Mixture of hidden Markov models for accelerometer data

Marie Du Roy de Chaumaray, Matthieu Marbac, Fabien Navarro

Mixture of hidden Markov models for accelerometer data

Marie du Roy de Chaumaray*, Matthieu Marbac†, Fabien Navarro‡

June 18, 2019

Abstract

This work is motivated by the analysis of accelerometer data. The analysis of such data consists in extracting statistics which characterize the physical activity of a subject (*e.g.,* the mean time spent at different activity levels and the probability of the transition between two levels). Therefore, we introduce a finite mixture model of hidden Markov chain to analyze accelerometer data by considering heterogeneity into the population. This approach does not specify activity levels in advance but estimates them from the data. In addition, it allows for the heterogeneity of the population to be taken into account and defines subpopulations having a homogeneous behavior regarding the physical activity. The main theoretical result is that, under mild assumptions, the probability of misclassifying an observation decreases at an exponential rate with its length. Moreover, we prove the model identifiability and we show how the model can handle missing values. Our proposition is illustrated using real data.

Keywords: Accelerometer data; Hidden Markov model; Longitudinal model; Missing data; Mixture models

1 Introduction

Inadequate sleep and physical inactivity affect physical and mental well-being while often exacerbating health problems. They are currently considered major risk factors for several health conditions (see for instance Kimm et al. (2005); Taheri et al. (2004); Lee et al. (2012); Grandner et al. (2013); McTiernan (2008)). Therefore, appropriate assessment of activity and sleep periods is essential in disciplines such as medicine and epidemiology. The use of accelerometers to evaluate physical activity—by measuring the acceleration of the part of the body to which they are attached—is a classic method that has become widespread in public health research. Indeed, the analysis of actigraphy data has been the subject of extensive studies over the past three decades.

Pioneer approaches have focused on automatic detection of the sleep and wake-up periods (Cole et al., 1992; Sadeh et al., 1994; Pollak et al., 2001; Van Hees et al., 2015). More recent developments are interested in the classification of different levels of activity (see Yang and Hsu (2010) for a review). These methods provide summary statistics like the mean time spent at different activity levels. In epidemiologic studies, the times spent by activity levels are often used as covariates in predictive models (see for instance the works of Noel et al. (2010); Palta et al. (2015); Innerd et al. (2018) where the links between physical activity and obesity are investigated). These statistics can be computed using deterministic cutoff levels (Freedson et al., 1998). However, with such an approach, the dependency in time is neglected and the cutoff levels are pre-specified and not estimated from the data.

Accelerometer data are characterized by a time dependency between the different measures. Thus, they can be analyzed by methods developed for functional data or by Hidden Markov Models (HMM).

*Marie du Roy de Chaumaray
CREST, ENSAI, France, E-mail: marie.du-roy-de-chaumaray@ensai.fr
†Matthieu Marbac
CREST, ENSAI, France, E-mail: matthieu.marbac-lourdelle@ensai.fr
‡Fabien Navarro
CREST, ENSAI, France, E-mail: fabien.navarro@ensai.fr

Methods for functional data need the observed data to be converted into functional data of time (Morris et al., 2006; Xiao et al., 2014; Gruen et al., 2017). For instance, Morris et al. (2006) use wavelet-basis for analyzing accelerometer profiles. The use of a function basis reduces the dimension of the data, and therefore the computing time. However, these methods do not define levels of activity and thus cannot directly provide the time spent at different activity levels.

If a discrete latent variable is considered to model time dependence, HMMs are appropriate for adjusting sequence data (Scott et al., 2005; Altman, 2007; Gassiat et al., 2016). Titsias et al. (2016) expand the amount of information which can be obtained from HMM including a procedure to find maximum *a posteriori* (MAP) of the latent sequences and to compute posterior probabilities of the latent states. Thus, HMM are used on activity data for monitoring circadian rythmicity (Huang et al., 2018b) or directly for estimating the sequence of activity levels from accelerometer data (Witowski et al., 2014). The approach of Witowski et al. (2014) assumes the homogeneity of the population and does not consider observations with missing values. However, heterogeneity in physical activity behaviors is often present (see for instance Geraci (2018)). In the following, we consider a population to be homogeneous if the average time spent per activity level and the probabilities of the transition between levels are similar for each individual of this population.

Clustering enables the heterogeneity of the population to be addressed by grouping observations into a few homogeneous classes. Finite mixture models (McLachlan and Peel, 2000; McNicholas, 2016) allow to cluster different types of data like: continuous (Banfield and Raftery, 1993), integer (Karlis and Meligkotsidou, 2007), categorical (Goodman, 1974), mixed (Hunt and Jorgensen, 2011; Kosmidis and Karlis, 2015), network (Hoff et al., 2002; Hunter et al., 2008; Matias et al., 2018) and sequence data (Wong and Li, 2000). Thus, recent methods use clustering for accelerometer data analysis. For instance, Wallace et al. (2018) use a specific finite mixture to identify novel sleep phenotypes, Huang et al. (2018a) perform a matrix-variate-based clustering on accelerometer data while Lim et al. (2019) use a functional data clustering.

In practice, the data collected may include missing intervals due to non-compliance by participants (*e.g.,* if the accelerometer is removed), making statistical analysis more challenging. Geraci and Farcomeni (2016) propose to identify different profiles of physical activity behaviors using a principal component analysis which allows missing values. Moreover, when the acceleration is measured each second, then many observations are zero. Thus, the use of zero-inflated distribution is quite common for modeling accelerometer data (Ae Lee and Gill, 2018; Bai et al., 2018).

In this paper, we present a finite mixture of HMM for analyzing accelerometer data. This model considers two latent variables: a categorical variables indicating the class membership of each individual and a sequence of categorical variable indicating the activity level of the individual at each time where its acceleration is measured. At time t, the acceleration is independent of the class membership conditionally on the state and follows a zero-inflated distribution. Thus, the definition of the activity levels are equal among the mixture components. This is a crucial point for using the summary statistics (*e.g.,* time spent at different activity levels, probabilities of transition between levels) in a future statistical study. Model identifiability is proved. Moreover, we show that, under mild assumptions, the probability of misclassifying an observation decreases at an exponential rate. Finally, we present a computationally efficient approach for dealing with missing values. This approach avoids the computation of large powers of the transition matrices in the algorithm used for parameter inference and thus reduces computation time.

This paper is organized as follows. Section 2 introduces the mixture of HMM. Section 3 presents the model properties (model identifiability, exponential decreasing of the probabilities of misclassification and a result for dealing with the non-wear periods). Section 4 discusses the inference and Section 5 illustrates the model properties on both simulated and real data. Section 6 illustrates the approach by analyzing the accelerometer data collected from 133 individuals by the NYC Department of Health and Mental Hygiene. Section 7 discusses some future developments. Proofs and technical lemmas are postponed in Appendix.

2 Mixture of hidden Markov models

In this section, we present the proposed model and the application context for which it has been defined.

2.1 The model

Observed data $\boldsymbol{y} = (\boldsymbol{y}_1^\top, \dots, \boldsymbol{y}_n^\top)$ are composed of n i.i.d sequences $\boldsymbol{y}_i = (y_{i(0)}, \dots, y_{i(T)})$. The sequence $\boldsymbol{y}_i$ corresponds to the measures of the accelerometer data at times $t \in \{0, 1, \dots, T\}$ for observation i, with $y_{i(t)} \in \mathbb{R}^+$. The diversity between the n observations can be considered by grouping the observations into K homogeneous classes. This is achieved by clustering that assesses a partition $\boldsymbol{z} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_n)$ among the observations $\boldsymbol{y}$, where $\boldsymbol{z}_i = (z_{i1}, \dots, z_{iK})$, $z_{ik} = 1$ if observation i belongs to class k and $z_{ik} = 0$ otherwise. Thus, each sequence $\boldsymbol{y}_i$ is assumed to independently arise from a mixture of K parametric distributions defined by the probability distribution function (pdf)

$$p(\boldsymbol{y}_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \delta_k \, p(\boldsymbol{y}_i; \boldsymbol{\pi}_k, \boldsymbol{A}_k, \boldsymbol{\lambda}, \boldsymbol{\varepsilon}), \tag{1}$$

where $\boldsymbol{\theta} = \{\boldsymbol{\lambda}, \boldsymbol{\varepsilon}\} \cup \{\delta_k, \boldsymbol{\pi}_k, \boldsymbol{A}_k; k = 1, \dots, K\}$ groups the model parameters, $\delta_k = \mathbb{P}(Z_{ik} = 1)$ is the proportion of components k with $\delta_k > 0$ and $\sum_{k=1}^{K} \delta_k = 1$, and $p(\cdot; \boldsymbol{\pi}_k, \boldsymbol{A}_k, \boldsymbol{\lambda}, \boldsymbol{\varepsilon})$ is the pdf of component k parametrized (*i.e.,* the pdf of $\boldsymbol{y}_i$ given $Z_{ik} = 1$) by $(\boldsymbol{\pi}_k, \boldsymbol{A}_k, \boldsymbol{\lambda}, \boldsymbol{\varepsilon})$ defined below.

Under component k, $\boldsymbol{y}_i$ follows a hidden Markov model where the hidden state sequence $\boldsymbol{x}_i = (\boldsymbol{x}_{i(0)}, \dots, \boldsymbol{x}_{i(T)}) \in \mathcal{X}$ takes M values for each observation $\boldsymbol{x}_{i(t)} = (x_{i(t)1}, \dots, x_{i(t)M})$ where $x_{i(t)h} = 1$ if observation i is at state h at time t and $x_{i(t)h} = 0$ otherwise. The model assumes that the distribution of the hidden state sequence depends on the cluster membership, and that $y_{i(t)}$ is drawn from a specific parametric distribution whose parameters depend on the state at time t but not on the component membership (*i.e.* $\boldsymbol{X}_i \not\perp \boldsymbol{Z}_i$ and $Y_{i(t)} \perp \boldsymbol{Z}_i \mid \boldsymbol{X}_{i(t)}$). The latter assumption is a crucial point. Indeed, one objective is to have summary statistics of $\boldsymbol{y}_i$ like the mean time spent at different intensity levels. In this model, each intensity level is defined by the distribution of $y_{i(t)}$ given a latent state. Therefore, it would be not useful to have different definitions of the intensity levels according to the cluster membership. The pdf of components k is

$$p(\boldsymbol{y}_i; \boldsymbol{\pi}_k, \boldsymbol{A}_k, \boldsymbol{\lambda}, \boldsymbol{\varepsilon}) = \sum_{\boldsymbol{x}_i \in \mathcal{X}} p(\boldsymbol{x}_i; \boldsymbol{\pi}_k, \boldsymbol{A}_k) \, p(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\lambda}, \boldsymbol{\varepsilon}). \tag{2}$$

The Markov assumption implies that

$$p(\boldsymbol{x}_i; \boldsymbol{\pi}_k, \boldsymbol{A}_k) = \prod_{h=1}^{\ell} \pi_{kh}^{x_{i(0)h}} \prod_{t=1}^{T} \prod_{h=1}^{M} \prod_{\ell=1}^{M} (\boldsymbol{A}_k[h, \ell])^{x_{i(t-1)h} x_{i(t)\ell}},$$

where $\boldsymbol{\pi}_k = (\pi_{k1}, \dots, \pi_{kM})$ defines the initial probabilities so that $\pi_{kh} = \mathbb{P}(X_{i(1)h} = 1 \mid Z_{ik} = 1)$ and $\boldsymbol{A}_k$ is the transition matrix so that $\boldsymbol{A}_k[h, \ell] = \mathbb{P}(X_{i(t)\ell} = 1 \mid X_{i(t-1)h} = 1)$. Finally, we have

$$p(\boldsymbol{y}_i \mid \boldsymbol{x}_i; \boldsymbol{\lambda}, \boldsymbol{\varepsilon}) = \prod_{t=0}^{T} \prod_{h=1}^{M} g(y_{i(t)}; \boldsymbol{\lambda}_h, \varepsilon_h)^{x_{i(t)h}},$$

where $g(\cdot; \boldsymbol{\lambda}_h, \varepsilon_h)$ is the pdf of a zero-inflated distribution defined by

$$g(y_{i(t)}; \boldsymbol{\lambda}_h, \varepsilon_h) = (1 - \varepsilon_h) g_c(y_{i(t)}; \boldsymbol{\lambda}_h) + \varepsilon_h \mathbf{1}_{\{y_{i(t)} = 0\}},$$

where $g_c(\cdot; \boldsymbol{\lambda}_h)$ is the density of a distribution defined on a positive space and parametrized by $\boldsymbol{\lambda}_h$. The choice of considering zero-inflated distributions arises from the application. Indeed, as the time lapse between two measures of the accelerometer is small, many $y_{i(t)}$ are zero. Different distributions can be chosen for $g_c(\cdot; \boldsymbol{\lambda}_h)$. Model properties and inference are discussed for any $g_c(\cdot; \boldsymbol{\lambda}_h)$. Because the model properties are obtained for a general distribution g_c, the discussion of its choice is postponed in Section 6.

2.2 Analysis of accelerometer data

Figure 1 presents an example of accelerometer data measured on one subject during a week. We observe that missingness occurs (the subject removes the accelerometer when he sleeps) and that missing values appear by sequence. The proposed model is designed to manage missing data and activity levels are not specified in advance. It provides an estimate of the latent state at each time and the probability of each state at each time (see the following sections for more details).

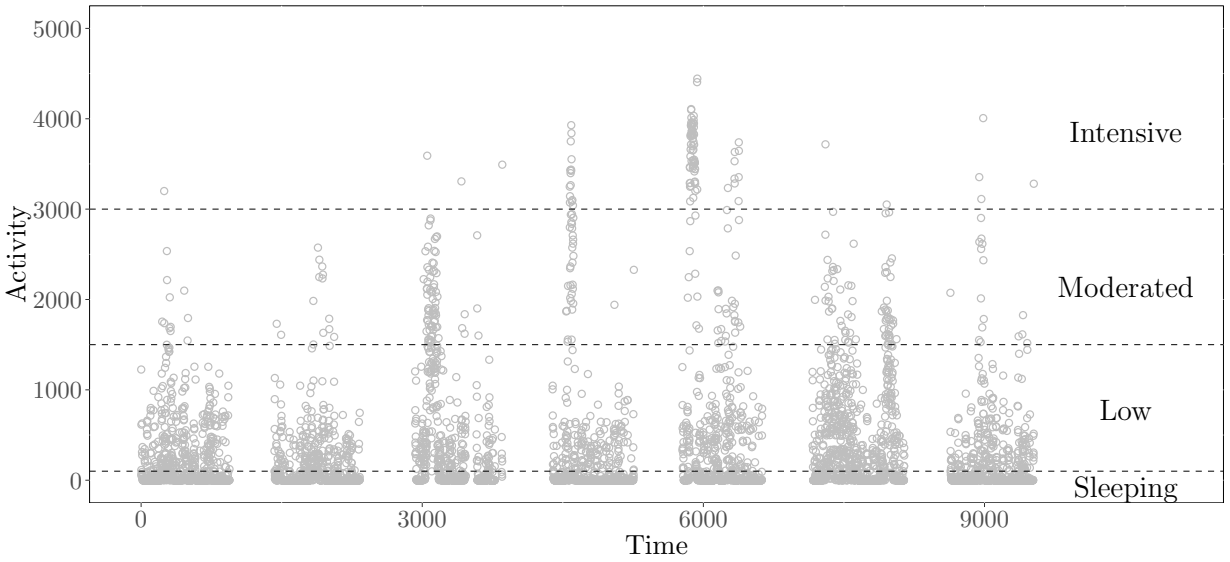


Figure 1: Example of accelerometer data.

3 Model properties

In this section, we present the properties of the mixture of parametric HMM. It starts with a discussion of three assumptions. Then, model identifiability is proved. It is shown that the probability of making an error in the partition estimation exponentially decreases with T, when the model parameters are known. Finally, the analysis of missing data is discussed.

