

MIXTURE OF HIDDEN MARKOV MODELS FOR ACCELEROMETER DATA

BY MARIE DU ROY DE CHAUMARAY^{*}, MATTHIEU MARBAC[†] AND FABIEN NAVARRO[‡]

CREST, ENSAI, ^{}marie.du-roy-de-chaumaray@ensai.fr; [†]matthieu.marbac-lourdelle@ensai.fr; [‡]fabien.navarro@ensai.fr*

Motivated by the analysis of accelerometer data taken across a population of individuals, we introduce a specific finite mixture of hidden Markov models with particular characteristics that adapt well to the specific nature of this type of longitudinal data. Our model allows for the computation of statistics that 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 activity levels) without specifying the activity levels in advance but by estimating them from the data. In addition, this approach allows the heterogeneity of the population to be taken into account and subpopulations with homogeneous physical activity behavior to be defined. We prove that, under mild assumptions, this model implies that the probability of misclassifying a subject decreases at an exponential decay with the length of its measurement sequence. Model identifiability is also investigated. We also report a comprehensive suite of numerical simulations to support our theoretical findings. The method is motivated by and applied to the Physical Activity and Transit Survey.

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, Grandner et al. (2013), Kimm et al. (2005), Lee et al. (2012), McTiernan (2008), Taheri et al. (2004)). 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, since the introduction in 2003 of the first objective assessment of physical activity using accelerometers, as part of the National Health and Nutrition Examination Survey (NHANES), the analysis of actigraphy data has been the subject of extensive studies over the past two decades. Recently, the New York City (NYC) Department of Health and Mental Hygiene conducted the 2010–2011 Physical Activity and Transit (PAT) Survey,¹ a random survey of adult New Yorkers that tracked levels of sedentary behavior and physical activity at work, at home and for leisure. A subset of interviewees was also invited to participate in a follow-up study to measure objectively their activity level using an accelerometer. One of the objectives of this study is to describe measured physical activity levels and to compare estimates of adherence to recommended physical activity levels, as assessed by accelerometer, with those from selfreports. In contrast to NHANES accelerometer data, PAT data still seem relatively unexplored in the statistical literature.

This paper is motivated by the analysis of the accelerometer data worn by 133 individuals over 65 years of age who responded to the PAT survey. Our objective is to propose a

Received March 2020; revised July 2020.

Key words and phrases. Accelerometer data, hidden Markov model, longitudinal data, missing data, mixture models.

¹NYC Department of Health and Mental Hygiene. Physical Activity and Transit Survey 2010–2011; public use datasets accessed on May 10, 2019. The data are freely accessible on this page: <https://www1.nyc.gov/site/doh/data/data-sets/physical-activity-and-transit-survey-public-use-data.page>.

model adapted to the specificities of these data and study its properties (positive and time-dependent measures, occurrence of missingness, heterogeneous population, need to estimate activity levels whose the definition is the same for all the individuals). In order to motivate the development of a new model, we present an overview of the literature on accelerometer data analysis.

Previous studies for analyzing accelerometer data have focused on automatic detection of the sleep and wake-up periods (Cole et al. (1992), Pollak et al. (2001), Sadeh, Sharkey and Carskadon (1994), 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 epidemiological studies, time spent by activity level is often used as a covariate in predictive models; see, for instance, the works of Innerd, Harrison and Coulson (2018), Noel et al. (2010), Palta et al. (2015) where the links between physical activity and obesity are investigated. These statistics can be computed using deterministic cutoff levels (Freedson, Melanson and Sirard (1998)). However, with such an approach the dependency in time is neglected and the cutoff levels are prespecified and not estimated from the data.

Accelerometer data are characterized by a time dependency between the different measures. They can be analyzed by methods developed for functional data or by hidden Markov models (HMM). Methods for functional data need the observed data to be converted into a function of time (Gruen et al. (2017), Morris et al. (2006), Xiao et al. (2015)). For instance, Morris et al. (2006) use wavelet basis for analyzing accelerometer profiles. The use of a basis function 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.

When considering a discrete latent variable to model time dependence, HMM are appropriate for adjusting sequence data (Altman (2007), Gassiat, Cleynen and Robin (2016), Scott, James and Sugar (2005)). Titsias, Holmes and Yau (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. HMM are used on activity data for monitoring circadian rhythmicity (Huang et al. (2018)) or, directly, for estimating the sequence of activity levels from accelerometer data (Witowski et al. (2014)). For simulated data, Witowski et al. (2014) established the superiority of different HMM models, in terms of classification error, over traditional methods based on a priori fixed thresholds. While the simplicity of implementing threshold-based methods is an obvious advantage, they have some significant disadvantages compared to the HMM methods, particularly for real data. Indeed, the variation in counts and the resulting dispersion is large, leading to considerable misclassification of counts recorded in erroneous activity ranges. The approach of Witowski et al. (2014) assumes homogeneity of the population and does not consider missingness within the observations. However, heterogeneity in physical activity behaviors is often present (see, for instance, Geraci (2019)) and the use of more than one HMM allows it to be taken into account (see, e.g., Van de Pol and Langeheine (1990)). 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 (2017)) permit 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 (2016)), network (Hoff, Raftery and Handcock (2002), Hunter, Goodreau and Handcock (2008), Matias, Rebafka and Villers (2018)) and sequence data (Wong and Li (2000)). 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. (2019) perform a matrix-variate-based clustering on accelerometer data while Lim, Oh and Cheung (2019) use a clustering

technique designed for functional data. Mixed hidden Markov models (MHMM) are a combination of HMM and generalized linear mixed models (Bartolucci, Farcomeni and Pennoni (2013), Van de Pol and Langeheine (1990)). These models consider one (or more) random effect(s) coming from either a continuous distribution (Altman (2007)) or a discrete distribution (Bartolucci, Pennoni and Vittadini (2011), Maruotti (2011)). Note that a MHMM with a single discrete random effect distribution, having a finite number of states, is a finite mixture of HMM. Such a model allows one to estimate a partition among the population and to consider the population heterogeneity. The impact of the random effect can be on the measurement model or on the latent model.

This paper focuses on the analysis of PAT data with a two-fold objective: obtain summary statistics about physical activity of the subjects without prespecifying cutoff levels, and obtain a partition which groups subjects into homogeneous classes. We define a class as homogeneous if its subjects have similar average times spent in the different activity levels and similar transition probabilities between activity levels. To achieve this goal, we introduce a specific finite mixture of HMM for analyzing accelerometer data. This model considers two latent variables: a categorical variable indicating each subject's class membership and a sequence of categorical variables indicating the subject's level of activity each time its acceleration is measured. At time t the measure is independent of the class membership conditionally on the activity level (i.e., the latent state) and follows a zero-inflated distribution—a distribution that allows for frequent zero-valued observations. The activity level defines the parameter of this distribution. The use of zero-inflated distribution is quite common for modeling accelerometer data (Bai et al. (2018), Lee and Gill (2018)), as the acceleration is measured every second, many observations are zero. Note that the definitions of the activity levels are equal among the mixture components. This is an important point for the use of the summary statistics (e.g., time spent at different activity levels, probabilities of transition between levels) in a future statistical study. The model we consider is thus a specific MHMM with a finite-states random effect that only impacts the distribution of latent physical activity levels. MHMM with a finite-states random effect have few developments in the literature (Bartolucci, Pennoni and Vittadini (2011), Maruotti (2011)), especially when the random effects only impact the latent model (and not the measurement model). We propose to theoretically study the model properties by showing that the probability of misclassifying an observation decreases at an exponential rate. In addition, since the distribution given the latent state is itself a bi-component mixture (due to the use of zero-inflated distributions), we investigate sufficient conditions for model identifiability.

