




Model selection for mixture hidden Markov models: an application to clickstream data

Furio Urso¹ · Antonino Abbuzzo¹  · Marcello Chiodi¹ · Maria Francesca Cracolici¹

Received: 21 August 2024 / Revised: 21 August 2024
© The Author(s) 2024

Abstract

In a clickstream analysis setting, Mixture Hidden Markov Models (MHMMs) can be used to examine categorical sequences assuming they evolve according to a mixture of latent Markov processes, each related to a different subpopulation. These models involve identifying both the number of subpopulations and hidden states. This study proposes a model selection criterion based on an integrated completed likelihood approach that accounts for the two latent classes in the model. We implemented a Monte Carlo simulation study to compare selection criteria performance. In scenarios characterised by categorical short length sequences, our proposed measure outperforms the most commonly used model selection criteria in identifying components and states. The paper presents a case study on clickstream data collected from the website of a company operating in the hospitality industry and modelled by an MHMM selected by the proposed score.

Keywords Model selection · Clusters · Hidden states · Clickstream data · Entropy-based scores · Information criteria

Mathematics Subject Classification 60J10 · 60J20 · 62B10 · 62H30 · 62J12 · 62M05 · 62P20

✉ Furio Urso
furio.urso@gmail.com

✉ Antonino Abbuzzo
antonino.abbuzzo@unipa.it

Marcello Chiodi
marcello.chiodi@unipa.it

Maria Francesca Cracolici
maria.francesca.cracolici@unipa.it

¹ Department of Economics, Business and Statistics, University of Palermo, Viale delle Scienze, 90146 Palermo, Sicily, Italy

1 Introduction

Nowadays, clickstream data are an essential source of information businesses employ to explore user behaviour on the web and determine prompt and effective business strategies (Cooley and Srivastava 2000; Liu and Kešelj 2007; Das and Turkoglu 2009). Clickstream data, collected in log files, enable us to track users' activity as they explore a website. A log file usually contains information on the pages visited, such as time spent on each page and additional web resources on that page, e.g. images and links. The user path can be obtained as a sequence whose elements correspond to a web page visited by the user at time t (a click). A severe limitation in using clickstream data is the lack of information about browsing behaviour. The data, for example, do not explain why users navigate from one page to another or precisely what they are looking for. Therefore, sequences that are similar in their order may represent different subpopulations of units. To extract this type of information from the data, allowing dissimilarities in browsing behaviour and subpopulations to be identified would be extremely useful in determining suitable online business strategies.

Mixture Hidden Markov Models (MHMMs) could be used to this end as they allow similarities to emerge among sequences of different sets of pages that satisfy similar user needs. Here, similarities refer to different unknown *mental states* governing the user's browsing behaviour; these are hidden states. MHMMs enable us to account for the evolution of a latent process, that is, why a user moves from one page to another and for a hidden variable related to the presence of clusters representing behaviour profiles. However, there are few applications of mixture hidden Markov models to identify subpopulations in clickstream data. One interesting study was that of Smyth (1997), who used these models as a clustering technique to classify generic sequences of observations. Smyth (1999) indicated web browsing behaviour as an interesting application. Later, Scott and Hann (2006) presented an application to multiple web sequences related to different web sessions for each user. Another application for clickstream data is that of Ypma and Heskes (2002), who classify users assuming prior information on the hidden states.

Our study applies time-discrete first-order mixture hidden Markov models to analyse clickstream data collected from the website of a company operating in the hospitality industry. Additionally, the paper aims to enrich the literature from a methodological perspective by proposing a model selection criterion based on an integrated completed likelihood approach that accounts for the two latent classes in the model: subpopulations (i.e. the mixture components) and hidden states.

Indeed, one of the main issues in applying MHMMs concerns selecting an unknown number of mixture components and states in grouping sequences. This selection is generally based on a priori information about the problem under analysis. However, when this is not known, model selection criteria are generally based on scores derived from Information Criteria (ICs), such as the Bayesian Information Criterion (BIC, Schwarz et al. 1978), the Akaike Information Criterion (AIC, Akaike 1974), and their variations, like the sample size adjusted BIC or ssBIC (Rissanen 1978). Although identifying the correct number of components and states is not a simple task, ICs have been widely employed for mixture models or (hidden) Markov models; their use has also been extended to MHMMs, although the limitations of their application have

received little attention in the literature (see e.g. Dias 2006; Celeux and Durand 2008; Helske et al. 2018).

ICs should be used carefully when we deal with a mixture of hidden Markov models, and their behaviour needs to be explored. The BIC is commonly used on the assumption that either the number of hidden states or clusters is known (Dias et al. 2009). However, it is also used in cases where both these numbers are unknown, even though its performance is not satisfactory and it is outperformed by other criteria if the target is to identify clusters in Mixture Markov models (e.g. Dias (2007)). In single-sequence HMMs, the BIC is proven to be consistent as the sequence length increases (see Boucheron and Gassiat 2007), but its performance in the MHMM context with multiple sequences of different (short) lengths has not been explored. It is worth noting that, ICs do not consider the degree of separation between latent classes, which may lead to selecting a model that identifies a low-quality data partition, thus making interpretation more difficult. For this reason, another approach to model selection, based on Classification Criteria (CCs), has been developed. These refer to complete-data log-likelihood and produce a model selection that accounts for the classification quality of the latent classes through a measure of entropy. This approach has received little attention in the literature on mixture hidden Markov models, although it has been used for Mixture Models (MMs) and Hidden Markov Models (HMMs).

As far as MMs are concerned, the Integrated Classification Likelihood criteria (ICL) (Biernacki and Govaert 1997), and an approximation of the integrated completed likelihood based on BIC (Biernacki et al. 2000) are used. Regarding HMMs, Celeux and Durand (2008) showed that AIC, BIC and the ICL behave similarly as they do in MMs (McLachlan and Peel 2004). AIC did show a tendency to select more complex models, while BIC and ICL behave similarly when sequence length increases. BIC performs better mainly if states are poorly separated. In the context of MHMMs, an interesting proposal along this line is the study presented by Volant et al. (2014). They suggested an HMM, presenting mixture components in the emission probabilities of the chain, which are added to a classic HMM for a continuous-valued sequence. The authors identified groups of observations by combining the components in the emission probabilities and, to determine the number of groups, they proposed an ICL based on BIC that involves only a measure of entropy related to the hidden states.¹

Our study aims to enrich this stream of literature on ICL by defining a model selection criterion for discrete-time first-order MHMMs, called BIC_H , adapting the entropy-based information criterion defined by Biernacki and Govaert (1999). The criterion allows us to group similar categorical sequences (in our case, web sessions) in order and time, assuming that they have been generated by a hidden Markov model (HMM) having a specific number of states. Therefore, we assume that sequences related to different HMMs (i.e., the mixture components) define well-separated clusters of sequences with high dissimilarities. In our study, we refer to time-constant categorical covariates, single-channel and single categorical response discrete-time first-order MHMMs (see Vermunt et al. 2008) because they are particularly appropriate in managing clickstream data. We assumed that the HMM components are

¹ Following Baudry et al. (2010), the authors outlined a procedure for combining components of a mixture model into clusters, measuring the classification quality through an entropy index.

first-order hidden Markov processes to simplify the structure of the model in this first phase of model selection analysis. We also consider a single web page sequence for each user, with the only available covariates being features related to IP addresses.

To summarize, the aim of this paper is twofold. First, it proposes an integrated completed likelihood based on BIC for MHMMs that simultaneously identifies components and states. Second, the study enriches the empirical literature on clickstream data by applying MHMMs to identify user profiles with similar browsing behaviour. Clickstream data from LovePanormus,² a Sicilian company operating in the hospitality sector, were used in the empirical analysis.

The paper is structured as follows: Sect. 2 introduces MHMMs for categorical observations with time-constant covariates. In Sect. 3, we present our proposal for model selection. Section 4 contains a simulation study. Finally, a case study is illustrated in Sect. 5, and the proposed criterion is used to identify profiles in clickstream data collected from the website of LovePanormus.

2 Statistical models

This section briefly introduces hidden Markov models and then moves on to their extension to MHMMs. Specifically, we focused on mixture of first-order hidden Markov models with mixture weights that depend on time-constant covariates and observed sequences generated by a discrete random variable. In these models, a discrete latent variable reflects different longitudinal patterns in the sequences. These patterns depend on the presence of unknown subpopulations in the data, i.e. latent classes, which, in turn, are obtained through sequences assigned to them with specific probabilities (Van de Pol and Langeheine 1990).

2.1 Hidden Markov models

Let $Y = (Y_1, Y_2, \dots, Y_T)$ be a discrete random vector of length T , and each element Y_t , $t = 1, 2, \dots, T$, assumes values in the discrete set $\mathcal{R} = \{1, \dots, R\}$, i.e. the observed states. A discrete first-order HMM comprises a Markov chain denoted by $U = (U_1, U_2, \dots, U_T)$, with state space of the chain $\mathcal{S} = \{1, \dots, S\}$; as a first-order Markov chain, the probability of the state at time t depends only on the state at previous time $t - 1$. The parameters that characterize a discrete-time first-order HMM are the probability of initial hidden state $\pi = \{\pi_s\}_{s \in \mathcal{S}} = \{\Pr(U_1 = s)\}_{s \in \mathcal{S}}$; the $S \times S$ transition probability matrix $A = \{a_{hj}\}_{h,j \in \mathcal{S}} = \{\Pr(U_t = h | U_{t-1} = j)\}_{h,j \in \mathcal{S}}$, whose element a_{hj} indicates the probability of making a transition from state h at time $t - 1$ to state j at time t . Finally, the $S \times R$ emission matrix $B = \{b_{sr}\}_{s \in \mathcal{S}, r \in \mathcal{R}} = \{\Pr(Y_t = r | U_t = s)\}_{s \in \mathcal{S}, r \in \mathcal{R}}$ connects the hidden states and the observed states, where $b_{sr} = b_s(r)$ is the probability of hidden state s emitting observed state r .

Let y_i be the i -th realization of Y , assume we collect n independent and identically distributed sequences $y = (y_1, \dots, y_i, \dots, y_n)$, with $y_i = (y_{i1}, y_{i2}, \dots, y_{iT})$, generated by the same hidden Markov model. Let us assume, without loss of generality,

² The company's name is a pseudonym.

that the n sequences have the same length T . The model parameters $\Theta = \{\pi, A, B\}$ are estimated by maximising the log-likelihood

$$\begin{aligned} \ell(\Theta; y) &= \sum_{i=1}^n \log P(y_i|\Theta) \\ &= \sum_{i=1}^n \log \left(\sum_{s, f \in \mathcal{S}} P(U_1 = s|\Theta)P(y_{i1}|U_1 = s, \Theta) \prod_{t=2}^T P(U_t = s|U_{t-1} = f, \Theta)P(y_{it}|U_t = s, \Theta) \right) \\ &= \sum_{i=1}^n \log \left(\sum_{s, f \in \mathcal{S}} \pi_s b_s(y_{i1}) \prod_{t=2}^T a_{s, f} b_s(y_{it}) \right), \end{aligned}$$

where the hidden sequence $U = (U_1, U_2, \dots, U_T)$ take all possible combinations of values in the hidden state space \mathcal{S} and where y_{it} is the element observed at time t of the i -th sequence y_i .