3.1 Assumptions

Assumption 1. *For each component k, the Markov chain is irreducible. Moreover, we assume that the sequence is observed at its stationary distribution (*i.e., *$\boldsymbol{\pi}_k$ is the stationary distribution so $\boldsymbol{\pi}_k^\top \boldsymbol{A}_k = \boldsymbol{\pi}_k^\top$). Therefore, there exists $0 \leq \nu < 1$ such that*

$$\forall k \in \{1, \ldots, K\},\ \nu_2(\boldsymbol{A}_k) \leq \nu,$$

where $\nu_2(\boldsymbol{A}_k)$ is the second-largest eigenvalue of $\boldsymbol{A}_k$. Finally, we denote by $\bar{\nu}_2(\boldsymbol{A}_k) = \max(0, \nu_2(\boldsymbol{A}_k))$.

Assumption 2. *The hidden states define different distributions for the observed sequence. Therefore, for $h \in \{1, \ldots, M\}$, $h' \in \{1, \ldots, M\} \setminus \{h\}$, we have $\boldsymbol{\lambda}_h \neq \boldsymbol{\lambda}_{h'}$. Moreover, the parametric family of distributions defining $g_c(\cdot; \boldsymbol{\lambda}_1), \ldots, g_c(\cdot; \boldsymbol{\lambda}_M)$ permits to consider an ordering such that for a fix value $\rho \in \mathbb{R}^+ \setminus \{0\}$, we have*

$$\forall h \in \{1, \ldots, M-1\},\ \lim_{y_{i(1)} \to \rho} \frac{g_c(y_{i(1)}; \boldsymbol{\lambda}_{h+1})}{g_c(y_{i(1)}; \boldsymbol{\lambda}_h)} = 0.$$

Assumption 3. *The transition probabilities are different over the mixture components and are not zero. Therefore, for $k \in \{1, \ldots, K\}$, $k' \in \{1, \ldots, K\} \setminus \{k\}$, we have $\forall (h, \ell)$, $\boldsymbol{A}_k[h, \ell] \neq \boldsymbol{A}_{k'}[h, \ell]$.*

Moreover, there exists $\zeta > 0$ *such that*

$$\forall k \in \{1, \dots, K\},\ \forall k' \in \{1, \dots, K\} \setminus \{k\},\ \sum_{h=1}^{M} \sum_{\ell=1}^{M} \pi_{kh} \log \frac{\boldsymbol{A}_k[h, \ell]}{\boldsymbol{A}_{k'}[h, \ell]} > \zeta.$$

Finally, without loss of generality, we assume that $A_k[1, 1] > A_{k+1}[1, 1]$.

Assumption 1 considers that the state at time 1 is drawn from the stationary distribution of the component that the observation belongs to. To obtain the model identifiability we do not need the assumption that the stationary distribution is different over the mixture components. As a result, two components having the same stationary distribution but different transition matrices can be considered. Assumption 2 and Assumption 3 are required to obtain the model identifiability. Assumption 3 can be interpreted as the Kullback-Leibler divergence between the distribution of the states under component k and their distribution under component k'. This constraint is required for model identifiability because it is related to the definition of the classes. Consequently, the matrices of the transition probability must be different among components.

3.2 Identifiability

Model identifiability is crucial for interpreting the estimators of the latent variables and of the parameters. It has been studied for some mixture models (Teicher, 1963, 1967; Allman et al., 2009; Celisse et al., 2012) and HMM (Gassiat et al., 2016), but not for the mixture of HMM. Generic identifiability (up to switching of the components and of the states) of the model defined in (1) implies that

$$\forall \boldsymbol{y}_i,\ p(\boldsymbol{y}_i; \boldsymbol{\theta}) = p(\boldsymbol{y}_i; \tilde{\boldsymbol{\theta}}) \Rightarrow \boldsymbol{\theta} = \tilde{\boldsymbol{\theta}}.$$

The following theorem states this property.

Theorem 1. *If Assumptions 1, 2 and 3 hold, then the model defined in* (1) *is generically identifiable (up to switching of the components and of the states) if* $T > 2K$.

Proof of Theorem 1 is given in Appendix A. The model defined by the marginal distribution of a single $y_{i(t)}$ is not identifiable. Indeed, the marginal distribution of $y_{i(t)}$ is a mixture of zero-inflated distributions and such mixture is not identifiable (*i.e.*, different class proportions and inflation proportions can define the same distribution). It is therefore this dependency over time that makes the proposed mixture generically identifiable. Note that such statement has been made by Gassiat et al. (2016) when they discuss the case where the emission distribution for an HMM follows a mixture model.

3.3 Probabilities of misclassification

In this section, we examine the probability that an observation will be misclassified when the model parameters are known. Thus, we consider the ratio between the probability that observation i belongs to class k given $\boldsymbol{y}_i$ and the probability that this observation belongs to the true class, and we quantify the probability of it being greater than some positive constant a. Let θ_0 be the true model parameter and $\mathbb{P}_0 = \mathbb{P}(\cdot \mid Z_{ik_0} = 1, \theta_0)$ denote the true conditional distribution (true label of observation i and parameters are known).

Theorem 2. *Assume that Assumptions 1 and 3 hold. If* $a > 0$ *is such that Assumption 4 (defined in Appendix B) holds, then for every* $k \neq k_0$

$$\mathbb{P}_0 \left[\frac{\mathbb{P}(Z_{ik} = 1 \mid \boldsymbol{y}_i)}{\mathbb{P}(Z_{ik_0} = 1 \mid \boldsymbol{y}_i)} > a \right] \leq \mathcal{O}(e^{-cT}),$$

where $c > 0$ *is a positive constant*

Therefore, by considering $a = 1$, Theorem 2 shows that the probability of misclassifying an observation $\boldsymbol{y}_i$, using the *maximum a posteriori* rule, tends to zero when T increases, if the model parameters are known. Proof of Theorem 2 and a sufficient condition that allows to consider $a = 1$ (value of interest when the partition is given by the MAP rule) are given in Appendix B. It should be noted that it is not so common to have an exponential rate of convergence for the ratio of the posterior probability of classification. Similar results are obtained for network clustering using the stochastic block model (Celisse et al., 2012) or for co-clustering (Brault and Mariadassou, 2015). For these two models, the marginal distribution of a single variable provides information about the class membership. For the proposed model, this is the dependency between the different observed variables which is the crucial point for recovering the true class membership.

3.4 Dealing with missing values

Due to the markovian character of the states, missing values can be handled by iterating the transition matrices. In our particular context, missing values appear when the accelerometer is not worn. Thus, we will not observe isolated missing values but rather wide ranges of missing values. Let d be the number of successive missing values, we thus have to compute the matrix A_k^{d+1} to obtain the distribution of the state at time $t+d$ knowing the state at time $t-1$. These powers of transition matrices should be computed many times during the algorithm used for inference (see Section 4). Moreover, after $d+1$ iterations with d large enough, the transition matrix can be considered sufficiently close to stationarity (*e.g.*, for any (h, ℓ), $A_k^{d+1}[h, \ell] \simeq \pi_{k\ell}$), which has actually been chosen as the initial distribution. Therefore, for numerical reasons, we will avoid computing the powers of the transition matrices and we will make the following approximation. An observation $\boldsymbol{y}_i$ with S_i observed sequences split with missing value sequences of size at least d are modeled as S_i independent observed sequences with no missing values, all belonging to the same component k. Namely, for each individual i, the pdf $p(\boldsymbol{y}_i; \boldsymbol{\pi}_k, \boldsymbol{A}_k, \boldsymbol{\lambda}, \boldsymbol{\varepsilon})$ of component k is approximated by the product of the pdf of the S_i observed sequences $\boldsymbol{y}_{i1}, \boldsymbol{y}_{i2}, \ldots, \boldsymbol{y}_{iS_i}$:

$$p(\boldsymbol{y}_i; \boldsymbol{\pi}_k, \boldsymbol{A}_k, \boldsymbol{\lambda}, \boldsymbol{\varepsilon}) \simeq \prod_{s=1}^{S_i} p(\boldsymbol{y}_{is}; \boldsymbol{\pi}_k, \boldsymbol{A}_k, \boldsymbol{\lambda}, \boldsymbol{\varepsilon}),$$

where, for each s, $\boldsymbol{y}_{is}$ is an observed sequence of length $T_{is}+1$: $\boldsymbol{y}_{is} = (y_{is(0)}, \ldots, y_{is(T_{is})})$ and $p(\boldsymbol{y}_{is}; \boldsymbol{\pi}_k, \boldsymbol{A}_k, \boldsymbol{\lambda}, \boldsymbol{\varepsilon})$ is defined as in (2). We note that the observation $\boldsymbol{y}_i$ can thus be rewritten as follows

$$\boldsymbol{y}_i = (y_{i1(0)}, \ldots, y_{i1(T_{i1})}, y_{i2(0)}, \ldots, y_{i2(T_{i2})}, \ldots, y_{iS_i(0)}, \ldots, y_{iS_i(T_{iS_i})}),$$

with $y_{i2(0)} = y_{i(T_{i1}+d_{i1}+1)}$ where the d_{i1} values $y_{i(T_{i1}+1)}, \ldots y_{i(T_{i1}+d_{i1})}$ correspond to the first sequence of missing values, and more generally, for each $s = 2, \ldots, S_i$, $y_{is(0)} = y_{i(\sum_{j=1}^{s-1}(T_{ij}+d_{ij}+1))}$, with d_{ij} being the number of missing values between the observed sequences $\boldsymbol{y}_{is_j}$ and $\boldsymbol{y}_{is_{j+1}}$.

Once the estimation of the parameters has been done, we make sure that this assumption was justified by verifying that the width of the smallest range $d_{min} = \min\{d_{i1}, \ldots, d_{i\,S_i-1}\}$ of missing values is sufficiently large to be greater than the mixing time of the obtained transition matrix. To do so, we use an upper bound for the mixing time given by Levin and Peres (2017, Theorem 12.4, p. 155). For each component k, we denote by ν_k^* the second maximal absolute eigenvalue of $\boldsymbol{A}_k$. For any positive η, if for each k

$$d_{min} \geq \frac{1}{1-\nu_k^*} \log \frac{1}{\eta \min_h \pi_{kh}},$$

then for any integer $D \geq d_{min}$, the maximum distance in total variation satisfies

$$\max_h \|A_k^D[h, \cdot] - \pi_k\|_{TV} \leq \eta.$$

4 Maximum likelihood inference

This section presents the methodology used to estimate the model parameters.

4.1 Inference

We proposed to estimate the model parameters by maximizing the log-likelihood function where missing values are managed as in Section 3.4 and we recall that the log-likelihood is also approximated for numerical reasons, to avoid computing large powers of the transition matrices. Therefore, we want to find $\hat{\boldsymbol{\theta}}$ which maximizes the following approximated log-likelihood function

$$\ell_K(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \delta_k \prod_{s=1}^{S_i} p(\boldsymbol{y}_{is}; \boldsymbol{\pi}_k, \boldsymbol{A}_k, \boldsymbol{\lambda}, \varepsilon) \right).$$

This maximization is achieved via an EM algorithm (Dempster, A.P. and Laird, N.M. and Rubin, D.B., 1977) which considers the complete-data log-likelihood defined by

$$\ell_K(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log \delta_k + \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left(\sum_{s=1}^{S_i} \log p(\boldsymbol{y}_{is}; \boldsymbol{\pi}_k, \boldsymbol{A}_k, \boldsymbol{\lambda}, \varepsilon) \right).$$

4.2 Conditional probabilities

Let $\alpha_{ikhs(t)}(\boldsymbol{\theta})$ be the probability of the partial sequence $y_{is(0)}, \ldots, y_{is(t)}$ and ending up in state h at time t under component k. Moreover, let $\beta_{ikhs(t)}(\boldsymbol{\theta})$ be the probability of the ending partial sequence $y_{is(t+1)}, \ldots, y_{is(T_{is})}$ given a start in state h at time t under component k. These probabilities can be easily obtained by the forward/backward algorithm (see Appendix C). We deduce that the probability $\gamma_{ikhs(t)}(\boldsymbol{\theta})$ of being in state h at time $t \in \{0, \ldots, T_{is}\}$ for $\boldsymbol{y}_i$ under component k is

$$\gamma_{ikhs(t)}(\boldsymbol{\theta}) = \mathbb{P}(X_{is(t)} = h \mid \boldsymbol{y}_{is}, Z_{ik} = 1; \boldsymbol{\theta}) = \frac{\alpha_{ikhs(t)}(\boldsymbol{\theta}) \beta_{ikhs(t)}(\boldsymbol{\theta})}{\sum_{\ell=1}^{M} \alpha_{ik\ell s(t)}(\boldsymbol{\theta}) \beta_{ik\ell s(t)}(\boldsymbol{\theta})}.$$

The probability $\xi_{ikh\ell s(t)}(\boldsymbol{\theta})$ of being in state ℓ at time $t \in \Omega_i$ and in state h at time $t-1$ for observation $\boldsymbol{y}_i$ under component k is

$$\begin{aligned} \xi_{ikh\ell s(t)}(\boldsymbol{\theta}) &= \mathbb{P}(X_{is(t)} = \ell, X_{is(t-1)} = h \mid \boldsymbol{y}_{is}, Z_{ik} = 1; \boldsymbol{\theta}) \\ &= \frac{\alpha_{ikhs(t)}(\boldsymbol{\theta}) \boldsymbol{A}_k[h, \ell] g(y_{is(t)}; \boldsymbol{\lambda}_\ell, \varepsilon_\ell) \beta_{ik\ell s(t)}(\boldsymbol{\theta})}{\sum_{h'=1}^{M} \sum_{\ell'=1}^{M} \alpha_{ikh's(t)}(\boldsymbol{\theta}) \boldsymbol{A}_k[h', \ell'] g(y_{is(t)}; \boldsymbol{\lambda}_{\ell'}, \varepsilon_{\ell'}) \beta_{ik\ell' s(t)}(\boldsymbol{\theta})}. \end{aligned}$$

The probability τ_{ik} that one observation arises from component k is

$$\tau_{ik}(\boldsymbol{\theta}) = \mathbb{P}(Z_{ik} = 1 \mid \boldsymbol{y}_i, \boldsymbol{\theta}) = \frac{\prod_{s=1}^{S_i} \sum_{h=1}^{M} \alpha_{ikhs(T_{is})}(\boldsymbol{\theta})}{\sum_{k'=1}^{K} \prod_{s=1}^{S_i} \sum_{h=1}^{M} \alpha_{ik'hs(T_{is})}(\boldsymbol{\theta})}.$$

The probability $\eta_{ihs(t)}$ that observation i is at state h at time t of sequence s is

$$\eta_{ihs(t)}(\boldsymbol{\theta}) = \mathbb{P}(X_{is(t)} = h \mid \boldsymbol{y}_i, \boldsymbol{\theta}) = \sum_{k=1}^{K} \tau_{ik}(\boldsymbol{\theta}) \gamma_{ikhs(t)}(\boldsymbol{\theta}).$$

4.3 EM algorithm

The EM algorithm is an iterative algorithm randomly initialized at the model parameter $\boldsymbol{\theta}^{[0]}$. It alternates between two steps: the Expectation step (E-step) consisting in computing the expectation of the complete-data likelihood under the current parameters, and the maximization step (M-step) consisting in maximizing this expectation over the model parameters. Iteration $[r]$ of the algorithm is defined by

E-step Conditional probability computation

$$\tau_{ik}(\boldsymbol{\theta}^{[r-1]}), \gamma_{ikhs(t)}(\boldsymbol{\theta}^{[r-1]}), \eta_{ihs(t)}(\boldsymbol{\theta}^{[r-1]}) \text{ and } \xi_{ikh\ell s(t)}(\boldsymbol{\theta}^{[r-1]}).$$

M-step Parameter updating

$$\delta_k^{[r]} = \frac{n_k(\boldsymbol{\theta}^{[r-1]})}{n},\ \pi_{kh}^{[r]} = \frac{n_{kh(0)}(\boldsymbol{\theta}^{[r-1]})}{n_k(\boldsymbol{\theta}^{[r-1]})},\ \boldsymbol{A}_k[h,\ell]^{[r]} = \frac{n_{kh\ell}(\boldsymbol{\theta}^{[r-1]})}{n_{kh}(\boldsymbol{\theta}^{[r-1]})},\ \varepsilon_h^{[r]} = \frac{w_h(\boldsymbol{\theta}^{[r-1]})}{n_{kh}(\boldsymbol{\theta}^{[r-1]})},$$

$$\boldsymbol{\lambda}_h^{[r]} = \underset{\lambda_h}{\operatorname{argmax}} \sum_{i=1}^{n}\sum_{s=1}^{S_i}\sum_{t=0}^{T_{is}} \eta_{ihs(t)}(\boldsymbol{\theta}^{[r-1]}) g_c(y_{is(t)};\boldsymbol{\lambda}_h),$$

where

$$n_k(\boldsymbol{\theta}) = \sum_{i=1}^{n}\tau_{ik}(\boldsymbol{\theta}),\quad n_{kh}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{s=1}^{S_i}\sum_{t=0}^{T_{is}}\tau_{ik}(\boldsymbol{\theta})\gamma_{ikhs(t)},\quad n_{kh(0)}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{s=1}^{S_i}\tau_{ik}(\boldsymbol{\theta})\gamma_{ikhs(0)}(\boldsymbol{\theta}),$$

$$n_{kh\ell}(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{s=1}^{S_i}\sum_{t=1}^{T_{is}}\tau_{ik}(\boldsymbol{\theta})\xi_{ikh\ell s(t)}(\boldsymbol{\theta}) \text{ and } w_h(\boldsymbol{\theta}) = \sum_{i=1}^{n}\sum_{s=1}^{S_i}\sum_{t=0}^{T_{is}}\eta_{ihs(t)}(\boldsymbol{\theta})\mathbf{1}_{\{y_{is(t)}=0\}}.$$

5 Numerical illustrations

This section illustrates the model properties on simulated data as well as on a subset of the real data studied by Huang et al. (2018b). First, we specify the design of the simulations and then we illustrate the evolution of the probability of misclassification (given by Theorem 2) and the convergence of estimators. We then study the robustness of the proposed method in the presence of missing data. Finally, we examine its performance on a real data set.