In practice, the data collected often include missing intervals due to noncompliance by participants (e.g., if the accelerometer is removed). Geraci and Farcomeni (2016) propose to identify different profiles of physical activity behaviors using a principal component analysis that allows for missing values. The PAT data contain three types of missing values corresponding to periods when the accelerometer is removed, making statistical analysis more challenging. First, missingness occurs at the beginning and at the end of the measured sequences due to the installation and the removing of the accelerometer. Second, subjects are asked to remove the accelerometer when they sleep at night. Third, missing values appear during the day (e.g., due to a shower period, nap, ...). We remove missing values which occur at the begin and at the end of the sequence. For missingness caused by night sleep, we consider that the different sequences describing different days of observations of a subject are independent and that the starting point (e.g., first observed measure of the accelerometer of the day) is drawn from the stationary distribution. For missing values measured during the day, the model and the estimation algorithm can handle these data. Moreover, we propose an approximation of the distribution that avoids the computation of large powers of the transition matrices in the algorithm used for parameter inference and thus reduces computation time. Theoretical guarantees and numerical experiments show the relevance of our proposition.

The R package **MHMM**, which implements the method introduced in this paper, is available on CRAN (Du Roy de Chaumaray, Marbac and Navarro (2020a)). It permits to analyze other accelerometer data, and, thus, it is complementary to existing packages for MHMM. Indeed, it takes into account the specificities of accelerometer data (the class membership only impacts the transition matrices; the emission distributions are zero-inflated gamma (ZIG) distributions). Among the R packages implementing MHMM methods, one can cite the R packages **LMest** (Bartolucci, Pandolfi and Pennoni (2017)) and **seqHMM** (Helske and Helske (2019)), which focus on univariate longitudinal categorical data, and the R package **mHMMbayes** (Aarts (2019)) which focuses on multivariate longitudinal categorical data.

This paper is organized as follows. Section 2 presents the PAT data and the context of the study. Section 3 introduces our specific mixture of HMM and its justification in the context of accelerometer data analysis. Section 4 presents the model properties (model identifiability, exponential decay of the probabilities of misclassification and a result for dealing with the nonwear periods). Section 5 discusses the maximum likelihood inference and Section 6 illustrates the model properties on both simulated and real data. Section 7 illustrates the approach by analyzing a subset of the PAT accelerometer data. Section 8 discusses some future developments. Proofs and technical lemmas are postponed in Supplementary Material, Du Roy de Chaumaray, Marbac and Navarro (2020b).

2. PAT data description. We consider a subset of the data from the PAT survey, the subjects who participated in the follow-up study to measure objectively their activity level using an accelerometer. A detailed methodological description of the study and an analysis of the data is provided in Immerwahr et al. (2012). Note that the protocols for accelerometer data for the PAT survey and NHANES were identical. One of the objectives of the PAT study is to investigate the relationships between selfreported physical activity and physical activity measured by the accelerometer in order to provide best practice recommendations for the use of selfreported data Wyker et al. (2013). Indeed, selfreported data may be subject to overreporting. This is particularly the case among less active people, due in particular to a social desirability bias or the cognitive challenge associated with estimating the frequency and duration of daily physical activity (see, e.g., Dyrstad et al. (2014), Lim et al. (2015), Slootmaker et al. (2009)). The results of Wyker et al. (2013) show that males tend to underreport their physical activity, while females and older adults (65 years and older) overreported (see also Troiano et al. (2008) for a detailed study of the differences between selfreported physical activity and accelerometer measurements in NHANES 2003–2004). Consequently, the study of data measured by accelerometer for these specific populations makes it possible to determine methods for correcting estimates from selfreported data, such as stratification by gender and/or age when comparing groups.

In this work we are particularly interested in the age category above 65 years old ($n = 133$). We present some characteristics related to PAT data and refer to Immerwahr et al. (2012) for a full description.² Accelerometers were worn for one week (beginning on Thursday and ending on Wednesday) and measured the activity minute-by-minute. The trajectory associated with each subject is therefore of length 10,080. In addition, a participant's day spans from 3 a.m.–3 a.m. (and not a calendar day) in order to record late-night activities and transit and contains missing data sequences of variable length at the beginning and end of the measurement period (these missing data sequences were excluded from the analysis). This length varies from one subject to another, and the mean and minimum trajectory length for

²Raw accelerometer data, covariates allowing the selection of the subset of the population as well as a detailed dictionary are freely accessible here: <https://www1.nyc.gov/site/doh/data/data-sets/physical-activity-and-transit-survey-public-use-data.page>.

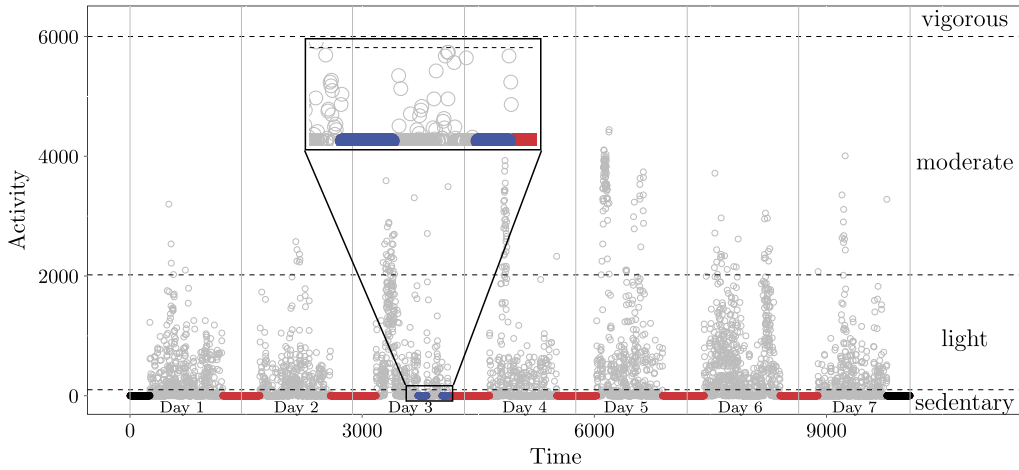


FIG. 1. Accelerometer data of subject Patcid:1200255 of the PAT study measured for one week (with a zoom on the afternoon of day 3): observed values (in gray), missing values during a daytime period (in blue), missing values during a period of night sleep (in red) and missing values at the start and end of the measurement period (in black). The dashed horizontal lines represent the four levels of physical activity based on the classification established by the US Department of Health and Human Services (2008).

the population under consideration (after excluding those missing at the edges) are 9474 and 5199, respectively (with a total number of observations equal to 1,259,981). The model of the accelerometer used is Actigraph GT3X; it is worn on the hips (which results in the fact that certain activities, such as lifting weights or biking, cannot be measured). In addition, participants were also asked to remove it when sleeping, swimming or bathing; hence, the data contains approximately 44% of missing values that appear mainly in sequence, appearing at night but also during the day. Figure 1 gives an example of accelerometer data measured on one subject (i.e., patcid:1200255) for one week where the three types of missing data can be seen. The four levels of physical activity, based on the classification established by the US Department of Health and Human Services (2008) in the Physical Activity Guidelines for Americans (PAGA) report, is also shown in Figure 1. Specifically, the PAT protocol for accelerometer data has established a classification according to PAGA, characterizing each minute of activity. Activity minutes with less than 100 activity counts were classified as “Sedentary,” minutes with 100–2019 counts were classified as “Light,” the class “Moderate” corresponds to 2020–5998 counts/minute and “Vigorous” 5999 and above counts/minute. A comparison between our method and this traditional threshold-based approach is provided in Section 7.3.

3. Mixture of hidden Markov models for accelerometer data. In this section we present the proposed model and the application context for which it has been defined.

3.1. The data. Observed data $\mathbf{y} = (\mathbf{y}_1^\top, \dots, \mathbf{y}_n^\top)$ are composed of n independent and identically distributed sequences \mathbf{y}_i . Each sequence $\mathbf{y}_i = (y_{i(0)}, \dots, y_{i(T)})$ which contains the values measured by the accelerometer at times $t \in \{0, 1, \dots, T\}$ for subject i , with $y_{i(t)} \in \mathbb{R}^+$. Throughout the paper, index i is related to the label of the subject, and index (t) is related to the time of measurement.