HMMs are particularly suitable for analysing the evolution of sequences. They are ideal for analysing behaviour (e.g., selecting products from a basket, animal movements, etc.) where the sequences identify a series of choices. They account for both the choices made from a set of alternatives (i.e. the observations) and the hidden goals and motivations (i.e. the hidden states) behind the observed phenomenon.³

If the heterogeneity related to the distribution of the observed sequences is considered, mixture of HMMs should be considered. There are several definitions of MHMM, which, in the context of continuous data, can be defined by introducing individual-specific random effects (see Humphreys 1998; Altman 2007). MHMMs used to cluster categorical sequences rely on the definition of mixture models and are also called mixture latent Markov models. A general model can be found in Vermunt et al. (2008).

In this paper, we refer to the model proposed in Helske and Helske (2019) which was first presented by Van de Pol and Langeheine (1990) as mixture Latent Markov models. The R package *seqHMM* (Helske and Helske 2019) was used to perform the simulation study and the empirical analysis.

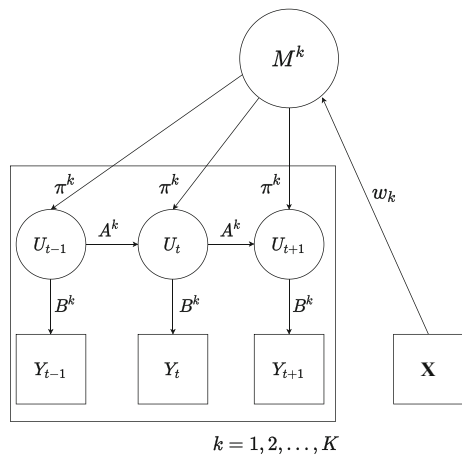
2.2 Mixture hidden Markov models

In addition to the observed variable Y , MHMM contains two different latent variables M and U . $M = \{M^k\}$ for $k \in \mathcal{K} = \{1, \dots, K\}$ is a time-constant latent variable, where M^k refers to the k -th cluster of the mixture. $U = \{U^k\}$ is a time-varying latent variable, where $U^k = \{U_1^k, \dots, U_T^k\}$ refers to a HMM whose state space is $\mathcal{S}^k = \{1, \dots, S^k\}$. Thus, each cluster k is characterized by a specific HMM.

The set of parameters that characterize the MHMM is $\Theta = (\Theta^1, \Theta^2, \dots, \Theta^K, \omega)$ where $\Theta^k = \{\pi^k, A^k, B^k\}$ and $\omega = \{\omega^k\}$ for $k \in \mathcal{K} = \{1, \dots, K\}$. Moreover,

³ For a more in-depth illustration of hidden Markov models, see Zucchini et al. (2017).

Fig. 1 Directed Acyclic Graph (DAG) of a mixture of hidden Markov models with covariates. The variables Y_t and the time-constant covariates X are observable, the U_t and M^k are hidden variables, k is the component label and M^k identifies the k -th component i.e. the k -th hidden Markov model, ω_k are the mixture coefficients, $\Theta^k = \{\pi^k, A^k, B^k\}$ are model parameters representing π^k initial probabilities, A^k transition matrix and B^k emission matrix for each component k



for each observed sequence y_i , given a set of Q time-constant covariates $X_i = (X_{i1}, \dots, X_{iQ})$, we can define the prior cluster probabilities as $\omega_i^k = \Pr(M^k|X_i)$, and $\omega^k = \{\omega_i^k\}_{i=1, \dots, n}$.

Fig. 1 schematizes the mixture of hidden Markov models with time-constant covariates. The log-likelihood for a mixture of HMMs is

$$\begin{aligned}
 \ell(\Theta; y, X) &= \sum_{i=1}^n \log P(y_i|\Theta, X_i) = \sum_{i=1}^n \log \left(\sum_{k=1}^K P(y_i, M^k|\Theta^k, \omega, X_i) \right) \\
 &= \sum_{i=1}^n \log \left(\sum_{k=1}^K \omega_i^k P(y_i|\Theta^k) \right) \tag{1} \\
 &= \sum_{i=1}^n \log \left(\sum_{k=1}^K \omega_i^k \sum_{s, f \in \mathcal{S}^k} \pi_s^k b_s^k(y_{i1}) \prod_{t=2}^T a_{s, f}^k b_f^k(y_{it}) \right),
 \end{aligned}$$

were $\omega = \{\omega^1, \dots, \omega^K\}$ and $\Theta = \{\Theta^1, \Theta^2, \dots, \Theta^K, \omega\}$. The cluster memberships ω_i^k of each sequence i are modelled according to the following multinomial logistic regression model:

$$\omega_i^k = P(M^k|X_i) = \frac{e^{\gamma_k X_i}}{1 + \sum_{j=2}^K e^{\gamma_j X_i}}, \tag{2}$$

where γ_k is the set of coefficients associated with the vector of covariates X_i for observation i and the k -th class, $\sum_{k=1}^K \omega_i^k = 1$, and $\gamma_1 = \{0, \dots, 0\}$.

According to Helske and Helske (2019), we can use standard HMM algorithms with a slight modification to estimate model parameters. The modification concerns the initial state probabilities π , which now vary between subjects, i.e., for subject i we have $\pi_i = (\omega_{i1}\pi^1, \dots, \omega_{iK}\pi^K)$, and the transition and emission matrices transformed into block diagonal matrices by considering a K -components MHMM as an HMM

with $S = \sum_{k=1}^K S^k$ hidden states. Then, parameters are estimated by adapting the EM algorithm for hidden Markov models and considering a K -components MHMM as an HMM with $S = \sum_{k=1}^K S^k$ hidden states and block diagonal transition and emission matrices (transition between components is not allowed). Vermunt et al. (2008) presented a forward-backward algorithm to estimate the parameters accounting for multiple categorical response variables, cluster probabilities depending on time-constant covariates and initial and transition probabilities depending on time-varying covariates. This algorithm can be adopted to estimate the model parameters used in this work since we are dealing with a more straightforward case with only time-constant covariates and a single categorical response. Indeed, we also need to estimate the regression coefficients $\gamma = (\gamma_2, \dots, \gamma_Q)$. In the M-step of the EM algorithm, to estimate the γ parameter, the package *seqHMM* uses the iterative Newton's method with analytic gradients and Hessian, which can be computed given all other model parameters (Helske and Helske 2019).⁴ The estimation process starts with some initial guesses on the parameters. As noted by Helske and Helske (2019), good starting values are needed to find the optimal solution in a reasonable time. Moreover, without good starting points, there is a high risk of being trapped in local maxima.⁵ Finally, the cluster posterior probabilities $P(M^k|y_i, X_i)$ are obtained as

$$P(M^k|y_i, X_i) = \frac{P(y_i|M^k, X_i)P(M^k|X_i)}{P(y_i|\Theta, X_i)},$$

where $P(M^k|X_i)$ are the cluster memberships defined in Eq. (2) and $P(y_i|M^k, X_i)$ are conditional probabilities of the observed sequences in cluster k .

Although several studies have been done on parameter estimation followed by different implementations in general statistical software, the essential task of model selection has received little attention.

Two directions can be followed in model selection for MHMMs. One focuses on estimating the chain order, i.e., the number of previous time steps in the chain required to predict the next state. The other infers the number of clusters and the states for each cluster. The latter direction has been followed in this work, which derives a score, the BIC_H , whose theoretical foundation is related to approximating the Integrated Completed Likelihood (ICL) by a Bayesian information criterion.

3 Model selection

As highlighted in the previous section, MHMMs can detect underlying latent structures and allow clustering sequences to be arranged into homogeneous subpopulations whose evolution follows the same hidden Markov process having specific parameters and some states. Therefore, model selection for MHMMs requires identifying the number of clusters and states related to each cluster. A model selection criterion based

⁴ The package *seqHMM* also implements global and local optimization routines such as the Stochastic Global Optimization method, Nelder-Mead or the Multilevel Single-linkage method and the L-BFGS.

⁵ As we note in the simulation study, the problem of local maximum also affects the model selection criteria. See Sect. 4 for more insights.

on a separability measure (i.e., an entropy measure) is the preferred choice to simultaneously identify the number of components and hidden states. Using an entropy-based criterion enables us to identify latent states so that the distribution of the observations—i.e. the elements in the sequences—conditionally to the latent states, will have a high degree of separability. Consequently, each latent state will be identified in groups of observations, seizing the unknown similarities among observations.

3.1 Proposed model selection criterion

From a model-based clustering perspective, the separability among possible clusters should be considered for selecting the correct model. In this regard, Biernacki et al. (2000) proposed an Integrated Completed Likelihood (ICL), approximated by a Bayesian information criterion, for selecting a mixture model in a cluster analysis setting. Following Biernacki et al. (2000), we define an ICL criterion in the context of MHMMs by maximizing the integrated completed likelihood instead of the observed one, to account for the degree of separation between latent classes. Specifically, the ICL criterion can be approximated as an entropy-penalized BIC (McLachlan and Peel 2004). This derivation was obtained by approximating the complete log-likelihood as the sum of two elements, the observed log-likelihood and the data entropy.⁶ To the best of our knowledge the only other similar proposal was made by Volant et al. (2014). However, they considered a different model structure characterized by a singular hidden Markov process with a mixture of the conditional distribution of emission probabilities.

Generally, in a Bayesian context, given a model $w \in \mathcal{W}$ where \mathcal{W} is the set of possible models, we select w by maximizing the posterior probability defined as

$$P(w|y) = \frac{P(y|w)P(w)}{P(y)},$$

where y identifies the data and $P(w)$ the model prior distribution. Under the assumption of a non-informative prior for w , we can identify the model by directly maximizing the integrated likelihood $P(y|w)$. As pointed out by Biernacki, when the target is to identify an unknown number of latent classes, we should consider the integrated completed likelihood instead, related to complete data.

