

# Markovian-Based Clustering of Internet Addiction Trajectories



Zhivko Taushanov and André Berchtold

## 1 Introduction

The clustering of trajectories has gained much interest in recent years from the scientific community, especially in the social sciences, because the number of longitudinal studies, as compared to cross-sectional ones, has been constantly increasing. As regards categorical data, the most common approach relies on the Optimal Matching (OM) to compute a distance between each pair of trajectories before clustering them, whereas the Growth Mixture Model (GMM) can be applied for continuous data. However, these two approaches suffer from some shortcomings, calling for the need to develop and apply alternative approaches. For instance, OM requires the choice of a substitution cost measure and other parameters. GMM gives a lot of importance to the shape of sequences. Therefore, there is a risk to overfit the data when nonlinear trajectories are considered on quite short sequences. The other issues of GMM include computational load, presence of local optima, missing data treatment, model selection criteria, the need for large sample size, and unclear Type I error rates (Wang and Bodner 2007).

In this paper, we study the use of a specific class of Markovian Models called the Hidden Mixture Transition Distribution (HMTD) model (Bolano and Berchtold 2016) for clustering purpose. Even if this model-based approach was developed as a tool for the analysis of continuous trajectories, it also allows for their clustering without a priori knowledge of cluster membership. Moreover, covariates can be easily included in the model.

The HMTD and GMM clustering approaches are applied and compared on a dataset of trajectories of the Internet Addiction Test (IAT). Excessive Internet use,

---

Z. Taushanov (✉) · A. Berchtold

Institute of Social Sciences and NCCR LIVES, University of Lausanne, Lausanne, Switzerland  
e-mail: [zhivko.taushanov@unil.ch](mailto:zhivko.taushanov@unil.ch); [andre.berchtold@unil.ch](mailto:andre.berchtold@unil.ch)

© The Author(s) 2018

G. Ritschard, M. Studer (eds.), *Sequence Analysis and Related Approaches*,

Life Course Research and Social Policies 10,

[https://doi.org/10.1007/978-3-319-95420-2\\_12](https://doi.org/10.1007/978-3-319-95420-2_12)

203

especially among youths, is an emerging health issue in the medical literature, with studies showing contrasting results. Surís et al. (2014) show a significant association between problematic Internet use and several somatic disorders, including back, weight, musculoskeletal, and sleep problems. Moreover, several chronic conditions are also significantly associated with problematic Internet use. In contrast, another study finds no significant Internet use effect on the development of overweight among youths (Barrense-Dias et al. 2015).

While several alternative approaches (e.g. Skarupova et al. 2015) have been introduced over the years, the tool most often used to quantify the degree of addiction to Internet is still the Internet Addiction Test (IAT), developed by Young (1998). However, since the test's scale is based on 20 items and is quite long, its psychometric properties are matters of controversy (Faraci et al. 2013) and the test is not considered suitable for the successive measurement of the same subjects (test-retest). Its use in longitudinal contexts remains problematic because of the difficulty to distinguish between the real evolution of subjects and changes due to behavior of the IAT itself.

To gain information on the behavior of the IAT in longitudinal studies, we need to compare the typical trajectories of the repeated IAT measurements with other characteristics of the subjects under study. Thus, we first cluster the IAT trajectories into a finite set of meaningful groups and then compare these groups with the known characteristics of subjects that are either time-invariant or evolve over time. Specifically, the goals of this study are (1) to separate the Internet addiction trajectories into an optimal number of meaningful categories using HMTD, (2) to explore how does the introduction of the covariates influence the previous optimal partition, and (3) to compare the HMTD clustering with an equivalent GMM clustering in order to gain information on the respective strengths of both approaches. We hypothesize that (1) the IAT scores computed for the same person can vary considerably over time, implying that the trajectories are difficult to classify; (2) a classification using covariates is easier to interpret than a classification without any additional information on the clustered variable itself; and (3) the HMTD approach can lead to more sound and easier-to-use solutions as compared to the solutions obtained using GMM. However, we must stress that it is impossible to conclude that one method is superior to another, especially using real data, without knowing the true cluster membership. So this work must be considered as a first step in the comparison of HMTD and GMM as clustering tools.

## 2 Data and Methods

### 2.1 Data

The data we considered are from *ado@Internet.ch* (Surís et al. 2012), a longitudinal study on the use of Internet among youths in the Swiss canton of Vaud (the largest

canton in the French-speaking part of Switzerland). The data were collected five times every six months from Spring 2012 ( $T_0$ , baseline) to Spring 2014 ( $T_4$ ) using an online questionnaire. The data for the first time were collected from schools during the computerlab periods. Then, the students who agreed to participate in the study were contacted again by email from  $T_1$  to  $T_4$  to answer follow-up questionnaires on their home computer. A convenience sample of  $n = 185$  adolescents who answered all five questionnaires is used for the present study (67% females; mean age at  $T_0$ : 14.1 years). For more details on the overall design of the study and data collection, (see Surfís et al. 2012; Pigué et al. 2016).

The main outcome is the IAT score measured at each wave for each subject. The IAT developed by Young (1998) and validated in French by Khazaal et al. (2008) is a scale ranging from 0 to 100, based on the answers to 20 items whose possible answers range from Never (coded 0) to Always (5). Examples of items are, *How often do you find yourself staying online longer than you intended?* and *How often do you fear that life without the Internet would be boring, empty, and joyless?*

In addition to the IAT, we also considered several important characteristics of the subjects, either fixed in time—gender, age at baseline, and education track at baseline (extended requirements vs. basic requirements)—or evolving over time—emotional well-being (measured by the WHO-5 index) and Body Mass Index (BMI, computed from auto-reported measures of height and weight). Note that the WHO-5 index was not evaluated on the third wave of the study, and so for the present paper, we imputed values as the simple mean between the values of the second and fourth waves. Similarly, we imputed the BMI for the second wave of the study as the mean between the values of the first and third waves.

## 2.2 Clustering Using the HMTD Model

We used a specific class of Markovian Models, the HMTD model, to cluster the longitudinal sequences of continuous data. This model combines a latent and an observed level (Bolano and Berchtold 2016). The visible level is a Mixture Transition Distribution (MTD) model that was first introduced by Raftery in 1985 as an approximation of high-order Markov chains Raftery (1985) and then developed by Berchtold (2001, 2003) and Berchtold and Raftery (2002). Here, we used a Gaussian version of the MTD model, where the mean of the Gaussian distribution is a function of past observations. Because of the small size of each sequence of the observed outcome (five data points, from  $T_0$  to  $T_4$ ), long dependencies between successive observations could not be considered, and therefore we fix the dependence order for the mean of the Gaussian distributions of each component to one:

$$\mu_{g,t} = \varphi_{g,0} + \varphi_{g,1} x_{t-1}$$

where  $\varphi_{g,0}$  is the constant for the mean for component  $g$  and  $\varphi_{g,1}$  is the autoregressive parameter indicating the dependence from the previous observation  $x_{t-1}$ . Similarly the variance of each component can be written as a function of the past periods variability:  $\sigma_{g,t}^2 = \theta_{g,0} + \sum_{s=1}^S \theta_{g,s} x_{t-s}^2$ . However given the small number of time periods in our dataset, and for the sake of simplicity, we decided to treat the variance as a constant:  $\sigma_{g,t}^2 = \theta_{g,0}$ .

