameters, both with a specific and useful interpretation: one controls the proportion of bad points and one specifies their degree of atypicality. A sufficient condition for the identifiability of the model is given, an expectation-conditional maximization algorithm is outlined for parameter estimation and various operational issues are discussed. Using a large-scale simulation study, but also an illustrative artificial dataset, we demonstrate the effectiveness of the proposed model in comparison with HMMs of different elliptical distributions, and we also evaluate the performance of some well-known information criteria in selecting the true number of latent states. The model is finally used to fit data on criminal activities in Italian provinces. Supplementary materials for this article are available online

Key Words: Atypical data; Elliptical distributions; Expected-conditional maximization (ECM) algorithm; Model selection; Robust model-based clustering.

1. INTRODUCTION

Hidden Markov models (HMMs) are the state of the art in the analysis of time-dependent data. HMMs have been applied in time series analysis for more than four decades

Antonio Punzo, Associate Professor of Statistics, Department of Economics and Business, University of Catania, 95129 Catania, Italy (E-mail: *antonio.punzo@unict.it*). Antonello Maruotti, Associate Professor of Statistics, Southampton Statistical Sciences Research Institute, University of Southampton, Southampton SO17 1BJ, UK and Department of Economic, Political Sciences and Modern Languages, Libera Università Maria Ss. Assunta, 00193 Roma, Italy (E-mail: *a.maruotti@lumsa.it*).

Journal of Computational and Graphical Statistics, Volume 25, Number 4, Pages 1097–1116
DOI: 10.1080/10618600.2015.1089776

(Baum and Petrie 1966) and, more recently, in the longitudinal setting (see Maruotti 2011; Bartolucci, Farcomeni, and Pennoni 2013, and the references therein). Serial dependence and heterogeneity in sample units characterize the longitudinal setting and can be properly investigated and accounted for in an HMM framework. Being dependent mixture models, HMMs allow the unambiguously recover of the structure of the data by rigorously defining homogenous latent subgroups and, simultaneously, provide meaningful interpretation of the inferred partition. For multivariate continuous data, attention is commonly focused on Gaussian HMMs (Bartolucci and Farcomeni 2010; Volant et al. 2014; Holzmann and Schwaiger 2015), with few notable exceptions (Bartolucci and Farcomeni 2009; Bulla et al. 2012; Lagona, Maruotti, and Padovano 2015).

Unfortunately, real data are often contaminated by outliers, spurious points, or noise (collectively referred to as *bad points* herein, as in Aitkin and Wilson 1980) that may affect parameter estimates and the recovering of the latent structure. The attempt of robustly estimating mixture models parameters has led to a heterogenous literature that includes: noise approaches (Banfield and Raftery 1993; Fraley and Raftery 2002), that is, methods aiming at identifying a noise component (modeled assuming a uniform component-specific distribution), while simultaneously clustering nonnoise observations; distance approaches (Rousseeuw and Leroy 2005; Cerioli 2010; Garcia-Escudero et al. 2015); and distribution-based robust approaches (Peel and McLachlan 2000; Andrews and McNicholas 2012). While all these methods offer important contributions to the topic, the last two methods do not allow for the direct detection of bad points. Approaches considering the uniform distribution, if used for discriminant analysis, cannot recognize a new bad observation (an observation that has not been used to fit the model) if it lies outside the support defined by the fitted uniform distribution(s). An alternative to these methods aiming at identifying outlying observations that deviate from the cluster-specific distribution has been recently proposed by Evans, Love, and Thurston (2015). Despite the wide literature on robust estimation of mixture models, there are not many articles dealing with robustness issues in HMMs. In the univariate case, Bulla (2011) introduced a structured HMM to account for outliers in financial time series, Humburg, Bulger, and Stone (2008) proposed the use of the *t*-distribution, and Maruotti (2014) considered a bi-square scale estimator in a regression setting. In the multivariate case, Farcomeni and Greco (2015) introduced a robust *S*-estimator and Bernardi, Maruotti, and Petrella (2014) proposed the use of the multivariate *t*-distribution for multivariate financial (time-series) data in an HMM framework.

In this article, we extend this branch of literature by introducing a joint approach to time-varying robust clustering and bad points detection under a longitudinal setting, extending the standard HMM framework (see Section 2.1). As emphasized by Davies and Gather (1993; see also Hennig 2002), bad observations should be defined with respect to a reference distribution. Accordingly, the region of bad points can be defined, for example, as a region where the density of the reference distribution is low. In analogy with other distribution-based approaches (as those based on the *t*-distribution), we choose the Gaussian distribution as the reference distribution but, differently from the *t* HMM, we replace the multivariate Gaussian state-dependent distribution with a two-component Gaussian mixture (Tukey 1960) where one (reference) component represents the data we would expect from the given state (i.e., good points) while the other component clusters the bad points; the latter component has a small prior probability, the same component-specific mean and

an inflated covariance matrix. Its investigation and use in a clustering framework is still in infancy, although some results have been recently obtained by Punzo and McNicholas (2014a, 2014b, 2016) in a cross-sectional setting. This change makes the model much more robust and allows for automatic detection of bad points. With respect to the latter issue, as it will be better explained later, once the contaminated Gaussian HMM is fitted to the observed longitudinal data, by means of maximum a posteriori probabilities, each observation can be first assigned to one of the states and then classified as good or bad; thus, we have a model for simultaneous robust clustering and detection of atypical observations in a longitudinal context. Of course, this is not the only attempt to deal with clustering under a longitudinal setting. Our proposal is somehow related to the models proposed by De la Cruz-Mesia, Quintana, and Marshall (2008), who introduced a (univariate) hierarchical mixture model, and by McNicholas and Murphy (2010), who considered a (univariate) mixture model in which a decomposed covariance structure is introduced to explicitly account for the relationship between measurements at different time points. However, none of the aforementioned approaches allows for time-varying clustering neither of bad points detection, and, moreover, both have been introduced in a univariate setting only.

After establishing a sufficient condition for the identifiability of the model (see Section 2.2), in Section 2.3 we outline an ad hoc version of the expectation-conditional maximization (ECM) algorithm to estimate model parameters, extending the Baum–Welch iterative procedure (Baum et al. 1970) to deal with contaminated Gaussian distributions. Further operational aspects are discussed in Section 3. In Section 4.1, we illustrate the proposal by a large-scale simulation study to investigate the empirical behavior of the proposed approach with respect to several factors—such as the number of observed units and times, and the nature of bad points—and in comparison with HMMs of different elliptical distributions. Indeed, different aspects of robustness are going to be described and analyzed. We will consider heavy tails (conditional) distributions as data-generation processes as well as distributions with contaminated units in the data. Furthermore, in Section 4.2, we provide insights on information criteria performances in this framework. At last, after an illustration on artificial data (see Section 5.1), we illustrate the proposal in Section 5.2 by analyzing a longitudinal dataset of Italian provinces on which four different crimes rates have been measured from 2005 to 2009, previously analyzed in a different context by Viroli (2011). Provinces are clustered and bad points are automatically detected.