The model considers M different activity levels (which are unobserved). These levels impact the distribution of the observed sequences of accelerometer data. The sequence of the hidden states \mathbf{x}_i indicates the activity level of subject i at the different times. Thus, $\mathbf{x}_i = (\mathbf{x}_{i(0)}, \dots, \mathbf{x}_{i(T)}) \in \mathcal{X}$ and the activity level (among the M possible levels) of subject i

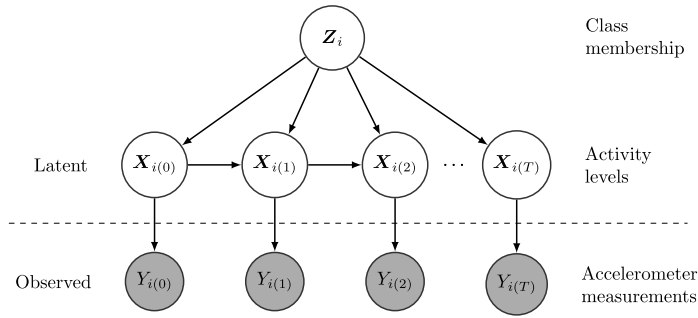


FIG. 2. Generative model of the specific mixture model of HMM used for the accelerometer data: an arrow between two variables indicates dependency and an absence of arrow indicates conditional independence.

at time t is defined by the binary vector $\mathbf{x}_{i(t)} = (x_{i(t)1}, \dots, x_{i(t)M})$ where $x_{i(t)h} = 1$ if subject i is at state h at time t and $x_{i(t)h} = 0$ otherwise.

The heterogeneity (in the sense of different physical activity behaviors) between the n subjects can be addressed by grouping subjects into K homogeneous classes. This is achieved by clustering that assesses a partition $\mathbf{z} = (z_1, \dots, z_n)$ among the n subjects based on their accelerometer measurements. Thus, the vector $\mathbf{z}_i = (z_{i1}, \dots, z_{iK})$ indicates the class membership of subject i , as $z_{ik} = 1$ if observation i belongs to class k and $z_{ik} = 0$ otherwise. Throughout the paper, index k refers to the label of a class grouping homogeneous subjects.

Each subject i is described by three random variables: one unobserved categorical variable \mathbf{z}_i (defining the membership of the class of homogeneous physical activity behaviors for subject i), one unobserved categorical longitudinal data \mathbf{x}_i (a univariate categorical discrete-time time series which defines the activity level of subject i at each time) and one observed positive longitudinal data \mathbf{y}_i (a univariate positive discrete-time time series which contains the values of the accelerometer measured on subject i at each time).

3.2. *Generative model.* The model described below considers that the observations are independent between the subjects and identically distributed. It is defined by the following generative model and summarized by Figure 2 (note that this figure is similar to Figure 6.2 of Bartolucci, Farcomeni and Pennoni (2013)):

1. Sample class membership \mathbf{z}_i from a multinomial distribution;
2. Sample the sequence of activity levels \mathbf{x}_i from a Markov model whose transition matrix depends on class membership;
3. Sample the accelerometer measurement sequence given the activity levels (each $\mathbf{Y}_{i(t)}$ follows a ZIG distribution whose parameters are defined only by $\mathbf{x}_{i(t)}$).

3.3. *Finite mixture model for heterogeneity.* The sequence of accelerometer measures obtained on each subject is assumed to independently arise from a mixture of K parametric distributions, so that the probability distribution function (pdf) of the sequence \mathbf{y}_i is

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

where $\boldsymbol{\theta} = \{\boldsymbol{\lambda}, \boldsymbol{\varepsilon}\} \cup \{\delta_k, \boldsymbol{\pi}_k, \mathbf{A}_k; k = 1, \dots, K\}$ groups the model parameters, δ_k is the proportion of components k with $\delta_k > 0$, $\sum_{k=1}^K \delta_k = 1$ and $p(\cdot; \boldsymbol{\pi}_k, \mathbf{A}_k, \boldsymbol{\lambda}, \boldsymbol{\varepsilon})$ is the pdf of component k parametrized by $(\boldsymbol{\pi}_k, \mathbf{A}_k, \boldsymbol{\lambda}, \boldsymbol{\varepsilon})$ defined below. Thus, δ_k is the marginal probability that a subject belongs to class k (i.e., $\delta_k = \mathbb{P}(Z_{ik} = 1)$). Moreover, $p(\cdot; \boldsymbol{\pi}_k, \mathbf{A}_k, \boldsymbol{\lambda}, \boldsymbol{\varepsilon})$ defines the distribution of a sequence of values measured by the accelerometer on a subject belonging to class k (i.e., $p(\cdot; \boldsymbol{\pi}_k, \mathbf{A}_k, \boldsymbol{\lambda}, \boldsymbol{\varepsilon})$ is the pdf of \mathbf{y}_i given $Z_{ik} = 1$).

3.4. *Hidden Markov model for activity levels.* The model assumes that the distribution of the hidden state sequence depends on the class membership and that the distribution of activity measurements depends on the state at time t but not on the component membership given the state (i.e., $\mathbf{X}_i \not\perp \mathbf{Z}_i$, $Y_{i(t)} \not\perp \mathbf{X}_{i(t)}$ and $Y_{i(t)} \perp \mathbf{Z}_i \mid \mathbf{X}_{i(t)}$). It is crucial that the distribution of $Y_{i(t)}$, given $\mathbf{X}_{i(t)}$, is independent of \mathbf{Z}_i . Indeed, each activity level is defined by the distribution of $Y_{i(t)}$ given the state. Therefore, to extract summary statistics on the whole population (as the average time spent by level of activity) the definition of the activity levels (and the distribution of $y_{i(t)}$ given the state) must be the same among the mixture components.

The pdf of y_i for components k (i.e., given $Z_{ik} = 1$) is

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

The Markov assumption implies that

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

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

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

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 is motivated by the large number of zeros in the accelerometer data (see Figure 1). For the application in Section 7, we use a gamma distribution for $g_c(\cdot; \boldsymbol{\lambda}_h)$. However, model properties and inference are discussed for a large family of densities $g_c(\cdot; \boldsymbol{\lambda}_h)$.

4. Model properties. In this section we present the properties of the mixture of parametric HMM. We start 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.

4.1. *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 $\boldsymbol{\pi}_k^\top \mathbf{A}_k = \boldsymbol{\pi}_k^\top$). Therefore, there exists $0 \leq \nu < 1$ such that

$$\forall k \in \{1, \dots, K\}, \quad \nu_2(\mathbf{A}_k) \leq \nu,$$

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

ASSUMPTION 2. The hidden states define different distributions for the observed sequence. Therefore, for $h \in \{1, \dots, M\}$, $h' \in \{1, \dots, M\} \setminus \{h\}$, we have $\lambda_h \neq \lambda_{h'}$. Moreover, the parametric family of distributions defining $g_c(\cdot; \lambda_1), \dots, g_c(\cdot; \lambda_M)$ permits to consider an ordering such that, for a fixed value $\rho \in \mathbb{R}^+ \setminus \{0\}$, we have

$$\forall h \in \{1, \dots, M - 1\}, \quad \lim_{y_{i(1)} \rightarrow \rho} \frac{g_c(y_{i(1)}; \lambda_{h+1})}{g_c(y_{i(1)}; \lambda_h)} = 0.$$

ASSUMPTION 3. The transition probabilities are different over the mixture components and are not zero. Therefore, for $k \in \{1, \dots, K\}$, $k' \in \{1, \dots, K\} \setminus \{k\}$, we have $\forall (h, \ell)$, $A_k[h, \ell] \neq A_{k'}[h, \ell]$. Moreover, there exists $\zeta > 0$ such that

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

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

Assumption 1 states 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.

4.2. *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, Matias and Rhodes (2009), Celisse, Daudin and Pierre (2012)) and HMM (Gassiat, Cleynen and Robin (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

$$\forall y_i, \quad p(y_i; \theta) = p(y_i; \tilde{\theta}) \quad \Rightarrow \quad \theta = \tilde{\theta}.$$

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 the Supplementary Material, Du Roy de Chaumaray, Marbac and Navarro (2020b), Section 1. The model defined by the marginal distribution of an accelerometer measure at a single time 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, Cleynen and Robin (2016) when they discuss the case where the emission distribution for an HMM follows a mixture model.