In the HMTD model, the latent level is a homogeneous Markov chain. Each state of the chain is associated with a different Gaussian component at the visible level, with the transition matrix used to determine which component best represents the current observation. To use the HMTD model as a clustering tool, we assume the hidden transition matrix to be the identity matrix. Consequently, each sequence of successive observations is associated with only one component of the model, thus generating a clustering of sequences into mutually exclusive groups. Notice that in this case, the resulting model is no more a hidden Markov model, but a mixture of Gaussian distributions. However, it is still interesting to view it as a HMTD, because it is then possible to compare the clustering model with other models, especially with semi-clustering models whose transition matrix is not the identity matrix, but a triangular matrix letting data trajectories move from one group to another in a specific order.

In addition to the clustering based on the IAT variable only, we performed a second clustering adding information from five covariates (gender, age at  $T_0$ , education track at  $T_0$ , WHO-5, and BMI). These covariates are introduced as additional terms in the specification of the mean of each visible component of the model, and the categorical variables are introduced as dummy variables. We then rewrite the mean of the  $g$ -th component as

$$\begin{aligned} \mu_{g,t} = & \varphi_{g,0} + \varphi_{g,1} x_{t-1} + \varphi_{g,2} \text{Gender}(male) + \varphi_{g,3} \text{Age} \\ & + \varphi_{g,4} \text{Education}(extended) + \varphi_{g,5} \text{WHO-5} + \varphi_{g,6} \text{BMI} \end{aligned}$$

with *female* and *basic requirements* used as reference modalities for Gender and Education, respectively.

In practice, continuous covariates are centered around the sample mean before computing the clustering model in order to allow for a better convergence of the estimation algorithm. A comparison of the two specifications of the mean, with and without covariates, illustrates whether the inclusion of covariates in the model helps to improve the clustering process. It must be mentioned that, in addition to these two HMTD models, many other specifications were tried, following a hierarchical approach (Bolano and Berchtold 2016), but none of these alternative specifications seemed to give a more useful clustering of IAT trajectories.

The HMTD model is estimated by maximizing its log-likelihood. When the variance of each component of the model is constant, the log-likelihood can be derived with respect to all parameters, but in the general case of time-varying variances (Berchtold 2003), the log-likelihood is generally not differentiable, and the solution space can be very complex. A specific heuristic is then applied to obtain

the solution (Taushanov and Berchtold 2017). Since this heuristic can accommodate to all possible specifications of the HMTD model, we used it for all computations. Regarding cluster assignment, we used the standard Viterbi algorithm which is able to find the best sequence of hidden states in function of the observed data and of the current model (Forney 1973). In the specific case of clustering, the Viterbi algorithm simply assign each observed trajectory to the most likely component.

We used a bootstrap procedure to obtain confidence intervals for each parameter, but since our goal here was to validate not the initial classification itself, but the parameters associated with the model describing each visible component of the model, we adopted the following approach: Instead of performing the bootstrap on the whole original sample, we divided the original sample into as many groups as can be retained in the final classification. We then applied a single-component version of the HMTD model to each sub-sample separately in order to estimate the coefficients using bootstrap. By applying the model on the sub-samples separately, instead of on the initial sample, we avoided the so-called label-switching problem that is very common in latent variable clustering. The inconvenient of separate bootstrapping is that since we rely on the validated clustering solution, we ignore the model uncertainty including the weights of each cluster. We computed the confidence intervals using 1000 bootstrap samples, and we used the results to evaluate the significance of the estimated parameters.

All computations were done using R, and a specific package should be released soon. In the meantime, a first version of the R syntaxes is available on <https://github.com/ztau/5352>.

### 2.3 *GMM as a Gold Standard Alternative*

To evaluate the HMTD approach as a tool for clustering sequences of continuous data, we need a gold standard alternative. We choose the Growth Mixture Model (GMM) approach for that purpose, since it is the only true longitudinal clustering tool used in the social sciences.

Growth modeling includes several similar frameworks aiming to model and discover the patterns of individual changes in a longitudinal data framework (Reinecke and Seddig 2011; McArdle and Epstein 1987). The basic growth model assumes that all trajectories belong to the same population and that they may be approximated by a single average growth trajectory using a single set of parameters. However, several models extend these assumptions; for example, the latent class growth analysis (LCGA) model, which assumes null variance-covariance for the growth trajectory within each class (Nagin 1999; Jung and Wickrama 2008), and the heterogeneity model (Verbeke and Lesaffre 1996), which goes a bit further but still imposes the same variance-covariance structure within each group of subjects. Therefore, we discuss the more flexible GMM in this section and use it in our analysis as gold standard.

The GMM developed by Muthén and Shedden (1999), Bauer and Curran (2003), and Wang and Bodner (2007) is designed to discover and describe unknown groups of sequences that share a similar pattern. This method may be represented as a mixture of mixed-effects models in which each of the unknown subpopulations follows a distinct linear mixed-effects model. Its main advantage over other similar models—like the heterogeneity model (Verbeke and Lesaffre 1996)—is that it allows for estimation of a specific variance-covariance structure within each class (Francis and Liu 2015). Within-class inter-individual variation is possible for latent variables via distinct intercept and slope variances, represented by a class-specific fixed-effects and random-effects distribution. In other words, the variation in an expected group-specific trajectory is distinct for each group (heterogeneity in growth trajectories). Because of these advantages, the model is a reference point in continuous longitudinal data modeling with various applications in criminology (Francis and Liu 2015; Reinecke and Seddig 2011), health and medicine (Muthén and Shedden 1999; Ram and Grimm 2009), psychology, and social science (Muthén 2001), among others.