5.1 Simulated data

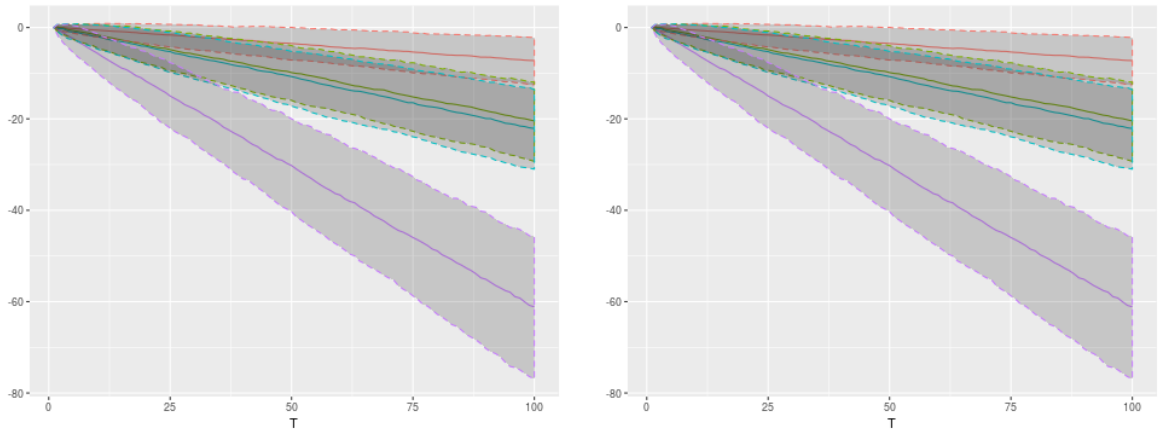
Simulation design Simulated data are sampled from a bi-component mixture of HMM with two states (*i.e.*, $K = M = 2$) and equal proportions (*i.e.*, $\delta_1 = \delta_2 = 1/2$). The distribution of $Y_{i(t)}$ conditionally on the state h is a zero-inflated gamma distribution denoted by $\mathcal{G}a(a_h, b_h)$. We have

$$\varepsilon_1 = \varepsilon_2 = 0.1,\ a_1 = 1,\ b_1 = b_2 = 1,\ \boldsymbol{A}_1 = \begin{bmatrix} e & 1-e \\ 1-e & e \end{bmatrix} \text{ and } \boldsymbol{A}_2 = \begin{bmatrix} 1-e & e \\ e & 1-e \end{bmatrix}.$$

The parameter $a_2 > 1$ controls the separation of the distribution of $Y_{i(t)}$ given the state and the parameter e controls the separation of the distribution of X given the class (when e increases, the constant c in Theorem 2 increases). We consider four cases (hard: $e = 0.75$ and $a_2 = 3$; medium-1: $e = 0.90$ and $a_2 = 3$; medium-2: $e = 0.75$ and $a_2 = 5$; easy: $e = 0.90$ and $a_2 = 5$).

Probabilities of misclassification For each of the four cases, the probability of misclassification is computed on 1000 observations for $T = 1, \ldots, 100$. Figure 2(a) shows the behavior of $\log \frac{\mathbb{P}(Z_{ik}=1|\boldsymbol{y}_i)}{\mathbb{P}(Z_{ik_0}=1|\boldsymbol{y}_i)}$ when k_0 is the true class and k the alternative. This quantity linearly decreases with T. Figure 2(b) presents the empirical probabilities of misclassification. As expected and predicted by our theoretical findings (Theorem 2), the probability of misclassification decreases at an exponential rate with T.

Convergence of the estimators For each case, 1000 samples composed of n sequences of length T are generated. Parameters are estimated by maximum likelihood. To investigate the accuracy of the estimation procedure, we computed the mean square error between the model parameters and their estimators. Moreover, we compute the adjusted Rand index (Hubert and Arabie (1985)) between the true partition and the partition given by the MAP rule, and between the true state sequences and the estimated state sequences given by the MAP rule (obtained with the Viterbi algorithm (Viterbi, 1967)). Table 1 shows the results obtained with two values of n and two values of T, considering the case medium-1. It can be seen that the latent variables and the model parameters are well estimated. Indeed, the MLE converge to the true parameters as T or n increases, except for the proportion in each component δ_k, which needs n to be sufficiently large to have observed enough individuals in each component. We notice that the partition obtained by our estimation procedure corresponds to the true partition (for n and T large enough) even if we are not under the true parameters but under the MLE, which is not an immediate consequence of Theorem 2. On the contrary, we do not find the



(a) Median and 90%-confidence region of $\log \frac{\mathbb{P}(Z_{ik}=1|\boldsymbol{y}_i)}{\mathbb{P}(Z_{ik_0}=1|\boldsymbol{y}_i)}$..

(b) Probability of misclassification.

Figure 2: Results obtained on 1000 observations for the hard (orange), medium-1 (green), medium-2 (blue) and easy (purple) cases.

true states sequences a.s. , as the number of states to be estimated is also growing with n and T. Results obtained for the three other cases are similar and are presented in Supplementary Material Section D.1.

		Adjusted Rand index		Mean square error				
n	T	partition	states	$\boldsymbol{A}_k$	ε_h	a_h	b_h	δ_k
10	100	0.994	0.652	0.024	0.001	0.277	0.032	0.054
10	500	0.999	0.659	0.005	0.000	0.054	0.006	0.049
100	100	0.998	0.660	0.002	0.000	0.025	0.003	0.005
100	500	1.000	0.660	0.000	0.000	0.005	0.001	0.005

Table 1: Maximum likelihood estimators convergence when data are sample from case medium-1.

Impact of the missing values To investigate the robustness of the proposed method with missingness, we add s sequences of q consecutive missing values to each observation. The location of the missing sequence is randomly sampled. Table 2 compares the results obtained with and without missingness, considering case medium-1. It shows that estimators are robust to missingness. Results obtained for the other three cases are similar and are reported in Supplementary Material Section D.1.

5.2 Using the approach on classical accelerometer data

We consider the accelerometer data measured on three subjects available from Huang et al. (2018b). The accelerometer measures the activity every five minutes for one week. Note that the first subject has 2% of missing values. The purpose of this section is to illustrate the differences between the method of Huang et al. (2018b) and the method proposed in this paper.

Huang et al. (2018b) consider one HMM per subject with three latent states. This model is used for monitoring the circadian rhythmicity, subject by subject. Because they fit one HMM per observation, the definition of the activity level is different for each observation (see Huang et al. (2018b, Figure 4)). This is not an issue for their study because the analysis is done subject by subject. However, the mean time spent by activity levels cannot be compared among the subjects. The method proposed here makes this comparison possible. Figure 3 depicts the activity data of the three subjects, the expected value of $Y_{i(t)}$ conditionally to the most likely state and on the most likely component and the probability of each state. Based on the QQ-plot (see Supplementary Material Section D.2), we consider $M = 4$ activity levels. These levels can be easily characterized with the model parameters presented in Table 3. Moreover, the transition matrices also make sense. For instance, class 1 (subjects

				Adjusted Rand index		Mean square error				
n	T	s	q	partition	states	$\boldsymbol{A}_k$	ε_h	a_h	b_h	δ_k
		0	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		1	10	0.993	0.996	1.107	1.119	1.101	1.166	1.005
10	100	2	10	0.990	0.995	1.222	1.251	1.159	1.313	1.002
		1	20	0.987	0.993	1.281	1.288	1.250	1.340	0.990
		2	20	0.967	0.986	1.587	1.657	1.551	1.727	0.994
		0	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		1	10	1.000	1.000	1.025	1.015	1.022	1.015	1.002
10	500	2	10	1.000	1.000	1.023	1.027	1.024	1.036	0.996
		1	20	1.000	1.000	1.010	1.032	1.036	1.048	1.006
		2	20	1.000	1.000	1.060	1.074	1.054	1.084	1.003
		0	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		1	10	0.998	0.998	1.148	1.089	1.166	1.169	0.997
100	100	2	10	0.996	0.997	1.277	1.203	1.286	1.233	1.002
		1	20	0.996	0.997	1.288	1.250	1.298	1.280	0.998
		2	20	0.983	0.994	1.632	1.561	1.638	1.585	1.003
		0	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		1	10	1.000	1.000	1.026	1.050	1.032	0.992	1.000
100	500	2	10	1.000	1.000	1.045	1.083	1.038	0.989	1.000
		1	20	1.000	1.000	1.037	1.071	1.008	0.942	1.000
		2	20	1.000	1.000	1.090	1.127	1.104	1.008	1.000

Table 2: Ratio between statistics obtained with and without missing data when data are sampled from case medium-1.

9 and 20) has an almost tri-diagonal transition matrix (by considering an order between the states given through the activity levels per state) and class-2 (subject 2) is composed of an individual with low-overall activity

$$\hat{\boldsymbol{A}}_1 = \begin{bmatrix} 0.86 & 0.14 & 0.00 & 0.00 \\ 0.12 & 0.81 & 0.06 & 0.01 \\ 0.00 & 0.07 & 0.79 & 0.14 \\ 0.00 & 0.00 & 0.13 & 0.87 \end{bmatrix}.$$

State name	ε_h	a_h	b_h	mean	sd
intensive-level	0.00	98.94	0.65	152.76	15.36
moderate-level	0.00	11.09	0.11	99.34	29.84
low-level	0.00	2.32	0.11	20.98	13.79
sleeping	0.22	1.48	0.72	2.06	1.70

Table 3: Parameters and mean time per states for the three subjects.

6 Analysis of the NYC accelerometer data

Data We consider the 133 individuals aged at least of 65 years who responded to the NYC Department of Health and Mental Hygiene study in 2010-2011.[1] Accelerometers were worn for one week and measured the activity minute by minute. The accelerometers were removed during sleep, hence the data contains 44% of missing values that appear in sequence. The analysis is conducted using four activity levels. The number of components is considered to be between one and six, and it has been selected with the BIC (Schwarz, 1978). For each number of components, 5000 random initializations of

[1]New York City Department of Health and Mental Hygiene. Physical Activity and Transit Survey 2010-2011; public use dataset accessed on May 10, 2019.

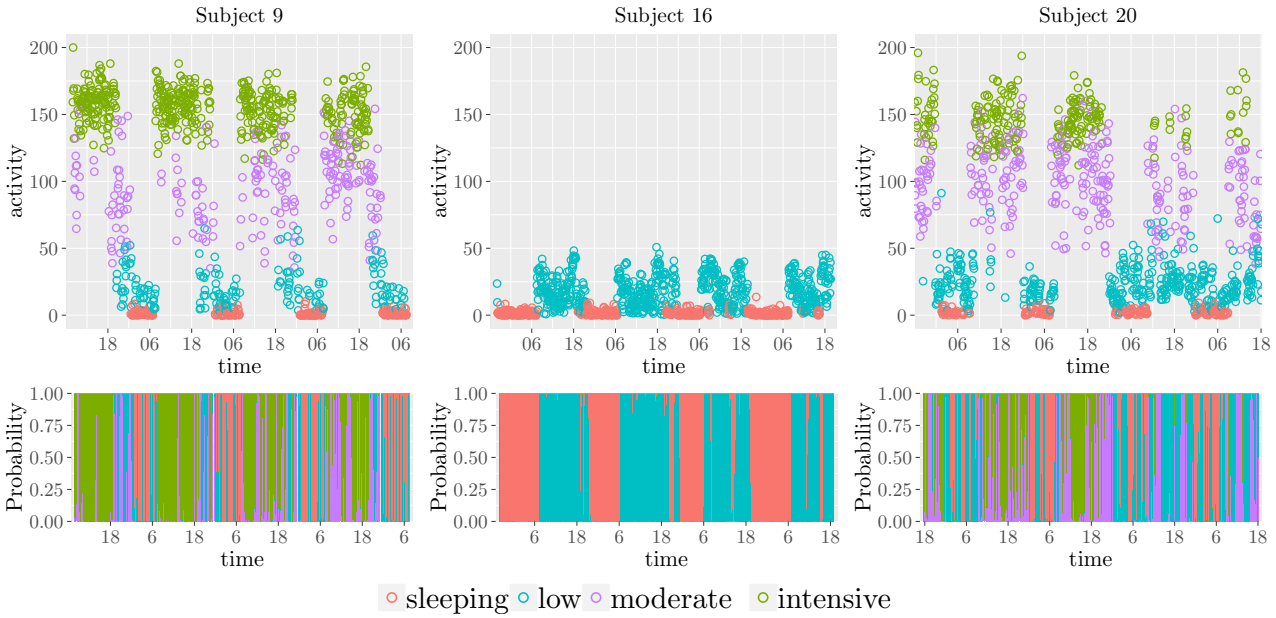


Figure 3: State estimation for the three subjects: (top) accelerometer data where color indicates the expected value of $Y_{i(t)}$ conditionally to the most likely state and to the most likely component; (bottom) probability of each state at each time.

the EM algorithm are performed. The analysis needs about one day of computation on a 32-Intel(R) Xeon(R) CPU E5-4627 v4 @ 2.60GHz.

Model selection Data are analyzed considering four latent states (*e.g.,* four activity levels). It is not easy to use information criteria for selecting the number of states in HMM (see the discussion in the conclusion). In addition, accelerometer data are generally analyzed with four activity levels. To select the number of components, we use two information criteria which are generally used in clustering: the BIC (Schwarz, 1978) and the ICL (Biernacki et al., 2000). Here, the ICL is defined according to the integrated complete-data likelihood computed on $(\boldsymbol{y}, \hat{\boldsymbol{z}})$ where $\hat{\boldsymbol{z}}$ is the partition given by the MAP rule with the MLE. These information criteria are defined as follows and their values (according to the number of clusters) are given in Table 4.

$$\text{BIC}(K) = \ell_K(\boldsymbol{\theta}; \boldsymbol{y}) - \frac{\nu_K}{2} \log(\sum_{i=1}^{n} \sum_{s=1}^{S_i} T_{is} + 1),$$

and

$$\text{ICL}(K) = \text{BIC}(K) + \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{z}_{ik} \log \tau_{ik}(\hat{\boldsymbol{\theta}}),$$

where $\nu_K = (K-1) + K(M + M^2) + 3M$ is the number of parameters with K components and M states. In practice, $\text{ICL}(K)$ is close to $\text{BIC}(K)$, because the entropy $\sum_{i=1}^{n} \sum_{k=1}^{K} \hat{z}_{ik} \log \tau_{ik}(\hat{\boldsymbol{\theta}}) \approx 0$. This is a consequence of Theorem 2 (see also numerical experiments in Section 5). On the NYC data, both criteria select five classes.