2. METHODOLOGY

2.1 THE MODEL

Let $\{\boldsymbol{Y}_{it}; i = 1, \ldots, I, t = 1, \ldots, T\}$ denote sequences of multivariate longitudinal observations recorded on I units and T times, where $\boldsymbol{Y}_{it} = (Y_{it1}, \ldots, Y_{itP})' \in \mathbb{R}^P$, and let $\{S_{it}; i = 1, \ldots, I, t = 1, \ldots, T\}$ be a first-order Markov chain defined on the state space $\{1, \ldots, k, \ldots, K\}$. An 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, that is,

$$\Pr(S_{it} = s_{it} \mid S_{i1} = s_{i1}, S_{i2} = s_{i2}, \ldots, S_{it-1} = s_{it-1}) = \Pr(S_{it} = s_{it} \mid S_{it-1} = s_{it-1}),$$

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

$$f\left(\boldsymbol{Y}_{it} = \boldsymbol{y}_{it} \mid \boldsymbol{Y}_{i1} = \boldsymbol{y}_{i1}, \ldots, \boldsymbol{Y}_{it-1} = \boldsymbol{y}_{it-1}, S_{i1} = s_{i1}, \ldots, S_{it} = s_{it}\right)$$
$$= f\left(\boldsymbol{Y}_{it} = \boldsymbol{y}_{it} \mid S_{it} = s_{it}\right),$$

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

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

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

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

$$\phi\left(\boldsymbol{y}_{it} \mid S_{it} = k; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \alpha_k, \eta_k\right) = \alpha_k \mathcal{N}_P\left(\boldsymbol{y}_{it} \mid S_{it} = k; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$$
$$+ (1 - \alpha_k)\mathcal{N}_P\left(\boldsymbol{y}_{it} \mid S_{it} = k; \boldsymbol{\mu}_k, \eta_k \boldsymbol{\Sigma}_k\right),$$

where $\mathcal{N}_P(\cdot; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ denotes the P-variate Gaussian distribution with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$, $\alpha_k \in (0, 1)$ is the proportion of good points in state k, and $\eta_k > 1$ is an inflation parameter in state k; the latter parameter can be also meant as a sort of "degree of atypicality" of the bad point(s), that is, as a measure of how different atypical observations are from the bulk of the (clustered) data.

2.2 IDENTIFIABILITY

An important issue in dealing with the proposed model is to establish its 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) showed that identifiability up to label switching follows from identifiability of the marginal finite mixtures (see Dannemann, Holzmann, and Leister 2014, sec. 2). In our case, the parametric family is constituted by contaminated Gaussian distributions and the marginal finite mixtures are represented by the finite mixtures of contaminated Gaussian distributions introduced by Punzo and McNicholas (2015). These authors also provide a sufficient condition for the identifiability of their mixture (see Punzo and McNicholas 2016, Proposition 1) that can be summarized as follows:

If $k \neq k_1$ implies

$$\left\|\boldsymbol{\mu}_k - \boldsymbol{\mu}_{k_1}\right\|_2^2 + \left\|\boldsymbol{\Sigma}_k - a\boldsymbol{\Sigma}_{k_1}\right\|_2^2 \neq 0$$

for all $a > 0$, where $\|\cdot\|_2$ is the Froebenius norm, then a finite mixture of contaminated Gaussian distributions is identifiable.

Accordingly, a finite mixture of contaminated Gaussian distributions is identifiable if two of the K Gaussian distributions representing the good observations have distinct component means and/or nonproportional component covariance matrices. Based on Leroux (1992), the same sufficient condition for identifiability is inherited by our contaminated Gaussian HMM.

2.3 MAXIMUM LIKELIHOOD ESTIMATION

To perform ML estimation of the parameters of the proposed model on the basis of the sample $\{\boldsymbol{y}_{it}; i = 1, \ldots, I, t = 1, \ldots, T\}$, the need arises of computing

$$\mathcal{L}(\boldsymbol{\vartheta}) = \prod_{i=1}^{I} \mathcal{L}_i(\boldsymbol{\vartheta}) = \prod_{i=1}^{I} \boldsymbol{\pi}' \boldsymbol{\phi}(\boldsymbol{y}_{i1}) \boldsymbol{\Pi} \boldsymbol{\phi}(\boldsymbol{y}_{i2}) \boldsymbol{\Pi} \cdots \boldsymbol{\phi}(\boldsymbol{y}_{iT-1}) \boldsymbol{\Pi} \boldsymbol{\phi}(\boldsymbol{y}_{iT}) \mathbf{1}_K, \tag{2}$$

where $\boldsymbol{\vartheta} = \{\boldsymbol{\pi}, \boldsymbol{\Pi}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \alpha_k, \eta_k; k = 1, \ldots, K\}$ corresponds to the set of all model parameters, $\mathbf{1}_K$ denotes a vector of K ones, and $\boldsymbol{\phi}(\boldsymbol{y}_{it})$ denotes a $K \times K$ diagonal matrix with conditional densities $\phi(\boldsymbol{Y}_{it} = \boldsymbol{y}_{it} \mid S_{it} = k)$ on the main diagonal. Finding the value of the parameters $\boldsymbol{\vartheta}$ that maximizes the log-transformation of (2) under the constraints $\sum_{k=1}^{K} \pi_k = 1$, $\sum_{k=1}^{K} \pi_{k|j} = 1$, $\alpha_k \in (0, 1)$, and $\eta_k > 1$, $k, j = 1, \ldots, K$, is not an easy problem since (2) is not available in an analytically convenient form. Efficient computation of (2) 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, an expectation-conditional maximization (ECM) algorithm (Meng and Rubin 1993) is used for fitting our contaminated model. The ECM is a variant of the classical EM algorithm (Baum et al. 1970; Dempster, Laird, and Rubin 1977), which is a natural approach for ML estimation when data are incomplete. In our setting, there are two sources of missing data: one arises from the fact that we do not know state membership and its evolution over time, and the other from the fact that we do not know whether an observation clustered in a specific state is good or bad.

Formally, let us define the unobserved state membership $\boldsymbol{z}_{it} = (z_{it1}, \ldots, z_{itk} \ldots, z_{itK})'$, the unobserved states transition

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

and the unobserved state-specific membership to the good points $\boldsymbol{v}_{it} = (v_{it1}, \ldots, v_{itk}, \ldots, v_{itK})'$, as missing data, with

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

and $v_{itk} = 1$ if observation i at time t in state k is a good point and $v_{itk} = 0$ if it is a bad point. Therefore, the complete data are given by $\mathcal{C} = \{\boldsymbol{y}_{it}, z_{it}, zz_{it}, \boldsymbol{v}_{it}; i = 1, \ldots, I, t = 1, \ldots, T\}$ and the complete-data log-likelihood can be written as

$$\ell_c(\boldsymbol{\vartheta} \mid \mathcal{C}) = \ell_{c_1}(\boldsymbol{\pi} \mid \mathcal{C}) + \ell_{c_2}(\boldsymbol{\Pi} \mid \mathcal{C}) + \ell_{c_3}(\boldsymbol{\alpha} \mid \mathcal{C}) + \ell_{c_4}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\eta} \mid \mathcal{C}),$$

where

$$\ell_{c_1}(\boldsymbol{\pi} \mid \mathcal{C}) = \sum_{i=1}^{I}\sum_{k=1}^{K} z_{i1k} \log(\pi_k)$$

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

$$\ell_{c_3}(\boldsymbol{\alpha} \mid \mathcal{C}) = \sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{k=1}^{K} z_{itk}\left[v_{itk}\log(\alpha_k) + (1 - v_{itk})\log(1 - \alpha_k)\right]$$

$$\ell_{c_4}(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\eta} \mid \mathcal{C}) = -\frac{1}{2}\sum_{i=1}^{I}\sum_{t=1}^{T}\sum_{k=1}^{K}\left[z_{itk}\log|\boldsymbol{\Sigma}_k| + Pz_{itk}(1 - v_{itk})\log(\eta_k) + z_{itk}\left(v_{itk} - \frac{1 - v_{itk}}{\eta_k}\right)(\boldsymbol{y}_{it} - \boldsymbol{\mu}_k)'\boldsymbol{\Sigma}_k^{-1}(\boldsymbol{y}_{it} - \boldsymbol{\mu}_k)\right],$$

where $\boldsymbol{\mu} = \{\boldsymbol{\mu}_k; k = 1, \ldots, K\}$, $\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_k; k = 1, \ldots, K\}$, $\boldsymbol{\eta} = \{\eta_k; k = 1, \ldots, K\}$, and $\boldsymbol{\alpha} = \{\alpha_k; k = 1, \ldots, K\}$.