The GMM approach uses both observed and latent variables. The observed variables consist of a  $p$ -dimensional vector of continuous dependent variables  $Y$  (often a variable with repeated measurements) and a  $q$ -dimensional vector of covariates  $X$ . The latent variables are represented as a continuous  $m$ -dimensional vector  $\eta$ . Finally, to indicate the group in which each subject is included, we use a dummy variable with multinomial distribution stored in a  $k$ -dimensional binary vector  $c$  (Muthén and Shedden 1999). The equation of the GMM approach for individual  $i$  then becomes

$$Y_i = \Lambda \eta_i + \epsilon_i, \quad (1)$$

where  $\Lambda$  is a  $p \times m$  parameter matrix (or matrix with basis vectors) that can be seen as a matrix of factor loadings,  $\eta_i$  is a vector of latent continuous variables, and  $\epsilon_i$  is an error term vector with zero mean.

In our case, the latent variable parameter matrix  $\Lambda$  has one column with parameters for the latent factor accounting for the intercept and another for the latent factor accounting for the slope. The general equation for every  $\eta$  is

$$\eta_i = A c_i + \Gamma x_i + \zeta_i, \quad (2)$$

where  $A$  is a matrix with columns of intercept parameters for each class,  $\Gamma$  is an  $m \times q$  parameter matrix and  $\zeta_i$  is an  $m \times 1$  vector of zero mean residuals (and covariance matrix  $\Psi$ ).

If we assume that some time-independent covariates  $z$  could influence the group membership  $c_i$ , a multinomial logistic regression can be considered (with parameters  $a$  and  $b$ ) as follows:

$$P(c_i = K | z_i) = \frac{\exp^{a_k + b_k z_i}}{\sum_{c=1}^K \exp^{a_c + b_c z_i}}$$

An alternative notation of the model for subject  $i$  as part of class  $k$  at time  $t$  is

$$Y_{i,t|c_i=k} = X_{1i}(t)^T \beta + X_{2i}(t)^T \gamma_k + V_i(t)^T u_{ik} + w_i(t) + \epsilon_{i,t}, \quad (3)$$

where  $X_{1i}$  is a vector of covariates with common fixed effects  $\beta$ ,  $X_{2i}$  is a vector of covariates with class-specific fixed effects  $\gamma_k$ , and  $V_i$  is a set of covariates with individual class-specific random effects  $u_{ik}$ . Finally,  $w_i(t)$  is an autocorrelated Gaussian process with null mean and covariance equal to  $cov(w_i(t)w_i(s)) = \sigma_w^2 \exp(-\rho|t-s|)$ .

The GMM is estimated by maximizing its likelihood using an ordinary EM algorithm. The continuous latent variables  $\eta$  and group membership variables  $c$  are considered missing data. The R package *lcmm* (Proust-Lima et al. 2017) was used to compute the GMM.

## 2.4 Statistical Analyses

To start with, we used the HMTD model to identify the best clustering of the IAT dataset without covariates, considering solutions from two to five groups. The best solution was selected on the basis of the Bayesian Information Criterion (BIC) (Raftery 1995). We then added covariates to this first model and analyzed the two resulting models, with and without covariates, particularly focusing on the IAT trajectories that did change group when covariates were added to the initial model. In order to isolate the impact of the covariates from any other computational issue or local optimum, we used the optimal solution obtained without covariates as a starting point for the full model. Therefore, we observe how this new model escapes the previous optimum.

We then computed the GMM models using the same dataset, and we compared the classifications obtained with the HMTD and GMM approaches. The usefulness of each covariate for discriminating between groups was evaluated using either a chi-square test for categorical covariates, or a single factor ANOVA for continuous ones. Notice that since it is not easy to compare two solutions with different number of clusters, we chose to compute a four-cluster GMM solution with all covariates instead of finding its own optimal number of clusters.

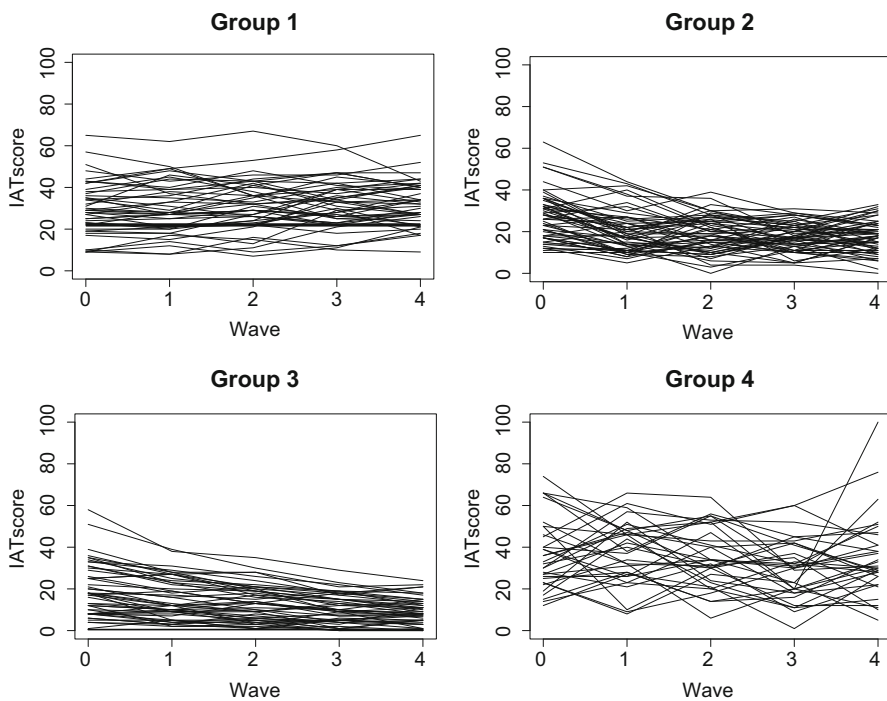
Our results are presented as figures displaying the IAT trajectories, and as tables describing the characteristics of subjects classified into groups and giving the HMTD model parameters.

### 3 Results

We provide here the results of the various clustering performed using the HMTD and GMM approaches, and we compare the resulting classifications. Notice however that given the iterative nature of the optimization algorithms, it is never possible to be sure that the final models are the best possible ones. Therefore, results should never be overinterpreted.

#### 3.1 HMTD Clustering

Without covariates, the best model identified by the BIC is a four-component model (model 1). Figure 1 shows the IAT trajectories in each group. We clearly differentiate a group with average volatility and IAT level (group 1), a group with relatively low scores and variability (group 2), a group with very low variability and a low and constantly diminishing IAT score (group 3), and a group with more complex trajectories and hence variability (group 4).