4.3. *Probabilities of misclassification.* In this section we examine the probability that an observation will be misclassified when the model parameters are known. We consider the ratio between the probability that subject i belongs to class k given \mathbf{y}_i and the probability that this subject belongs to its 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 subject i and parameters are known).

THEOREM 2. *Assume that Assumptions 1 and 3 hold. If $a > 0$ is such that Assumption 1 (defined in the Supplementary Material, Du Roy de Chaumaray, Marbac and Navarro (2020b), Section 2) holds, then, for every $k \neq k_0$,*

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

where $c > 0$ is a positive constant

Moreover, the exponential bounds of Theorem 2 allows to use the Borel–Cantelli lemma to obtain the almost sure convergence.

COROLLARY 1. *Assume that Assumptions 1 and 3 hold. If \mathbf{y}_i is generated from component k_0 (i.e., $Z_{ik_0} = 1$), then, for every $k \neq k_0$,*

$$\frac{\mathbb{P}(Z_{ik} = 1 \mid \mathbf{y}_i)}{\mathbb{P}(Z_{ik_0} = 1 \mid \mathbf{y}_i)} \xrightarrow[T \rightarrow +\infty]{a.s.} 0, \quad \mathbb{P}(Z_{ik_0} = 1 \mid \mathbf{y}_i) \xrightarrow[T \rightarrow +\infty]{a.s.} 1 \quad \text{and} \quad \mathbb{P}(Z_{ik} = 1 \mid \mathbf{y}_i) \xrightarrow[T \rightarrow +\infty]{a.s.} 0.$$

Therefore, by considering $a = 1$, Theorem 2 and Corollary 1 show that the probability of misclassifying the subject i based on the observation \mathbf{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 the Supplementary Material, Du Roy de Chaumaray, Marbac and Navarro (2020b), Section 2. 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, Daudin and Pierre (2012)) or for coclustering (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.

4.4. *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 (see Section 2 for explanations of the reasons of missingness). 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 5). 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 \mathbf{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 $p(\mathbf{y}_i; \boldsymbol{\pi}_k, \mathbf{A}_k, \boldsymbol{\lambda}, \boldsymbol{\epsilon})$ of component k is approximated by the product of the pdf of the S_i observed sequences $\mathbf{y}_{i1}, \mathbf{y}_{i2}, \dots, \mathbf{y}_{iS_i}$,

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

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

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

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

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 $d_{\min} = \min\{d_{i1}, \dots, d_{iS_i-1}\}$ 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 \mathbf{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

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

5. Maximum likelihood inference. This section presents the methodology used to estimate the model parameters.

5.1. *Inference.* We proposed to estimate the model parameters by maximizing the log-likelihood function where missing values are managed as in Section 4.4, and we recall that the log-likelihood is also approximated for numerical reasons to avoid computing large powers of the transition matrices. We want to find $\hat{\boldsymbol{\theta}}$ which maximizes the following approximated log-likelihood function:

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

This maximization is achieved via an EM algorithm ([Dempster, Laird and Rubin \(1977\)](#)) which considers the complete-data log-likelihood defined by

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

5.2. *Conditional probabilities.* Let $\alpha_{ikhs(t)}(\boldsymbol{\theta})$ be the probability of the partial sequence $y_{is(0)}, \dots, y_{is(t)}$ that ends 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 $y_{is(t+1)}, \dots, y_{is(T_{is})}$ given a start in state h at time t under component k . These probabilities can be easily obtained by the forward/backward algorithm (see the Supplementary Material, Du Roy de Chaumaray, Marbac and Navarro (2020b), Section 3). We deduce that the probability $\gamma_{ikhs(t)}(\boldsymbol{\theta})$ of being in state h at time $t \in \{0, \dots, T_{is}\}$ for y_i under component k is

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

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 y_i under component k is

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

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

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

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

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

5.3. *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, updating of

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

M-step parameter updating

$$\begin{aligned} \delta_k^{[r]} &= \frac{n_k(\boldsymbol{\theta}^{[r-1]})}{n}, & \pi_{kh}^{[r]} &= \frac{n_{kh(0)}(\boldsymbol{\theta}^{[r-1]})}{n_k(\boldsymbol{\theta}^{[r-1]})}, \\ \mathbf{A}_k[h, \ell]^{[r]} &= \frac{n_{kh\ell}(\boldsymbol{\theta}^{[r-1]})}{n_{kh}(\boldsymbol{\theta}^{[r-1]})}, & \varepsilon_h^{[r]} &= \frac{w_h(\boldsymbol{\theta}^{[r-1]})}{n_{kh}(\boldsymbol{\theta}^{[r-1]})} \quad \text{and} \\ \lambda_h^{[r]} &= \operatorname{argmax}_{\lambda_h} \sum_{i=1}^n \sum_{s=1}^{S_i} \sum_{t=0}^{T_{is}} \eta_{ihs(t)}(\boldsymbol{\theta}^{[r-1]})g_c(y_{is(t)}; \lambda_h), \end{aligned}$$

where

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

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

6. Numerical illustrations. This section aims to highlight the main properties of the model on numerical experiments. First, simulated data are used to illustrate the exponential decay of the probabilities of misclassification (given by Theorem 2), the convergence of estimators and the robustness of the approach to missingness. Second, our approach is applied to the data from the PAT study. All the experiments are conducted with the R package **MHMM** available on CRAN.

6.1. *Simulated data.*

Simulation design. All the simulations are performed according to the same model. This model is a bicomponents 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 ZIG distribution. We have

$$\varepsilon_1 = \varepsilon_2 = 0.1, \quad a_1 = 1, \quad b_1 = b_2 = 1,$$

$$A_1 = \begin{bmatrix} e & 1 - e \\ 1 - e & e \end{bmatrix} \quad \text{and} \quad A_2 = \begin{bmatrix} 1 - e & e \\ e & 1 - e \end{bmatrix}.$$

The parameter $a_2 > 1$ controls the separation of the distribution of $Y_{i(t)}$ given the state. 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 $a_2 = 3$), hard-medium ($e = 0.90$ and $a_2 = 3$), medium-easy ($e = 0.75$ and $a_2 = 5$) and easy ($e = 0.90$ and $a_2 = 5$).

Illustrating the exponential rate of the probabilities of misclassification. Theorem 2 states that the probabilities of misclassification decrease at an exponential rate with T . To illustrate this property, 1000 sequences are generated for $T = 1, \dots, 100$ and the four cases. For each sequence \mathbf{y}_i , we compute $\log(\mathbb{P}(Z_{ik} = 1 | \mathbf{y}_i) / \mathbb{P}(Z_{ik_0} = 1 | \mathbf{y}_i))$ when k_0 is the true class, k the alternative and the true model parameters are used. Figure 3(a) shows the behavior of $\log(\mathbb{P}(Z_{ik} = 1 | \mathbf{y}_i) / \mathbb{P}(Z_{ik_0} = 1 | \mathbf{y}_i))$ (the median of this log ratio is plotted in plain, and a 90% confidence interval is plotted in gray). Note that this log ratio of probabilities linearly decreases with T which illustrates the exponential decay of the probabilities of misclassification. Moreover, Figure 3(b) presents the empirical probabilities of misclassification and thus also illustrates Theorem 2. As expected, this shows that the decay of the probabilities of misclassification is faster as the overlaps between class decrease.