The E-step, at the $(r + 1)$th iteration, computes the conditional expectations of ℓ_c with respect to $\boldsymbol{\vartheta}$, given the observed data and the current estimates of the parameters. To do this, we replace z_{itk} and z_{itjk} with their conditional expectations, namely, $\tilde{z}_{itk}^{(r)}$ and $\tilde{z}_{itjk}^{(r)}$ (for computational details, see Section A.1 in the supplementary material) and v_{itk} with

$$\tilde{v}_{itk}^{(r)} = E\left(V_{itk} \mid \boldsymbol{y}_{it}, \boldsymbol{\vartheta}^{(r)}\right) = \frac{\alpha_k \mathcal{N}_P\left(\boldsymbol{y}_{it} \mid S_{it} = k; \boldsymbol{\mu}_k^{(r)}, \boldsymbol{\Sigma}_k^{(r)}\right)}{\phi\left(\boldsymbol{y}_{it} \mid S_{it} = k; \boldsymbol{\mu}_k^{(r)}, \boldsymbol{\Sigma}_k^{(r)}, \alpha_k^{(r)}, \eta_k^{(r)}\right)}, \tag{3}$$

where V_{itk} is the random variable related to v_{itk}.

At the first CM-step of the $(r + 1)$th iteration, maximizing with respect to π_k, $\boldsymbol{\Pi}$, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}_k$, and α_k yields

$$\pi_k^{(r+1)} = \frac{\sum_{i=1}^{I}\tilde{z}_{i1k}^{(r)}}{I}, \qquad \pi_{k|j}^{(r+1)} = \frac{\sum_{i=1}^{I}\sum_{t=2}^{T} zz_{itjk}^{(r)}}{\sum_{i=1}^{I}\sum_{t=2}^{T}\sum_{k=1}^{K} zz_{itjk}^{(r)}},$$

$$\alpha_k^{(r+1)} = \frac{\sum_{i=1}^{I}\sum_{t=1}^{T}\tilde{z}_{itk}^{(r)}\tilde{v}_{itk}^{(r)}}{\sum_{i=1}^{I}\sum_{t=1}^{T}\tilde{z}_{itk}^{(r)}}, \boldsymbol{\mu}_k^{(r+1)} = \frac{\sum_{i=1}^{I}\sum_{t=1}^{T}\tilde{z}_{itk}^{(r)}\left(\tilde{v}_{itk}^{(r)} + \frac{1 - \tilde{v}_{itk}^{(r)}}{\eta_k^{(r)}}\right)\boldsymbol{y}_{it}}{\sum_{i=1}^{I}\sum_{t=1}^{T}\tilde{z}_{itk}^{(r)}\left(\tilde{v}_{itk}^{(r)} + \frac{1 - \tilde{v}_{itk}^{(r)}}{\eta_k^{(r)}}\right)}, \tag{4}$$

$$\boldsymbol{\Sigma}_k^{(r+1)} = \frac{\sum_{i=1}^{I}\sum_{t=1}^{T}\tilde{z}_{itk}^{(r)}\left(\tilde{v}_{itk}^{(r)} + \dfrac{1-\tilde{v}_{itk}^{(r)}}{\eta_k^{(r)}}\right)(\boldsymbol{y}_{it}-\boldsymbol{\mu}_k)(\boldsymbol{y}_{it}-\boldsymbol{\mu}_k)'}{\sum_{i=1}^{I}\sum_{t=1}^{T}\tilde{z}_{itk}^{(r)}}. \tag{5}$$

At the second CM-step of the $(r+1)$th iteration, we maximize the expectation of the complete-data log-likelihood with respect to η_k, fixing all other parameters to their updated values at the first CM-step. In particular, we have to maximize

$$-\frac{P}{2}\sum_{i=1}^{I}\sum_{t=1}^{T}\tilde{z}_{itk}^{(r)}\left(1-\tilde{v}_{itk}^{(r)}\right)\log(\eta_k) - \frac{1}{2}\sum_{i=1}^{I}\sum_{t=1}^{T}\tilde{z}_{itk}^{(r)}\frac{1-\tilde{v}_{itk}^{(r)}}{\eta_k}\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_k^{(r+1)}\right)'$$
$$\times\left(\boldsymbol{\Sigma}_k^{(r+1)}\right)^{-1}\left(\boldsymbol{y}_{it}-\boldsymbol{\mu}_k^{(r+1)}\right)$$

with respect to η_k, under the constraint $\eta_k > 1, k = 1,\ldots,K$. The `optimize()` function in the `stats` package of the R software (R Core Team 2013) is used to perform numerical search of the maximum of the previous expression.

3. OPERATIONAL ASPECTS

3.1 NOTE ON ROBUSTNESS

Let us focus on the weights $(\tilde{v}_{itk} + \frac{1-\tilde{v}_{itk}}{\eta_k})$ in (4) and (5). We can rewrite $\tilde{v}_{itk}$ as an explicit function of the squared Mahalanobis distance, say δ, as

$$h(\delta;\alpha_k,\eta_k) = \frac{\alpha_k\exp\left(-\frac{\delta}{2}\right)}{\alpha_k\exp\left(-\frac{\delta}{2}\right)+\frac{(1-\alpha_k)}{\sqrt{\eta_k}}\exp\left(-\frac{\delta}{2\eta_k}\right)} = \frac{1}{1+\frac{(1-\alpha_k)}{\alpha_k}\frac{1}{\sqrt{\eta_k}}\exp\left[\frac{\delta}{2}\left(1-\frac{1}{\eta_k}\right)\right]},$$

with $\delta \geq 0$.

Due to $\eta_k > 1$, $h(\delta;\alpha_k,\eta_k)$ is a decreasing function of δ. Accordingly,

$$w(\delta;\alpha_k,\eta_k) = \left(\tilde{v}_{itk} + \frac{1-\tilde{v}_{itk}}{\eta_k}\right) = h(\delta;\alpha_k,\eta_k) + \frac{1-h(\delta;\alpha_k,\eta_k)}{\eta_k}$$
$$= \frac{1}{\eta_k}\left[1+(\eta_k-1)\,h(\delta;\alpha_k,\eta_k)\right],$$

which is an increasing function of $h(\delta;\alpha_k,\eta_k)$ and, thus, a decreasing function of δ. Therefore, $w(\delta;\alpha_k,\eta_k)$ reduces the effect of bad points in the estimation of $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$, $k = 1,\ldots,K$, so providing a robust way to estimate these parameters (see also Punzo and McNicholas 2016).