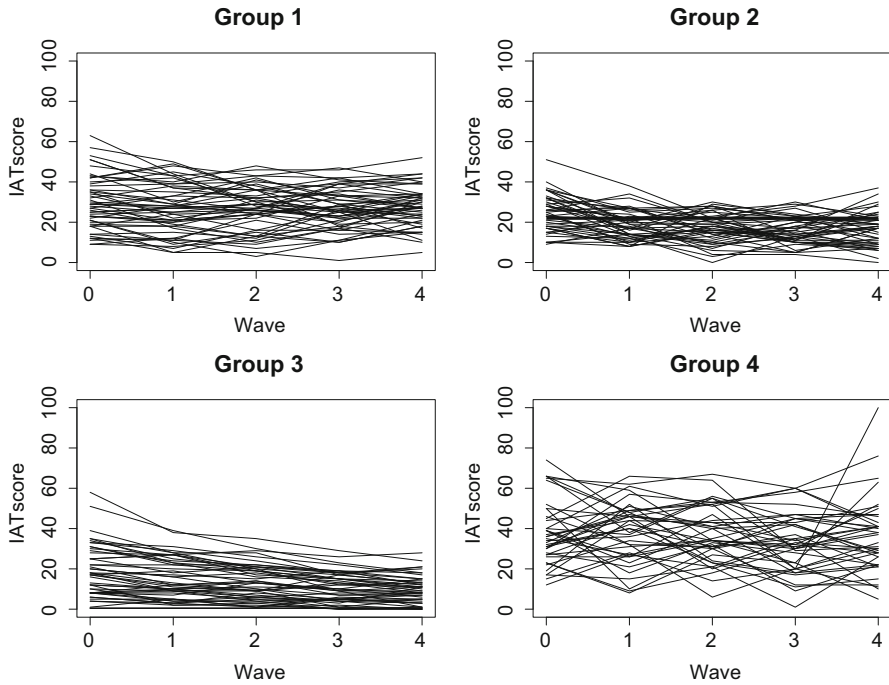
**Fig. 1** IAT trajectories associated with each group in the four-group HMTD solution without covariates (model 1)



When we include the covariates in model 1 (Fig. 2) and relabel the four groups of the solution in order to match the groups of model 1, we obtain a similar four-group structure (model 2). As a comparison of the two figures might show, the most important difference is with the first two groups: group 2 of model 2 lost its higher-valued trajectories and focused more on a low IAT-level and stable trajectories. This change will be explored in more details later.

Table 1 provides the parameter estimation for both models. In addition to the point estimates, we also provide the 95% bootstrap confidence intervals.

As regards the first model without covariates, the  $\theta_0$  parameters giving the variance of each component of the model confirm the first impression given by Fig. 1: Group 4 is characterized by a much higher variability than the three other groups, and group 3 has the lowest variance, indicating less variation among the successive observations of a single individual. Parameters  $\varphi_0$  corresponding to the constant in the modeling of the mean of each component also take expected values, with higher values associated with groups showing higher average IAT level. Finally, the autoregressive parameter  $\varphi_1$  takes a value closer to one for the groups with trajectories showing smoother evolutions from one wave to the next, that is groups 1 and 3. All parameters of this first model are significant at the 95% level, as demonstrated by the confidence intervals.



**Fig. 2** IAT trajectories associated with each group in the four-group HMTD solution with five covariates (model 2)

**Table 1** Estimated coefficients for the two HMTD models. For each parameter, we also provide the minimum and maximum values, and the 95% bootstrap confidence intervals obtained from 1000 bootstrap samples. Significant parameters at the 95% level are indicated with an asterisk

Model 2	$\theta_0$	$\phi_0$	$\phi_1$	WHO-5	BMI	Gender (male)	Age	Education (extended req.)
<b>Model 1</b>								
<b>group 1</b>	coefficient	41.098*	2.798*	0.891*				
$n = 46$	min, max	23.898, 108.133	1.881, 30.144	-0.070, 0.964				
	95% interval	(26.312; 38.317)	(3.417; 8.162)	(0.752; 0.919)				
<b>group 2</b>	coefficient	49.712*	7.228*	0.572*				
$n = 56$	min, max	0.001 89.683	5.721 22.126	-0.036 0.615				
	95% interval	(36.871; 54.309)	(6.507; 10.735)	(0.375; 0.563)				
<b>group 3</b>	coefficient	11.570*	1.072*	0.753*				
$n = 50$	min, max	7.535, 19.014	-0.102, 2.518	0.685, 0.814				
	95% interval	(8.715; 13.091)	(0.456; 1.737)	(0.715; 0.787)				
<b>group 4</b>	coefficient	186.582*	15.511*	0.514*				
$n = 33$	min, max	0.000, 384.934	6.951, 39.902	-0.186, 0.743				
	95% interval	(0.002; 291.178)	(14.906; 28.189)	(0.226; 0.560)				
<b>Model 2</b>								
<b>group 1</b>	coefficient	38.587*	-1.148	0.656*	1.331*	-2.663*	0.941	-0.468
$n = 52$	min, max	24.369 39.084	-27.000 47.072	0.371 0.714	-0.079 3.269	-7.317 0.084	-2.232 2.883	-3.271 4.786
	95% interval	(27.009 37.035)	(-9.082 31.695)	(0.507 0.672)	(0.596 2.535)	(-5.433 -1.672)	(-1.116 1.675)	(-2.107 1.928)
<b>group 2</b>	coefficient	43.371*	-1.444	0.539*	-1.025	0.549	0.638	-2.705
$n = 45$	min, max	0.530 86.137	-15.527 55.269	0.100 0.601	-3.692 0.427	-5.127 4.760	-3.290 1.838	-8.183 1.675
	95% interval	(34.931 52.838)	(-1.607 18.077)	(0.217 0.540)	(-1.950 0.073)	(-0.319 2.851)	(-0.545 0.888)	(-4.887 0.237)
<b>group 3</b>	coefficient	9.162*	0.256	0.723*	-1.424*	-0.757	0.142	1.060*
$n = 48$	min, max	5.987 11.771	-20.810 29.836	0.627 0.780	-3.110 -0.036	-4.049 2.320	-1.995 1.870	-0.547 5.000
	95% interval	(7.222 10.013)	(-10.036 16.612)	(0.666 0.758)	(-2.069 -0.900)	(-2.151 0.458)	(-1.030 0.959)	(0.350 2.288)
<b>group 4</b>	coefficient	153.507*	22.274	0.307*	-4.014*	11.303*	-1.135	2.063
$n = 40$	min, max	99.659 242.775	-48.500 96.927	-0.099 0.518	-9.000 -0.593	4.427 20.000	-6.000 4.656	-6.000 8.000
	95% interval	(108.832 198.571)	(-48.500 54.129)	(0.011 0.378)	(-7.542 -2.135)	(7.541 18.251)	(-3.493 4.095)	(-3.450 7.002)

As regards model 2, even if the first three parameters ( $\theta_0$ ,  $\varphi_0$ , and  $\varphi_1$ ) take values different from those of model 1,  $\theta_0$  and  $\varphi_1$  take values in the same range as of model 1. On the other hand, important differences are found for the constant parameter  $\varphi_0$ , and this parameter is no more significant in any group. Note that  $\theta_0$  and  $\varphi_1$  tend to take smaller values in model 2. This can be interpreted as the first proof of interest of the covariates included in model 2: the groups are now more homogeneous (lower intra-group variance) and the explanation of a specific trajectory relies less on the immediately preceding observation. As regards the covariates, Age is never significant and could be eventually removed from the model. This could be due to the lack of a real age difference between participants (from 13 to 15 years old at baseline). However, the four other covariates remain significant for at least one of the groups.