K	1	2	3	4	5	6	7
BIC	-2953933	-2952313	-2951809	-2951705	-2951308	-2951364	-2951696
ICL	-2953933	-2952313	-2951810	-2951707	-2951309	-2951364	-2951697

Table 4: Information criteria obtained on the NYC accelerometer data with four activity levels.

Description of the activity levels The parameters of the zero-inflated gamma distributions are presented in Table 5. The four distributions are ordered by the value of their means. The *sleeping state* is characterized by a large probability of observing zero (*i.e.,* ε_h is close to one). However, ε_h is not zero for the other states but the more active the state is, the smaller ε_h is.

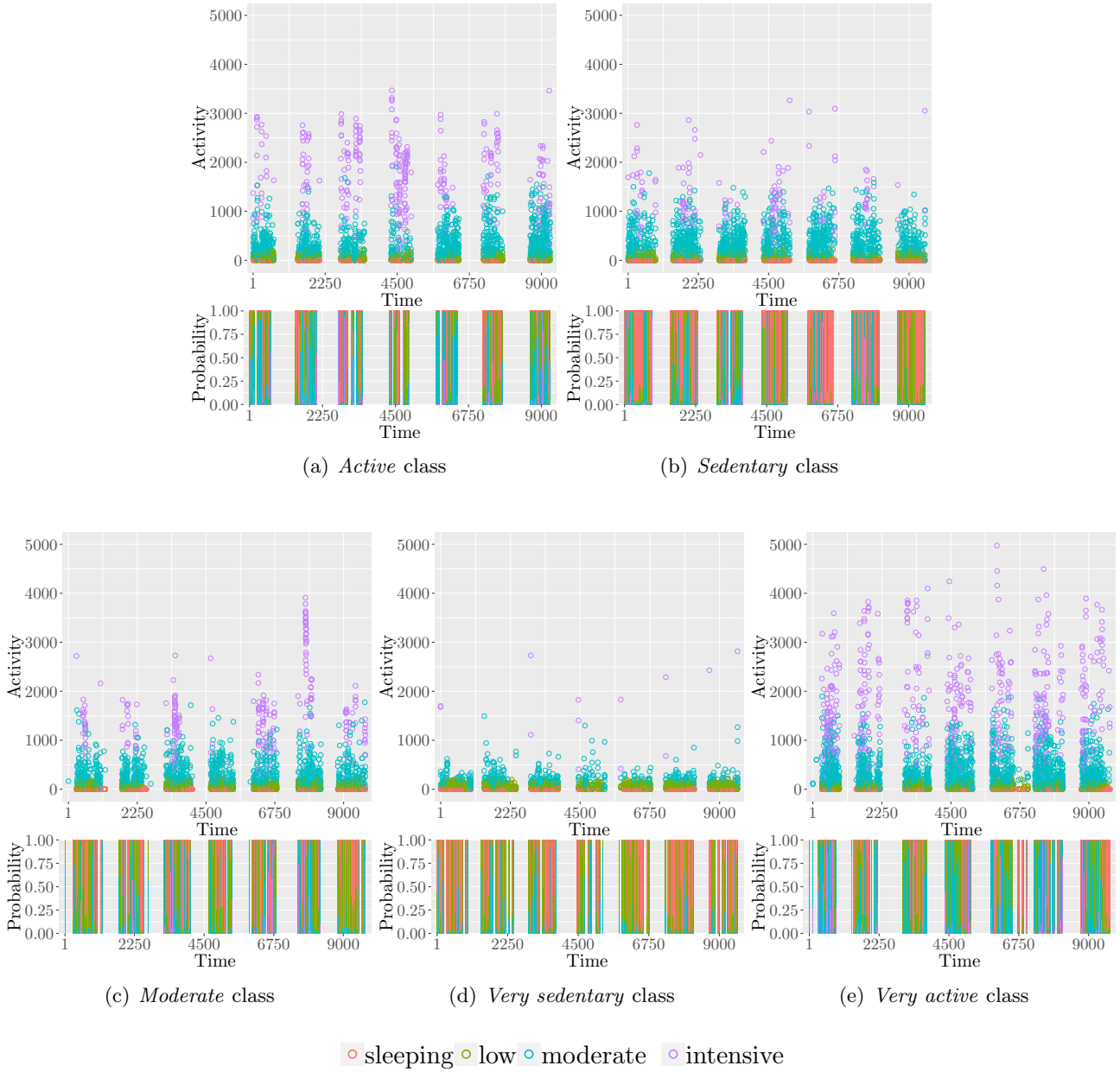
Name of the activity level	ε_h	a_h	b_h	mean
sleeping	0.988	7.470	7.470	0.012
low-level	0.260	0.974	0.020	36.926
moderate-level	0.025	1.408	0.004	329.249
intensive-level	0.007	2.672	0.002	1696.935

Table 5: Parameters describing the four activity levels for the NYC accelerometer data.

Description of the classes Classes can be described using their proportions and their associated parameters presented in Table 5. The data are composed of a majority class ($\delta_1 = 0.518$). Three other classes are composed of more sedentary individuals (*e.g,* their marginal probabilities of being in states 1 and 2 are higher). Finally, there is a small class ($\delta_5 = 0.045$) which contains the most active subjects (*e.g.,* $\pi_{k4} = 0.143$). Figure 4 presents one observation which characterizes each class and the probabilities of the activity levels. Finally, the transition matrices presented in Table 6 are almost tri-diagonal. This could be expected because it seems relevant to obtain a low probability of jumping between the sleeping state and the intensive state. Additionally, the approximation made for efficiently handling the missingness (see Section 3.4) turns out to be relevant. The minimal range of missing values is indeed equal to $d_{\min} = 60$ which leads to a distance in total variation between the $d_{\min}$-power of the transition matrices and the stationary distribution being less than 5.10^{-4} for any component.

	Class *active*			
	sleeping	low-level	moderate-level	intensive-level
sleeping	0.87	0.12	0.01	0.00
low-level	0.17	0.73	0.10	0.00
moderate-level	0.04	0.30	0.66	0.01
intensive-level	0.08	0.08	0.18	0.66
	Class *sedentary*			
	sleeping	low-level	moderate-level	intensive-level
sleeping	0.79	0.16	0.05	0.00
low-level	0.17	0.66	0.16	0.01
moderate-level	0.05	0.14	0.79	0.03
intensive-level	0.01	0.02	0.15	0.82
	Class *moderate*			
	sleeping	low-level	moderate-level	intensive-level
sleeping	0.76	0.21	0.03	0.00
low-level	0.16	0.73	0.11	0.00
moderate-level	0.03	0.20	0.73	0.04
intensive-level	0.01	0.04	0.16	0.80
	Class *very sedentary*			
	sleeping	low-level	moderate-level	intensive-level
sleeping	0.85	0.08	0.06	0.00
low-level	0.20	0.67	0.13	0.01
moderate-level	0.10	0.11	0.76	0.03
intensive-level	0.01	0.04	0.14	0.82
	Class *very active*			
	sleeping	low-level	moderate-level	intensive-level
sleeping	0.80	0.14	0.05	0.01
low-level	0.08	0.74	0.17	0.01
moderate-level	0.03	0.18	0.69	0.10
intensive-level	0.01	0.05	0.21	0.74

Table 6: Transition matrix for the five classes.



(a) *Active* class (b) *Sedentary* class

(c) *Moderate* class (d) *Very sedentary* class (e) *Very active* class

Figure 4: Examples of observation assigned into the five classes with the probabilities of the states.

7 Conclusion

A specific mixture of HMM has been introduced to analyze accelerometer data. It avoids the traditional cut-off point method for analyzing such data. Furthermore, this model can take into account the heterogeneity within the population. Properties (identifiability, probability of misclassifying an observation) have been proved. Applications on real data are promising.

In the application, the number of activity levels was not estimated but fixed at a common value for accelerometer data. Estimating the number of states for a mixture of HMM is an interesting but complex topic. Indeed, the use of BIC is criticized (see Cappé et al. (2005, Chapter 15)). Thus, pseudo-likelihood based criteria could be used (Gassiat, 2002; Csiszr and Talata, 2006) but the fact that the marginal distribution of one $Y_{i(t)}$ is not identifiable limits this approach. A more promising approach could be to use cross-validated likelihood (Celeux and Durand, 2008) but it would be computationally intensive because accelerometer data provide a large amount of observations.

A Model identifiability

The proof of Theorem 1 is split in two parts:

1. Identifiability of the parameters of the specific distribution per state is obtained using the approach of Teicher (1963). Hence $\forall h = 1, \ldots, M$

$$\boldsymbol{\lambda}_h = \tilde{\boldsymbol{\lambda}}_h \text{ and } \sum_{k=1}^{K} \delta_k \pi_{kh}(1-\varepsilon_h) = \sum_{k=1}^{K} \tilde{\delta}_k \tilde{\pi}_{kh}(1-\tilde{\varepsilon}_h).$$

2. Identifiability of the transition matrices and of the ε is shown using properties of Vandermonde matrices. Hence,

$$\forall k = 1, \ldots, K,\ \delta_k = \tilde{\delta}_k,\ \boldsymbol{A}_k = \tilde{\boldsymbol{A}}_k,\ \pi_k = \tilde{\pi}_k, \text{ and } \boldsymbol{\varepsilon} = \tilde{\boldsymbol{\varepsilon}}.$$

A.1 Identifiability of the parameters of the specific distribution per state

Considering the marginal distribution at time $t = 0$, we have

$$\sum_{k=1}^{K}\sum_{h=1}^{M} \delta_k \pi_{kh} g(y_{i(0)}; \boldsymbol{\lambda}_h, \varepsilon_h) = \sum_{k=1}^{K}\sum_{h=1}^{M} \tilde{\delta}_k \tilde{\pi}_{kh} g(y_{i(0)}; \tilde{\boldsymbol{\lambda}}_h, \tilde{\varepsilon}_h).$$

Note that $g(y_{i(0)}; \boldsymbol{\lambda}_h, \varepsilon_h) = (1-\varepsilon_h) g_c(y_{i(0)}; \boldsymbol{\lambda}_h) + \varepsilon_h \mathbf{1}_{\{y_{i(0)}=0\}}$ is a pdf of a zero-inflated distribution, so it is a pdf of a bi-component mixture. We now use the same reasoning as Teicher (1963). We have

$$1 + \sum_{h=2}^{M} \frac{g(y_{i(0)}; \boldsymbol{\lambda}_h, \varepsilon_h) \sum_{k=1}^{K} \delta_k \pi_{kh}}{g(y_{i(0)}; \boldsymbol{\lambda}_1, \varepsilon_1) \sum_{k=1}^{K} \delta_k \pi_{k1}} = \frac{g(y_{i(0)}; \tilde{\boldsymbol{\lambda}}_1, \tilde{\varepsilon}_1) \sum_{k=1}^{K} \tilde{\delta}_k \tilde{\pi}_{k1}}{g(y_{i(0)}; \boldsymbol{\lambda}_1, \varepsilon_1) \sum_{k=1}^{K} \delta_k \pi_{k1}} + \sum_{h=2}^{M} \frac{g(y_{i(0)}; \tilde{\boldsymbol{\lambda}}_h, \tilde{\varepsilon}_h) \sum_{k=1}^{K} \tilde{\delta}_k \tilde{\pi}_{kh}}{g(y_{i(0)}; \boldsymbol{\lambda}_1, \varepsilon_1) \sum_{k=1}^{K} \delta_k \pi_{k1}}.$$

Considering $y_{i(0)} \to \rho$, by Assumption 2, we have

$$\boldsymbol{\lambda}_1 = \tilde{\boldsymbol{\lambda}}_1 \text{ and } (1-\varepsilon_1) \sum_{k=1}^{K} \delta_k \pi_{k1} = (1-\tilde{\varepsilon}_1) \sum_{k=1}^{K} \tilde{\delta}_k \tilde{\pi}_{k1}.$$

Repeating the previous argument with $h = 2, \ldots, M$, we conclude that, for $h \in \{1, \ldots, M\}$,

$$\boldsymbol{\lambda}_h = \tilde{\boldsymbol{\lambda}}_h \text{ and } (1-\varepsilon_h) \sum_{k=1}^{K} \delta_k \pi_{kh} = (1-\tilde{\varepsilon}_h) \sum_{k=1}^{K} \tilde{\delta}_k \tilde{\pi}_{kh}.$$

A.2 Identifiability of the transition matrices

First, we introduce two technical lemmas of which proofs are discussed in the next subsection. Second, we show that $\boldsymbol{A}_k[1,1] = \tilde{\boldsymbol{A}}_k[1,1]$ then we extend the results to the whole transition matrices.

Lemma 1. *Let N_0, N_1, $\tilde{N}_0$ and $\tilde{N}_1$ be four definite positive matrices of size $K \times K$ such that for $u \in \{1,\dots,K\}$ and $k \in \{1,\dots,K\}$,*

$$N_0[u,k] = a_k^{u-1},\ N_1[u,k] = a_k^{K+u-1},\ \tilde{N}_0[u,k] = \tilde{a}_k^{u-1},\ \tilde{N}_1[u,k] = \tilde{a}_k^{K+u-1},$$

with $a_k > a_{k+1} > 0$, $\tilde{a}_k > \tilde{a}_{k+1} > 0$ and $a_1 \geq \tilde{a}_1$. If for any $\tilde{w} \in \mathbb{R}_+^K$ there exists $w \in \mathbb{R}_+^K$ $N_0 w = \tilde{N}_0 \tilde{w}$ and $N_1 w = \tilde{N}_1 \tilde{w}$ then for $k \in \{1,\dots,K\}$ $a_k = \tilde{a}_k$ and $w = \tilde{w}$.

Lemma 2. *Let N_0, $\tilde{N}_0$ be two definite positive matrices of size $K \times K$ such that for $u \in \{1,\dots,K\}$ and $k \in \{1,\dots,K\}$,*

$$N_0[u,k] = a_k^{u-1},\ N_1[u,k] = a_k^{K+u-1},$$

with $a_k > a_{k+1} > 0$, $\tilde{a}_k > \tilde{a}_{k+1} > 0$ and $a_1 \geq \tilde{a}_1$. Let $D_u = \operatorname{diag}(a_1^{Ku},\dots,a_K^{Ku})$ and $\tilde{D}_u = \operatorname{diag}(\tilde{a}_1^{Ku},\dots,\tilde{a}_K^{Ku})$. If there exist $\alpha \in]0,1[$, $\tilde{\alpha} \in]0,1[$, $w \in \mathbb{R}_+^K$ and $\tilde{w} \in \mathbb{R}_+^K$ such that for $u \in \{0,\dots,K-1\}$, we have

$$\alpha N_0 D_u w = \tilde{\alpha} \tilde{N}_0 \tilde{D}_u \tilde{w},$$

then for $k \in \{1,\dots,K\}$ $a_k = \tilde{a}_k$ and $w = \tilde{w}$.