Illustrating the convergence of the estimators. We illustrate the convergence of the estimators (partition, latent states and parameters) when the model parameters are estimated by maximum likelihood (see Section 5). We compute the mean square error (MSE) between the model parameters and their estimators. Moreover, we compute the adjusted Rand index (ARI; Hubert and Arabie (1985)) between the true partition and the partition given by the maximum a posteriori probability (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 different sample sizes n and two different lengths of sequences T , considering the case hard-medium. It can be seen that the partition and the model parameters are well estimated. Indeed, the MLE converge to the true parameters as T or n increases, except for the proportion of each component δ_k . The convergence of the estimator of the proportions depends mainly on the sample size n . We notice that the partition obtained by our estimation procedure corresponds to the true partition (for n and T large

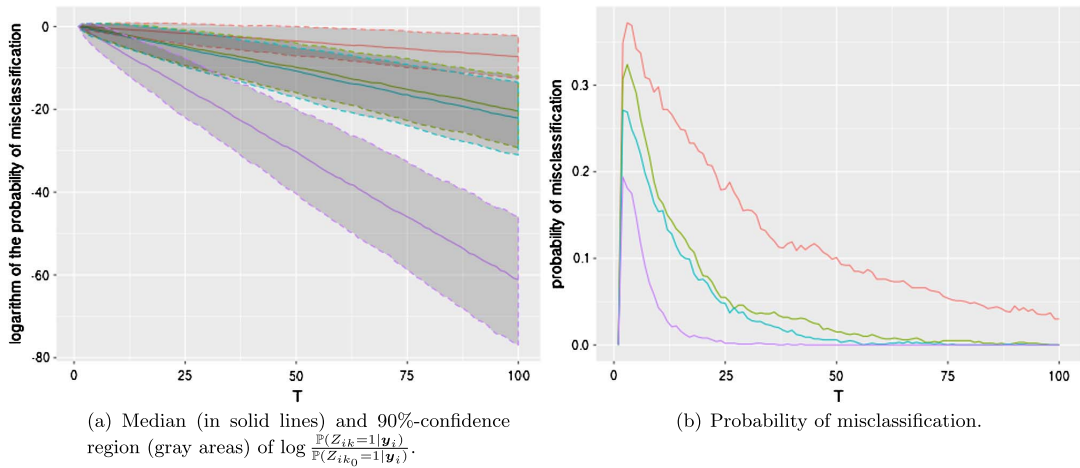


FIG. 3. Results obtained on 1000 observations for the hard (orange), hard-medium (green), medium-easy (blue) and easy (purple) cases.

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 true state sequences a.s., as the number of states to be estimated is also growing with n and T . This result was expected because the number of latent states increases with T and n while the number of parameters and the dimension of the partition do not increase with T . Results obtained for the three other cases are similar and are presented in the Supplementary Material, Du Roy de Chaumaray, Marbac and Navarro (2020b), Section 4.1.

Illustrating the robustness to missingness. We now investigate the robustness of the proposed method with missingness. We compare the accuracy of the estimators (ARI for the latent variables and MSE for the parameters) obtained on samples without missingness to the accuracy of the estimators obtained when missingness is added to the samples. Three situations of missingness are considered: missing completely at random-1 (MCAR-1) (i.e., one sequence of 10 missing values is added to each sequence y_i , the location of the sequence follows a uniform distribution), MCAR-2 (i.e., two sequences of 20 missing values are added for each sequence y_i , the location of the sequences follows a uniform distribution) and missing not at random (MNAR) (i.e., the probability to observe the value $y_{i(t)}$ is equal to $e^{y_{i(t)}} / (1 + e^{y_{i(t)}})$). Note that the last situation adds many missing values when the true value of $y_{i(t)}$ is close to zero, so the occurrence of missing values depends on the latent states. Table 2 compares the results obtained with and without missingness, considering case hard-medium. It shows that estimators are robust to missingness. Results obtained for the other three cases

TABLE 1

Convergence of estimators when 1000 replicates are drawn from case hard-medium: ARI between estimated and true partition, ARI between estimated and true latent states and MSE between the MLE and the true parameters

n	T	ARI (latent variables)		MSE (model parameters)				
		Partition	States	A_k	ε_h	a_h	b_h	δ_k
10	100	0.995	0.621	0.021	0.001	0.088	0.024	0.047
10	500	1.000	0.632	0.007	0.000	0.020	0.005	0.048
100	100	0.996	0.630	0.004	0.000	0.011	0.003	0.005
100	500	1.000	0.634	0.003	0.000	0.005	0.002	0.005

TABLE 2

Convergence of estimators obtained over 1000 replicates with and without missing data when data are sampled from case hard-medium: ARI between estimated and true partition, ARI between estimated and true latent states and MSE between the MLE and the true parameters

n	T	Missingness	Adjusted Rand index		Mean square error				
			Partition	States	A_k	ε_h	a_h	b_h	δ_k
10	100	no missingness	0.995	0.621	0.021	0.001	0.088	0.024	0.047
		MCAR-1	0.991	0.613	0.024	0.001	0.102	0.028	0.047
		MCAR-2	0.987	0.605	0.028	0.001	0.113	0.032	0.047
		MNAR	0.934	0.497	0.051	0.003	0.398	0.050	0.050
10	500	no missingness	1.000	0.632	0.007	0.000	0.020	0.005	0.048
		MCAR-1	1.000	0.631	0.007	0.000	0.020	0.005	0.048
		MCAR-2	1.000	0.631	0.007	0.000	0.019	0.005	0.048
		MNAR	0.999	0.516	0.021	0.003	0.233	0.028	0.048
100	100	no missingness	0.996	0.630	0.004	0.000	0.011	0.003	0.005
		MCAR-1	0.994	0.624	0.004	0.000	0.013	0.003	0.005
		MCAR-2	0.989	0.618	0.005	0.000	0.014	0.004	0.005
		MNAR	0.951	0.512	0.014	0.002	0.200	0.026	0.005
100	500	no missingness	1.000	0.634	0.003	0.000	0.005	0.002	0.005
		MCAR-1	1.000	0.633	0.002	0.000	0.006	0.002	0.005
		MCAR-2	1.000	0.632	0.002	0.000	0.005	0.002	0.005
		MNAR	1.000	0.520	0.011	0.002	0.198	0.026	0.005

are similar and are reported in the Supplementary Material, [Du Roy de Chaumaray, Marbac and Navarro \(2020b\)](#), Section 4.1.

6.2. *Using the approach on classical accelerometer data.* We consider the accelerometer data measured on three subjects available from [Huang et al. \(2018\)](#). 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. \(2018\)](#) and the method proposed in this paper.

[Huang et al. \(2018\)](#) 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 sequence measured by the accelerometer of a subject, the definition of the activity level is different for each subject (see [Huang et al. \(2018\)](#), 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 4 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, [Du Roy de Chaumaray, Marbac and Navarro \(2020b\)](#), Section 4.2), we consider $M = 4$ activity levels. These levels can be easily characterized with the model parameters presented in Table 3. The model permits to describe the physical activity of each subject by computing the repartition of the time spent in each state (see Table 4). Moreover, the transition matrices also make sense. For instance, class 1 (subjects 9 and 20) has an almost tridiagonal transition matrix (by considering an order between the states given through the activity levels per state) and class-2 (subject 2) is composed of a subject with low-overall activity. Note that the state that we call *sleeping* is characterized by a very low activity, so we assume that subjects are sleeping when they are in this state. However, this assumption could be verified if polysomnography (PSG)

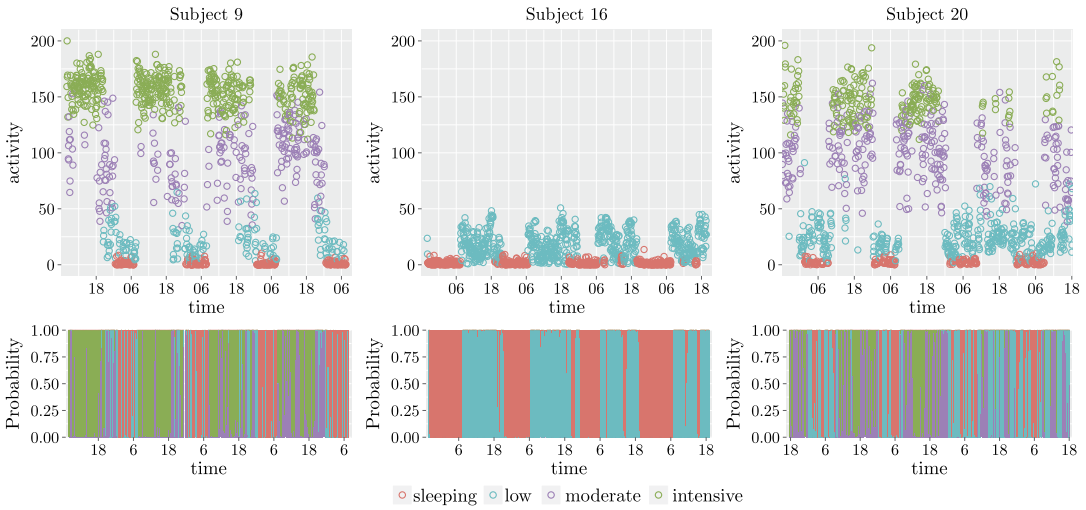


FIG. 4. *State estimation for the three subjects: (top) accelerometer data where color indicates the expected value of $Y_{i(t)}$ conditionally to the most likely state and to the most likely component; (bottom) probability of each state at each time.*

data would be available.