3.2 DETECTION OF BAD POINTS AND FURTHER CONSTRAINTS

For the proposed model, the classification of an observation $\boldsymbol{y}_{it}$ is a two-step procedure

Step 1. determine state membership via local or global decoding procedures (see Section A.2 in the supplementary material);

Step 2. establish if it is either a good or a bad observation in that state.

Once the hidden path is inferred, for each observation, we look at $\tilde{v}_{itk}$ for the inferred state and $\boldsymbol{y}_{it}$ is good if $\tilde{v}_{itk} > 0.5$ and bad otherwise.

Bearing in mind that $(1-\alpha_k)$ represents the proportion of bad points we would require that in the kth hidden state, $k = 1, \ldots, K$, the proportion of good data is at least equal to a fixed value α_k^*. In this case, the `optimize()` function is also used for a numerical search of the maximum $\alpha_k^{(r+1)}$, over the interval $(\alpha_k^*, 1)$, of the function

$$\sum_{i=1}^{I}\sum_{t=1}^{T}\tilde{z}_{itk}^{(r)}\left[\tilde{v}_{itk}^{(r)}\log\alpha_k + \left(1-\tilde{v}_{itk}^{(r)}\right)\log\left(1-\alpha_k\right)\right].$$

In both the simulation study and the empirical application, we use this approach to update α_k and we take $\alpha_k^* = 0.5$, $k = 1, \ldots, K$. The value 0.5 is justified because, in robust statistics, it is usually assumed that at least half of the points are good (see Hennig 2002, p. 250). Note that it is also possible to fix α_k and/or η_k a priori. This is somewhat analogous to the trimmed clustering approach, where one must to specify the proportion of outliers (the so-called trimming proportion) in advance (see Fritz, García-Escudero, and Mayo-Iscar 2012).

4. SIMULATION STUDIES

In this section, we investigate various aspects of the proposed model through large-scale simulation studies performed using R (R Core Team 2013).

4.1 COMPARISON BETWEEN HMMS OF ELLIPTICAL DISTRIBUTIONS

The first simulation study aims to demonstrate the effectiveness of the proposed model in comparison with HMMs of some elliptical distributions. A general feedback on advantages and drawbacks of each model is also given. We compare: the HMM of Gaussian distributions (NHMM); the HMM of t-distributions (tHMM); the HMM of contaminated Gaussian distributions (CNHMM). To generate the data, we consider the following five data-generation processes with bivariate ($P = 2$) state-specific distributions and $K = 2$ hidden states:

(a) NHMM.

(b) tHMM with $\nu_1 = 4$ and $\nu_2 = 10$ degrees of freedom.

(c) CNHMM with $\alpha_1 = 0.9$, $\alpha_2 = 0.8$, $\eta_1 = 2$, and $\eta_2 = 20$.

(d) NHMM with 1% of points randomly substituted by high atypical points with coordinates $\left(0, y_{it2}^*\right)$, where y_{it2}^* is generated from a uniform distribution over the interval (10, 15).

(e) NHMM with 5% of points randomly substituted by noise points generated from a uniform distribution over the interval $(-10, 10)$ on each dimension.

All of these data-generation processes share the following common parameters

$$\pi_1 = 0.3,\ \boldsymbol{\Pi} = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix},\ \boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ -3 \end{pmatrix},\ \boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1,$$

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix},\ \text{and } \boldsymbol{\Sigma}_2 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

The five scenarios above cover different situations, which may arise dealing with real-world data: no bad points for scenario (a), heavy-tails conditional distributions for scenarios (b) and (c), and two different types of bad points for scenarios (d) and (e). Under each scenario, we simulate 100 samples considering two experimental factors: the number of analyzed units I (50, 100, and 200) and the number of repeated measurements T (5, 10, and 20). This yields a total of 4500 generated datasets. On each generated dataset, the EM-based algorithm of the three competing models is directly run with $K = 2$, is initialized according to the partition provided by the K-means method, and is stopped when the difference between the updated parameter estimates of two consecutive iterations is less than 10^{-4}.

For comparison's sake, we report the bias (BIAS) and the standard deviation (STD) of the estimates for the initial weight π_1, the transition probabilities $\pi_{1|1}$ and $\pi_{2|2}$ (diagonal elements of $\boldsymbol{\Pi}$), the univariate means μ_{11} and μ_{21} (elements of $\boldsymbol{\mu}_1$), and the univariate means μ_{12} and μ_{22} (elements of $\boldsymbol{\mu}_2$). We would remark that HMMs, and mixture models in general, are affected by label switching issues (see, e.g., Yao 2012), which render estimators evaluation using simulations more complex. There are no generally accepted labeling methods. In our simulation study, because of true values $\pi_1 = 0.3$ and $\pi_2 = 0.7$, we simply attribute the label 1 to the state with the lowest estimated initial probability.

The obtained results are reported in Tables B1–B9 in the supplementary material. As concerns the estimates of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, we note the following general findings. Under scenario (a), that is, when there are no bad points, all the approaches perform comparably, as expected since, in this situation, the tHMM and the CNHMM tend to the NHMM. Under scenarios (b) and (c), the robust approaches, tHMM and CNHMM, are better than the traditional NHMM, especially when data are generated by the CNHMM (see, e.g., Table B7 in the supplementary material). Moreover, the fitted tHMM and CNHMM perform comparably in both scenarios; such a comparable behavior agrees with the simulation results of Little (1988) about the t and the contaminated Gaussian distributions. Even under scenarios (d) and (e), the tHMM and the CNHMM perform comparably and much better than the NHMM. In particular, under scenario (d), we note how the NHMM-estimates of the means μ_{21} and μ_{22} for the second dimension are mainly affected by the bad points; this is a natural result if we recall that these points are bad due to the value of the second dimension.

Referring to hidden parameters, as expected, under scenario (a) all the considered approaches perform well, providing unbiased estimates for both the transition probability matrix $\boldsymbol{\Pi}$ and the initial probabilities π_1 and π_2. Even under scenario (b), the three approaches performs well and almost comparably, with a slight worse performance for the NHMM. Under scenario (c), the traditional NHMM underestimates the first element $\pi_{1|1}$ on the main diagonal of the transition probability matrix of the hidden chain. In other words, this approach estimates a higher number of transitions, from state 1 to state 2, than the ones assumed by the model used to generate the data. Moreover, the initial probability

Table 1. Values of TPRs and FPRs; they refer to rates across 100 replications

	I	T	TPR	FPR		I	T	TPR	FPR
Scenario (d)	50	5	1.000	0.003	Scenario (e)	50	5	0.860	0.003
		10	1.000	0.002			10	0.841	0.003
		20	1.000	0.000			20	0.844	0.002
	100	5	1.000	0.001		100	5	0.838	0.002
		10	1.000	0.000			10	0.844	0.002
		20	1.000	0.000			20	0.835	0.002
	200	5	1.000	0.000		200	5	0.839	0.002
		10	1.000	0.000			10	0.845	0.002
		20	1.000	0.000			20	0.839	0.002

π_1 of state 1 is slightly underestimated too. Of course, these could represent issues if the underlying latent structure is of interest. On the contrary, under the same scenario, the tHMM and the CNHMM perform well, providing unbiased estimates for $\boldsymbol{\Pi}$, π_1, and π_2. The findings about scenario (d) are similar to those under scenario (b). Finally, scenario (e), that is, the noise case, is the most problematic for the traditional NHMM. In fact, being the noisy observations drawn at random from a Bernoulli distribution with parameter equal to 0.05, they may produce sudden changes in the latent structure, altering its dynamics. In particular, the initial probability π_1, as well as the first element $\pi_{1|1}$ on the main diagonal of the transition probability matrix of the hidden chain, are strongly underestimated. An a posteriori analysis, as well as the magnitude of the bias for π_1, reveals that the NHMM identifies a state (i.e., the persistent state) where all the good observations are clustered and another state (i.e., the nonpersistent state) where all the noisy observations are grouped. The resulting hidden structure is, thus, characterized by sudden changes toward the nonpersistent state followed by successive changes to the persistent state.