In the MHMM context, for each observation i , complete data are represented by three discrete random variables: (Y_i, M_i, U_i) . Specifically, the first one can be observed, and the observed sequence is denoted as $y_i = (y_{i1}, y_{i2}, \dots, y_{iT})$. The second one is not observable and represents the i -th observation's membership to one of the K clusters. Therefore, we define a set of binary variables $m_i = (m_i^1, m_i^2, \dots, m_i^K)$ where $m_i^k = 1$ if sequence i was generated by the k -th HMM and $m_i^k = 0$ otherwise. The third one, denoted by u_i , is also not observable. It represents the hidden sequence for y_i , where for each time step t , $u_i = (u_{i1}, u_{i2}, \dots, u_{iT})$ and $u_{it} = (u_{it1}, u_{it2}, \dots, u_{itS})$ is a vector of binary variables such that $u_{its} = 1$ if the i -th observation at time t takes the hidden state s and $u_{its} = 0$ otherwise.

⁶ For its derivation, see Biernacki et al. (2000).

The ICL for MHMMs is defined as

$$P(y, m, u|w) = \int P(y, m, u|w, \Theta)P(\Theta|w)d\Theta, \tag{3}$$

where $P(\Theta|w)$ is the parameter prior distribution and w is a MHMM, $y = \{y_i\}$, $m = \{m_i\}$ and $u = \{u_i\}$ with $i \in \{1, 2, \dots, n\}$. Following Biernacki et al. (2000), to define our selection criterion, we consider a BIC-like approximation for the integral in Eq. (3). Thus, the quantity $-2 \log P(y, m, u|w)$ can be approximated as

$$-2 \log P(y, \hat{m}, \hat{u}|w, \hat{\Theta}) + (\log N)df, \tag{3.1}$$

where $N = n \times T$ with n number of sequences and T the sequence length, \hat{m} and \hat{u} are posterior modes given observations y and parameter estimates $\hat{\Theta}$ and df the model's degrees of freedom.

The definition of ICL was obtained in a Bayesian context; that is not our case. Thus, \hat{m} and \hat{u} are replaced by considering the conditional expectation for latent variables (M, U) (see McLachlan and Peel (2004)). We define the BIC_H as

$$BIC_H = -2E_{M,U}[\log P(y, m, u|X, w, \hat{\Theta})] + (\log N)df. \tag{4}$$

The model assumes that the hidden sequences depend on the clusters, the observed sequences depend on the hidden sequences, and the observed sequences are independent by the clusters given the hidden sequences, i.e., $U_i \perp\!\!\!\perp M_i, Y_i \perp\!\!\!\perp U_i, Y_i \perp\!\!\!\perp M_i|U_i$. The joint probability of y, m and u can be decomposed as follows

$$P(y, m, u|X, w, \hat{\Theta}) = P(y|u, m, X, w, \hat{\Theta})P(u|m, X, w, \hat{\Theta})P(m|X, w, \hat{\Theta}). \tag{4.1}$$

Referring to i -th item, since the observed sequences $y_i, i = 1, 2, \dots, n$, are independent and identically distributed given u and m ; the hidden sequences u_i are independent and identically distributed given cluster membership m and cluster memberships m_i are independent and identically distributed given covariates X . Thus, the expression on the right side of equation (4.1) can be written as follows:

$$\begin{aligned} P(y|u, m, X, w, \hat{\Theta})P(u|m, X, w, \hat{\Theta})P(m|X, w, \hat{\Theta}) \\ = \prod_i \{P(y_i|m_i, u_i, X_i, w, \hat{\Theta})P(u_i|m_i, X_i, w, \hat{\Theta})P(m_i|X_i, w, \hat{\Theta})\}, \end{aligned} \tag{4.2}$$

It is worth noting that, each y_i is identified by its cluster membership m_i and its hidden sequence u_i . The hidden sequence u_i is identified by its cluster membership m_i , and the cluster membership m_i is identified by its covariates' values X_i . On the basis of the multiplication rule, Eq. (4.2) can be written as product of joint probabilities:

$$\prod_i \{P(y_i|m_i, u_i, X_i, w, \hat{\Theta})P(u_i|m_i, X_i, w, \hat{\Theta})P(m_i|X_i, w, \hat{\Theta})\} \tag{4.3}$$

$$= \prod_i^n P(y_i, m_i, u_i | X_i, w, \hat{\Theta}),$$

That said, on the basis of Eq. (4.3), we can rewrite Eq. (4)

$$\text{BIC}_H = -2\mathbb{E}_{M,U}[\log \prod_i^n P(y_i, m_i, u_i | X_i, w, \hat{\Theta})] + (\log N)\text{df}.$$

then, by decomposing each joint probability, we obtain

$$\begin{aligned} \text{BIC}_H &= -2\mathbb{E}_{M,U}[\log \prod_i^n P(y_i, m_i, u_i | X_i, w, \hat{\Theta})] + (\log N)\text{df} \\ &= -2\mathbb{E}_{M,U} \left[\sum_i^n \log \left\{ P(y_i, m_i, u_i | X_i, w, \hat{\Theta}) \right\} \right] + (\log N)\text{df} \\ &= -2\mathbb{E}_{M,U} \left[\sum_i^n \log \{ P(y_i | X_i, w, \hat{\Theta}) P(m_i, u_i | y_i, X_i, w, \hat{\Theta}) \} \right] + (\log N)\text{df} \\ &= -2\mathbb{E}_{M,U} \left[\sum_i^n \log \{ P(y_i | X_i, w, \hat{\Theta}) \} \right] \\ &\quad - 2\mathbb{E}_{M,U} \left[\sum_i^n \log \{ P(m_i, u_i | y_i, X_i, w, \hat{\Theta}) \} \right] + (\log N)\text{df} \\ &= -2 \sum_i^n \log P(y_i | X_i, w, \hat{\Theta}) - 2 \sum_i^n \mathbb{E}_{M,U} \left[\log P(m_i, u_i | y_i, X_i, w, \hat{\Theta}) \right] \\ &\quad + (\log N)\text{df} \\ &= -2\ell(\Theta; y, X) + 2 \sum_i^n H(m_i, u_i | y_i, X_i, w, \hat{\Theta}) + (\log N)\text{df}. \end{aligned} \tag{5}$$

The BIC_H comprises three quantities, namely the log-likelihood (see Eq. (1)), a sum of entropies and the degrees of freedom.

The joint entropy for a single sequence $H(m_i, u_i | y_i, X_i, w, \hat{\Theta})$,⁷ which comprises two levels of entropies, is

$$H(m_i, u_i | y_i, X_i, \hat{\Theta}) = H(m_i | y_i, X_i, \hat{\Theta}) + H(u_i | m_i, y_i, X_i, \hat{\Theta}).$$

$H(m_i | y_i, X_i, \hat{\Theta})$ is the entropy measure related to the number of components, and $H(u_i | m_i, y_i, X_i, \hat{\Theta})$ is the entropy related to the sequence of hidden states. The first-

⁷ Hereafter, we will omit the model term w to simplify the notation, e.g. we will write $H(m_i, u_i | y_i, X_i, \hat{\Theta})$ instead of $H(m_i, u_i | y_i, X_i, w, \hat{\Theta})$.

level entropy measure, related to the cluster, is given by

$$H(m_i|y_i, X_i, \hat{\Theta}) = - \sum_{k=1}^K P(m_i^k|y_i, X_i, \hat{\Theta}) \log P(m_i^k|y_i, X_i, \hat{\Theta}), \tag{6}$$

where $P(m_i^k|y_i, X_i, \hat{\Theta})$ are component posterior probabilities, i.e. the probability that given the i -th observed sequence, the latter was generated by the k -th hidden Markov model.

The entropy related to the sequence of hidden states is given by

$$\begin{aligned} &H(u_i|m_i, y_i, X_i, \hat{\Theta}) \\ &= - \sum_{k=1}^K P(m_i^k|y_i, X_i, \hat{\Theta}) H(u_i|m_i^k, y_i, \hat{\Theta}) \\ &= - \sum_{k=1}^K P(m_i^k|y_i, X_i, \hat{\Theta}) \left[H(u_{i1}|y_i, m_i^k, \hat{\Theta}) + \sum_{t=2}^T H(u_{it}|u_{i,t-1}, y_i, m_i^k, \hat{\Theta}) \right], \end{aligned} \tag{7}$$

where $H(u_{i1}|y_i, m_i^k, \hat{\Theta})$ for $t = 1$, and $H(u_{it}|u_{i,t-1}, y_i, m_i^k, \hat{\Theta})$ for $t = 2, \dots, T$, i.e. the conditional entropy profiles, are computed as proposed by (Durand and Guédon 2016) based on the Hernando et al. (2005) conditional entropy definition for HMM. Specifically, for each i , given the k -th HMM

$$H(u_{i1}|y_i, m_i^k, \hat{\Theta}) = \sum_{s \in \mathcal{S}^k} P(u_{i1s}|y_i, m_i^k, \hat{\Theta}) \log P(u_{i1s}|y_i, m_i^k, \hat{\Theta}),$$

and

$$\begin{aligned} &H(u_{it}|u_{i,t-1}, y_i, m_i^k, \hat{\Theta}) \\ &= \sum_{s, f \in \mathcal{S}^k} P(u_{its}, u_{i,t-1,f}|y_i, m_i^k, \hat{\Theta}) \log P(u_{its}|u_{i,t-1,f}, y_i, m_i^k, \hat{\Theta}), \end{aligned}$$

where

$$P(u_{its}, u_{i,t-1,f}|y_i, m_i^k, \hat{\Theta}) = L_{it}^k(s) a_{fs}^k \hat{\alpha}_{i,t-1}^k(f) / G_{it}^k(s),$$

and

$$P(u_{its}|u_{i,t-1,f}, y_i, m_i^k, \hat{\Theta}) = L_{it}^k(s) a_{fs}^k \hat{\alpha}_{i,t-1}^k(f) / \{G_{it}^k(s) L_{i,t-1}^k(f)\}.$$

Here, a_{fs}^k are transition probabilities from state f to state s in cluster k , $\alpha_{it}^k(s)$ are the forward probabilities, the posterior state probabilities $L_{it}^k(s) = P(u_{its}|y_i, m_i^k, \hat{\Theta})$

being obtained in the forward-backward algorithm. The $G_{it}^k(s)$ are called predicted probabilities and are computed from the forward probabilities as

$$G_{i,t+1}^k(s) = P(u_{i,t+1,s} | y_{i1}, y_{i2}, \dots, y_{it}, m_i^k, \hat{\Theta}) = \sum_{f=1}^{S^k} a_{fs}^k \hat{\alpha}_{it}^k(f).$$

Finally, the number of free parameters in Eq. (5), df , depend on a set of time-static covariates through a logistic regression model, and is computed as

$$df = \sum_{k=1}^K \left[S^k (R - 1) + S^k - 1 + S^k (S^k - 1) \right] + (K - 1) \left\{ 1 + \left[2^{Q^c} - 1 \right] + \left[\prod_{q=1}^{Q^d} x_q^d - 1 \right] \right\}. \quad (8)$$

Here, K is the number of components, S^k is the number of hidden states in component k , R is the number of observed states, Q^c is the number of continuous covariates, Q^d is the number of categorical covariates and x_q^d is the number of values of categorical covariates q . Specifically, we considered the model as having interactions among either continuous or categorical covariates. The product in the computation between braces is the total number of coefficients in the multinomial regression model in Eq. (2) considering Q^d categorical covariates and all the possible interactions between these covariates. For more details, see Appendix A.