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

7. Analysis of PAT data. In this section we analyze the data presented in Section 2.

7.1. Experimental conditions. In order to compare our approach to the cuts defined a priori in the PAT study (see Section 2), the model was fitted with four activity levels. Note that selecting the number of states in HMM stays a challenging problem (see the discussion in the Conclusion). However, approaches considering four activity levels are standard for accelerometer data. The number of components (i.e., the number of classes) is estimated, using an information criterion, unlike the PAT study where it is arbitrarily set at 3 or 4. For each number of components, 5000 random initializations of the EM algorithm are performed. The analysis needs about one day of computation on a 32-Intel(R) Xeon(R) CPU E5-4627 v4 @ 2.60 GHz.

TABLE 3

Description of the states: Estimator of the parameters, mean and standard deviation (sd) of the count per five minutes measured by the accelerometer conditionally on the state

State name	ε_h	a_h	b_h	Count per five minutes	
				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 4

Repartition of the times spent at the different states for each subject (sum for each subject is equal to one)

State name	Subject 9	Subject 16	Subject 20
Intensive-level	0.42	0.00	0.20
Moderate-level	0.21	0.00	0.27
Low-level	0.15	0.55	0.34
Sleeping	0.22	0.45	0.19

7.2. *Model selection.* 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, Celeux and Govaert (2000)) defined by

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

and

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

where $\nu_K = (K - 1) + K(M + M^2) + 3M$ is the number of parameters for a model with K components and M states and $\hat{z}_{ik}(\hat{\boldsymbol{\theta}})$ defines the partition by the MAP rule associated to the MLE such that

$$\hat{z}_{ik}(\hat{\boldsymbol{\theta}}) = \begin{cases} 1 & \text{if } \tau_{ik}(\hat{\boldsymbol{\theta}}) = \operatorname{argmax}_{\ell=1, \dots, K} \tau_{i\ell}(\hat{\boldsymbol{\theta}}), \\ 0 & \text{otherwise.} \end{cases}$$

The ICL is defined according as the integrated complete-data likelihood computed with the partition given by the MAP rule with the MLE. The values of the information criteria are given in Table 5 for different number of classes. Both criteria select five components. The values of $\text{ICL}(K)$ are close to those of the $\text{BIC}(K)$, implying that 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 6). In the following we interpret the results obtained with $M = 4$ activity levels and $K = 5$ classes.

7.3. *Description of the activity levels.* The parameters of the ZIG distributions are presented in Table 6. 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). Again, note that we call this state *sleeping state* due to the low activity level but that PSG data should be used to verify that subject are really sleeping within this state. However, ε_h is not equal to zero for the other states, but the more active the state is, the smaller ε_h is. We also compute the marginal cutoffs (i.e., the cutoffs by considering the MAP of $\mathbb{P}(X_{i(t)} | Y_{i(t)})$). These

TABLE 5

Information criteria obtained on PAT data with four levels of activity (minima are in bold)

K	1	2	3	4	5	6	7
BIC	-2,953,933	-2,952,313	-2,951,809	-2,951,705	-2,951,308	-2,951,364	-2,951,696
ICL	-2,953,933	-2,952,313	-2,951,810	-2,951,707	-2,951,309	-2,951,364	-2,951,697

TABLE 6

Parameters describing the four activity levels for PAT data and statistics on the distribution of the counts per five minutes per state

Name of the activity level	Parameters			Statistics	
	ε_h	a_h	b_h	Mean	Marginal cutoffs
Sleeping	0.988	7.470	7.470	0.012	[0, 0]
Low-level	0.260	0.974	0.020	36.926]0, 97.7]
Moderate-level	0.025	1.408	0.004	329.249]97.7, 614.4]
Intensive-level	0.007	2.672	0.002	1696.935]614.4, +∞[

cutoffs neglect the time dependency due to the Markov structure but can be compared to the cutoffs proposed by the PAT study. Indeed, according to the PAT study, minutes with < 100 counts are assigned to “Sedentary” activity, minutes with 100–2019 counts were classified as “Light,” the class Moderate corresponds to 2020–5998 counts/minute and “Vigorous” 5999 and above counts/minute. The marginal cutoff associated with the *low-level* state is very close to that of the “Sedentary” class of the PAT. We find, however, that our marginal cutoffs are more accurate for higher levels of activity. PAT cutoffs do not adequately characterize the activity level of the study population. Finally, contrary to classical thresholds, our modeling approach allows to capture and characterize the variability associated with the different levels of activity, variability which seems important (see Figure 5 and Table 6).

7.4. *Description of the classes.* Classes can be described using their proportions and their associated parameters presented in Table 6. 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 (i.e., $\pi_{k4} = 0.143$). For three of the five classes, Figure 5 presents a characteristic subject of each class and the probabilities of the activity levels (the associated graphs for the two remaining classes are given in the Supplementary Material, Du Roy de Chaumaray, Marbac and Navarro (2020b), Section 4.3). Classes can be interpreted from the mean time spent at different activity levels presented in Table 7 and from transition matrices presented in Table 8 which are almost tridiagonal. This could be expected

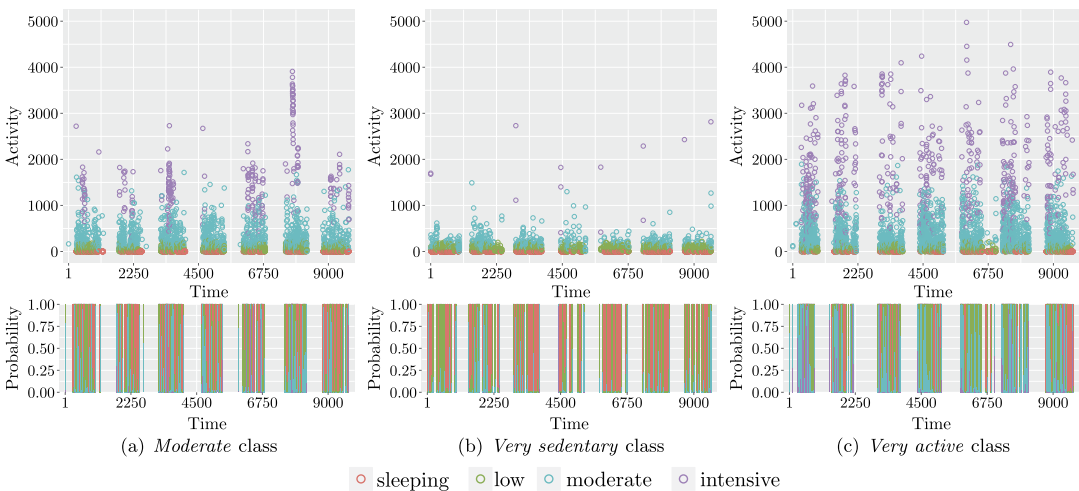


FIG. 5. Examples of observations assigned into the five classes with the probabilities of the states.

TABLE 7
Repartition of the times spent at the different states for each class (sum for each class is equal to one)

	<i>Active</i>	<i>Sedentary</i>	<i>Moderate</i>	<i>Very sedentary</i>	<i>Very active</i>
	Class				
Sleeping	0.306	0.467	0.304	0.504	0.189
Low-level	0.284	0.209	0.411	0.366	0.351
Moderate-level	0.338	0.263	0.225	0.124	0.316
Intensive-level	0.072	0.061	0.060	0.006	0.143

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 4.4) 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 $5 \cdot 10^{-4}$ for any component.