When we consider each component of model 2 separately, the changes occurring in the trajectories associated with the first component are found related to the well-being of the concerned adolescents: a higher well-being such as measured by the WHO-5 index is significantly associated with a lower IAT-level. Males tend to have a lower IAT level than females, and a higher BMI is associated with higher IAT level. In group 3, a higher WHO-5 or BMI is associated with reduced IAT level, but being in the extended requirement school track is associated with a higher IAT level. Finally, in group 4, a higher WHO-5 or BMI is associated with reduced IAT level, and males tend to show a much higher IAT level than females.

Table 2 provides the main characteristics of the subjects classified into each group. For time-dependent variables, we considered the average value of each individual. A comparison is performed for each variable separately to test whether the groups are significantly different with regard to the variable. Considering only the two HMTD models, we observe that in addition to the expected differences in IAT level, the only other variable with significantly different values across groups is the WHO-5 measure of well-being. For both models, we observe two groups (2 and 3) with lower average IAT scores. The same two groups also display higher emotional well-being, as compared to the other groups, confirming previous results (Surís et al. 2014). No differences are observed for the other covariates, even if Gender comes close to significance in model 1. Even if not significant at the 95% level, probably because of the reduced sample size, we find a gender separation at the sample level; groups 2 and 4 contain a higher proportion of boys compared to the other two groups. The education track also shows a difference at the sample level: the first two groups contain more individuals following the highest education track as compared to groups 3 and 4. On the other hand, no notable difference is observed between the groups for Age and BMI, even if BMI, used as a covariate in model 2, is statistically significant in the modeling of the mean of each component.

**Table 2** Characteristics of subjects classified into groups for different clustering. The *p*-value gives the result of the test comparing the different groups for each variable. The number of sequences classified into each group is provided in brackets after the group number

	IAT mean (sd)	WHO-5 mean (sd)	BMI mean (sd)	Gender % male	Age at $T_0$ mean (sd)	Educ. at $T_0$ % extended req.
<b>HMTD model 1</b>						
Group 1 (46)	30.94 (11.7)	63.43 (15.6)	19.97 (2.35)	24	14.13 (0.50)	80.5
Group 2 (56)	20.29 (9.78)	71.01 (15.6)	20.02 (3.30)	45	14.05 (0.59)	67.9
Group 3 (50)	13.31 (9.88)	72.28 (13.6)	20.45 (2.57)	24	14.14 (0.67)	64.0
Group 4 (33)	34.69 (16.1)	63.49 (16.8)	20.06 (3.03)	39	14.27 (0.45)	60.6
<i>p</i>	<0.001	<0.001	0.764	0.055	0.381	0.214
<b>HMTD model 2</b>						
Group 1 (52)	27.43 (11.3)	67.35 (16.6)	20.12 (2.40)	31	14.19 (0.60)	71.2
Group 2 (45)	18.57 (8.41)	70.85 (15.2)	19.96 (3.53)	40	14.02 (0.45)	73.3
Group 3 (48)	13.62 (9.97)	70.64 (14.0)	20.46 (2.54)	21	14.10 (0.69)	64.6
Group 4 (40)	36.36 (15.6)	63.06 (16.4)	19.96 (2.86)	43	14.22 (0.48)	65.0
<i>p</i>	<0.001	0.015	0.741	0.113	0.331	0.746
<b>GMM 2</b>						
Group 1 (169)	22.08 (13.2)	68.79 (15.8)	20.20 (2.90)	32	14.15 (0.57)	0.68
Group 2 (16)	39.90 (14.8)	61.13 (13.9)	19.40 (2.12)	43	14.00 (0.52)	0.75
<i>p</i>	<0.001	0.022	0.210	0.496	0.322	0.771
<b>GMM 4</b>						
Group 1 (76)	13.35 (8.97)	73.32 (14.2)	20.69 (2.75)	32	14.09 (0.61)	0.63
Group 2 (31)	38.98 (11.2)	58.48 (16.2)	20.15 (2.40)	29	14.16 (0.52)	0.74
Group 3 (75)	26.46 (10.2)	67.09 (15.4)	19.62 (2.98)	33	14.17 (0.55)	0.73
Group 4 (3)	54.06 (18.3)	62.40 (9.66)	18.78 (3.30)	100	14 (0)	2/3
<i>p</i>	<0.001	<0.001	0.043	0.094	0.802	0.593
<b>GMM 4 cov</b>						
Group 1 (98)	18.79 (10.6)	69.64 (14.6)	20.06 (2.90)	29	13.91 (0.32)	77.9
Group 2 (44)	18.58 (10.5)	68.88 (17.8)	20.85 (2.95)	24	15.16 (0.55)	44.0
Group 3 (28)	39.38 (12.9)	64.95 (18.2)	20.14 (2.73)	48	14.24 (0.44)	48.3
Group 4 (15)	41.98 (14.8)	60.00 (13.7)	19.36 (2.05)	54	14.00 (0.41)	76.9
<i>p</i>	<0.001	0.032	0.321	0.058	<0.001	<0.001

### 3.2 Usefulness of the Covariates

From the results of the previous section, we find that the inclusion of covariates in the first classification obtained with the HMTD model helped us better differentiate the four groups, but without entirely changing their interpretation. We would like to better understand the changes in trajectory classification that occurred between these two models. Table 3 indicates how many subjects changed groups between the initial model without covariates and model 2 with covariates. As noted earlier, most of these changes occurred between groups 1 and 2. In particular, 19 second-group subjects of model 1 were transferred to the first group in model 2, and the steady

**Table 3** Number of IAT trajectories associated with each group in HMTD models 1 (without covariates, rows) and 2 (including covariates, columns)

Model 1	Model 2			
	Group 1	Group 2	Group 3	Group 4
Group 1	31	6	2	7
Group 2	19	34	1	2
Group 3	2	3	45	0
Group 4	0	2	0	31