4 Simulation study

The performance of our proposed model selection criterion (see Eq. (5)) was assessed and compared with the most widely used IC for MHMMs, i.e. the AIC, BIC and the ssBIC (see Table 9), through a Monte Carlo study. We used the R package *seqHMM* (Helske and Helske 2019) to carry out the simulation study. Specifically, we simulated datasets from six different MHMMs with a known number of components K and latent states (S^1, S^2, \dots, S^K) . For details on the parameters generating the MHMMs, see Appendix B.1.

We investigated 24 scenarios by varying the number of sequences $n \in \{300, 500\}$, and the length of the sequences $T \in \{10, 20\}$. An element of the sequence can take one of four observed states. We generated 70 datasets for each scenario and evaluated which information and classification criteria performed better in identifying the numbers of states and components.

We considered the success rate—i.e. the rate of identifying the correct number of clusters and states—for BIC, AIC, ssBIC and the BIC_H for MHMM to assess criteria performance. Preliminary simulations, not reported here, show how difficult it is to identify the correct number of components and hidden states. Indeed, the success rate was very low or close to zero. So, we investigated whether the criteria could identify

Table 1 The six MHMMs used in the Monte Carlo are denoted by O_j

Model	K	(S^1, S^2, \dots, S^K)
O_2	2	(3,4)
O_3	3	(2,3,4)
O_4	4	(2,3,4,2)
O_5	5	(2,3,4,2,4)
O_6	6	(2,3,4,2,4,3)
O_7	7	(2,3,4,2,4,3,4)

K represents the number of components, and (S^1, S^2, \dots, S^K) the number of hidden states within each component

the correct number of components K and the numbers of hidden states equal or *close enough* to the correct ones. We refer to this measure as an *approximate success rate*. To further investigate the performance of the criteria, we considered the *failure rate* related to under and overestimating the number of components. For details on the measures see Appendix B.3.⁸

The results, which report the approximate success rate, are displayed in Tables 2 and 4. Tables 3 and 5 show the failure rates. Each row in Tables 2 and 4 refers to one of the six data generation models illustrated in Table 1, the best among the selection criteria is indicated in bold.

For $T = 10$ and $n = 300$, the proposed BIC_H outperforms the other information criteria with AIC being its closest competitor, while BIC and ssBIC struggle the most to identify even an approximation of the model (Table 2). The ICs seem to systematically underestimate the number of components when the model is O_2 and select models with no mixtures if that option is available. When the number of components increases, all criteria struggle to identify the numbers of subpopulations and states. Still, the BIC_H performs best with short length sequences (Table 3).

For $T = 20$ and $n = 300$, we note that AIC and BIC_H have similar results. ICs perform better (Table 2), but still tend to underestimate the number of components (Table 3).

For $T = 20$ and $n = 500$, the ICs do not appear to have improved overall, as reported in Table 4. Again, the ICs seem to systematically underestimate the number of components when the model is O_2 and select models with no mixtures if that option is available (Table 5).

Furthermore, the results of the Z-tests presented in Appendix B.5 comparing the performance of the BIC_H criterion against AIC, BIC, and ssBIC provide interesting insights into its effectiveness across different scenarios. When considering a shorter time series length ($T = 10$), the evidence suggests a significant advantage for the BIC_H criterion over the other information criteria. However, when the time series length is increased to $T = 20$, the evidence becomes less conclusive. While the BIC_H

⁸ Note that we assumed a certain amount of a priori knowledge of the number of components and a range of hidden states. Specifically, we defined alternatives by considering $\{K - 1, K, K + 1\}$ for the number of components and $S^j \in [2; 6]$, for $j = 1, \dots, K$, for the hidden states. For details on the investigated models see Appendix B.4.

Table 2 Results of the Monte Carlo study for $n = 300$, $T = (10, 20)$

$n = 300$		AIC	BIC	ssBIC	BIC_H
$T = 10$	O_2	0.10 (0.030)	0.03 (0.017)	0.03 (0.017)	0.39 (0.049)
	O_3	0.29 (0.045)	0.18 (0.038)	0.18 (0.038)	0.50 (0.050)
	O_4	0.51 (0.050)	0.50 (0.050)	0.51 (0.050)	0.69 (0.046)
	O_5	0.42 (0.049)	0.40 (0.049)	0.40 (0.049)	0.68 (0.047)
	O_6	0.62 (0.049)	0.60 (0.049)	0.61 (0.049)	0.80 (0.040)
	O_7	0.42 (0.049)	0.40 (0.049)	0.40 (0.049)	0.56 (0.050)
	$T = 20$	O_2	0.29 (0.045)	0.02 (0.014)	0.05 (0.022)
O_3		0.42 (0.049)	0.23 (0.042)	0.28 (0.045)	0.56 (0.050)
O_4		0.58 (0.049)	0.49 (0.050)	0.49 (0.050)	0.77 (0.042)
O_5		0.52 (0.050)	0.47 (0.050)	0.48 (0.050)	0.67 (0.047)
O_6		0.50 (0.050)	0.52 (0.050)	0.52 (0.050)	0.60 (0.049)
O_7		0.52 (0.050)	0.49 (0.050)	0.51 (0.050)	0.58 (0.049)

The *approximate success rates* of identifying the generating models are reported. Standard errors are shown in parentheses. See Appendix B.3 for the definition of the approximate success rate, and Appendix B.5 Table 11 for the Z-tests

criterion still has some advantages over the other criteria, the Z-tests yield higher p values, suggesting a reduced level of statistical significance.

It should be pointed out that one of the challenges of model selection in mixture hidden Markov models is related to parameter estimation via the EM algorithm. The inherent complexity of the likelihood, characterized by multiple local maxima, poses a risk of the EM algorithm getting trapped in suboptimal solutions, particularly when random starting points are used as initial parameter values. Thus, the error of the estimates will remain even when increasing sample sizes and sequence length. To address this issue, we repeated the simulation study employing initial values for the EM algorithm near the true parameters (reported in Appendix B.1). We introduced a disturbance term $\epsilon = 0.05$ to generate these initial values to one element of each initial probability vector π^k , with $k = 1, 2, \dots, K$. The vectors were normalized to ensure that their elements sum to one. This transformation was applied to each row of the transition and emission probability matrices A^k and B^k . Thus, initial values close to the correct parameters served as starting points for the EM algorithm in the model with the correct number of clusters and states. For competitor models (with the number of clusters and states as defined in the main text), we generated starting values from a flat Dirichlet distribution with the concentration parameter $\alpha = 1$. The results indicated a substantial improvement in parameter estimation and, consequently, the performance of all selection criteria as n and T increased (Table 6).

It is noteworthy that increasing the number of sequences and sequence length, assuming initial values close to the correct ones, ensures criteria consistency. However, it is essential to acknowledge that these results were obtained explicitly by fixing the initial values close to the correct ones. Therefore, the simulation results reported earlier (Tables 2, 3, 4 and 5) should be interpreted in light of these considerations, as

Table 3 Results of the Monte Carlo study for $n = 300$, $T = (10, 20)$

$n = 300$			AIC	BIC	ssBIC	BIC_H	
$T = 10$	O_2	U	0.90 (0.030)	0.97 (0.017)	0.97 (0.017)	0.40 (0.049)	
		O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.21 (0.041)	
	O_3	U	0.66 (0.047)	0.80 (0.040)	0.80 (0.040)	0.31 (0.046)	
		O	0.05 (0.022)	0.02 (0.014)	0.02 (0.014)	0.29 (0.045)	
	O_4	U	0.49 (0.050)	0.50 (0.050)	0.49 (0.050)	0.23 (0.042)	
		O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.08 (0.027)	
	O_5	U	0.58 (0.049)	0.60 (0.049)	0.60 (0.049)	0.30 (0.046)	
		O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.02 (0.014)	
	O_6	U	0.38 (0.049)	0.40 (0.049)	0.39 (0.049)	0.17 (0.038)	
		O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.03 (0.017)	
	O_7	U	0.58 (0.049)	0.60 (0.049)	0.60 (0.049)	0.43 (0.050)	
		O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.01 (0.010)	
	$T = 20$	O_2	U	0.65 (0.048)	0.98 (0.014)	0.95 (0.022)	0.00 (0.000)
			O	0.06 (0.024)	0.00 (0.000)	0.00 (0.000)	0.64 (0.048)
O_3		U	0.53 (0.050)	0.75 (0.043)	0.69 (0.046)	0.00 (0.000)	
		O	0.05 (0.022)	0.02 (0.014)	0.03 (0.017)	0.07 (0.026)	
O_4		U	0.42 (0.049)	0.51 (0.050)	0.51 (0.050)	0.20 (0.040)	
		O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.03 (0.017)	
O_5		U	0.48 (0.050)	0.53 (0.050)	0.52 (0.050)	0.31 (0.046)	
		O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.02 (0.014)	
O_6		U	0.50 (0.050)	0.48 (0.050)	0.48 (0.050)	0.32 (0.047)	
		O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.08 (0.027)	
O_7		U	0.48 (0.050)	0.51 (0.050)	0.49 (0.050)	0.36 (0.048)	
		O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.06 (0.024)	

Failure rate related to underestimate (U) and overestimate (O) the number of components. Standard errors are shown in parentheses. See Appendix B.3 for the definition of failure rate

consistency may not be achievable even with an increase in the length and number of sequences.

5 Case study

5.1 Dataset

We applied the proposed BIC_H to analyse the clickstream data collected from the website of a firm called LovePanormus, which operates in the hospitality sector.⁹ The company provides tourism services such as accommodation, the booking of an experiential holiday and information on cultural events.

⁹ The dataset is not publicly available due to confidential company data. However, it is available from the authors upon reasonable request and with the permission of LovePanormus.