Table B10 in the supplementary material summarizes the obtained average misclassification rates. Misclassification rates are computed via the `classError()` function of the `mclust` package for R (Fraley et al. 2015). Note that, under scenarios (d) and (e), misclassification rates are computed only with respect to the true good observations. The results in Table B10 corroborate the previous simulation findings; in particular, the robust approaches have a similar behavior and they, apart from scenario (a), are better than the traditional NHMM, especially under scenarios (c) and (e).

Thus far, the tHMM and the CNHMM have shown a similar behavior; however, as previously emphasized, the CNHMM has the advantage to allow for the automatic detection of bad points. For the purpose of evaluation of this aspect, we report the true positive rate (TPR), measuring the proportion of bad points that are correctly identified as bad points, and the false positive rate (FPR), corresponding to the proportion of good points incorrectly classified as bad points.

Table 1 reports these measures for scenarios (d) and (e). We note almost optimal results under scenario (d) and for the FPRs under scenario (e). Furthermore, the fact that the TPRs do not approach at one under scenario (e) is not necessarily an error: the way the noisy points are inserted into the data makes possible that some of them will have values related to good points and, as such, these points will be detected as good points by our model.

Table 2. Average elapsed time (in seconds over 100 replications) to fit a CNHMM as a function of I and T

	T			
I	5	10	15	20
50	0.918	1.897	3.955	7.215
100	2.252	8.522	16.039	24.153
150	3.627	14.779	29.357	40.727
200	5.662	24.205	51.474	82.731

Finally, to have an idea of the computational burden required by our ECM algorithm, Table 2 shows the average elapsed time (in seconds over 100 replications) to fit a single CNHMM under scenario (c). Computation is performed on a Windows 8.1 PC, with Intel i7 3.50GHz CPU, 16.0 GB RAM, using R 32 bit, and the elapsed time is computed via the `proc.time()` function of the `base` package. To make the analysis finer, simulations with $I = 150$ and $T = 15$ have been added.

The minimum average elapsed time of 0.918 sec is obtained in correspondence of the pair $(I = 50, T = 5)$, while the maximum (82.731 sec) is obtained for the pair $(I = 200, T = 20)$. Furthermore, the average elapsed time seems to be a function of the overall size IT of the data, although the time-length of the panel affects the elapsed time slightly more than the sample size; this conjecture is corroborated by the plot in Figure 1.

4.2 SELECTING THE NUMBER OF HIDDEN STATES

The performance of the information criteria illustrated in Section A.3 in the supplementary material is here investigated for the CNHMM. To generate the data, we consider the following two CNHMMs with $P = 2$ dimensions:

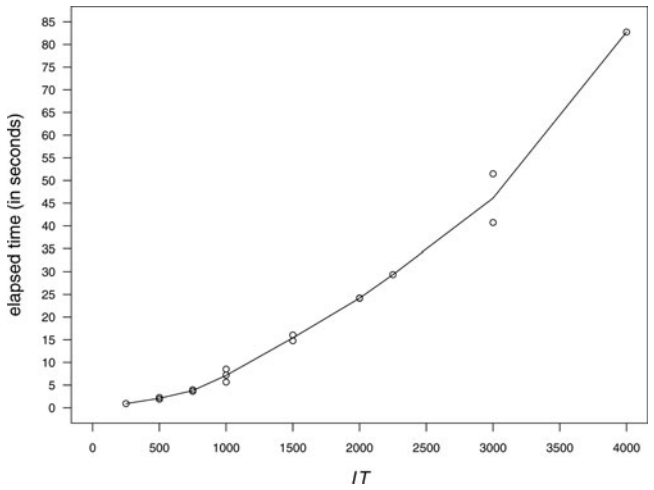


Figure 1. Average elapsed time (in seconds over 100 replications) to fit a CNHMM as a function of IT.

(f) the same two-state model considered under scenario (c);
(g) a three-state model with the following parameters

$$\pi_1 = 0.16, \quad \pi_2 = 0.34, \quad \boldsymbol{\Pi} = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}, \quad \boldsymbol{\mu}_1 = \begin{pmatrix} 0 \\ -8 \end{pmatrix}, \quad \boldsymbol{\mu}_2 = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

$$\boldsymbol{\mu}_3 = \begin{pmatrix} 0 \\ 8 \end{pmatrix}, \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & -0.5 \\ -0.5 & 1 \end{pmatrix}, \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}, \quad \boldsymbol{\Sigma}_3 = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix},$$

$$\alpha_1 = 0.9, \quad \alpha_2 = 0.8, \quad \alpha_3 = 0.9, \quad \eta_1 = 2, \quad \eta_2 = 5, \quad \text{and} \quad \eta_3 = 10.$$

Under each scenario, we simulate 500 samples considering two experimental factors: the number of analyzed units I (50 and 200) and the number of repeated measurements T (5 and 10). This yields a total of 4000 generated datasets. On each generated dataset, the ECM algorithm for the CNHMM is run for $K \in \{1, 2, 3, 4, 5\}$, is initialized according to the partition provided by the K-means method, and is stopped when the difference between the updated parameter estimates of two consecutive iterations is less than 10^{-4}.

Table B11 in the supplementary material summarizes the obtained results in terms of selection rate (over the 500 replications); the selection rate is defined here as the proportion of times each value of K is selected by the corresponding criterion shown on the top of the column. The rows related to the true value of K are highlighted in gray; to facilitate performance evaluation, the last row of Table B11 gives the mean selection rate of each criterion, computed over the true values of K (i.e., computed over the gray rows).

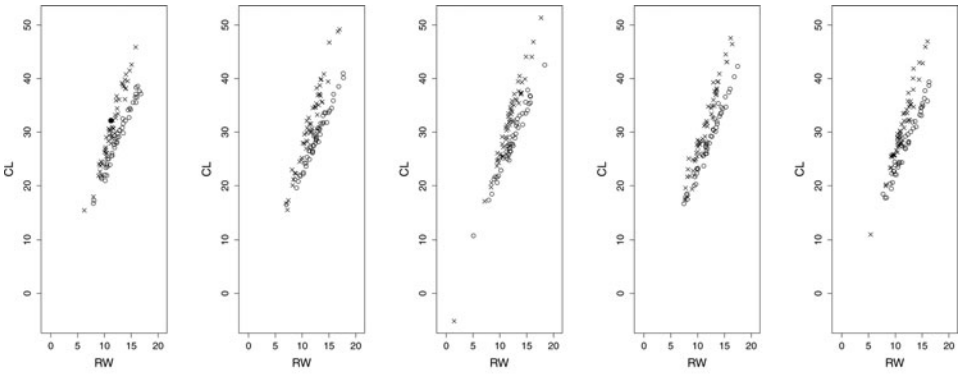
The Bayesian information criterion (BIC) and the integrated completed likelihood (ICL) perform comparably and much better than the Akaike information criterion (AIC) that, especially under scenario (g), tends to overestimate the number of states.