**Table 4** The characteristics of 19 subjects moving from group 2 to group 1 (group 2→1) as compared to subjects staying in the same group (either 1 or 2) in both HMTD classifications. The means (numerical variables) or proportions (categorical variables) are provided, and differences with the subjects remaining in the same group (either 1 or 2) are assessed using *t*-tests and  $\chi^2$ -tests with continuity correction: ns: non-significant

	IAT	WHO-5	BMI	Sex (% male)	Age	Education
<b>Group 2→1</b>	22.76	72.93	19.69	57.9	14.26	52.6
<b>vs group 1</b>	31.26**	62.77**	20.41 ns	9.70***	14.16 ns	80.6 ns
<b>vs group 2</b>	18.72 ns	71.42 ns	20.30 ns	38.2 ns	13.97 ns	73.5 ns

\**p* < 0.05; \*\**p* < 0.01; \*\*\**p* < 0.001

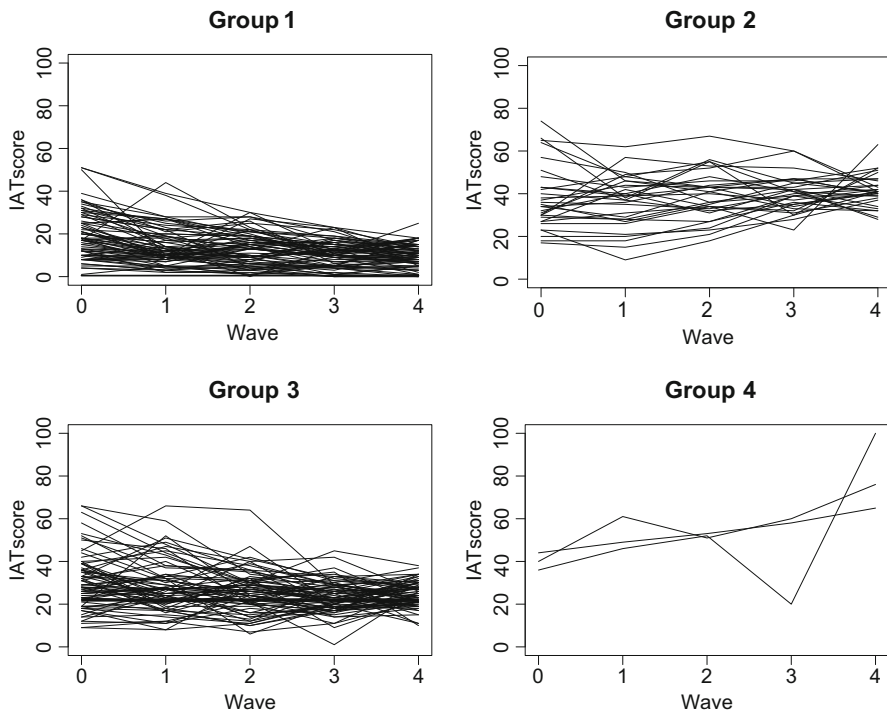
low Internet addiction profile of the second group became even more pronounced, with the higher Internet addiction subjects joining the first group. However, since some trajectories simultaneously left group 1 for the three other groups, the average IAT level of group 1 also decreased. Overall, the inclusion of covariates appears beneficial for the differentiation of trajectory features among groups.

The 19 individuals who switched from group 2 to group 1 represent the main difference between the two models, with all the other changes concerning at the most seven subjects. Thus, it is interesting to explore how these individuals differed from those who remained in the first or second group in both classifications. Table 4 summarizes our findings using *t*-tests and  $\chi^2$ -tests to compare the different variables. The average IAT scores are quite different between the three considered sub-groups, and, as expected, the “moving” sub-group shows an Internet dependence level between the two “stable” sub-groups. Thus, the moving individuals were among the most Internet-problematic members of the full second group of model 1, and even if the average IAT score is not the only indicator of group affiliation, a visualization of the trajectories would confirm the ambiguous nature of these individuals. The moving subgroup is also significantly different from the group of individual staying in group 1 as regards the WHO-5 index of well-being and the gender ratio, with a higher well-being and higher proportion of males among the moving subgroup. No other significant differences are observed.

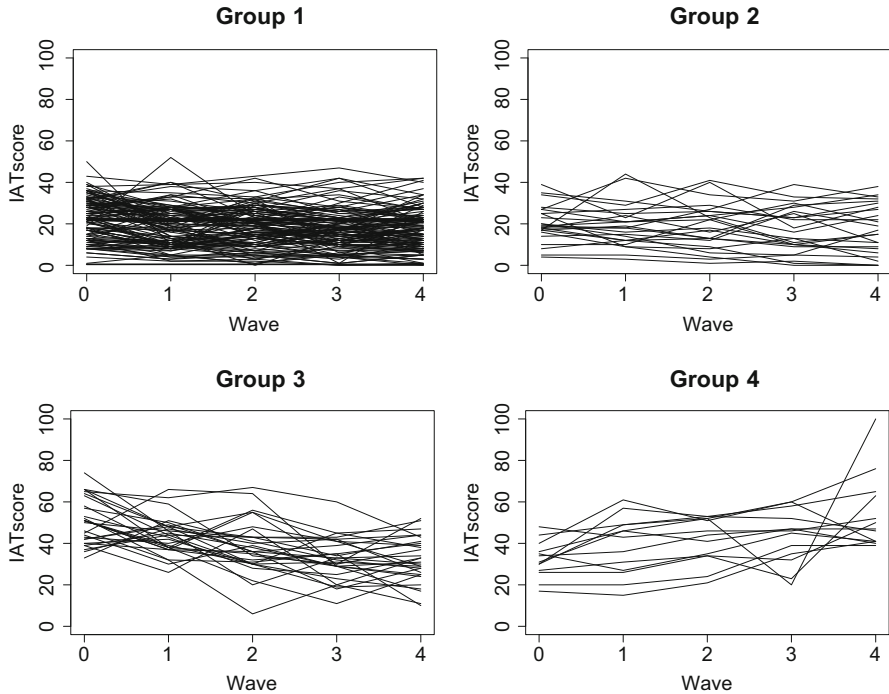
### 3.3 GMM Clustering

Without covariates, the best GMM solution in terms of BIC is a two-group solution, but given the high difference in number of trajectories associated to each group (169 vs 16), this solution is not really interpretable and hence less useful than the four-group solution given by the HMTD approach. Therefore, we also estimated a four-group GMM (Fig. 3).