We consider the marginal distribution of $(y_{i(0)},\dots,y_{i(t-1)})$ with $t = 1,\dots,2K$, where $y_{i(0)} = y_{i(t')}$ for each $t' = 1,\dots,t-3$, $y_{i(t-2)} = y_{i(0)}^{\tau_1}$, $y_{i(t-1)} = y_{i(0)}^{\tau_2}$ and $y_{i(t)} = y_{i(0)}^{\tau_3}$. Therefore, taking $\tau_1 = \tau_2 = \tau_3 = 1$ and letting $y_{i(0)}$ tend to ρ (see Assumption A), we obtain, for $t = 1,\dots,2K$, that

$$(1-\varepsilon_1)\sum_{k=1}^{K} \delta_k \pi_{k1} \left(\boldsymbol{A}_k[1,1](1-\varepsilon_1)\right)^{t-1} = (1-\tilde{\varepsilon}_1)\sum_{k=1}^{K} \tilde{\delta}_k \tilde{\pi}_{k1} \left(\tilde{\boldsymbol{A}}_k[1,1](1-\tilde{\varepsilon}_1)\right)^{t-1}.$$

Because, we consider $2K$ marginal distributions, we can use Lemma 2 by setting $\alpha = (1-\varepsilon)$, $\tilde{\alpha} = (1-\tilde{\varepsilon})$, $a_k = \boldsymbol{A}_k[1,1](1-\varepsilon_1)$, $\tilde{a}_k = \tilde{\boldsymbol{A}}_k[1,1](1-\tilde{\varepsilon}_1)$, $w_k = \delta_k \pi_{k1}$ and $w_k = \tilde{\delta}_k \tilde{\pi}_{k1}$. Therefore, we have $\varepsilon = \tilde{\varepsilon}$, $\boldsymbol{A}_k[1,1] = \tilde{\boldsymbol{A}}_k[1,1]$ and $\delta_k \pi_{k1} = \tilde{\delta}_k \tilde{\pi}_{k1}$. Using the previous approach, with $\tau_1 = \tau_2 = 1$ and $\tau_3 < 1$, with $h = 2,\dots,M$, we have for $t = 1,\dots,K$

$$(1-\varepsilon_h)(1-\varepsilon_1)\sum_{k=1}^{K} \delta_k \pi_{k1} \left(\boldsymbol{A}_k[1,1](1-\varepsilon_1)\right)^{t-2} \boldsymbol{A}_k[1,h] =$$
$$(1-\tilde{\varepsilon}_h)(1-\varepsilon_1)\sum_{k=1}^{K} \delta_k \pi_{k1} \left(\boldsymbol{A}_k[1,1](1-\varepsilon_1)\right)^{t-2} \tilde{\boldsymbol{A}}_k[1,h],$$

and thus $\boldsymbol{A}_k[1,h] = \tilde{\boldsymbol{A}}_k[1,h]$ and $\varepsilon_h = \tilde{\varepsilon}_h$. Similarly, taking $\tau_2 < 1$ and $\tau_1 = \tau_3 = 1$, we have $\boldsymbol{A}_k[h,1] = \tilde{\boldsymbol{A}}_k[h,1]$. Finally, we have $\boldsymbol{A}_k[h,h'] = \tilde{\boldsymbol{A}}_k[h,h']$ by increasing h and h', by noting that with suitable choices of τ_1, τ_2 and τ_3, we have for $t = 1,\dots,K$

$$\sum_{k=1}^{K} \delta_k \pi_{k1} \left(\boldsymbol{A}_k[1,1](1-\varepsilon_1)\right)^{t-2} \boldsymbol{A}_k[1,h]\boldsymbol{A}_k[h,h']\boldsymbol{A}_k[h',1] =$$
$$\sum_{k=1}^{K} \tilde{\delta}_k \tilde{\pi}_{k1} \left(\boldsymbol{A}_k[1,1](1-\varepsilon_1)\right)^{t-2} \boldsymbol{A}_k[1,h]\tilde{\boldsymbol{A}}_k[h,h']\boldsymbol{A}_k[h',1].$$

A.3 Proofs of the two technical lemmas

Proof of Lemma 1. Since $a_k \neq a_{k'}$ and $\tilde{a}_k \neq \tilde{a}_{k'}$, then N_0, N_1, $\tilde{N}_0$ and $\tilde{N}_1$ are Vandermonde matrices and thus are invertible. Therefore, we have $w = N_0^{-1}\tilde{N}_0\tilde{w} = N_1^{-1}\tilde{N}_1\tilde{w}$, and thus

$$(N_0^{-1}\tilde{N}_0 - N_1^{-1}\tilde{N}_1)\tilde{w} = 0,$$

or similarly for $u \in \{1, \dots, K\}$,

$$\sum_{k=1}^{K} a_k^u w_k = \sum_{k=1}^{K} \tilde{a}_k^u \tilde{w}_k.$$

Since the previous equation holds for any $\tilde{w}$, we have $N_0^{-1}\tilde{N}_0 = N_1^{-1}\tilde{N}_1$. Moreover, we have $N_1 = N_0 D$ and $\tilde{N}_1 = \tilde{N}_0\tilde{D}$ where $D = \text{diag}(a_1^K, \dots, a_K^K)$ and $\tilde{D} = \text{diag}(\tilde{a}_1^K, \dots, \tilde{a}_K^K)$. Denoting $R = N_0^{-1}\tilde{N}_0$, $DR = R\tilde{D}$ and then for $u \in \{1, \dots, K\}$ and $k \in \{1, \dots, K\}$

$$a_u^K R[u,k] = \tilde{a}_k^K R[u,k]. \tag{3}$$

We now show that $D = \tilde{D}$ and, $w = \tilde{w}$, and hence $R = I_K$ and $\tilde{N}_0 = N_0$, where I_K is the identity matrix of size K. First we show that $a_1 = \tilde{a}_1$ and $w_1 = \tilde{w}_1$.

- If $R[1,j] \neq 0$, (3) implies that $a_1^K R[1,j] = \tilde{a}_j^K R[1,j]$ and thus $a_1 = \tilde{a}_j$. However, this is impossible because $a_1 \geq \tilde{a}_1 > a_j$ for $j \in \{2, \dots, K\}$. Hence, we have $R[1,j] = 0$ for $j = 2, \dots K$.
- Noting that R is a product of two invertible matrices, R is invertible. Therefore, $R[1,1] \neq 0$ because $R[1,j] = 0$ for $j = 2, \dots K$. Hence, we have $a_1 = \tilde{a}_1$.
- Note that $R[1,1] = \sum_{k=1}^{K}(N_0^{-1})[1,k]\tilde{N}_0[k,1]$ and that $\tilde{N}_0[k,1] = \tilde{a}_1^k = a_1^k = N_0[k,1]$. Therefore, we have $R[1,1] = \sum_{k=1}^{K}(N_0^{-1})[1,k]N_0[k,1] = (N_0^{-1}N_0)[1,1] = 1$.
- For $j = 2, \dots K$, $a_1 > a_j$ so we have $R[j,1] = 0$, because $a_1 = \tilde{a}_1$.
- Because $w = R\tilde{w}$, we have $w_1 = \tilde{w}_1$.

Equality of $a_k = \tilde{a}_k$ and $w_k = \tilde{w}_k$ can be shown recursively for $k = 2, \dots, K$ using the same reasoning. □

B Probabilities of misclassification

B.1 Technical lemmas

This section presents some notations and three lemmas which are used for the proof of Theorem 2. The technical lemmas discuss the concentration of the frequency of the observation $y_{i(t)}$ in a region of interest W, give an upper bound of $p(\boldsymbol{y}_i \mid Z_{ik} = 1)$ and a concentration result of the ratio of $\frac{p(\tilde{\boldsymbol{x}}_{ik}, \boldsymbol{y}_i \mid Z_{ik}=1)}{p(\tilde{\boldsymbol{x}}_{ik_0}, \boldsymbol{y}_i \mid Z_{ik_0}=1)}$, where $\tilde{x}_{ik} = \text{argmax}_{\boldsymbol{x}_i \in \mathcal{X}}\, p(\boldsymbol{x}_i, \boldsymbol{y}_i \mid Z_{ik} = 1)$ is the estimator of the latent states conditionally on the observation $\boldsymbol{y}_i$ and on component k obtained by applying the *maximum a posteriori* rule with the true parameter $\boldsymbol{\theta}$. The proof of the lemmas uses two concentration results for hidden Markov chains given by Kontorovich and Weiss (2014) and by León and Perron (2004).

Preliminaries Let $\boldsymbol{v}_{ik(t)} = (v_{ik(t)h\ell}; h = 1, \dots, M; \ell = 1, \dots, M)$ with $v_{ik(t)h\ell} = x_{ik(t-1)h}x_{ik(t)\ell}$ and $\tilde{\boldsymbol{v}}_{ik(t)} = (\tilde{v}_{ik(t)h\ell}; h = 1, \dots, M; \ell = 1, \dots, M)$ with $\tilde{v}_{ik(t)h\ell} = \tilde{x}_{ik(t-1)h}\tilde{x}_{ik(t)\ell}$. In the following, $\mathbb{P}_0(\cdot) = \mathbb{P}(\cdot \mid Z_{ik_0} = 1)$ by considering the true parameters.

Remark 1. *For any $k = 1, \dots, g$, $\boldsymbol{V}_{ik(t)}$ is a finite, ergodic and reversible Markov chain with M^2 states and transition matrix $\boldsymbol{P}_k$ with general term defined for any $(h_1, h_2, h_3, h_4) \in M^4$ by*

$$\boldsymbol{P}_k[(h_1-1)M + h_2, (h_3-1)M + h_4] = \mathbb{P}(V_{ik(t)h\ell} = 1) = \begin{cases} 0 & \text{if } h_2 \neq h_3 \\ \boldsymbol{A}_k[h_2, h_4] & \text{otherwise} \end{cases}.$$

Moreover, the non-zero eigenvalues of $\boldsymbol{P}_k$ are the non-zero eigenvalues of $\boldsymbol{A}_k$ and the eigenvectors of $\boldsymbol{P}_k$ are obtained from the eigenvectors of $\boldsymbol{A}_k$.

Theorem 3 (Kontorovich and Weiss (2014))**.** *Let $U_{(1)}, U_{(2)}, \dots$ be a stationary $\mathbb{N}$-valued (G, η)-geometrically ergodic Markov or hidden Markov chain, and consider the occupation frequency*

$$\hat{\rho}(E) = \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}_{\{U_{(t)} \in E\}}, \quad E \subset \mathbb{N}.$$

When $\sum_{u\in\mathbb{N}}\sqrt{\rho_u}<\infty$ *with* $\rho_u=\mathbb{P}(U_{(1)}=u)$*, then for any* $\varepsilon>0$

$$\mathbb{P}\left(sup_{E\subset\mathbb{N}}\mid\rho(E)-\hat{\rho}(E)\mid>\varepsilon+\gamma_T(G,\eta)\sum_{y\in\mathbb{N}}\sqrt{\rho_y}\right)\leq e^{-\frac{T}{2G^2}(1-\eta)^2\varepsilon^2},$$

where

$$\gamma_T(G,\eta)=\frac{1}{2}\sqrt{\frac{1+2G\eta}{T(1-\eta)}}.$$

Theorem 4 (León and Perron (2004))**.** *For all pairs* (V,f)*, such that* $V=(V_{(1)},\ldots,V_{(T)})$ *is a finite, ergodic and reversible Markov chain in stationary state with the second-largest eigenvalue* λ *and* f *is a function taking values in* $[0,1]$ *such that* $\mathbb{E}[f(V_{(t)})]<\infty$*, the following bounds, with* $\lambda_0=\max(0,\lambda)$*, hold for all* $s>0$ *such that* $\mathbb{E}[f(V_{(t)})]+s<1$ *and all time* T

$$\mathbb{P}\left(\sum_{t=1}^{T}f(V_{(t)})\geq(\mathbb{E}[f(V_{(1)})]+s)T\mid Z_{ik_0}=1\right)\leq\exp\left(-2\frac{1-\lambda_0}{1+\lambda_0}Ts^2\right).$$

Concentration of the frequency of the observations in W Let $W\subset\mathbb{R}^+$ be the subset of $\mathbb{R}^+$ where the estimator of $x_{i(t)}$ obtained by the *maximum a posteriori* rule is sensitive to $x_{i(t-1)}$ and $x_{i(t)}$ conditionally on $y_{i(t)}$ and component k. Thus, we define

$$W=\{u\in\mathbb{R}^+:\ \text{card}(\cup_{k=1}^{g}E_k(u))\geq 2\},$$

where

$$E_k(u)=\{h_2:\ \exists(h_1,h_3),\ h_2=\text{argmax}\,e_k(u;h_1,h_2,h_3)\},$$

and

$$e_k(u;h_1,h_2,h_3)=\boldsymbol{A}_k[h_1,h_2]\boldsymbol{A}_k[h_2,h_3]g(u;\boldsymbol{\lambda}_{h_2}).$$

Lemma 3. *Let* $\rho_{k_0}=\mathbb{P}_0(Y_{i(2)}\in W)$ *and* $\hat{\rho}_{k_0}=\sum_{t=1}^{T}\mathbf{1}_{\{y_{i(t)}\in W\}}$*. For any* $\delta_1>\frac{1}{\sqrt{2T}}$*,*

$$\mathbb{P}_0(\hat{\rho}_{k_0}<\rho_{k_0}-\delta_1)\leq e^{-Tc_1},$$

$\delta_1=\varepsilon+\frac{1}{\sqrt{2T}}$ *and* $c_1=\frac{1}{2}(\delta_1-\frac{1}{\sqrt{2T}})^2>0$*. Moreover,* $\hat{\rho}_{k_0}$ *is a consistent estimate of* ρ_{k_0} *because the marginal distribution of* $Y_{i(t)}$ *is the same for any* t*, and thus* $\rho_{k_0}=\mathbb{P}_0(Y_{i(t)}\in W)$ *for any* t*.*

Proof of Lemma 3. We have,

$$\mathbb{P}\left(\mid\rho_{k_0}-\hat{\rho}_{k_0}\mid>\varepsilon+\frac{1}{\sqrt{2T}}\right)\leq\mathbb{P}\left(\mid\rho_{k_0}-\hat{\rho}_{k_0}\mid>\varepsilon+\frac{1}{2\sqrt{T}}(\sqrt{\rho_{k_0}}+\sqrt{1-\rho_{k_0}})\right).$$

Let $U_{(t)}=\mathbf{1}_{\{y_{i(t)}\in W\}}$. Then, for any $k=1,\ldots,g$, $U_{(1)},\ldots,U_{(T)}$ is a stationary $\{0,1\}$-valued $(1,0)$-geometrically ergodic hidden Markov chain conditionally on component k. Hence, by Theorem 3,

$$\mathbb{P}\left(\mid\rho_{k_0}-\hat{\rho}_{k_0}\mid>\varepsilon+\frac{1}{2\sqrt{T}}(\sqrt{\rho_{k_0}}+\sqrt{1-\rho_{k_0}})\right)\leq e^{-\frac{T}{2}\varepsilon^2}.$$

We can conclude that

$$\mathbb{P}\left(\hat{\rho}_{k_0}<\rho_{k_0}-\delta_1\right)\leq e^{-Tc_1},$$

$\delta_1=\varepsilon+\frac{1}{\sqrt{2T}}$ and $c_1=\frac{1}{2}(\delta_1-\frac{1}{\sqrt{2T}})^2$. □

Upper-bound of the conditional probability of $\boldsymbol{y}_i$ **given** $Z_{ik}=1$ Let γ and $\bar{\gamma}$ be upper-bounds of the ratio $\frac{p(\tilde{x}_{i(t-1)},x_{i(t)},\tilde{x}_{i(t+1)},y_{i(t)}|Z_{ik}=1)}{p(\tilde{x}_{i(t-1)},\tilde{x}_{i(t)},\tilde{x}_{i(t+1)},y_{i(t)}|Z_{ik}=1)}$ when $y_{i(t)}\in W$ and $y_{i(t)}\notin W$ respectively. Thus, $\gamma=\max_k\max_{h_1,h_2,h_3,h_4}\frac{A_k[h_1,h_2]}{A_k[h_3,h_4]}$ and $\bar{\gamma}$ permit to upper bound the ratio between the likelihood computed

for any $(\boldsymbol{x}_i, \boldsymbol{y}_i)$ given $Z_{ik}=1$ and the likelihood computed with $(\tilde{\boldsymbol{x}}_{ik}, \boldsymbol{y}_i)$ given $Z_{ik}=1$. We have, if $y_{i(t)} \in W$,

$$\frac{p(\tilde{x}_{i(t-1)}, x_{i(t)}, \tilde{x}_{i(t+1)}, y_{i(t)} \mid Z_{ik}=1)}{p(\tilde{x}_{i(t-1)}, \tilde{x}_{i(t)}, \tilde{x}_{i(t+1)}, y_{i(t)} \mid Z_{ik}=1)} \leq \max_{u\in W} \max_{h_2\in E_k(u), h_{2'}\in E_k(u), h_2\neq h_{2'}} \frac{\max_{(h_1,h_3)} e_k(u;h_1,h_2,h_3)}{\min_{(h_{1'},h_{3'})\in \mathbf{e}_k(u;h_{2'})} e_k(u;h_{1'},h_{2'},h_{3'})} \leq \gamma,$$

where $\mathbf{e}_k(u;h_2) = \{(h_1,h_3) : h_2 = \operatorname{argmax} e_k(u;h_1,h_2,h_3)\}$. Moreover, we have, if $y_{i(t)} \notin W$,

$$\frac{p(\tilde{x}_{i(t-1)}, x_{i(t)}, \tilde{x}_{i(t+1)}, y_{i(t)} \mid Z_{ik}=1)}{p(\tilde{x}_{i(t-1)}, \tilde{x}_{i(t)}, \tilde{x}_{i(t+1)}, y_{i(t)} \mid Z_{ik}=1)} \leq \max_{u\notin W} \frac{\max_{h_2} \max_{(h_1,h_3)\notin \mathbf{e}_k(u;h_2)} e_k(u;h_1,h_2,h_3)}{\max_{h_2\in E_k(u)} \min_{(h_{1'},h_{3'})\in \mathbf{e}_k(u;h_{2'})} e_k(u;h_{1'},h_{2'},h_{3'}))} = \bar{\gamma}.$$

Note that $\gamma \geq 1$ and $\bar{\gamma} < 1$.