8. Conclusion. A specific mixture of HMM has been introduced to analyze accelerometer data. It avoids the traditional cutoff point method and provides a better characterization of activity levels for the analysis of these data while adapting to the population. The proposed model could be applied to a population with different characteristics (e.g., younger) which would lead to different definitions of activity levels. In addition, the use of several HMMs make to take into account dependency over time and thus improve the traditional method based on cutoff points (Witowski et al. (2014)). This approach also allows us to take into account the heterogeneity of the population (in the sense of physical activity).

An interesting perspective is to consider adjusting for covariates (e.g., gender or age). These confusing factors could impact the probabilities of transition between the latent spaces (e.g., using a generalized linear model approach) and/or the definition of the accelerometer measurement given a state (e.g., linear regression on some parameters of the ZIG distribution).

TABLE 8
Transition matrix for the five classes

	Sleeping	Low-level	Moderate-level	Intensive-level
	Class <i>moderate</i>			
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 <i>very sedentary</i>			
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 <i>very active</i>			
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

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: for instance, the use of BIC is criticized (see Cappé, Moulines and Rydén (2005), Chapter 15). This makes the study of relevant information criteria for selecting the number of states an interesting topic for future work. Pseudo-likelihood based criteria could be used (Csiszár and Talata (2006), Gassiat (2002)), 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 number of observations.

Acknowledgements. The authors thank Samuel Danthine and Todd Donahue for their careful proofreading, along with the constructive and helpful comments of the Editor, Associate Editor and three anonymous reviewers.

SUPPLEMENTARY MATERIAL

Mixture of hidden Markov models for accelerometer data (DOI: [10.1214/20-AOAS1375SUPP](https://doi.org/10.1214/20-AOAS1375SUPP); .pdf). Supplementary material contains technical proofs and complements to the numerical experiments.

REFERENCES

- AARTS, E. (2019). mHMMbayes: Multilevel hidden Markov models using Bayesian estimation. R package version 0.1.1.
- ALLMAN, E. S., MATIAS, C. and RHODES, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.* **37** 3099–3132. [MR2549554](https://doi.org/10.1214/09-AOS689) <https://doi.org/10.1214/09-AOS689>
- ALTMAN, R. M. (2007). Mixed hidden Markov models: An extension of the hidden Markov model to the longitudinal data setting. *J. Amer. Statist. Assoc.* **102** 201–210. [MR2345538](https://doi.org/10.1198/016214506000001086) <https://doi.org/10.1198/016214506000001086>
- 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** 744–752. [MR3825361](https://doi.org/10.1111/biom.12781) <https://doi.org/10.1111/biom.12781>
- BANFIELD, J. D. and RAFTERY, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* **49** 803–821. [MR1243494](https://doi.org/10.2307/2532201) <https://doi.org/10.2307/2532201>
- BARTOLUCCI, F., FARCOMENI, A. and PENNONI, F. (2013). *Latent Markov Models for Longitudinal Data. Statistics in the Social and Behavioral Sciences Series*. CRC Press, Boca Raton, FL. [MR3184304](https://doi.org/10.1214/13-BS004)
- BARTOLUCCI, F., PANDOLFI, S. and PENNONI, F. (2017). Lmest: An R package for latent Markov models for longitudinal categorical data. *J. Stat. Softw.* **81** 1–38.
- BARTOLUCCI, F., PENNONI, F. and VITTADINI, G. (2011). Assessment of school performance through a multi-level latent Markov rasch model. *J. Educ. Behav. Stat.* **36** 491–522.
- 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** 719–725.
- BRAULT, V. and MARIADASSOU, M. (2015). Co-clustering through latent bloc model: A review. *J. SFdS* **156** 120–139. [MR3432606](https://doi.org/10.1007/s00180-007-0097-1)
- 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. [MR2159833](https://doi.org/10.1214/12-EJS729)
- CELEUX, G. and DURAND, J.-B. (2008). Selecting hidden Markov model state number with cross-validated likelihood. *Comput. Statist.* **23** 541–564. [MR2448181](https://doi.org/10.1007/s00180-007-0097-1) <https://doi.org/10.1007/s00180-007-0097-1>
- CELISSE, A., DAUDIN, J.-J. and PIERRE, L. (2012). Consistency of maximum-likelihood and variational estimators in the stochastic block model. *Electron. J. Stat.* **6** 1847–1899. [MR2988467](https://doi.org/10.1214/12-EJS729) <https://doi.org/10.1214/12-EJS729>
- 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** 461–469.
- CSISZÁR, I. and TALATA, Z. (2006). Consistent estimation of the basic neighborhood of Markov random fields. *Ann. Statist.* **34** 123–145. [MR2275237](https://doi.org/10.1214/009053605000000912) <https://doi.org/10.1214/009053605000000912>

- DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B* **39** 1–38. [MR0501537](#)
- DU ROY DE CHAUMARAY, M., MARBAC, M. and NAVARRO, F. (2020b). Supplement to “Mixture of hidden Markov models for accelerometer data.” <https://doi.org/10.1214/20-AOAS1375SUPP>
- DU ROY DE CHAUMARAY, M., MARBAC, M. and NAVARRO, F. (2020a). *MHMM: Finite mixture of hidden Markov model*. R package version 1.0.0.
- DYRSTAD, S. M., HANSEN, B. H., HOLME, I. M. and ANDERSSSEN, S. A. (2014). Comparison of self-reported versus accelerometer-measured physical activity. *Med. Sci. Sports Exerc.* **46** 99–106. <https://doi.org/10.1249/MSS.0b013e3182a0595f>
- FREEDSON, P. S., MELANSON, E. and SIRARD, J. (1998). Calibration of the computer science and applications. *Inc. Accelerometer. Med. Sci. Sports Exerc.* **30** 777–781.
- GASSIAT, E. (2002). Likelihood ratio inequalities with applications to various mixtures. *Ann. Henri Poincaré (B)* **38** 897–906. [MR1955343](#) [https://doi.org/10.1016/S0246-0203\(02\)01125-1](https://doi.org/10.1016/S0246-0203(02)01125-1)
- GASSIAT, E., CLEYNEN, A. and ROBIN, S. (2016). Inference in finite state space non parametric hidden Markov models and applications. *Stat. Comput.* **26** 61–71. [MR3439359](#) <https://doi.org/10.1007/s11222-014-9523-8>
- GERACI, M. (2019). Additive quantile regression for clustered data with an application to children’s physical activity. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **68** 1071–1089. [MR4002384](#)
- 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. R. Stat. Soc. Ser. C. Appl. Stat.* **65** 51–75. [MR3438238](#) <https://doi.org/10.1111/rssc.12105>
- GOODMAN, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61** 215–231. [MR0370936](#) <https://doi.org/10.1093/biomet/61.2.215>
- GRANDNER, M. A., SANDS-LINCOLN, M. R., PAK, V. M. and GARLAND, S. N. (2013). Sleep duration, cardiovascular disease, and proinflammatory biomarkers. *Nat. Sci. Sleep* **5** 93–107. <https://doi.org/10.2147/NSS.S31063>
- 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** e0169576.
- HELSCHE, S. and HELSCHE, J. (2019). Mixture hidden Markov models for sequence data: The seqHMM package in R. *J. Stat. Softw.* **88** 1–32.
- HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *J. Amer. Statist. Assoc.* **97** 1090–1098. [MR1951262](#) <https://doi.org/10.1198/016214502388618906>
- HUANG, Q., COHEN, D., KOMARZYNSKI, S., LI, X.-M., INNOMINATO, P., LÉVI, F. and FINKENSTÄDT, B. (2018). Hidden Markov models for monitoring circadian rhythmicity in telemetric activity data. *J. R. Soc. Interface* **15** 20170885.
- HUANG, L., BAI, J., IVANESCU, A., HARRIS, T., MAURER, M., GREEN, P. and ZIPUNNIKOV, V. (2019). Multilevel matrix-variate analysis and its application to accelerometry-measured physical activity in clinical populations. *J. Amer. Statist. Assoc.* **114** 553–564. [MR3963162](#) <https://doi.org/10.1080/01621459.2018.1482750>
- HUBERT, L. and ARABIE, P. (1985). Comparing partitions. *J. Classification* **2** 193–218.
- HUNT, L. and JORGENSEN, M. (2011). Clustering mixed data. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **1** 352–361.
- HUNTER, D. R., GOODREAU, S. M. and HANDCOCK, M. S. (2008). Goodness of fit of social network models. *J. Amer. Statist. Assoc.* **103** 248–258. [MR2394635](#) <https://doi.org/10.1198/016214507000000446>
- IMMERWAHR, S., WYKER, B., BARTLEY, K. and EISENHOWER, D. (2012). The physical activity and transit survey device follow-up study: Methodology report. The New York City Department of Health and Mental Hygiene.
- 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** 543. <https://doi.org/10.1186/s12889-018-5215-1>
- KARLIS, D. and MELIGKOTSIDOU, L. (2007). Finite mixtures of multivariate Poisson distributions with application. *J. Statist. Plann. Inference* **137** 1942–1960. [MR2323875](#) <https://doi.org/10.1016/j.jspi.2006.07.001>
- 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 multi-centre longitudinal study. *Lancet* **366** 301–307.
- KOSMIDIS, I. and KARLIS, D. (2016). Model-based clustering using copulas with applications. *Stat. Comput.* **26** 1079–1099. [MR3535046](#) <https://doi.org/10.1007/s11222-015-9590-5>
- LEE, J. A. and GILL, J. (2018). Missing value imputation for physical activity data measured by accelerometer. *Stat. Methods Med. Res.* **27** 490–506. [MR3761118](#) <https://doi.org/10.1177/0962280216633248>