Table 4 Results of the Monte Carlo study for $n = 500$, $T = (10, 20)$

$n = 500$		AIC	BIC	ssBIC	BIC_H
$T = 10$	O_2	0.25 (0.043)	0.06 (0.024)	0.08 (0.027)	0.42 (0.049)
	O_3	0.30 (0.046)	0.17 (0.038)	0.18 (0.038)	0.58 (0.049)
	O_4	0.49 (0.050)	0.40 (0.049)	0.41 (0.049)	0.72 (0.045)
	O_5	0.43 (0.050)	0.42 (0.049)	0.43 (0.050)	0.61 (0.049)
	O_6	0.52 (0.050)	0.50 (0.050)	0.51 (0.050)	0.69 (0.046)
	O_7	0.47 (0.050)	0.47 (0.050)	0.46 (0.050)	0.67 (0.047)
	$T = 20$	O_2	0.38 (0.049)	0.06 (0.024)	0.13 (0.034)
O_3		0.47 (0.050)	0.25 (0.043)	0.29 (0.045)	0.57 (0.050)
O_4		0.46 (0.050)	0.42 (0.049)	0.45 (0.050)	0.60 (0.049)
O_5		0.45 (0.050)	0.47 (0.050)	0.49 (0.050)	0.54 (0.050)
O_6		0.54 (0.050)	0.55 (0.050)	0.56 (0.050)	0.60 (0.049)
O_7		0.48 (0.050)	0.50 (0.050)	0.51 (0.050)	0.53 (0.050)

Approximate success rate of identifying the generating models. Standard errors are shown in parentheses. See Appendix B.3 for the definition of the approximate success rate, and Appendix B.5 Table 11 for the Z-tests

The original data consisted of 2,487,802 observations (i.e. web resources)¹⁰ arranged in chronological order. The analysis was carried out on data collected from September to December 2017. Data were cleaned and pre-processed by removing all the irrelevant log lines and suspected bots. The new dataset presented only the resources with *html* extension (i.e. the pages viewed) and comprised 95,201 lines identifying web pages accessed by users. We also extracted information about user browsers, software and devices using the *uaparserjs* R package.

The clickstream data refer to anonymous access to the site. Therefore, it was not possible to recognize whether two different accesses to the site were by the same person. In this case, it is common practice to make some assumptions and identify an individual user's activity by performing a session identification procedure. This approach is necessary as IP addresses are assigned by the server to different users and we need to recognize when the same IP is assigned to a new user by considering additional information in clickstream data and criteria based on time (e.g. average reading time for pages). Specifically, clickstream data may record for each IP both the current accessed page and the *previous step* in the same line, enabling us to recognise if the same IP was assigned to different users as the path is not connected. Unfortunately, since this information was not present in our dataset, we decided to differentiate the navigation path of each IP address into sessions according to the time spent on each individual page. Taking into consideration the structure of the site and the content of the individual pages, a time threshold of 10 min was chosen. Thus, if an IP-address remained on the same page for more than 10 min the current session y_1 was considered concluded and the pages visited by the IP-address in subsequent clicks were attributed to a new session y_2 representing a potential new user. In other words, a statistical unit

¹⁰ Web resources consist of every element of a web page such as image files, java codes, links etc.

Table 5 Results of the Monte Carlo study for $n = 500, T = (10, 20)$

$n = 500$			AIC	BIC	ssBIC	BIC_H	
$T = 10$	O_2	U	0.74 (0.044)	0.94 (0.024)	0.92 (0.027)	0.27 (0.044)	
		O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.33 (0.047)	
	O_3	U	0.66 (0.047)	0.83 (0.038)	0.82 (0.038)	0.10 (0.030)	
		O	0.04 (0.020)	0.00 (0.000)	0.00 (0.000)	0.32 (0.047)	
	O_4	U	0.51 (0.050)	0.60 (0.049)	0.59 (0.049)	0.07 (0.026)	
		O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.21 (0.041)	
	O_5	U	0.57 (0.050)	0.58 (0.049)	0.57 (0.050)	0.13 (0.034)	
		O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.26 (0.044)	
	O_6	U	0.48 (0.050)	0.50 (0.050)	0.49 (0.050)	0.27 (0.044)	
		O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.04 (0.020)	
	O_7	U	0.53 (0.050)	0.53 (0.050)	0.54 (0.050)	0.30 (0.046)	
		O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.03 (0.017)	
	$T = 20$	O_2	U	0.62 (0.049)	0.94 (0.024)	0.87 (0.034)	0.00 (0.000)
			O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.55 (0.050)
O_3		U	0.23 (0.042)	0.73 (0.044)	0.68 (0.047)	0.34 (0.047)	
		O	0.30 (0.046)	0.02 (0.014)	0.03 (0.017)	0.09 (0.029)	
O_4		U	0.54 (0.050)	0.58 (0.049)	0.55 (0.050)	0.17 (0.038)	
		O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.23 (0.042)	
O_5		U	0.55 (0.050)	0.53 (0.050)	0.51 (0.050)	0.35 (0.048)	
		O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.11 (0.031)	
O_6		U	0.46 (0.050)	0.45 (0.050)	0.44 (0.050)	0.32 (0.047)	
		O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.08 (0.027)	
O_7		U	0.52 (0.050)	0.50 (0.050)	0.49 (0.050)	0.40 (0.049)	
		O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.07 (0.026)	

Rates related to under (U) and overestimating (O) the number of components. See Appendix B.3 for the definition of failure rate

Table 6 Results of the Monte Carlo study for $n = (500, 5000), T = (20, 50)$, assuming that the initial values used in the EM algorithm are close to the correct one by adding a disturbance term $\epsilon = 0.05$ as illustrated in Sect. 4

			AIC	BIC	ssBIC	BIC_H	
$n = 500$	$T = 20$	O_4	S	0.70 (0.046)	0.48 (0.050)	0.50 (0.050)	0.78 (0.041)
			U	0.30 (0.046)	0.52 (0.050)	0.50 (0.050)	0.07 (0.026)
			O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.15 (0.036)
$n = 5000$	$T = 50$	O_4	S	1.00 (0.000)	0.95 (0.022)	0.97 (0.017)	0.88 (0.032)
			U	0.00 (0.000)	0.05 (0.022)	0.03 (0.017)	0.02 (0.014)
			O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.10 (0.030)

Approximate success rate (S) of identifying the generating model and failure rate related to under (U) and overestimating (O) the number of components. Standard errors are shown in parentheses. See Appendix B.3 for the definition of the approximate success rate and failure rate

is represented by a session. Using this methodology, we were able to obtain 43,182 user sessions. Those sessions having $T = 3$ clicks (i.e. short length sessions) were removed since they were likely to be accidental users. After all these refinements, the dataset was reduced to $n = 10,252$ user sessions.

Each session enabled us to follow a user's path within the site through a sequence of pages viewed. However, instead of considering web pages as elements of the sequence, we decided to view page categories. This was due to the high number of pages and also because in the reference period, they were removed, added or modified. These categories correspond to the thematic areas on the site and are as follows. The "Homepage" area contains several pages that have image links to pages in other areas. The "Attractions" area contains information on tourist attractions or general information on Sicily. In the "Accommodation" area, the user can view pages of apartments for rent and search by a period of the year or number of guests. The "Events" area provides a calendar of the leading seasonal concerts and festivals. The "Experiences" area contains additional bookable tourist activities. In the "Services" area, we find bookable transport or food delivery services. Finally, the "Info" area provides general information on the company and its staff and business partners.

To summarise, the dataset contained sequences of pages visited by users, where a sequence derived from a combination of seven basic states: Homepage, Attractions, Accommodation, Events, Experiences, Services and Info. This dataset is an example of single-channel sequence data with seven categorical states.¹¹ The maximum sequence length among the considered data was $T = 20$ (the few sequences exceeding this length were removed as outliers).

5.2 Empirical analysis

We apply MHMMs to explore if and how browsing behaviour differs across users. Analysis was performed through R package seqHMM.¹² MHMMs are particularly suitable for this analysis as they capture the underlying variability in users' movements by incorporating a latent process. This process reveals hidden "mental states" that influence the choices users make as they explore a website. Mixture Markov models identify clusters based solely on observed sequences, which often require a large number of clusters to achieve relatively homogeneous groups. In contrast, MHMMs leverage hidden sequences to identify clusters. Each cluster corresponds to a particular hidden behaviour, identified by a latent Markovian process whose dynamic behaviour represents changes in people's goals and underlying "mental states" (i.e the hidden states) during their navigation in the website. By considering the hidden aspects of user sequences, MHMMs can capture latent patterns that may not be apparent in

¹¹ Multi-channel sequence data refer to the presence of multiple interdependent sequences for the same subject, see Helske et al. (2018) as an example of three-channel data (partnership, parenthood, labour market participation).

¹² seqHMM allows us to account for sequences having different lengths T_i by considering sample size as $N = \sum_{i=1}^n \sum_{t=1}^{T_i} I(y_{it})$, where I is the indicator function equal to 1 for "observed" y_{it} and zero if y_{it} is "missing". If all sequences have the same length T , then this summation is equal to $n \times T$. If each sequence has length T_i , then let $T = \max(T_1, \dots, T_n)$, and sequences with shorter lengths than T are augmented with $T - T_i$ NA.

Table 7 Results of the model selection procedure: different candidate MHMMs and their BIC_H value

(S^1, S^2, \dots, S^K)	BIC_H
(5,4)	122041.8
(2,3)	124267.5
(2,5)	131526.6
(4,6)	123685.2
(3,3,3)	129330.9
(3,4,3)	130747.1
(4,3,4)	129125.4
(4,4,4)	128768.4
<i>(3,2,4)</i>	116936.1
(4,5,4)	128661.4
(5,5,5)	131517.3
(3,2,3)	122801.7
(3,3,3,3)	128452.2
(4,4,4,4)	127663.5
(4,5,5,4)	132606.7
(2,4,3,3,2)	127621.7
(2,5,3,3,2)	119308.8
(2,2,3,6,2)	118674.1
(2,3,3,6,2)	129943.2

The models are identified by the vector (S^1, S^2, \dots, S^K) that represents the number of hidden states within each component k in a K components MHMM. The selected model is in italic. The best among the selection criteria are indicated in bold.

observed data alone. Thus, more meaningful and informative clusters can be identified, resulting in a more detailed analysis of the clickstream data and a more comprehensive understanding of user behaviour and preferences.

Prior cluster membership was estimated through a multinomial logistic model (see Eq. (2)) with three time-constant covariates: IP address geographic area (i.e., Africa, Asia, Eastern Europe, Italy, Latin America, the Middle East, North America, Northern Europe, Oceania, Russia and Southern Europe), access device (PC and mobile) and access month (September, October, November, December).¹³

Different MHMMs are considered by varying the number of components $K \in \{2, 3, 4, 5\}$, and hidden states for each component $S^k \in \{2, 3, 4, 5\}$. Using the BIC_H to take into account two levels of uncertainty and identify clusters and states, we have selected the MHMM consisting of three components and hidden states (3,2,4) (i.e. identifying three clusters/browsing profiles). Results of the model selection procedure are shown in Table 7.

These three clusters contain 22%, 19% and 59% of web sequences, respectively.

Figure 2 shows the empirical distributions of each covariate (geographical area, access device, and access month) within each cluster. The i -th subject is assigned to

¹³ Due to the significant percentage of Italian users in the dataset (44.9%), we decided to separate Italy from the rest of Europe.