Lemma 4. *We have, for any $k = 1, \ldots, g$,*

$$\log p(\boldsymbol{y}_i \mid Z_{ik}=1) \leq \log p(\tilde{\boldsymbol{x}}_{ik}, \boldsymbol{y}_i \mid Z_{ik}=1) + T\log(\tilde{\gamma}+\bar{\gamma}) + T\hat{\rho}_{k_0} c_2 + \log\tilde{\gamma} + \log\left(2M\gamma \max_{h,\ell} \frac{\pi_{kh}}{\pi_{k\ell}}\right),$$

where $c_2 = 1 + \frac{\gamma}{1+\bar{\gamma}}$ and $\tilde{\gamma} = \max(2,\gamma)$.

Proof of Lemma 4. By definition, we have

$$p(\boldsymbol{y}_i \mid Z_{ik}=1) = p(\tilde{\boldsymbol{x}}_{ik}, \boldsymbol{y}_i \mid Z_{ik}=1) \sum_{\boldsymbol{x}\in\mathcal{X}} \frac{p(\boldsymbol{x}, \boldsymbol{y}_i \mid Z_{ik}=1)}{p(\tilde{\boldsymbol{x}}_{ik}, \boldsymbol{y}_i \mid Z_{ik}=1)}.$$

Let $B_p(\tilde{\boldsymbol{x}}_{ik}) = \{\boldsymbol{x} : \| \boldsymbol{x} - \boldsymbol{x}_{ik} \|_0 = p\}$, then

$$\sum_{\boldsymbol{x}\in\mathcal{X}} \frac{p(\boldsymbol{x}, \boldsymbol{y}_i \mid Z_{ik}=1)}{p(\tilde{\boldsymbol{x}}_{ik}, \boldsymbol{y}_i \mid Z_{ik}=1)} = \sum_{p=0}^{T+1} \sum_{\boldsymbol{x}\in B_p(\tilde{\boldsymbol{x}}_{ik})} \frac{p(\boldsymbol{x}, \boldsymbol{y}_i \mid Z_{ik}=1)}{p(\tilde{\boldsymbol{x}}_{ik}, \boldsymbol{y}_i \mid Z_{ik}=1)}.$$

Remark that

$$\frac{p(\boldsymbol{x}_{i(0)}, \boldsymbol{y}_{i(0)} \mid Z_{ik}=1, \boldsymbol{x}_{i(1)}, \boldsymbol{y}_{i(1)})}{p(\tilde{\boldsymbol{x}}_{i(0)}, \boldsymbol{y}_{i(0)} \mid Z_{ik}=1, \tilde{\boldsymbol{x}}_{i(1)}, \boldsymbol{y}_{i(1)})} < \gamma \max_{h,\ell} \frac{\pi_{kh}}{\pi_{k\ell}}.$$

Moreover, we observe $T_W = T\hat{\rho}_{k_0}$ elements of the sequence $y_{i(1)}, \ldots, y_{i(T)}$ which belongs to W. Thus, we have

$$\begin{aligned}
\sum_{\boldsymbol{x}\in\mathcal{X}} \frac{p(\boldsymbol{x}, \boldsymbol{y}_i \mid Z_{ik}=1)}{p(\tilde{\boldsymbol{x}}_{ik}, \boldsymbol{y}_i \mid Z_{ik}=1)} &\leq \left(M\gamma \max_{h,\ell} \frac{\pi_{kh}}{\pi_{k\ell}}\right) \sum_{p=0}^{T} \sum_{r=0}^{p} \binom{T_W}{\min(r, T_W)} \binom{T-T_W}{\min(u, T-T_W)} \gamma^r \bar{\gamma}^u \\
&= \left(M\gamma \max_{h,\ell} \frac{\pi_{kh}}{\pi_{k\ell}}\right) \left(\sum_{r=0}^{T_W} \binom{T_W}{r} \gamma^r \sum_{u=0}^{T-r} \binom{T-T_W}{u} \bar{\gamma}^u \right. \\
&\quad \left. + \sum_{r=1+T_W}^{T} \sum_{u=0}^{T-r} \binom{T-T_W}{\min(u, T-T_W)} \gamma^r \bar{\gamma}^u \right).
\end{aligned}$$

We have

$$\sum_{r=0}^{T_W} \binom{T_W}{r} \gamma^r \sum_{u=0}^{T-r} \binom{T-T_W}{u} \bar{\gamma}^u = (1+\bar{\gamma})^T \left(1 + \frac{\gamma}{1+\bar{\gamma}}\right)^{T_W},$$

and

$$\sum_{r=1+T_W}^{T} \sum_{u=0}^{T-r} \binom{T-T_W}{\min(u, T-T_W)} \gamma^r \bar{\gamma}^u \leq (\tilde{\gamma}+\bar{\gamma})^T \tilde{\gamma} \left(\frac{\tilde{\gamma}}{\bar{\gamma}+\tilde{\gamma}}\right)^{T_W},$$

where $\tilde{\gamma} = \max(2,\gamma)$. Noting that $1+\bar{\gamma} < \tilde{\gamma}+\bar{\gamma}$ and $1 + \frac{\gamma}{1+\bar{\gamma}} > \frac{\tilde{\gamma}}{\bar{\gamma}+\tilde{\gamma}}$, we have

$$\log p(\boldsymbol{y}_i \mid Z_{ik}=1) \leq \log p(\tilde{\boldsymbol{x}}_{ik}, \boldsymbol{y}_i \mid Z_{ik}=1) + T\log(\tilde{\gamma}+\bar{\gamma}) + T\hat{\rho}_{k_0} c_2 + \log\tilde{\gamma} + \log\left(2M\gamma \max_{h,\ell} \frac{\pi_{kh}}{\pi_{k\ell}}\right),$$

where $c_2 = 1 + \frac{\gamma}{1+\bar{\gamma}}$. Note that $\gamma + 1 > c_2 > 1$.

□

Concentration of the ratio of complete-data likelihood

Lemma 5. *For any $k \neq k_0$ and for any δ_3 such that $-\zeta < \delta_3 < u_{kk_0}$, we have*

$$\mathbb{P}_0\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{h=1}^{M}\sum_{\ell=1}^{M} v_{i(t)h\ell}\log\left(\frac{A_k[h,\ell]}{A_{k_0}[h,\ell]}\right) > \delta_3\right) \leq \exp\left(-Tc_3\right),$$

where $c_3 = 2\frac{1-\bar{\nu}_2(\boldsymbol{A}_{k_0})}{1+\bar{\nu}_2(\boldsymbol{A}_{k_0})}s^2 > 0$ and $s = \frac{\delta_3}{\omega_{kk_0}} + \frac{1}{\omega_{kk_0}}\sum_{h=1}^{M}\sum_{\ell=1}^{M}\pi_{k_0h}A_{k_0}[h,\ell]\log\left(\frac{A_{k_0}[h,\ell]}{A_k[h,\ell]}\right)$.

Proof of Lemma 5. Let $f(\cdot) \in [0,1]$ defined by

$$f(\boldsymbol{v}_{i(t)}) = \frac{1}{\omega_{kk_0}}\left(\sum_{h=1}^{M}\sum_{\ell=1}^{M} v_{i(t)h\ell}\log\left(\frac{A_k[h,\ell]}{A_{k_0}[h,\ell]}\right) + u_{k_0k}\right),$$

where $\omega_{kk_0} = u_{kk_0} + u_{k_0k}$, $u_{kk_0} = \max_{(h,\ell)}\log\left(\frac{A_k[h,\ell]}{A_{k_0}[h,\ell]}\right)$. Denoting $\mathbb{E}_0[\cdot] = \mathbb{E}[\cdot \mid Z_{ik_0} = 1]$ the conditional expectation computed with the true parameters, we have, for $t = 1, \ldots, T$,

$$\mathbb{E}_0\left[f(\boldsymbol{V}_{i(t)})\right] = \frac{1}{\omega_{kk_0}}\sum_{h=1}^{M}\sum_{\ell=1}^{M}\pi_{k_0h}A_{k_0}[h,\ell]\left(\log\left(\frac{A_k[h,\ell]}{A_{k_0}[h,\ell]}\right) + u_{k_0k}\right).$$

Therefore, we have

$$\begin{aligned}\mathbb{P}_0\left(\sum_{t=1}^{T}\sum_{h=1}^{M}\sum_{\ell=1}^{M} v_{i(t)h\ell}\log\left(\frac{A_k[h,\ell]}{A_{k_0}[h,\ell]}\right) > \delta_2\right) &= \mathbb{P}_0\left(\sum_{t=1}^{T} f(\boldsymbol{v}_{i(t)}) > \frac{\delta_2 + Tu_{k_0k}}{\omega_{kk_0}}\right)\\ &= \mathbb{P}_0\left(\sum_{t=1}^{T} f(\boldsymbol{v}_{i(t)}) > T(\mathbb{E}[f(\boldsymbol{V}_{i(1)})] + s)\right),\end{aligned}$$

where $s = \frac{\delta_2}{T\omega_{kk_0}} + \frac{1}{\omega_{kk_0}}\sum_{h=1}^{M}\sum_{\ell=1}^{M}\pi_{k_0h}A_{k_0}[h,\ell]\log\left(\frac{A_{k_0}[h,\ell]}{A_k[h,\ell]}\right)$.

Note that $\omega_{kk_0} > 0$ and that, by Assumption 3, $\sum_{h=1}^{M}\sum_{\ell=1}^{M}\pi_{k_0h}A_{k_0}[h,\ell]\log\left(\frac{A_{k_0}[h,\ell]}{A_k[h,\ell]}\right) > \zeta > 0$ because it is a weighted sum of M Kullback-Leibler divergences. Thus, if $-T\zeta < \delta_2$ then $s > 0$. Moreover, if $\delta_2 < Tu_{kk_0}$, then $\mathbb{E}[f(\boldsymbol{V}_{i(1)})] + s < 1$. Assumption 1 and Remark 1 imply that $\bar{\nu}_2(\boldsymbol{A}_{k_0})$ is the maximum between zero and the second-largest eigenvalue of reversible Markov chain of $\boldsymbol{V}_{i(t)}$. Therefore, using Theorem 4, we have for any δ_3 such that $-\zeta < \delta_3 < u_{kk_0}$,

$$\mathbb{P}_0\left(\frac{1}{T}\sum_{t=1}^{T}\sum_{h=1}^{M}\sum_{\ell=1}^{M} v_{i(t)h\ell}\log\left(\frac{A_k[h,\ell]}{A_{k_0}[h,\ell]}\right) > \delta_3\right) \leq \exp\left(-Tc_3\right),$$

where $c_3 = 2\frac{1-\bar{\nu}_2(\boldsymbol{A}_{k_0})}{1+\bar{\nu}_2(\boldsymbol{A}_{k_0})}s^2$ and $s = \frac{\delta_3}{\omega_{kk_0}} + \frac{1}{\omega_{kk_0}}\sum_{h=1}^{M}\sum_{\ell=1}^{M}\pi_{k_0h}A_{k_0}[h,\ell]\log\left(\frac{A_{k_0}[h,\ell]}{A_k[h,\ell]}\right)$. □

B.2 Proof of Theorem 2

Noting that $\mathbb{P}(Z_{ik} = 1 \mid \boldsymbol{y}_i) \propto \delta_k p(\boldsymbol{y}_i \mid Z_{ik} = 1)$ and using Lemma 4, we have

$$\begin{aligned}\mathbb{P}_0\left(\frac{\mathbb{P}(Z_{ik} = 1 \mid \boldsymbol{y}_i)}{\mathbb{P}(Z_{ik_0} = 1 \mid \boldsymbol{y}_i)} > a\right) \leq \mathbb{P}_0\Bigg(\log\frac{p(\tilde{\boldsymbol{x}}_{ik}, \boldsymbol{y}_i \mid Z_{ik} = 1)}{p(\tilde{\boldsymbol{x}}_{ik_0}, \boldsymbol{y}_i \mid Z_{ik_0} = 1)} > -\log\frac{\delta_k}{a\delta_{k_0}} - \log\left(2M\tilde{\gamma}\gamma\max_{h,\ell}\frac{\pi_{kh}}{\pi_{k\ell}}\right)\\ -T\log(\tilde{\gamma} + \bar{\gamma}) - T\hat{\rho}_{k_0}c_2\Bigg).\end{aligned}$$

Moreover,

$$\log\frac{p(\tilde{\boldsymbol{x}}_{ik}, \boldsymbol{y}_i \mid Z_{ik} = 1)}{p(\tilde{\boldsymbol{x}}_{ik_0}, \boldsymbol{y}_i \mid Z_{ik_0} = 1)} = \sum_{t=1}^{T}\left(d_{k1(t)} + d_{k2(t)}\right) + \sum_{h=1}^{M}\tilde{x}_{ik(1)h}\log\pi_{kh} - \tilde{x}_{ik_0(1)h}\log\pi_{k_0h},$$

where

$$d_{k1(t)} = \sum_{h=1}^{M}\sum_{\ell=1}^{M}(\tilde{v}_{ik(t)h\ell} - \tilde{v}_{ik_0(t)h\ell}) \log\left(\boldsymbol{A}_{k_0}[h,\ell] g_\ell(y_{i(t)};\boldsymbol{\lambda}_\ell,\varepsilon_\ell)\right),$$

and

$$d_{k2(t)} = \sum_{h=1}^{M}\sum_{\ell=1}^{M}\tilde{v}_{ik(t)h\ell} \log\frac{\boldsymbol{A}_k[h,\ell]}{\boldsymbol{A}_{k_0}[h,\ell]}.$$

Therefore, we have

$$\mathbb{P}_0\left(\frac{\mathbb{P}(Z_{ik}=1 \mid \boldsymbol{y}_i)}{\mathbb{P}(Z_{ik_0}=1 \mid \boldsymbol{y}_i)} > a\right) \leq \mathbb{P}_0\left(\frac{1}{T}\sum_{t=1}^{T}\left(d_{k1(t)} + d_{k2(t)}\right) > -\frac{c_4}{T} - \log(\tilde{\gamma}+\bar{\gamma}) - \hat{\rho}_{k_0}c_2\right),$$

with $c_4 = \log\frac{\delta_k}{a\delta_{k_0}} + \log\left(2M\tilde{\gamma}\gamma \max_{h,\ell}\frac{\pi_{kh}}{\pi_{k\ell}}\right) + \max_{k,k_0,h,\ell}\log\frac{\pi_{kh}}{\pi_{k_0\ell}}$. By definition of W, we have $\tilde{v}_{ik(t)h\ell} = \tilde{v}_{ik_0(t)h\ell}$ if $y_{i(t)} \notin W$. Moreover, because $\tilde{\boldsymbol{v}}_{ik_0}$ is the maximum *a posteriori* rule, if $y_{i(t)} \in W$, then $d_{k1(t)} < \gamma$. Therefore, we have

$$\mathbb{P}_0\left(\frac{\mathbb{P}(Z_{ik}=1 \mid \boldsymbol{y}_i)}{\mathbb{P}(Z_{ik_0}=1 \mid \boldsymbol{y}_i)} > a\right) \leq \mathbb{P}_0\left(\frac{1}{T}\sum_{t=1}^{T} d_{k2(t)} > -\frac{c_4}{T} - (\gamma+c_2)\hat{\rho}_{k_0} - \log(\tilde{\gamma}+\bar{\gamma})\right).$$

Hence, we have,

$$\mathbb{P}_0\left(\frac{\mathbb{P}(Z_{ik}=1 \mid \boldsymbol{y}_i)}{\mathbb{P}(Z_{ik_0}=1 \mid \boldsymbol{y}_i)} > a\right) \leq \mathbb{P}_0\left(\hat{\rho}_{k_0} > \rho_{k_0} + \delta_1\right) + \mathbb{P}_0\left(\frac{1}{T}\sum_{t=1}^{T} d_{k2}(t) > -\frac{c_4}{T} - \log(\tilde{\gamma}+\bar{\gamma}) - (\gamma+c_2)(\rho_{k_0}+\delta_1)\right).$$

Using Lemma 3, if $\delta_1 > \frac{1}{\sqrt{2T}}$, the first term of the right side of the previous equation can be upper bounded by e^{-Tc_1} with $c_1 = \frac{1}{2}(\delta_1 - \frac{1}{\sqrt{2T}})^2$.