5. ILLUSTRATIVE EXAMPLES

5.1 ARTIFICIAL LONGITUDINAL BLUE CRAB DATA

This section is based on an artificial longitudinal version of the very popular crab dataset of Campbell and Mahon (1974). 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 $P = 2$ measurements (in millimeters), namely, the rear width (RW) and the length along the midline of the carapace (CL). Mardia's test suggests that it is reasonable to assume that the two group-conditional distributions are bivariate normal (see Greselin, Ingrassia, and Punzo 2011; Greselin and Punzo 2013; Bagnato, Greselin, and Punzo 2014 for details). The ML estimates of the parameters $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, $\boldsymbol{\Sigma}_1$, and $\boldsymbol{\Sigma}_2$ are given in Greselin, Ingrassia, and Punzo (2011, p. 158); based on these estimates, and further introducing a transition probabilities matrix

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



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

Figure 2. Scatterplots of the artificial data (× and ∘ denote state 1 and group 2, respectively; • denotes the observation perturbed for the analysis of Section 5.1).

we randomly generate a longitudinal version of this dataset on $T = 5$ times, based on the NHMM; the dataset is available at *http://www.economia.unict.it/punzo/Data.htm*. The scatterplots of the generated data, for each $t \in \{1, \ldots, 5\}$, are displayed in Figure 2.

In the fashion of Peel and McLachlan (2000), eight "perturbed" datasets are created by substituting the original value of CL for the 17th point at time 1 (highlighted by a bullet in Figure 2(a)) with eight atypical values shown in the first column of Table 3. We recall that, in the cross-section setting, the aim of Peel and McLachlan (2000) was to show that, unlike Gaussian mixtures, mixtures of t-distributions are robust to these perturbations when applied for clustering.

Ceteris paribus with Peel and McLachlan (2000), we directly fit the NHMM, the tHMM, and the CNHMM, with $K = 2$. For each of the three competing models, Table 3 reports the number of misallocated observations for each perturbed dataset.

It can be seen that, as expected, the tHMM and the CNHMM clusterings are more robust to these perturbations than the NHMM clustering. However, the CNHMM is systematically the most robust to these perturbations, with the number of misallocated observations remaining fixed at 3 regardless of the particular value perturbed; this is in contrast to the

Table 3. Number of misallocated artificial blue crabs ($I = 100$ and $T = 5$). The last two columns report the estimated value of η in the state containing the outlier and its posterior probability to be a good point, respectively

Value of CL	NHMM	tHMM	CNHMM	$\hat{\eta}$	$\hat{v}$
−15	13	5	3	220.966	0.005
−10	12	5	3	174.162	0.006
−5	12	5	3	133.014	0.008
0	11	5	3	97.475	0.010
5	8	5	3	67.539	0.015
10	7	5	3	43.231	0.023
15	7	5	3	24.608	0.041
20	5	5	3	11.729	0.085

NHMM where the number of misclassifications changes as the extent of the perturbation increases. Interestingly, the CNHMM always detects the perturbed value as a bad point regardless from its magnitude. Furthermore, by recalling that the original value of CL for the 17th point at time $t = 1$ was 32.158, it is also interesting to note that the estimated value of η_k (in the group containing the outlier, which is always the first state) increases as the value of this point further departs from its true value (refer to the fifth column of Table 3). In connection with this aspect, we also report the estimated posterior probability to be a good point (see Equation (3)) for the 17th observation at time 1 (refer to the sixth column of Table 3); as we can see, farther the perturbed value is from its state, lower is its probability to be a good point. Such a low probability is also related to the down-weighting of this bad point in the estimation of $\boldsymbol{\mu}_1$ and $\boldsymbol{\Sigma}_1$, and this is an important aspect for the robust estimation of these parameters (see Section 3.1).

Finally note that, in all the considered cases, the CNHMM detects a false positive bad point, which can be easily seen at the bottom-left corner of Figure 2(c).

5.2 CRIMINAL ACTIVITIES IN ITALIAN PROVINCES

In this section, we analyze data on criminal activities in Italy. Data are taken from an Italian financial newspaper (*Il Sole 24 Ore*, *www.ilsole24ore.com*), and have been previously analyzed by Viroli (2011). Italian crime has specific features. First, criminal patterns may vary across times and types of activity; second, *organized crime* has often territorial roots in specific Italian areas. Bearing this in mind, we would capture differences in (nonviolent) criminal activities across time, types of crimes, and territorial units, aiming at identifying different levels of safety conditions (represented by the hidden states).

Our analysis focuses on 103 NUTS3 (European Nomenclature of Territorial Units for Statistics) units in Italy, on which we recorded $P = 4$ criminal indicators: home-invasion robberies (per 100,000 residents; HOME); teenage crime rate (per 1,000 residents; TEEN); reported robberies (per 100,000 residents; ROB); rate of muggings and pickpockets (per 100,000 residents; PICK) over 5 years, from 2005 to 2009. Summaries of the evolution over time of these variables are reported in Figure C1 in the supplementary material. We observe an increase in home-invasion robberies over time, while all other indicators do not show, at first glance, any significant temporal variations. Moreover, it is clear, even from simple boxplots, that some units show *unusual values*.

In analyzing the dataset, the most interesting scientific question concerns the existence of areas with similar criminal rates in Italy. Also of interest is the strength of time dependence as measured by the transition probability matrix. Indeed, a strong time-dependence implies no improvements in hindering criminal activities. As described in previous sections, we jointly allow for time-varying clustering as well as for atypical data detections, which may affect the resulting clustering if not properly accounted for. The central idea is that the hidden states cope with the temporal and the spatial structure of the data, and that the contaminated Gaussian distributions can account for atypical data.

On these data, we fit the proposed model with a number of hidden states ranging from 1 to 10. For completeness, we also fit the NHMM and the *t*HMM. The results are reported in Table 4 in terms of AIC, BIC, and ICL. For each value of K, we adopt a K-means approach, with 20 random starting points, to initialize the ECM algorithm, and we report the results

Table 4. Model selection

		Number of hidden states (K)									
		1	2	3	4	5	6	7	8	9	10
AIC	CNHMM	−20495.69	−19649.58	−19093.36	−18737.51	−18584.11	−18443.71	−18341.15	−18298.57	−18222.24	−18237.17
	NHMM	−21507.87	−19768.81	−19106.63	−18792.78	−18597.50	−18471.70	−18402.28	−18322.74	−18251.39	−18173.76
	*t*HMM	−20442.14	−19647.40	−19088.17	−18750.09	−18595.55	−18498.36	−18375.11	−18274.80	−18200.13	−18129.92
BIC	CNHMM	−20537.84	−19741.79	−19240.90	−18945.65	−18858.12	−18788.86	−18762.71	−18801.81	−18812.42	−18919.56
	NHMM	−21544.75	−19850.49	−19238.37	−18979.85	−18845.17	−18785.23	−18786.95	−18783.81	−18794.01	−18803.46
	*t*HMM	−20481.67	−19734.35	−19227.81	−18947.70	−18856.39	−18827.70	−18778.22	−18756.95	−18766.60	−18785.97
ICL	CNHMM	−20537.84	−19743.67	−19247.93	−18949.27	−18866.21	−18794.69	−18768.71	−18811.05	−18819.57	−18925.40
	NHMM	−21544.75	−19853.99	−19244.35	−18986.99	−18856.12	−18791.33	−18798.01	−18795.69	−18797.59	−18811.22
	*t*HMM	−20481.67	−19735.65	−19232.32	−18953.05	−18864.92	−18833.48	−18785.28	−18770.45	−18775.57	−18793.86

corresponding to the best solution in terms of likelihood. On the basis of these results, we conclude that $K = 7$ is a suitable number of hidden states for the considered dataset. This value of K corresponds to the maximum value of both the BIC and ICL criteria, whereas the AIC selects nine states. However, in the simulation study of Section 4.2, we show that BIC and ICL perform well in recovering the *true* number of hidden states, whereas the AIC may overestimate this number.