In the two-group solution, a large majority of trajectories are associated with group 1, and only 16 sequences are associated with group 2. The average IAT level is higher in group 2, but both groups exhibit an important variability, as indicated in Table 2. Moreover, in terms of interpretation, one can only say that IAT sequences with a clear increasing trend are separated from the other sequences. In the four-group solution, even if the number of groups is the same as in the HMTD models, there is no a priori correspondence between the HMTD and GMM groups. In the four-group GMM solution, the number of subjects per group shows much more variability than that observed with the HMTD group, with the majority of individuals classified in groups 1 or 3, and only three subjects in group 4.



**Fig. 3** IAT sequences associated with each group in the four-group GMM solution without covariates



**Fig. 4** IAT sequences associated with each group in the four-group GMM solution with four covariates

Finally, as with the HMTD approach, we enhanced the four-group solution by adding covariates. Four of the five covariates used in the HMTD approach appeared useful in the GMM solution as well. Figure 4 displays the resulting groups obtained after adding Gender and Education as predictors for group membership (multinomial regression on  $c_i$ ), and WHO5 and BMI as fixed effect. On the other hand, Age was not included in the final model because the estimation process would then lead to a one-group solution. Another important issue with the GMM approach is the results' sensitivity to the order in which the covariates are included in the model. Various covariate combinations were tested before we chose the above-mentioned combination as the best one in terms of clustering results. For instance,  $classmb = gender + education\ sector$  does not give the same results as  $classmb = education\ sector + gender$ . This surprising result may be due to a bug in the *lcmm* R package, but in our opinion the reason could rather be the optimization procedure. It is well known that EM-type algorithms converge to the nearest local optimum, and that this optimum is not always the global one. Therefore, the solution depends on the initial values of the parameters, especially when the solution space is complex, which is the case here.

As Fig. 4 shows, the number of trajectories associated with each group is quite variable, with the large majority assigned to group 1. The first two groups are

characterized by low variability and an overall low IAT level. The trajectories in these two groups seem very similar, but since this four-group solution might be suboptimal and is computed only for the purpose of comparison with HMTD models 1 and 2, a three-group solution could merge these two groups into one group. The last two groups have a higher average IAT level, both exhibiting a general linear trend over time, decreasing in group 3 and increasing in group 4.

Table 2 gives the characteristics of individuals classified in each group of the GMM models and compares the groups for each variable. Note that given the large differences in group size, the test results for the GMM models should be interpreted with caution. As observed earlier in the HMTD case, significant differences exist between groups for both the IAT and WHO-5 variables. A significant difference exists also for BMI in the four-group GMM model without covariates. More interestingly, the Age and Education track at baseline also show significantly different values across groups, with one of the variables (Education track) being included in the model as covariable, but not the other. This difference between the HMTD and GMM clustering points to the fact that the solutions provided by both approaches are not identical or interchangeable, and that the two models used information in a different manner to provide usable data sequence clusterings.

## 4 Comparison of HMTD and GMM

When used for clustering purposes, the HMTD and GMM models share some characteristics: They both represent a kind of mixture model, they can include covariates of any type at the visible level, and they can also include covariates at the latent level and use them to estimate the initial probability of each cluster. However, HMTD and GMM also have several differences. First of all, since GMM is a mixture of mixed models, it is able to accept both fixed and random effects. Another difference is the possibility of HMTD to include an autoregressive specification for the variance and thus to allow for the clustering of longitudinal sequences whose variance evolves in time. For instance, sequences becoming more instable over time can more easily be grouped together. However, to exploit this feature, it is necessary to work with long data sequences, what was not the case here with the IAT example.

Another feature of HMTD that is worth stressing is the possibility of using it to perform different kind of clustering (Bolano and Berchtold 2016). The transition between components is driven by the hidden transition matrix  $A$ . In this paper,  $A$  was constrained to be a diagonal identity matrix, implying that each sequence was assigned to one and only one group, and all sequences assigned to the same group were described by the same visible model. However, there are several alternatives. For instance, different latent states may be required to alternate over time in order to find the optimal modeling of a given sequence. If  $A$  is constrained to have the following structure:



$$A = \begin{pmatrix} a_1 & 1 - a_1 & 0 & 0 \\ a_2 & 1 - a_2 & 0 & 0 \\ 0 & 0 & a_3 & 1 - a_3 \\ 0 & 0 & a_4 & 1 - a_4 \end{pmatrix}$$

where  $a_1, a_2, a_3$  and  $a_4$  are transition probabilities, then one performs at the same time a modeling and a clustering of the data sequences. The first two states are used to model the first cluster, and states 3 and 4 are used to model the second cluster. In other words, data sequences are clustered into two groups, but inside each group there are two different visible models allowing for a better representation of these sequences when their behavior evolves over time.

Another specification of  $A$  would allow some sequences to remain always in the same cluster, whereas other ones could transit at some point in time from the first to the second cluster:

$$A = \begin{pmatrix} a_1 & 1 - a_1 & 0 & 0 \\ a_{21} & a_{22} & 1 - a_{21} - a_{22} & 0 \\ 0 & 0 & a_3 & 1 - a_3 \\ 0 & 0 & a_4 & 1 - a_4 \end{pmatrix}$$

## 5 Conclusion

Hidden Markovian models are known to be valuable tools to analyze the dynamics in longitudinal continuous data and in life course data (e.g. Helske et al. 2018). The present study demonstrates that the sequences of continuous longitudinal data can also be classified into as many groups as required, and that the HMTD model can be used as a valid alternative to GMM. The inclusion of covariates has beneficial effects on clustering, because the resulting groups have lower intra-variability compared to the solution without covariates.

In a comparative study involving the use of GMM for clustering, our first finding is that the HMTD approach is a good alternative to GMM, because in terms of interpretability its results are at least as interesting as the results given by GMM. However, on the basis of just one practical example, we obviously cannot conclude that one approach is better than the other; moreover, this is not the purpose of this study. What we can conclude is that the HMTD approach is not only theoretically, but also practically useful to classify sequences of continuous data in mutually excluding groups.

In the literature, excessive Internet use has been found to be highly related to several somatic conditions, sleep disturbance in particular. However, in this paper, our main objective is not to explain IAT trajectories, but to find ways to classify such trajectories into meaningful groups. Moreover, there is still an ongoing debate on the direction of the relationship between Internet use and sleep disturbance, not

to speak of causality. Therefore, we chose not to consider sleep disturbance in this analysis, but to concentrate on other covariates that are more neutral to IAT scores. Nevertheless, even with this restriction, the results obtained with the HMTD model are highly significant and allow for a sound interpretation. The four resulting groups differ in terms of average value and variability. The relationship observed between IAT and the emotional well-being of subjects suggests that both concepts are linked and that a higher risk of Internet addiction is related to poorer well-being. Gender is also a discriminating factor between groups, with a lower proportion of males in the first and third groups, but, given the small sample size, the differences are not significant at the population level.