Using Lemma 5, the second term of the right side of the previous equation can be upper bounded by e^{-Tc_3} with $c_3 = 2\frac{1-\bar{\nu}_2(\boldsymbol{A}_{k_0})}{1+\bar{\nu}_2(\boldsymbol{A}_{k_0})}s^2$, where $s = \frac{\delta_3}{\omega_{kk_0}} + \frac{1}{\omega_{kk_0}}\sum_{h=1}^{M}\sum_{\ell=1}^{M}\pi_{k_0h}A_{k_0}[h,\ell]\log\left(\frac{A_{k_0}[h,\ell]}{A_k[h,\ell]}\right)$ and $\delta_3 = -\frac{c_4}{T} - \log(\tilde{\gamma}+\bar{\gamma}) - (\gamma+c_2)(\rho_{k_0}+\delta_1)$, if δ_3 is such that $-\zeta < \delta_3 < u_{kk_0}$. Thus, we have the following condition on δ_1

$$\frac{\zeta - \frac{c_4}{T} - \log(\tilde{\gamma}+\bar{\gamma})}{\gamma+c_2} - \rho_{k_0} > \delta_1 > -\frac{u_{kk_0} + \frac{c_4}{T} + \log(\tilde{\gamma}+\bar{\gamma})}{\gamma+c_2} - \rho_{k_0}.$$

Noting that $\gamma + c_2 < 1 + 2\gamma$, the previous upper bound can be satisfied under the following assumption

Assumption 4. *It holds that*

$$\frac{\zeta - c_4 - \log(\tilde{\gamma}+\bar{\gamma})}{1+2\gamma} - \rho_{k_0} - \frac{1}{\sqrt{2}} > 0,$$

with $c_4 = \log\frac{\delta_k}{a\delta_{k_0}} + \log\left(2M\tilde{\gamma}\gamma\max_{h,\ell}\frac{\pi_{kh}}{\pi_{k\ell}}\right) + \max_{k,k_0,h,\ell}\log\frac{\pi_{kh}}{\pi_{k_0\ell}}$.

Thus, for any a such that Assumption 4 holds and for any δ_1 with $\frac{1}{\sqrt{2T}} < \delta_1 < \frac{\zeta - c_4 - \log(\tilde{\gamma}+\bar{\gamma})}{\gamma+c_2} - \rho_{k_0}$, we have

$$\mathbb{P}_0\left(\hat{\rho}_{k_0} > \rho_{k_0} + \delta_1\right) \leq \mathcal{O}(e^{-Tc_1}),$$

and

$$\mathbb{P}_0\left(\frac{1}{T}\sum_{t=1}^{T} d_{k2}(t) > \delta_3\right) \leq \mathcal{O}(e^{-Tc_3}),$$

with $\delta_3 = -\frac{c_4}{T} - \log(\tilde{\gamma}+\bar{\gamma}) - (\gamma+c_2)(\rho_{k_0}+\delta_1)$. Therefore, there exists $c > 0$ such that

$$\mathbb{P}_0\left(\frac{\mathbb{P}(Z_{ik}=1 \mid \boldsymbol{y}_i)}{\mathbb{P}(Z_{ik_0}=1 \mid \boldsymbol{y}_i)} > a\right) \leq \mathcal{O}(e^{-Tc}).$$

If the misclassification error is studied, we should consider $a = 1$. Then, a sufficient condition to have the exponential decreasing of the probability of misclassifying an observation is obtained on the basis of Assumption 4 with $a = 1$.

C Details about the conditional distribution

Forward formula We define

$$\alpha_{ikhs(t)}(\boldsymbol{\theta}) = \mathbb{P}(X_{is(t)} = h \mid Z_{ik} = 1; \boldsymbol{\theta})\, p(y_{is(0)}, \ldots, y_{is(t)} \mid X_{is(t)} = h, Z_{ik} = 1; \boldsymbol{\theta}),$$

which measures the probability of the partial sequence $y_{is(0)}, \ldots, y_{is(t)}$ and ending up in state h at time t under component k. For any (i, k, h, s), we can define $\alpha_{ikhs(t)}$ recursively, as follows,

$$\alpha_{ikhs()}(\boldsymbol{\theta}) = \pi_{kh}\, p(y_{is()}; \boldsymbol{\lambda}_h)$$
$$\forall t \in \{0, \ldots, T_{is} - 1\} \quad \alpha_{ik\ell s(t+1)}(\boldsymbol{\theta}) = \left(\sum_{h=1}^{M} A_k[h, \ell] \alpha_{ikhs(t)}(\boldsymbol{\theta}) \right) p(y_{is(t+1)}; \boldsymbol{\lambda}_h).$$

Considering independence between the S_i sequences $\boldsymbol{y}_{is}$, the pdf of $\boldsymbol{y}_i$ under component k is

$$p(\boldsymbol{y}_i \mid Z_{ik} = 1; \boldsymbol{\theta}) = \prod_{s=1}^{S_i} \sum_{h=1}^{M} \alpha_{ikhs(T_{is})}(\boldsymbol{\theta}).$$

Therefore,

$$p(\boldsymbol{y}_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \delta_k \left(\prod_{s=1}^{S_i} \sum_{h=1}^{M} \alpha_{ikhs(T_{is})}(\boldsymbol{\theta}) \right).$$

Backward formula We define

$$\beta_{ikhs(t)}(\boldsymbol{\theta}) = p(y_{is(t+1)}, \ldots, y_{is(T_{is})} \mid X_{is(t)} = h, Z_{ik} = 1; \boldsymbol{\theta}),$$

which measures the probability of the ending partial sequence $y_{is(t+1)}, \ldots, y_{is(T_{is})}$ given a start in state h at time t under component k. We can define $\beta_{ikhs(t)}(\boldsymbol{\theta})$ recursively, for any (i, k, h, s), as

$$\beta_{ikhs(T_{is})}(\boldsymbol{\theta}) = 1$$
$$\forall t \in \{0, \ldots, T_{is} - 1\} \quad \beta_{ikhs(t)}(\boldsymbol{\theta}) = \sum_{\ell=1}^{M} A_k[h, \ell] p(y_{i(t+1)}; \boldsymbol{\lambda}_\ell) \beta_{ik\ell s(t+1)}(\boldsymbol{\theta}).$$

Considering independence between the S_i sequences $\boldsymbol{y}_{is}$, the pdf of $\boldsymbol{y}_i$ under component k is

$$p(\boldsymbol{y}_i \mid Z_{ik} = 1; \boldsymbol{\theta}) = \prod_{s=1}^{S_i} \sum_{h=1}^{M} \pi_{kh} \beta_{ikhs(0)}(\boldsymbol{\theta}) p(y_{i(0)}; \boldsymbol{\lambda}_h).$$

$$p(\boldsymbol{y}_i; \boldsymbol{\theta}) = \sum_{k=1}^{K} \delta_k \left(\prod_{s=1}^{S_i} \sum_{h=1}^{M} \pi_{kh} \beta_{ikhs(0)}(\boldsymbol{\theta}) p(y_{i(0)}; \boldsymbol{\lambda}_h) \right).$$

References

Ae Lee, J. and Gill, J. (2018). Missing value imputation for physical activity data measured by accelerometer. *Stat. Methods Med. Res.*, 27(2):490–506.

Allman, E., Matias, C., and Rhodes, J. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, 37(6A):3099–3132.

Altman, R. M. (2007). Mixed hidden markov models: an extension of the hidden markov model to the longitudinal data setting. *J. Am. Statist. Assoc.*, 102(477):201–210.

Bai, J., Sun, Y., Schrack, J. A., Crainiceanu, C. M., and Wang, M.-C. (2018). A two-stage model for wearable device data. *Biometrics*, 74(2):744–752.

Banfield, J. and Raftery, A. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, pages 803–821.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(7):719–725.

Brault, V. and Mariadassou, M. (2015). Co-clustering through latent bloc model: A review. *Journal de la Société Française de Statistique*, 156(3):120–139.

Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in hidden Markov models*. Springer Series in Statistics. Springer, New York. With Randal Douc's contributions to Chapter 9 and Christian P. Robert's to Chapters 6, 7 and 13, With Chapter 14 by Gersende Fort, Philippe Soulier and Moulines, and Chapter 15 by Stéphane Boucheron and Elisabeth Gassiat.

Celeux, G. and Durand, J.-B. (2008). Selecting hidden Markov model state number with cross-validated likelihood. *Comput. Stat.*, 23(4):541–564.

Celisse, A., Daudin, J.-J., Pierre, L., et al. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electron. J. Stat.*, 6:1847–1899.

Cole, R. J., Kripke, D. F., Gruen, W., Mullaney, D. J., and Gillin, J. C. (1992). Automatic sleep/wake identification from wrist activity. *Sleep*, 15(5):461–469.

Csiszr, I. and Talata, Z. (2006). Consistent estimation of the basic neighborhood of markov random fields. *Ann. Statist.*, 34(1):123–145.

Dempster, A.P. and Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Stat. Soc. B*, 39(1):1–38.

Freedson, P. S., Melanson, E., and Sirard, J. (1998). Calibration of the computer science and applications, inc. accelerometer. *Med. Sci. Sports Exerc.*, 30(5):777–781.

Gassiat, E. (2002). Likelihood ratio inequalities with applications to various mixtures. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 38, pages 897–906. Elsevier.

Gassiat, E., Cleynen, A., and Robin, S. (2016). Inference in finite state space non parametric hidden Markov models and applications. *Stat. Comput.*, 26(1-2):61–71.

Geraci, M. (2018). Additive quantile regression for clustered data with an application to children's physical activity. *J. Royal Stat. Soc. C*.

Geraci, M. and Farcomeni, A. (2016). Probabilistic principal component analysis to identify profiles of physical activity behaviours in the presence of non-ignorable missing data. *J. Royal Stat. Soc. C*, 65(1):51–75.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, 61(2):215–231.

Grandner, M. A., Sands-Lincoln, M. R., Pak, V. M., and Garland, S. N. (2013). Sleep duration, cardiovascular disease, and proinflammatory biomarkers. *Nature and science of sleep*, 5:93.

Gruen, M. E., Alfaro-Córdoba, M., Thomson, A. E., Worth, A. C., Staicu, A.-M., and Lascelles, B. D. X. (2017). The use of functional data analysis to evaluate activity in a spontaneous model of degenerative joint disease associated pain in cats. *PloS one*, 12(1):e0169576.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *J. Am. Statist. Assoc.*, 97(460):1090–1098.

Huang, L., Bai, J., Ivanescu, A., Harris, T., Maurer, M., Green, P., and Zipunnikov, V. (2018a). Multilevel matrix-variate analysis and its application to accelerometry-measured physical activity in clinical populations. *J. Am. Statist. Assoc.*, pages 1–12.

Huang, Q., Cohen, D., Komarzynski, S., Li, X.-M., Innominato, P., Lévi, F., and Finkenstädt, B. (2018b). Hidden markov models for monitoring circadian rhythmicity in telemetric activity data. *J. Royal Soc. Interface*, 15(139):20170885.

Hubert, L. and Arabie, P. (1985). Comparing partitions. *Journal of classification*, 2(1):193–218.

Hunt, L. and Jorgensen, M. (2011). Clustering mixed data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(4):352–361.

Hunter, D. R., Goodreau, S. M., and Handcock, M. S. (2008). Goodness of fit of social network models. *J. Am. Statist. Assoc.*, 103(481):248–258.

Innerd, P., Harrison, R., and Coulson, M. (2018). Using open source accelerometer analysis to assess physical activity and sedentary behaviour in overweight and obese adults. *BMC public health*, 18(1):543.

Karlis, D. and Meligkotsidou, L. (2007). Finite mixtures of multivariate poisson distributions with application. *J. Stat. Plan. Inference*, 137(6):1942–1960.

Kimm, S. Y., Glynn, N. W., Obarzanek, E., Kriska, A. M., Daniels, S. R., Barton, B. A., and Liu, K. (2005). Relation between the changes in physical activity and body-mass index during adolescence: a multicentre longitudinal study. *The Lancet*, 366(9482):301–307.

Kontorovich, A. and Weiss, R. (2014). Uniform Chernoff and Dvoretzky-Kiefer-Wolfowitz-type inequalities for Markov chains and related processes. *J. Appl. Probab. Stat.*, 51(4):1100–1113.

Kosmidis, I. and Karlis, D. (2015). Model-based clustering using copulas with applications. *Stat .Comput.*, pages 1–21.

Lee, I.-M., Shiroma, E. J., Lobelo, F., Puska, P., Blair, S. N., Katzmarzyk, P. T., Group, L. P. A. S. W., et al. (2012). Effect of physical inactivity on major non-communicable diseases worldwide: an analysis of burden of disease and life expectancy. *The lancet*, 380(9838):219–229.

León, C. A. and Perron, F. (2004). Optimal Hoeffding bounds for discrete reversible Markov chains. *Ann. Appl. Probab.*, 14(2):958–970.

Levin, D. A. and Peres, Y. (2017). *Markov chains and mixing times*. American Mathematical Society, Providence, RI. Second edition of [MR2466937], With contributions by Elizabeth L. Wilmer, With a chapter on "Coupling from the past" by James G. Propp and David B. Wilson.

Lim, Y., Oh, H.-S., and Cheung, Y. K. (2019). Functional clustering of accelerometer data via transformed input variables. *J. Royal Stat. Soc. C*, 68(3):495–520.

Matias, C., Rebafka, T., and Villers, F. (2018). A semiparametric extension of the stochastic block model for longitudinal networks. *Biometrika*, 105(3):665–680.

McLachlan, G. and Peel, D. (2000). *Finite mixture models*. Wiley Series in Probability and Statistics: Applied Probability and Statistics. Wiley-Interscience, New York.

McNicholas, P. (2016). *Mixture Model-Based Classification*. Chapman & Hall/CRC Press, Boca Raton, FLk.

McTiernan, A. (2008). Mechanisms linking physical activity with cancer. *Nature Reviews Cancer*, 8(3):205.

Morris, J. S., Arroyo, C., Coull, B. A., Ryan, L. M., Herrick, R., and Gortmaker, S. L. (2006). Using wavelet-based functional mixed models to characterize population heterogeneity in accelerometer profiles: a case study. *J. Am. Stat. Assoc.*, 101(476):1352–1364.