On the basis of the estimates of the parameters under the selected CNHMM with $K = 7$ (see Table 5), we conclude that Italian provinces show territorial-specific characteristics and heterogenous criminal-related situations. As it is clear from the estimated state-specific mean vectors (Table 5) and the inferred clustering structure (Figure C2 in the supplementary material), States 1, 4, and 7 are characterized by similar home-invasion robberies rates. They differ in the other indicators. State 1 identifies high teen-crime and low robberies rates, whereas in State 7 home-invasions arise along with reported robberies and pickpockets. State 4, instead, is characterized by home robberies only. With few exceptions, these three states are observed in the most industrialized areas, for example, in North-North-West provinces. State 5 characterizes big cities (e.g., Rome, Turin, and Milan) and touristic places (e.g., Rimini). These are the provinces with the highest values of home robberies, teenage crimes, and reported muggings, and therefore the most dangerous ones in terms of the crime indicators considered in this analysis. Safer provinces are clustered in State 3, whereas unsafe southern provinces, which are notoriously and particularly unsafe in terms of robberies and muggings are clustered in State 6. At last, North-East and Center-North provinces with high rates of pickpockets are clustered in State 2. As we can note by the estimated transition probability matrix

$$\boldsymbol{\Pi} = \begin{pmatrix} 0.949 & 0.021 & 0.030 & 0.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.984 & 0.000 & 0.000 & 0.000 & 0.000 & 0.016 \\ 0.000 & 0.014 & 0.980 & 0.000 & 0.000 & 0.006 & 0.000 \\ 0.000 & 0.000 & 0.000 & 1.000 & 0.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 1.000 & 0.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.000 & 0.000 \\ 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 0.000 & 1.000 \end{pmatrix},$$

transitions between states are rare, persistence is the norm (as expected) and, as a consequence, the global and local decoding procedures provide the same inferred clustering structure. As also discussed by Bulla (2011), the strong persistence is a consequence of the robustification of the HMM. Indeed the NHMM with $K = 7$ has less persistent states. Furthermore, the few transitions are a clear indication that the political action to reduce these criminal activities has not achieved any improving results across the provinces. On the other hand, this very high persistence may indicate the absence of a time-varying clustering structure. The time-varying clustering structure is thus investigated. The data are fitted assuming an additional constraint on the selected model: the identity transition probability matrix is imposed to check for time-constant clustering. The AIC, BIC, and ICL values obtained under this model are -18486.29, -18797.19, and -18799.64, respectively; all of them are lower than the values reported in Table 4. Therefore, the time-dependence for the clustering structure is supported by the data.

Table 5. State-specific parameters

		Hidden states						
		1	2	3	4	5	6	7
μ	HOME	308.965	224.503	152.547	301.355	281.040	146.659	293.847
	TEEN	21.814	15.317	9.545	8.707	22.807	8.672	13.838
	ROB	25.145	27.223	22.736	39.350	94.426	81.336	47.521
	PICK	110.239	160.512	49.022	100.677	651.558	130.300	261.593
α		0.999	0.965	0.999	0.999	0.832	0.775	0.951
η		1.001	5.266	1.014	1.001	2.421	12.370	3.575

Other considerations arising from Table 5 are that atypical data can be detected under specific states only. Indeed, the probability of having good data in States 1, 3, and 4 is substantially 1, that is, all data are estimated as good points. Bad points identified by the CNHMM are displayed in Figure C3 in the supplementary material. Naples and Caserta are estimated as atypical in State 6, the one clustering most of Southern provinces, over all the time periods. In Viroli (2011), a state is devoted to cluster these two provinces only (a similar result is obtained from the *t*HMM, in which a further state is devoted to cluster Naples and Caserta only, that is, $K = 8$, and a fuzzier clustering is estimated). Few other provinces are also identified as possible bad points in other states. The η parameters provide the degree of atypicality of these bad points (see Section 5.1).

6. DISCUSSION

We have presented a new model for clustering multivariate longitudinal data in a hidden Markov framework, which is robust to outlying observations, spurious observations, or noise, which we collectively referred to as *bad points*. On real and simulated data, we have demonstrated that our model works better than the standard hidden Markov model based on the Gaussian distribution and comparably well with respect to the hidden Markov model based on the heavy-tails *t*-distribution. The main advantage of our model lies in automatic *bad points* detection that is performed by using a maximum a posteriori rule. In addition to these advantages, the choice of this approach is motivated by considerable conceptual and computational simplicity in the attempt to generalize the classical Gaussian HMM in terms of robustness and automatic detection of bad points; indeed, only minor modifications to the standard EM algorithm for the Gaussian HMM are involved (see Section 2.3). We also investigated information criteria behavior in this framework and observed good BIC and ICL performances in recovering the hidden structure.

There are different possibilities for further work, three of which are worth mentioning. First of all, our approach should be extended to large dimensions. This can be done for instance constraining the covariance matrices of the states, at the price of more complex estimation strategies. As often in the longitudinal setting, covariates information are also available along with multiple response variables. A straightforward extension would deal with the regression framework, in which contamination may arise in the covariate part of the regression model. Similarly, the homogenous assumption on the hidden Markov chain can be easily relaxed, allowing for time and/or individual-specific Markov chains, as well

as it is possible to allow for partially missing observations without too much effort, in a missing at random setting.

SUPPLEMENTARY MATERIALS

In the supplementary material we provide useful details for the implementation of the proposed approach and collect some tables related to the simulation studies discussed in Section 4. Finally, we summarize the data considered in the empirical application of Section 5.2 (Figure C1) and provide the inferred partition and evidence on outliers identification (see Figures C2–C3) according to the preferred CNHMM.

[Received January 2015. Revised May 2015.]

REFERENCES

Aitkin, M., and Wilson, G. T. (1980), "Mixture Models, Outliers, and the EM Algorithm," *Technometrics*, 22, 325–331. [1098]

Andrews, J. L., and McNicholas, P. D. (2012), "Model-Based Clustering, Classification, and Discriminant Analysis via Mixtures of Multivariate *t*-Distributions," *Statistics and Computing*, 22, 1021–1029. [1098]

Bagnato, L., Greselin, F., and Punzo, A. (2014), "On the Spectral Decomposition in Normal Discriminant Analysis," *Communications in Statistics—Simulation and Computation*, 43, 1471–1489. [1108]

Banfield, J. D., and Raftery, A. E. (1993), "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics*, 49, 803–821. [1098]

Bartolucci, F., and 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, 816–831. [1098]

——— (2010), "A Note on the Mixture Transition Distribution and Hidden Markov Models," *Journal of Time Series Analysis*, 31, 132–138. [1098]

Bartolucci, F., Farcomeni, A., and Pennoni, F. (2013), *Latent Markov Models for Longitudinal Data*, London: Taylor & Francis. [1098]