Fig. 2 Empirical distributions of each covariate: **a** geographical area, **b** access device, and **c** access month within each cluster

cluster j according to the posterior cluster probability, which is calculated conditioned on the sequence y_i and the covariates X_i . In Fig. 2a), we observe that Italians are mostly assigned to cluster 3 (51% of the cluster) and cluster 2 (43%). Users from the North of Europe are mainly in clusters 2 and 3 too (22% and 21%), while North Americans are primarily found in cluster 2 (25% of that cluster) and are less present in the other two (10% and 13%). Asian and Eastern European users, sparsely featured in the data, are present in Cluster 1 (10% and 30% of this cluster, respectively), while their presence in other Clusters is below 3% and 5%, respectively. In Fig. 2b), we note similar distributions as regards access by PC or Mobile. In Fig. 2c), in Clusters 1 and 3 user access is almost uniformly from September till December, whereas in Cluster 2 it is mostly in November and December.

Figure 3 sketches the path of users in each profile (i.e. cluster). Each pie graph represents a hidden state; the edges are the transitions between states (transition probabilities displayed on the edges). The colour and size of the pie slices represent the emission probabilities of the observed states (the thematic area of the page). Emission probabilities lower than 0.05 are classified as “others”. Initial and emission probabilities are reported in Appendix C. As shown in Fig. 3a), Cluster 1 users start their web navigation from state 1, which emits the observed state “Homepage” with a probability of 0.98. They remain on pages from this area with a probability of 0.73 or transition to state 2 with a probability of 0.15 and state 3 with a probability of 0.12 and remain in these states (with probabilities of 0.9 and 0.85 respectively). Specifically, state 2 tends to emit “Services” with a probability of 0.34, “Attractions” with a probability of 0.22 and “Accommodation” with a probability of 0.19. In Cluster 2, Fig. 3b), users have a higher probability of starting from state 1 (0.66) and a 0.99 probability of staying there. State 1 emits “Attractions” with a probability of 0.91. If they start from state 2 with a probability of 0.34 and remain there with a probability of 0.96, they have a high probability (0.61) of accessing “Events” pages. Finally, in Cluster 3, Fig. 3c), users start from state 1 with a probability of 0.56 and move to state 2 with a probability of 0.43. Once users reach state 2, their probability of staying there is 0.85. State 2 emits “Attractions” with a probability of 0.74. Another path is moving from state 1 to state 3 with a probability of 0.25 and staying in this state with a probability of 0.94. State 3 emits “Accommodation” with a probability equal to 0.91. In summary, by crossing the probability of belonging to a cluster with transaction probabilities, we can identify three user profiles of browsing, namely the casual explorer or potential partner (Cluster 1), the Information seeker (Cluster 2), and the potential tourist (Cluster 3).

Concerning the *casual explorer or potential partner (Cluster 1)*, hidden states in this cluster identify three different “mental-states” and three sub-paths. The first one is related to a lack of interest in the website or an interest in general tourist information, as this state emits pages in the “Homepage” area. A second hidden state emits different areas and seems to indicate an exploratory attitude. Finally, there is a third hidden state that emits “Info”, indicating that there is a specific subgroup of users which seems interested in accessing information about the company and its partners. This group includes most Asians and Eastern Europeans and has the lowest percentage of access via mobile. This cluster makes up 22% of users.

The *information seekers (Cluster 2)* are looking for tourist information. Here the hidden states identify two different sub-paths. Some users start from state 2 accessing only “Attractions” pages, while others start from state 1 and follow a mixed exploratory path with a focus on “Events” pages. These users seem not to be interested in accessing purchase-oriented pages. However, the second “mental state” may actually be related to the desire to schedule a trip to see one of the seasonal events listed. This cluster has the highest percentage of mobile access (although PC access is still the preferred one) and includes 19% of users.

The *potential tourists (Cluster 3)* view the website to search for both tourist information and tourist products (59% of the sessions). Hidden states represent preliminary states related to browsing the “Homepage” area, an exploratory state looking at “Attractions” and a buying state going through “Accommodation” pages. There is also a fourth state that accounts for the other thematic areas which, although rarely accessed, indi-

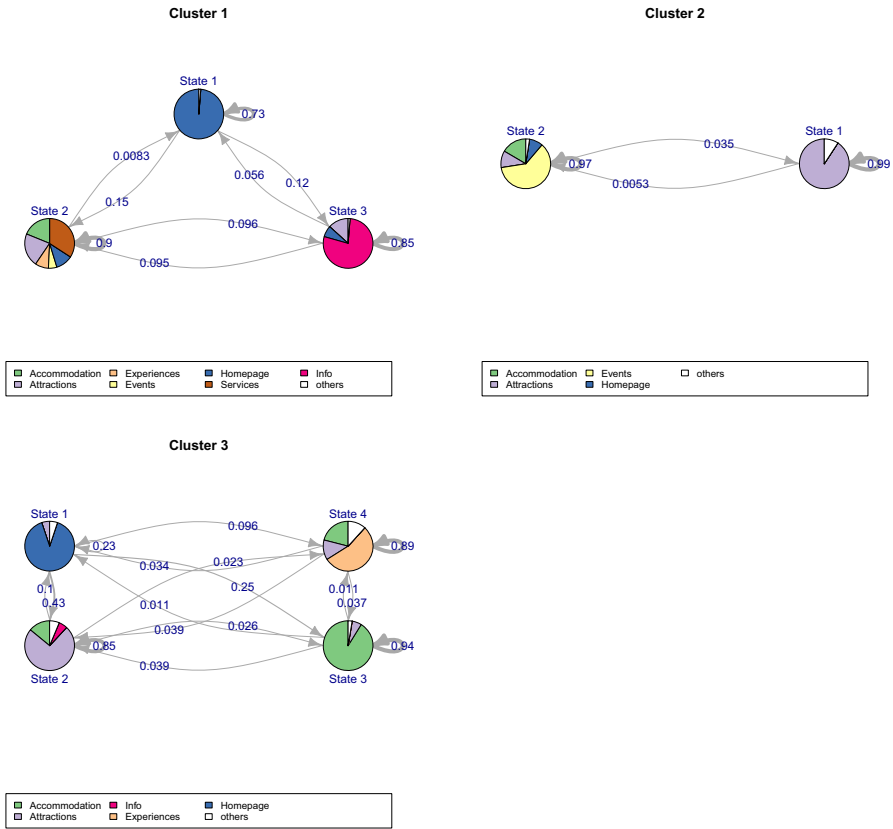


Fig. 3 Hidden Markov process structures for clusters 1,2, and 3. Vertices represent hidden states. The slices show emission probabilities, and the edges show the transition. Details on initial, transition and emission matrices probabilities are reported in Appendix C—Table D10–D18 probabilities

cates an interest in accessing “Experiences” pages. Users in this cluster begin by selecting something from “Homepage” before moving on to tourist attractions pages or bookable apartments. There is still a greater interest in information-oriented pages, but the cluster does identify two sub-behaviours related to purchase-oriented pages. One is a group of users who start at “Homepage” and then move to “Accommodation” (state 3) and stay there; the other moves from “Homepage” to a mixed path mainly concentrated in “Experiences” (state 4). This is a cluster of Italian, North American and Northern European users. Users with this profile preferred exploring the site using their PC and logging in during November and December.

Our findings show that the website reaches two types of primary users, involving two business models. One concerns the last two clusters and consists of potential tourists interested in knowing about or buying tourist products or services. This result is expected since it reflects the business model B2C adopted by the firm, i.e. the so-called business-to-consumer model. The second user target refers to cluster 1, which, based on its characteristics, likely includes Asian and Eastern European businesses

potentially interested in promoting outgoing tourism toward Italy. This second group of users should give LovePanormus pause for thought as to whether its B2C business model is too limited and that they may also need to focus on selling products and services to other companies (i.e. the business model of B2B, business-to-business).

6 Conclusion

Following the literature on model selection in mixture models and hidden Markov models, this study proposes a score based on classification criteria, i.e. BIC_H , for mixture hidden Markov models. The main contribution of our work is to enrich the literature from a methodological perspective by proposing a model selection criterion based on an integrated completed likelihood approach that accounts for the two latent classes in the model: subpopulations (i.e. mixture components) and hidden states.

We used an approximation of the integrated completed likelihood based on BIC, the BIC_H , by defining a new entropy penalization, i.e. a joint entropy obtained as the sum of cluster-level entropy and state-level entropy. The former is the mixture model entropy and the latter is based on the Hernando et al. (2005) conditional entropy definition.

The most suitable model was selected from a set of candidates by minimizing the BIC_H . We employed the proposed criterion to identify the number of states and components with the best degree of class separation. We implemented a Monte Carlo simulation study to compare selection criteria (BIC, AIC and ssBIC) with the new entropy-based criterion BIC_H by varying sample size and sequence length. Simulations demonstrated that with all the selection criteria, it is not straightforward to identify the correct number of components and states. However, it is worth noting that the proposed BIC_H , although it requires a range of possibilities for the number of clusters and states to be defined, seems to outperform the other criteria.

The BIC, commonly used in literature, struggles to identify clusters and states, particularly when the sequence length is short, which is a common scenario in web sequences. Further research is needed to improve the measure of entropy and the procedure of modelling selection of MHMMs because of their increasing use, especially by businesses in analysing clickstream data from their websites.

We used MHMMs to analyse web sequences related to a Sicilian hospitality company's website. To select a model and the number of clusters and states for each hidden Markov model, we used the proposed BIC_H that accounts for hidden heterogeneity in clickstream data. The company had recently had its website modified by enriching the tourist information provided and requested an analysis of the clickstream data to understand better its users' browsing behaviour: how many users selected purchase-oriented pages, how they behaved before accessing the "Accommodation" pages, and the characteristics of potential buyers.

At first, we explored the web sequences by estimating simple transition probabilities that highlighted a "single track" navigation behaviour. We noticed that users who viewed "Accommodation" or "Attractions" continued browsing without changing areas, while those in other areas began to vary their route. The exploratory analysis

carried out gave us insights into user behaviour and suggested the presence of different behavioural profiles.

By adopting a MHMM selected through our proposed model selection score, we discovered three user profiles of website browsing: the Casual explorer/Potential partner, the Information seeker and the Potential tourist. The discriminant factors in these profiles are the users' movements and the geographical location of the IP address, the access device and the access month.

The MHMM allows us to identify, through hidden states, similarities between observed states (the thematic areas) accessed by users to achieve the same goals. These hidden states identified the most accessed areas with hidden states and isolated the least commonly accessed areas into one state. They also provided a better understanding of user movements by highlighting sub-behaviours inside clusters.