Noel, S. E., Mattocks, C., Emmett, P., Riddoch, C. J., Ness, A. R., and Newby, P. (2010). Use of accelerometer data in prediction equations for capturing implausible dietary intakes in adolescents. *The American journal of clinical nutrition*, 92(6):1436–1445.

Palta, P., McMurray, R., Gouskova, N., Sotres-Alvarez, D., Davis, S., Carnethon, M., Castaeda, S., Gellman, M., Hankinson, A., Isasi, C., Schneiderman, N., Talavera, G., and Evenson, K. (2015). Self-reported and accelerometer-measured physical activity by body mass index in us hispanic/latino adults: Hchs/sol. *Preventive Medicine Reports*, 2:824 – 828.

Pollak, C. P., Tryon, W. W., Nagaraja, H., and Dzwonczyk, R. (2001). How accurately does wrist actigraphy identify the states of sleep and wakefulness? *Sleep*, 24(8):957–965.

Sadeh, A., Sharkey, M., and Carskadon, M. A. (1994). Activity-based sleep-wake identification: an empirical test of methodological issues. *Sleep*, 17(3):201–207.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.*, 6(2):461–464.

Scott, S. L., James, G. M., and Sugar, C. A. (2005). Hidden markov models for longitudinal comparisons. *J. Am. Statist. Assoc.*, 100(470):359–369.

Taheri, S., Lin, L., Austin, D., Young, T., and Mignot, E. (2004). Short sleep duration is associated with reduced leptin, elevated ghrelin, and increased body mass index. *PLoS medicine*, 1(3):e62.

Teicher, H. (1963). Identifiability of finite mixtures. *Ann. Math. Stat.*, pages 1265–1269.

Teicher, H. (1967). Identifiability of mixtures of product measures. *Ann. Math. Stat.*, 38:1300–1302.

Titsias, M. K., Holmes, C. C., and Yau, C. (2016). Statistical inference in hidden markov models using k-segment constraints. *EEE Trans. Inf. Theory*, 111(513):200–215.

Van Hees, V. T., Sabia, S., Anderson, K. N., Denton, S. J., Oliver, J., Catt, M., Abell, J. G., Kivimäki, M., Trenell, M. I., and Singh-Manoux, A. (2015). A novel, open access method to assess sleep duration using a wrist-worn accelerometer. *PloS one*, 10(11):e0142533.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *J. Am. Statist. Assoc.*, 13(2):260–269.

Wallace, M. L., Buysse, D. J., Germain, A., Hall, M. H., and Iyengar, S. (2018). Variable selection for skewed model-based clustering: application to the identification of novel sleep phenotypes. *J. Am. Statist. Assoc.*, 113(521):95–110.

Witowski, V., Foraita, R., Pitsiladis, Y., Pigeot, I., and Wirsik, N. (2014). Using hidden markov models to improve quantifying physical activity in accelerometer data–a simulation study. *PloS one*, 9(12):e114089.

Wong, C. S. and Li, W. K. (2000). On a mixture autoregressive model. *J. Royal Stat. Soc. B*, 62(1):95–115.

Xiao, L., Huang, L., Schrack, J. A., Ferrucci, L., Zipunnikov, V., and Crainiceanu, C. M. (2014). Quantifying the lifetime circadian rhythm of physical activity: a covariate-dependent functional approach. *Biostatistics*, 16(2):352–367.

Yang, C.-C. and Hsu, Y.-L. (2010). A review of accelerometry-based wearable motion detectors for physical activity monitoring. *Sensors*, 10(8):7772–7788.

Case hard ($e = 0.75$ and $a_2 = 3$)

		Adjusted Rand index		Mean square error				
n	T	partition	states	$\boldsymbol{A}_k$	ε_h	a_h	b_h	δ_k
10	100	0.805	0.329	0.092	0.002	0.263	0.064	0.061
10	500	0.998	0.359	0.031	0.000	0.070	0.022	0.050
100	100	0.872	0.348	0.030	0.000	0.082	0.027	0.006
100	500	1.000	0.359	0.019	0.000	0.044	0.018	0.005

Case medium-2 ($e = 0.75$ and $a_2 = 5$)

		Adjusted Rand index		Mean square error				
n	T	partition	states	$\boldsymbol{A}_k$	ε_h	a_h	b_h	δ_k
10	100	0.992	0.605	0.025	0.001	0.086	0.025	0.050
10	500	1.000	0.614	0.010	0.000	0.018	0.005	0.054
100	100	0.997	0.612	0.004	0.000	0.011	0.003	0.005
100	500	1.000	0.615	0.002	0.000	0.005	0.001	0.005

Case easy ($e = 0.90$ and $a_2 = 5$)

		Adjusted Rand index		Mean square error				
n	T	partition	states	$\boldsymbol{A}_k$	ε_h	a_h	b_h	δ_k
10	100	1.000	0.827	0.009	0.000	0.159	0.016	0.048
10	500	0.999	0.831	0.011	0.000	0.030	0.003	0.048
100	100	1.000	0.829	0.001	0.000	0.017	0.002	0.005
100	500	1.000	0.831	0.000	0.000	0.003	0.000	0.005

Table 7: Convergence of the maximum likelihood estimator when data are sample in the other three cases considered.

				Adjusted Rand index		Mean square error				
n	T	s	q	partition	states	$\boldsymbol{A}_k$	ε_h	a_h	b_h	δ_k
		0	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		1	10	0.927	0.977	1.153	1.315	1.307	1.229	1.059
10	100	2	10	0.865	0.955	1.342	1.411	1.355	1.346	1.142
		1	20	0.845	0.954	1.308	1.703	1.230	1.312	1.129
		2	20	0.705	0.906	1.813	2.565	2.462	2.336	1.389
		0	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		1	10	1.001	1.001	1.003	1.120	1.041	0.959	1.009
10	500	2	10	1.001	0.997	1.039	1.074	1.031	1.025	1.006
		1	20	1.002	0.997	1.071	1.152	1.006	1.036	1.014
		2	20	1.002	0.999	1.005	1.064	1.006	1.001	1.014
		0	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		1	10	0.963	0.989	1.123	1.177	1.063	1.088	1.123
100	100	2	10	0.919	0.977	1.159	1.308	1.179	1.194	1.158
		1	20	0.915	0.975	1.196	1.299	1.178	1.204	1.219
		2	20	0.813	0.953	1.280	1.647	1.439	1.330	1.514
		0	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		1	10	1.000	0.999	0.990	1.091	0.918	1.014	1.000
100	500	2	10	1.000	1.002	0.942	0.987	1.027	0.948	1.000
		1	20	1.000	1.002	0.938	1.095	1.072	0.966	1.000
		2	20	1.000	1.001	0.912	1.042	0.996	0.967	0.999

Table 8: Ratio between the statistics obtained with missingness and without missingness when data are sample from case hard.

				Adjusted Rand index		Mean square error				
n	T	s	q	partition	states	$\boldsymbol{A}_k$	ε_h	a_h	b_h	δ_k
		0	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		1	10	0.997	0.987	1.074	1.103	1.133	1.144	1.004
10	100	2	10	0.992	0.978	1.298	1.289	1.325	1.303	1.011
		1	20	0.986	0.982	1.320	1.286	1.333	1.320	1.006
		2	20	0.964	0.958	1.870	1.727	1.759	1.783	1.005
		0	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		1	10	1.000	0.999	1.053	1.020	1.062	0.987	1.000
10	500	2	10	0.999	0.997	1.199	1.026	1.101	1.011	0.996
		1	20	1.000	0.999	0.963	1.052	1.080	1.043	1.000
		2	20	1.000	0.997	1.047	1.065	1.170	1.052	1.000
		0	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		1	10	0.997	0.992	1.084	1.160	1.058	1.040	0.995
100	100	2	10	0.993	0.984	1.184	1.323	1.215	1.205	0.997
		1	20	0.993	0.989	1.187	1.316	1.277	1.191	0.995
		2	20	0.974	0.974	1.478	1.723	1.492	1.479	1.003
		0	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		1	10	1.000	0.999	0.970	1.069	0.969	1.061	1.000
100	500	2	10	1.000	0.998	0.995	1.050	1.011	1.074	1.000
		1	20	1.000	0.999	1.016	1.054	0.972	1.081	1.000
		2	20	1.000	0.997	1.016	1.116	0.956	1.045	1.000

Table 9: Ratio between the statistics obtained with missingness and without missingness when data are sample from case medium-2.

				Adjusted Rand index		Mean square error				
n	T	s	q	partition	states	$\boldsymbol{A}_k$	ε_h	a_h	b_h	δ_k
		0	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		1	10	1.000	0.998	1.104	1.120	1.150	1.128	1.000
10	100	2	10	1.000	0.995	1.227	1.254	1.277	1.243	0.999
		1	20	1.000	0.996	1.238	1.233	1.253	1.273	1.000
		2	20	1.000	0.991	1.682	1.557	1.572	1.573	1.000
		0	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		1	10	0.999	1.000	0.986	1.028	1.015	1.024	0.997
10	500	2	10	0.999	0.999	0.986	1.045	1.035	1.041	0.992
		1	20	0.999	0.999	0.977	1.037	1.022	1.032	0.999
		2	20	0.999	0.999	0.983	1.065	1.064	1.068	0.999
		0	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		1	10	1.000	0.997	1.119	1.125	1.127	1.118	1.000
100	100	2	10	1.000	0.994	1.280	1.263	1.248	1.237	1.000
		1	20	1.000	0.996	1.246	1.279	1.260	1.242	1.000
		2	20	1.000	0.991	1.593	1.608	1.560	1.524	1.000
		0	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000
		1	10	1.000	0.999	1.033	1.027	1.033	1.035	1.000
100	500	2	10	1.000	0.999	1.055	1.046	1.060	1.070	1.000
		1	20	1.000	0.999	1.053	1.047	1.052	1.030	1.000
		2	20	1.000	0.999	1.108	1.088	1.102	1.096	1.000

Table 10: Ratio between the statistics obtained with missingness and without missingness when data are sample from case easy.

D Supplementary material

D.1 Supplementary tables for the analysis of simulated data

D.2 Supplementary figures for the analysis of the accelerometer data of Huang et al. (2018b)

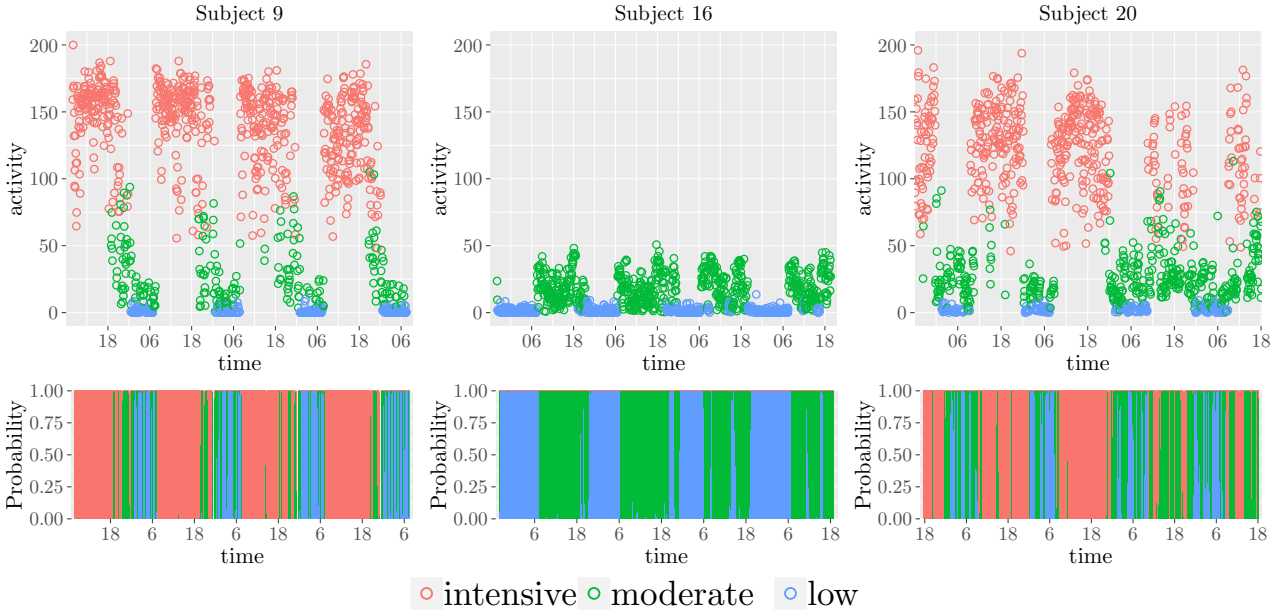


Figure 5: State estimation for the three subjects: (top) accelerometer data where color indicates the expected value of $Y_{i(t)}$ conditionally the most likely state and on the most likely component; (bottom) probability of each state at each time.

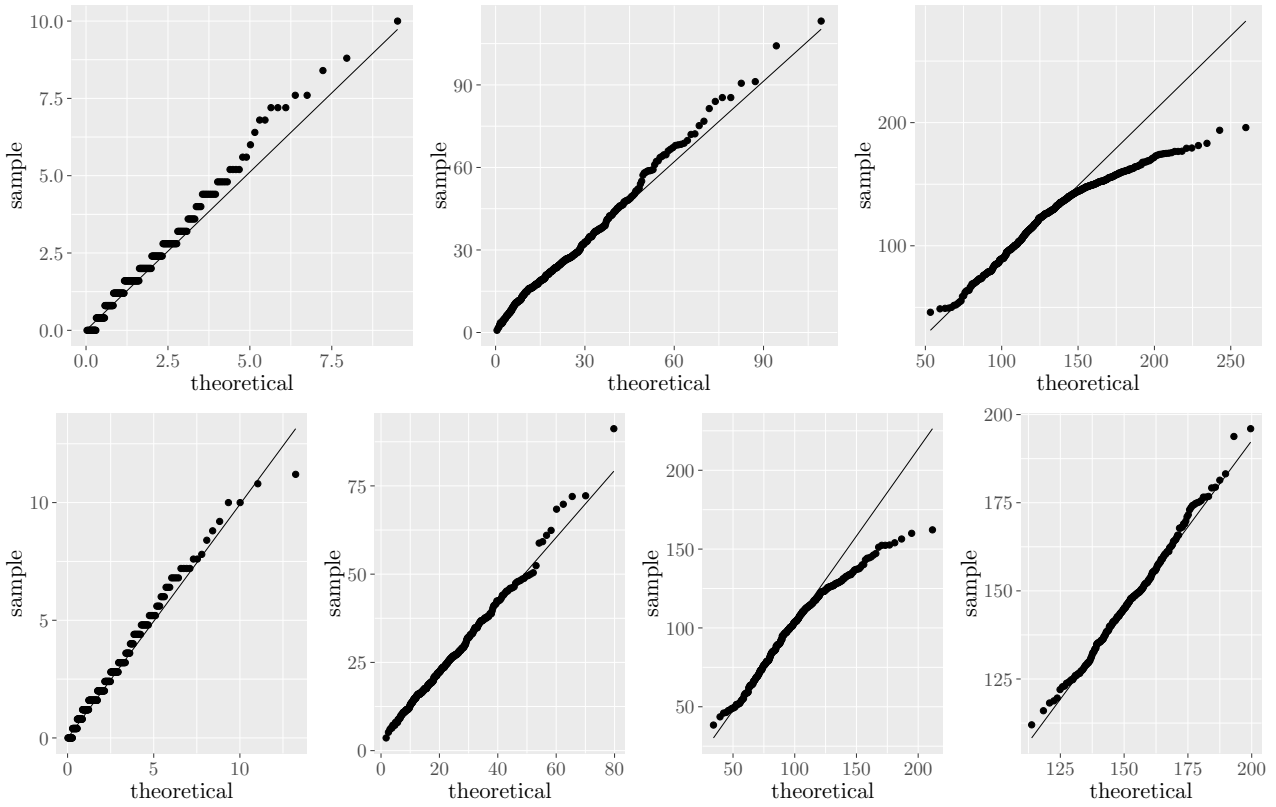


Figure 6: QQ-plots of subject 20. Top: from left to right, state 1 to 3. Bottom: Top: from left to right, state 1 to 4.