- LEE, I.-M., SHIROMA, E. J., LOBELO, F., PUSKA, P., BLAIR, S. N., KATZMARZYK, P. T. et al. (2012). Effect of physical inactivity on major non-communicable diseases worldwide: An analysis of burden of disease and life expectancy. *Lancet* **380** 219–229.
- LEVIN, D. A. and PERES, Y. (2017). *Markov Chains and Mixing Times*. Amer. Math. Soc., Providence, RI. [MR3726904](#)
- LIM, Y., OH, H.-S. and CHEUNG, Y. K. (2019). Functional clustering of accelerometer data via transformed input variables. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **68** 495–520. [MR3937460](#)
- LIM, S., WYKER, B., BARTLEY, K. and EISENHOWER, D. (2015). Measurement error of self-reported physical activity levels in New York city: Assessment and correction. *Am. J. Epidemiol.* **181** 648–655.
- MARUOTTI, A. (2011). Mixed hidden Markov models for longitudinal data: An overview. *Int. Stat. Rev.* **79** 427–454.
- MATIAS, C., REBAFKA, T. and VILLERS, F. (2018). A semiparametric extension of the stochastic block model for longitudinal networks. *Biometrika* **105** 665–680. [MR3842891](#) <https://doi.org/10.1093/biomet/asy016>
- MCLACHLAN, G. and PEEL, D. (2000). *Finite Mixture Models*. *Wiley Series in Probability and Statistics: Applied Probability and Statistics*. Wiley, New York. [MR1789474](#) <https://doi.org/10.1002/0471721182>
- MCNICHOLAS, P. D. (2017). *Mixture Model-Based Classification*. CRC Press, Boca Raton, FL. [MR3642443](#)
- MCTIERNAN, A. (2008). Mechanisms linking physical activity with cancer. *Nat. Rev. Cancer* **8** 205–211. <https://doi.org/10.1038/nrc2325>
- 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. Amer. Statist. Assoc.* **101** 1352–1364. [MR2307570](#) <https://doi.org/10.1198/016214506000000465>
- 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. *Am. J. Clin. Nutr.* **92** 1436–1445.
- PALTA, P., MCMURRAY, R., GOUSKOVA, N., SOTRES-ALVAREZ, D., DAVIS, S., CARNETHON, M., CASTAÑEDA, S., GELLMAN, M., HANKINSON, A. L. et al. (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** 957–965.
- SADEH, A., SHARKEY, M. and CARSKADON, M. A. (1994). Activity-based sleep-wake identification: An empirical test of methodological issues. *Sleep* **17** 201–207.
- SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. [MR0468014](#)
- SCOTT, S. L., JAMES, G. M. and SUGAR, C. A. (2005). Hidden Markov models for longitudinal comparisons. *J. Amer. Statist. Assoc.* **100** 359–369. [MR2170459](#) <https://doi.org/10.1198/016214504000001592>
- SLOOTMAKER, S. M., SCHUIT, A. J., CHINAPAW, M. J., SEIDELL, J. C. and VAN MECHELEN, W. (2009). Disagreement in physical activity assessed by accelerometer and self-report in subgroups of age, gender, education and weight status. *Int. J. Behav. Nutr. Phys. Act.* **6** 17.
- 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 Med.* **1** e62.
- TEICHER, H. (1963). Identifiability of finite mixtures. *Ann. Math. Stat.* **34** 1265–1269. [MR0155376](#) <https://doi.org/10.1214/aoms/1177703862>
- TEICHER, H. (1967). Identifiability of mixtures of product measures. *Ann. Math. Stat.* **38** 1300–1302. [MR0216635](#) <https://doi.org/10.1214/aoms/1177698805>
- TITSIAS, M. K., HOLMES, C. C. and YAU, C. (2016). Statistical inference in hidden Markov models using k -segment constraints. *J. Amer. Statist. Assoc.* **111** 200–215. [MR3494653](#) <https://doi.org/10.1080/01621459.2014.998762>
- TROIANO, R. P., BERRIGAN, D., DODD, K. W., MASSE, L. C., TILERT, T. and MCDOWELL, M. (2008). Physical activity in the United States measured by accelerometer. *Med. Sci. Sports Exerc.* **40** 181–188.
- US DEPARTMENT OF HEALTH AND HUMAN SERVICES (2008). 2008 physical activity guidelines for americans: Be active, healthy, and happy! Available at <http://www.health.gov/paguidelines>.
- VAN HEES, V. T., SABIA, S., ANDERSON, K. N., DENTON, S. J., OLIVER, J., CATT, M., ABELL, J. G., KIVIMÄKI, M., TRENELL, M. I. et al. (2015). A novel, open access method to assess sleep duration using a wrist-worn accelerometer. *PLoS ONE* **10** e0142533. <https://doi.org/10.1371/journal.pone.0142533>
- VAN DE POL, F. and LANGEHEINE, R. (1990). Mixed Markov latent class models. *Sociol. Method.* 213–247.
- VITERBI, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *J. Amer. Statist. Assoc.* **13** 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. Amer. Statist. Assoc.* **113** 95–110. MR3803442 <https://doi.org/10.1080/01621459.2017.1330202>
- 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** e114089. <https://doi.org/10.1371/journal.pone.0114089>
- WONG, C. S. and LI, W. K. (2000). On a mixture autoregressive model. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **62** 95–115. MR1747398 <https://doi.org/10.1111/1467-9868.00222>
- WYKER, B., BARTLEY, K., HOLDER-HAYES, E., IMMERWAHR, S., EISENHOWER, D. and HARRIS, T. (2013). Self-reported and accelerometer-measured physical activity: a comparison in new york city. New York (NY): New York City Department of Health and Mental Hygiene: Epi Research Report, 1–12.
- XIAO, L., HUANG, L., SCHRACK, J. A., FERRUCCI, L., ZIPUNNIKOV, V. and CRAINICEANU, C. M. (2015). Quantifying the lifetime circadian rhythm of physical activity: A covariate-dependent functional approach. *Biostatistics* **16** 352–367. MR3365433 <https://doi.org/10.1093/biostatistics/kxu045>
- YANG, C.-C. and HSU, Y.-L. (2010). A review of accelerometry-based wearable motion detectors for physical activity monitoring. *Sensors* **10** 7772–7788.