Our findings have raised concerns for the company management. The homepage offers a personalized search menu, links to different site areas, and image links related to "Attractions" at the foot of the page to encourage further exploration. However, once users start exploring attractions, there is no encouragement to switch areas, leading them to navigate along "separate tracks".

One way of overcoming these problems would be to diversify the image links provided in each area, thus improving site links. However, in view of the company's main mission, it would be better to focus on the "Accommodation" area and include image links relating to apartments adjacent to the tourist attraction described on each "Attractions" page.

Furthermore, the profile identification results indicate two primary user types and business models. The last two clusters represent potential tourists interested in buying products or services (B2C model). Cluster 1, however, likely includes potential Asian and Eastern European business partners interested in promoting outgoing tourism (B2B model). LovePanormus needs to focus on targeting both user groups and adapting its business model accordingly.

This would involve selling to other companies and creating business partnerships to promote Sicily as a tourist destination.

In conclusion, we have shown that the use of entropy-based measures should be encouraged since, in the case of short length sequences, they perform better than classic IC measures. Importantly, our empirical findings have also demonstrated that MHMMs, although not commonly used for clickstream data, are very useful in identifying user profiles with similar browsing behaviour.

Appendix A Proposed criterion

In this section, we describe the calculation of the degree of freedom for the BIC_H (see (4)).

The df is the number of free parameters given by

$$df = \sum_{k=1}^K [S^k (R - 1) + S^k - 1 + S^k (S^k - 1)]$$

$$+(K - 1)\{1 + [2^{Q^c} - 1] + [(\prod_{q=1}^{Q^d} x_q^d) - 1]\}. \tag{A1}$$

Here, K is the number of components, S^k is the number of hidden states for each component k , R is the number of observed states, Q^c is the number of continuous covariates, Q^d is the number of categorical covariates and x_q^d is the cardinality of the categorical covariate q . In other words, the $\prod_q x_q^d$ inside braces is the total number of coefficients in the multinomial regression model (Sect. 2.2, Eq. 2) considering Q^d categorical covariates and all the possible interactions between covariates.

Let consider a model with no continuous covariates and three categorical covariates, i.e. $Q^c = 0$ and $Q^d = 3$, with the latter having number of values x_1^d, x_2^d and x_3^d ; then, the total number of coefficients in the multinomial regression model is

$$\begin{aligned} &(K - 1)\{1 + (x_1^d - 1) + (x_2^d - 1) + (x_3^d - 1) \\ &\quad + (x_1^d - 1) \times (x_2^d - 1) + (x_1^d - 1) \times (x_3^d - 1) \\ &\quad + (x_2^d - 1) \times (x_3^d - 1) + (x_1^d - 1) \times (x_2^d - 1) \times (x_3^d - 1)\} \\ &= (K - 1)\{[(x_1^d - 1) + 1][(x_2^d - 1) + 1][(x_3^d - 1) + 1]\} \\ &= (K - 1)\{1 + [(\prod_q x_q^d) - 1]\}. \end{aligned}$$

Appendix B Simulation study

In this section, we give some details of the simulation study.

B.1 Details on the MHMM parameters for the data generating process

Next, we show the parameters $\Theta^k = \{\pi^k, A^k, B^k\}$ used to generate the simulated data, varying the number of components and hidden states. Firstly, we recall the models in Table 8. Then, we report initial probability vectors, transition and emission matrices necessary to generate the data.

Table 8 MHMMs component and hidden state numbers

Model	K	(S^1, S^2, \dots, S^K)
O_2	2	(3,4)
O_3	3	(2,3,4)
O_4	4	(2,3,4,2)
O_5	5	(2,3,4,2,4)
O_6	6	(2,3,4,2,4,3)
O_7	7	(2,3,4,2,4,3,4)

1. Hidden states initial probabilities for each component π^k :

$$\begin{aligned}\pi^1 &= (0.30, 0.70); & \pi^2 &= (0.40, 0.20, 0.40); & \pi^3 &= (0.30, 0.30, 0.20, 0.20); \\ \pi^4 &= (0.40, 0.60); & \pi^5 &= (0.20, 0.30, 0.15, 0.35); & \pi^6 &= (0.10, 0.40, 0.50); \\ & & \pi^7 &= (0.40, 0.10, 0.20, 0.30).\end{aligned}$$

2. Hidden states Transition Matrices for each component A^k :

$$\begin{aligned}A^1 &= \begin{bmatrix} 0.80 & 0.20 \\ 0.20 & 0.80 \end{bmatrix}; & A^2 &= \begin{bmatrix} 0.10 & 0.80 & 0.10 \\ 0.70 & 0.10 & 0.20 \\ 0.30 & 0.10 & 0.60 \end{bmatrix}; \\ A^3 &= \begin{bmatrix} 0.50 & 0.20 & 0.10 & 0.20 \\ 0.00 & 0.60 & 0.20 & 0.20 \\ 0.10 & 0.30 & 0.00 & 0.60 \\ 0.20 & 0.00 & 0.20 & 0.60 \end{bmatrix}; & A^4 &= \begin{bmatrix} 0.20 & 0.80 \\ 0.80 & 0.20 \end{bmatrix}; \\ A^5 &= \begin{bmatrix} 0.10 & 0.20 & 0.10 & 0.60 \\ 0.20 & 0.10 & 0.50 & 0.20 \\ 0.10 & 0.10 & 0.60 & 0.20 \\ 0.70 & 0.00 & 0.10 & 0.20 \end{bmatrix}; & A^6 &= \begin{bmatrix} 0.35 & 0.50 & 0.15 \\ 0.40 & 0.30 & 0.30 \\ 0.50 & 0.20 & 0.30 \end{bmatrix}; \\ A^7 &= \begin{bmatrix} 0.30 & 0.35 & 0.05 & 0.30 \\ 0.30 & 0.20 & 0.00 & 0.50 \\ 0.00 & 0.30 & 0.60 & 0.10 \\ 0.10 & 0.50 & 0.30 & 0.10 \end{bmatrix}.\end{aligned}$$

3. Observed states emission matrices for each component, B^k :

$$\begin{aligned}B^1 &= \begin{bmatrix} 0.10 & 0.20 & 0.10 & 0.60 \\ 0.50 & 0.05 & 0.20 & 0.25 \end{bmatrix}; & B^2 &= \begin{bmatrix} 0.10 & 0.10 & 0.60 & 0.20 \\ 0.20 & 0.50 & 0.20 & 0.10 \\ 0.15 & 0.15 & 0.10 & 0.60 \end{bmatrix}; \\ B^3 &= \begin{bmatrix} 0.00 & 0.10 & 0.30 & 0.60 \\ 0.20 & 0.60 & 0.20 & 0.00 \\ 0.55 & 0.00 & 0.10 & 0.35 \\ 0.25 & 0.00 & 0.60 & 0.15 \end{bmatrix}; & B^4 &= \begin{bmatrix} 0.10 & 0.60 & 0.10 & 0.20 \\ 0.20 & 0.10 & 0.45 & 0.25 \end{bmatrix}; \\ B^5 &= \begin{bmatrix} 0.20 & 0.65 & 0.15 & 0.00 \\ 0.50 & 0.00 & 0.25 & 0.25 \\ 0.05 & 0.00 & 0.35 & 0.60 \\ 0.15 & 0.00 & 0.70 & 0.15 \end{bmatrix}; & B^6 &= \begin{bmatrix} 0.30 & 0.40 & 0.10 & 0.20 \\ 0.00 & 0.40 & 0.30 & 0.30 \\ 0.35 & 0.00 & 0.25 & 0.40 \end{bmatrix}; \\ B^7 &= \begin{bmatrix} 0.00 & 0.35 & 0.35 & 0.30 \\ 0.10 & 0.30 & 0.60 & 0.00 \\ 0.30 & 0.00 & 0.30 & 0.40 \\ 0.20 & 0.40 & 0.10 & 0.30 \end{bmatrix}.\end{aligned}$$

Table 9 Information criteria

AIC	$-2\hat{\ell} + 2df$
BIC	$-2\hat{\ell} + (\log N)df$
ssBIC	$-2\hat{\ell} + \log\left(\frac{2+N}{24}\right)df$

B.2 Most commonly used information criteria

Most commonly used ICs are presented in Table 9.

We considered a comparison of our proposed BIC_H with these ICs since, to the best of our knowledge, model selection criteria related to these MHMMs are only based on information criteria (Du et al. 2011; Marino and Alfo 2020) and have the same limitations as model selection for MMs and HMMs. Helske et al. (2018) suggested a BIC score to select the number of clusters and states analyzing short length categorical sequences, but the information criterion “kept suggesting models with more and more states” (Helske et al. 2018). Dias et al. (2009) applied an MHMM to analyze continuous financial sequences and identified the number of clusters using BIC, with the authors assuming they knew the number of states a priori. In a later work they relaxed that assumption and used MHMM to classify financial series identifying clusters and states by using BIC (Dias et al. 2015).

B.3 Measure of success for the model selection procedure

To define the *approximate success rate*, we first verified whether the number of components selected by the criterion, denoted as K_{sl} , was equal to the number of elements fixed for generating the sequences, i.e. K . If that condition was satisfied and the criterion selected a model with K components, we checked whether or not the number of hidden states picked, denoted as $s_{sl} = (S_{sl}^1, S_{sl}^2, \dots, S_{sl}^{K_c})$, was an approximation of the number of hidden states used to generate sequences, $s = (S^1, S^2, \dots, S^K)$.

The model selection is a success if $K_{sl} = K$ and the Manhattan distance between s and s_{sl} ,

$$d(s_{sl}, s) = \min_{\eta[s_{sl}^{(k)}]} \sum_{k=1}^K \left| s_{sl}^{(\eta(k))} - s^{(k)} \right| < 3, \tag{B2}$$

where $\eta[s_{sl}^{(k)}]$ is a permutation of the selected numbers of states, is satisfied. Setting the threshold at three means that we considered the model selection a success even if we over or underestimated the number of states by, at best, two states. For example, let us assume the number of clusters and the number of states used to generate the data were $K = 4$ and $s = (3, 4, 5, 5)$; thus, if a selection criterion led to choosing a model with $K_{sl} = 4$ clusters, the following selection for the number of states $\{(3, 4, 4, 5), (2, 4, 5, 5), (4, 5, 5, 5), (2, 5, 5, 5)\}$, is considered a success. Note that the threshold was selected by trying options with one and two. When the threshold was one (i.e. the correct number identification), all the criteria failed, as they did when the

threshold was two. A threshold of three was the minimum value which had satisfactory success rates.

Moreover, we considered any permutation of the hidden states vector in defining Eq. (B2) to avoid the label-switching problem that may occur during estimation. This means that the estimated components will not respect the order of generation, and we would need to identify the correct order before comparing results.