Baum, L. E., and Petrie, T. (1966), "Statistical Inference for Probabilistic Functions of Finite State Markov Chains," *The Annals of Mathematical Statistics*, 37, 1554–1563. [1098]

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970), "A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains," *The Annals of Mathematical Statistics*, 41, 164–171. [1099,1101]

Bernardi, M., Maruotti, A., and Petrella, L. (2014), "Multivariate Markov-Switching Models and Tail Risk Interdependence," arXiv:1312.6407v3. [1098]

Bulla, J. (2011), "Hidden Markov Models With *t* Components Increased persistence and other aspects," *Quantitative Finance*, 11, 459–475. [1098,1112]

Bulla, J., Lagona, F., Maruotti, A., and Picone, M. (2012), "A Multivariate Hidden Markov Model for the Identification of Sea Regimes From Incomplete Skewed and Circular Time Series," *Journal of Agricultural, Biological and Environmental Statistics*, 17, 544–567. [1098]

Campbell, N. A., and Mahon, R. J. (1974), "A Multivariate Study of Variation in Two Species of Rock Crab of Genus Leptograpsus," *Australian Journal of Zoology*, 22, 417–425. [1108]

Cerioli, A. (2010), "Multivariate Outlier Detection With High-Breakdown Estimators," *Journal of the American Statistical Association*, 105, 147–156. [1098]

Dannemann, J., Holzmann, H., and Leister, A. (2014), "Semiparametric Hidden Markov Models: Identifiability and Estimation," *Wiley Interdisciplinary Reviews: Computational Statistics*, 6, 418–425. [1100]

Davies, L., and Gather, U. (1993), "The Identification of Multiple Outliers," *Journal of the American Statistical Association*, 88, 782–792. [1098]

De la Cruz-Mesia, R., Quintana, F. A., and Marshall, G. (2008), "Model-Based Clustering for Longitudinal Data," *Computational Statistics & Data Analysis*, 52, 1441–1457. [1099]

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society,* Series B, 39, 1–38. [1101]

Evans, K., Love, T., and Thurston, S. (2015), "Outlier Identification in Model-Based Cluster Analysis," *Journal of Classification*, 32, 63–84. [1098]

Farcomeni, A., and Greco, L. (2015), "S-estimation of Hidden Markov Models," *Computational Statistics*, 30, 57–80. [1098]

Fraley, C., and Raftery, A. E. (2002), "Model-Based Clustering, Discriminant Analysis, and Density Estimation," *Journal of the American Statistical Association*, 97, 611–631. [1098]

Fraley, C., Raftery, A. E., Scrucca, L., Murphy, T. B., and Fop, M. (2015), *Mclust: Normal Mixture Modelling for Model-Based Clustering, Classification, and Density Estimation*, R package version 5.0.1. [1106]

Fritz, H., García-Escudero, L. A., and Mayo-Iscar, A. (2012, 5), "Tclust: An R Package for a Trimming Approach to Cluster Analysis," *Journal of Statistical Software*, 47, 1–26. [1104]

Garcia-Escudero, L. A., Gordaliza, A., Matran, C., and Mayo-Iscar, A. (2015), "Avoiding Spurious Local Maximizers in Mixture Modeling," *Statistics and Computing*, 25, 619–633. [1098]

Greselin, F., Ingrassia, S., and Punzo, A. (2011), "Assessing the Pattern of Covariance Matrices via an Augmentation Multiple Testing Procedure," *Statistical Methods & Applications*, 20, 141–170. [1108]

Greselin, F., and Punzo, A. (2013), "Closed Likelihood Ratio Testing Procedures to Assess Similarity of Covariance Matrices," *The American Statistician*, 67, 117–128. [1108]

Hennig, C. (2002), "Fixed Point Clusters for Linear Regression: Computation and Comparison," *Journal of Classification*, 19, 249–276. [1098,1104]

Holzmann, H., and Schwaiger, F. (2015), "Hidden Markov Models With State-Dependent Mixtures: Minimal Representation, Model Testing and Applications to Clustering," *Statistics and Computing*, 25, 1185–1200. doi: 10.1007/s11222-014-9481-1. [1098]

Humburg, P., Bulger, D., and Stone, G. (2008), "Parameter Estimation for Robust HMM Analysis of Chip-chip Data," *BMC Bioinformatics*, 9, 343. [1098]

Lagona, F., Maruotti, A., and 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. [1098]

Leroux, B. G. (1992), "Maximum-likelihood Estimation for Hidden Markov Models," *Stochastic Processes and Their Applications*, 40, 127–143. [1100,1101]

Little, R. J. A. (1988), "Robust Estimation of the Mean and Covariance Matrix From Data With Missing Values," *Applied Statistics*, 37, 23–38. [1105]

Maruotti, A. (2011), "Mixed Hidden Markov Models for Longitudinal Data: An Overview," *International Statistical Review*, 79, 427–454. [1098]

——— (2014), "Robust Fitting of Hidden Markov Regression Models Under a Longitudinal Setting," *Journal of Statistical Computation and Simulation*, 84, 1728–1747. [1098]

Maruotti, A., and Rocci, R. (2012), "A Mixed Non-homogeneous Hidden Markov Model for Categorical Data, With Application to Alcohol Consumption," *Statistics in Medicine*, 31, 871–886. [1100]

McNicholas, P. D., and Murphy, T. B. (2010), "Model-Based Clustering of Longitudinal Data," *The Canadian Journal of Statistics*, 38, 153–168. [1099]

Meng, X.-L., and Rubin, D. B. (1993), "Maximum Likelihood Estimation via the ECM Algorithm: A General Framework," *Biometrika*, 80, 267–278. [1101]

Peel, D., and McLachlan, G. J. (2000), "Robust Mixture Modelling Using the t Distribution," *Statistics and Computing*, 10, 339–348. [1098,1109]

Punzo, A., and McNicholas, P. D. (2014a), "Robust Clustering in Regression Analysis via the Contaminated Gaussian Cluster-Weighted Model," available at *http://arxiv.org/abs/1409.6019*. [1099]

——— (2014b), "Robust High-Dimensional Modeling With the Contaminated Gaussian Distribution," available at *http://arxiv.org/abs/1408.2128*. [1099]

——— (2016), "Parsimonious Mixtures of Contaminated Gaussian Distributions With Application to Allometric Studies," *Biometrical Journal*. doi: 10.1002/bimj.201500144. [1099,1100,1103]

R Core Team (2013), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. [1103,1104]

Rousseeuw, P. J., and Leroy, A. M. (2005), *Robust Regression and Outlier Detection*, New York: Wiley. [1098]

Tukey, J. W. (1960), "A Survey of Sampling From Contaminated Distributions," in *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling, Stanford Studies in Mathematics and Statistics*, ed. I. Olkin, Standford, CA: Stanford University Press, pp. 448–485. [1098]

Viroli, C. (2011), "Model Based Clustering for Three-Way Data Structures," *Bayesian Analysis*, 6, 573–602. [1099,1110,1113]

Volant, S., Berard, C., Martin-Magniette, M. L., and Robin, S. (2014), "Hidden Markov Models With Mixtures as Emission Distributions," *Statistics and Computing*, 24, 493–504. [1098]

Yao, W. (2012), "Model Based Labeling for Mixture Models," *Statistics and Computing*, 22, 337–347. [1105]

Zucchini, W., and MacDonald, I. L. (2009), *Hidden Markov Models for Time Series: An Introduction Using R*, Boca Raton, FL: Chapman & Hall. [1101]