The main strength of this study is the demonstration of the usefulness of the HMTD approach as a valuable alternative to the GMM approach for clustering continuous data sequences. Researchers would be advised to consider both approaches to take full advantage of the information in their data. However, some weaknesses of this study are to be mentioned. At the theoretical level, we include covariates in the HMTD model only at the visible level, but it is also possible to include them at the latent level as well in order to enhance the prior probabilities of each cluster. As regards the application of the model to IAT trajectories, we used a rather small convenience sample; this is not representative of the population of adolescents living in the canton of Vaud. More analyses need to be conducted with larger databases to define a real typology of IAT trajectories.

Overall, in spite of some shortcomings, the HMTD model can be considered as a complete framework for the analysis of continuous data sequences. It is an explanatory tool as well as a clustering tool, and by adding covariates, constraints on the transition matrix, and autoregressive modeling of the mean and variance of each component, the model goes well beyond the traditional Markovian models such as homogeneous Markov chains or hidden Markov models.

**Acknowledgements** This publication benefited from the support of the Swiss National Centre of Competence in Research LIVES – Overcoming vulnerability: Life course perspectives, which is financed by the Swiss National Science Foundation (grant number: 51NF40-160590). The ado@internet.ch study was financed by the Public Health Department of the Vaud canton and by the Swiss National Science Foundation (grant number: 105319\_140354). The authors are grateful to both institutions for their financial support. The funding bodies had no role in the design and conduct of the study; collection, analysis and interpretation of data; or preparation, review, or approval of the manuscript.

## References

- Barrense-Dias, Y., Berchtold, A., Akre, C., & Suris, J. C. (2015). The relation between internet use and overweight among adolescents: A longitudinal study in Switzerland. *International Journal of Obesity*, 40, 45–50.
- Bauer, D., & Curran, P. (2003). Distributional assumptions of growth mixture models: Implications for overextraction of latent trajectory classes. *Psychological Methods*, 8(3), 338–363.

- Berchtold, A. (2001). Estimation in the mixture transition distribution model. *Journal of Time Series Analysis*, 22, 379–397.
- Berchtold, A. (2003). Mixture transition distribution (MTD) modeling of heteroscedastic time series. *Computational Statistics and Data Analysis*, 41, 399–411.
- Berchtold, A., & Raftery, A. (2002). The mixture transition distribution model for high-order Markov chains and non-Gaussian time series. *Statistical Science*, 17, 328–356.
- Bolano, D., & Berchtold, A. (2016). General framework and model building in the class of hidden mixture transition distribution models. *Computational Statistics and Data Analysis*, 93, 131–145.
- Faraci, P., Craparo, G., Messina, R., & Severino, S. (2013). Internet addiction test (IAT): Which is the best factorial solution? *Journal of Medical Internet Research*, 15(10), e225.
- Forney, G. D. (1973). The Viterbi algorithm. *Proceedings of the IEEE*, 61, 268–278.
- Francis, B., & Liu, J. (2015). Modelling escalation in crime seriousness: A latent variable approach. *Metron*, 73(2), 277–297.
- Helske, S., Helske, J., & Eerola, M. (2018). Analysing complex life sequence data with hidden Markov modelling. In G. Ritschard & M. Studer (Eds.), *Sequence analysis and related approaches: Innovative methods and applications*. Cham: Springer (this volume).
- Jung, T., & Wickrama, K. A. S. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass*, 2(1), 302–317.
- Khazaal, Y., Billieux, J., Thorens, G., Khan, R., Scarlatti, E., Theintz, F., Lederrey, J., Van Der Linden, M., & Zullino, D. (2008). French validation of the internet addiction test. *Cyberpsychology Behavior*, 11(6), 703–706.
- McArdle, J. J., & Epstein, D. (1987). Latent growth curves within developmental structural equation models. *Child Development*, 58(1), 110–133.
- Muthén, B. O. (2001). Latent variable mixture modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 1–33). Mahawa: LEA.
- Muthén, B. O., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics*, 55(2), 463–469.
- Nagin, D. (1999). Analyzing developmental trajectories: A semiparametric, group-based approach. *Psychological Methods*, 4(2), 139–157.
- Piguet, C., Berchtold, A., Zimmermann, G., & Surís, J. C. (2016). Rapport final de l'étude longitudinale AdoInternet.ch. Lausanne: Raisons de santé.
- Proust-Lima, C., Philipps, V., & Liqueur, B. (2017). Estimation of extended mixed models using latent classes and latent processes: the R package lcmm. *Journal of Statistical Software*, 78(2), 1–56.
- Raftery, A. (1985). A model for high-order Markov chains. *Journal of the Royal Statistical Society, Series B*, 47(3), 528–539.
- Raftery, A. (1995). Bayesian model selection in social research. *Sociological Methodology*, 25, 111–163.
- Ram, N., & Grimm, K. J. (2009). Growth mixture modeling: A method for identifying differences in longitudinal change among unobserved groups. *International Journal of Behavioral Development*, 33(6), 565–576.
- Reinecke, J., & Seddig, D. (2011). Growth mixture models in longitudinal research. *AStA Advances in Statistical Analysis*, 95(4), 415–434.
- Skarupova, K., Olafsson, K., & Blinka, L. (2015). Excessive internet use and its association with negative experiences: Quasi-validation of a short scale in 25 European countries. *Computers in Human Behavior*, 53, 118–123.
- Surís, J. C., Akre, C., Berchtold, A., Fleury-Schubert, A., Michaud, P. A., & Zimmermann, G. (2012). Ado@Internet.ch: Usage d'internet chez les adolescents vaudois. Raisons de santé 208. Lausanne: Institut universitaire de médecine sociale et préventive.
- Surís, J. C., Akre, C., Piguet, C., Ambresin, A. E., Zimmermann, G., & Berchtold, A. (2014). Is internet use unhealthy? A cross-sectional study of adolescent internet overuse. *Swiss Medical Wkly*, 144, w14061.

- Taushanov, Z., & Berchtold, A. (2017). A direct local search method and its application to a Markovian model. *Statistics, Optimization and Information Computing*, 5(1), 19–34.
- Verbeke, G., & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association*, 91(433), 217–221.
- Wang, M., & Bodner, T. E. (2007). Growth mixture modeling identifying and predicting unobserved subpopulations with longitudinal data. *Organizational Research Methods*, 10(4), 635–656.
- Young, K. S. (1998). Internet addiction: The emergence of a new clinical disorder. *CyberPsychology & Behavior*, 1(3), 237–244.

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