As regards the results presented in Tables 3 and 5, we refer to the *failure rates* which help in distinguishing between the over and underestimation of the number of clusters and states. Overestimation rises if $K_{sl} > K$ and underestimation if $K_{sl} < K$. In the case of $K_{sl} = K$ but when the success threshold in Eq. (B2) is not satisfied, i.e. $d(s_{sl}, s_c) \geq 3$, it is considered an overestimation if the sum of errors between selected and correct hidden states in each cluster is

$$\sum_{k=1}^K (s_{sl}^{(k)}(\eta) - s^{(k)}) \geq 3,$$

where $s_{sl}(\eta)$ is the vector of selected hidden states considering the permutation that minimizes the Manhattan distance in B2; it is considered an underestimation if

$$\sum_{k=1}^K (s_{sl}^{(k)}(\eta) - s^{(k)}) \leq -3.$$

B.4 Details on the explored scenarios

We restrict the simulation study to a scenario where a certain amount of prior information about the problem under analysis is available. These restrictions are due to the fact that all criteria struggle to identify the correct number of components and states and to the high number of models that should be considered in the simulation, e.g. if $K \in \{1, 2, \dots, 7\}$ and $S^k \in \{2, 3, \dots, 6\}$ for each component k , the number of models can be computed as a sum of K combinations with repetition:

$$\sum_{k=1}^K \frac{(s + k - 1)!}{k!(s - 1)!},$$

where $s = 5$ represents the S^k alternatives. Consequently, we have 791 combinations of components and states. Thus, once the sequences are generated for a specific scenario, we fitted to the simulated data a model having the correct number of components and states and 19 other models with a number of components close to the correct one (i.e. $K - 1$ or $K + 1$) and the number of hidden states for each component S^k from a discrete uniform distribution taking values between 2 and 6, i.e. $S^k \sim \mathcal{U}(2, 6)$.

Table 10 Results of the Monte Carlo study for $n = 300, T = 10$

				AIC	BIC	ssBIC
$n = 300$	$T = 10$	O_2	Statistic	21.192	36.92	36.92
			p value	2.077e-06	6.154e-10	6.154e-10
			Conf.int	[0.186; 1]	[0.265; 1]	[0.265; 1]
		O_3	Statistic	8.369	21.413	21.413
			p value	0.001908	1.852e-06	1.852e-06
			Conf.int	[0.089; 1]	[0.206; 1]	[0.206; 1]
		O_4	Statistic	6.021	6.723	6.021
			p value	0.007069	0.00476	0.007069
			Conf.int	[0.058; 1]	[0.068; 1]	[0.058; 1]
		O_5	Statistic	12.626	14.674	14.674
			p value	0.0001902	6.39e-05	6.39e-05
			Conf.int	[0.138; 1]	[0.159; 1]	[0.159; 1]
		O_6	Statistic	7.018	8.595	7.789
			p value	0.004035	0.001685	0.002628
			Conf.int	[0.067; 1]	[0.086; 1]	[0.076; 1]
		O_7	Statistic	4.502	5.789	5.125
			p value	0.01693	0.008062	0.01179
			Conf.int	[0.035; 1]	[0.056; 1]	[0.045; 1]

Z-tests related to BIC_H success rate against other information criteria. The results reported are the statistic, p value and 95% confidence interval

B.5 Statistical tests on simulation results

The simulation study in the main text highlights that BIC_H has the best performance when it comes to identifying the number of clusters and states in MHMMs for short length sequences. In Tables 10, 11, 12 and 13, we present Z-tests performed to assess whether the BIC_H success rate is higher than that of AIC, BIC and ssBIC.

B.6 Increasing the number of sequences and sequence lengths

Due to the computational time required for each scenario, the simulation study results reported in the main text are obtained from a number of sequences smaller than real clickstream data, i.e. we considered $n \in \{300, 500\}$. As an example, in Table 14 we report a scenario with a number of sequences $n = 10,000$, specifically with $T = 10$ and the two-component model O_2 . It should be noted that all criteria perform better even if BIC still selects the HMM over the MHMM. Furthermore, we report the results obtained when $n = 300$ and $T \in \{40, 60\}$ in Tables 15 and 16. Results still confirm what is seen in the main text. However, if $K = 2$ AIC outperforms the BIC_H while BIC and ssBIC still select the lack of clusters.

Table 11 Results of the Monte Carlo study for $n = 300$, $T = 20$

				AIC	BIC	ssBIC
$n = 300$	$T = 20$	O_2	Statistic	0.821	35.38	27.612
			p value	0.1825	1.356e-09	7.414e-08
			Conf.int	[-0.049; 1]	[0.248; 1]	[0.213; 1]
		O_3	Statistic	3.381	21.425	14.963
			p value	0.03297	1.84e-06	5.482e-05
			Conf.int	[0.015; 1]	[0.213; 1]	[0.16; 1]
		O_4	Statistic	7.385	15.637	15.637
			p value	0.003289	3.837e-05	3.837e-05
			Conf.int	[0.073; 1]	[0.163; 1]	[0.163; 1]
	O_5	Statistic	4.067	6.629	7.364	
		p value	0.02187	0.005016	0.003326	
		Conf.int	[0.027; 1]	[0.067; 1]	[0.077; 1]	
	O_6	Statistic	1.636	0.994	0.994	
		p value	0.1004	0.1593	0.1593	
		Conf.int	[-0.025; 1]	[-0.045; 1]	[-0.045; 1]	
	O_7	Statistic	0.505	1.286	0.726	
		p value	0.2386	0.1284	0.1971	
		conf.int	[-0.066; 1]	[-0.036; 1]	[-0.056; 1]	

Z-tests related to BIC_H success rate against other information criteria. Results reported are the statistic, p value and 95% confidence interval

Table 12 Results of the Monte Carlo study for $n = 500$, $T = 10$

				AIC	BIC	ssBIC
$n = 500$	$T = 10$	O_2	Statistic	5.746	33.58	29.04
			p value	0.008264	3.42e-09	3.545e-08
			Conf.int	[0.052; 1]	[0.26; 1]	[0.237; 1]
		O_3	Statistic	14.793	34.133	32.279
			p value	5.999e-05	2.573e-09	6.676e-09
			Conf.int	[0.159; 1]	[0.298; 1]	[0.287; 1]
		O_4	Statistic	10.127	19.501	18.309
			p value	0.0007307	5.028e-06	9.389e-06
			Conf.int	[0.109; 1]	[0.201; 1]	[0.19; 1]
	O_5	Statistic	5.789	6.486	5.789	
		p value	0.008062	0.005437	0.008062	
		Conf.int	[0.056; 1]	[0.066; 1]	[0.056; 1]	

Table 12 continued

		AIC	BIC	ssBIC
O_6	Statistic	5.356	6.723	6.021
	p value	0.01032	0.00476	0.007069
	Conf.int	[0.048; 1]	[0.068; 1]	[0.058; 1]
O_7	Statistic	7.768	7.768	8.561
	p value	0.002659	0.002659	0.001717
	Conf.int	[0.082; 1]	[0.082; 1]	[0.093; 1]

Z-tests related to BIC_H success rate against other information criteria. Results reported are the statistic, p value and 95% confidence interval

Table 13 Results of the Monte Carlo study for $n = 500, T = 20$

		AIC	BIC	ssBIC		
$n = 500$	$T = 20$	O_2	Statistic	0.741	38.005	23.337
			p value	0.1946	3.528e-10	6.799e-07
			Conf.int	[-0.054; 1]	[0.289; 1]	[0.211; 1]
	O_3	Statistic	1.623	19.864	14.871	
		p value	0.1014	4.158e-06	5.754e-05	
		Conf.int	[-0.026; 1]	[0.202; 1]	[0.16; 1]	
	O_4	Statistic	3.392	5.782	3.93	
		p value	0.03275	0.008094	0.02372	
		Conf.int	[0.015; 1]	[0.056; 1]	[0.025; 1]	
	O_5	Statistic	1.28	0.72	0.32	
		p value	0.1289	0.1981	0.2857	
		Conf.int	[-0.036; 1]	[-0.056; 1]	[-0.076; 1]	
	O_6	Statistic	0.51	0.327	0.185	
		p value	0.2376	0.2836	0.3337	
		Conf.int	[-0.065; 1]	[-0.075; 1]	[-0.085; 1]	
	O_7	Statistic	0.32	0.08	0.02	
		p value	0.2858	0.3886	0.4437	
		Conf.int	[-0.076; 1]	[-0.096; 1]	[-0.106; 1]	

Z-tests related to BIC_H success rate against other information criteria. Results reported are the statistic, p value and 95% confidence interval

Table 14 Results of the Monte Carlo study for $n = 10,000, T = 10$

		AIC	BIC	ssBIC	BIC_H		
$n = 10000$	$T = 10$	O_2	S	0.81 (0.039)	0.50 (0.050)	0.57 (0.050)	0.88 (0.032)
			U	0.00 (0.000)	0.50 (0.050)	0.43 (0.050)	0.00 (0.000)
			O	0.19 (0.039)	0.00 (0.000)	0.00 (0.000)	0.12 (0.032)

Approximate success rate (S) of identifying the correct model or an approximation of the model and *Failure rate* related to underestimate (U) and overestimate (O) the number of components. Standard errors are shown in parentheses. See Appendix B.3 for the definition of the approximate success or failure rate. The best among the selection criteria are indicated in bold

Table 15 Results of the Monte Carlo study for $n = 300$, $T \in \{40, 60\}$

$n = 300$		AIC	BIC	ssBIC	BIC_H
$T = 40$	O_2	0.34 (0.047)	0.10 (0.030)	0.16 (0.037)	0.58 (0.049)
	O_3	0.50 (0.050)	0.32 (0.047)	0.41 (0.049)	0.57 (0.050)
	O_4	0.49 (0.050)	0.48 (0.050)	0.51 (0.050)	0.63 (0.048)
	O_5	0.50 (0.050)	0.35 (0.048)	0.36 (0.048)	0.60 (0.049)
	O_6	0.57 (0.050)	0.55 (0.050)	0.56 (0.050)	0.72 (0.045)
	O_7	0.46 (0.050)	0.42 (0.049)	0.42 (0.049)	0.56 (0.050)
	$T = 60$	O_2	0.44 (0.050)	0.20 (0.040)	0.34 (0.047)
O_3		0.50 (0.050)	0.34 (0.047)	0.42 (0.049)	0.58 (0.049)
O_4		0.30 (0.046)	0.44 (0.050)	0.44 (0.050)	0.64 (0.048)
O_5		0.36 (0.048)	0.46 (0.050)	0.45 (0.050)	0.59 (0.049)
O_6		0.57 (0.050)	0.56 (0.050)	0.59 (0.049)	0.67 (0.047)
O_7		0.50 (0.050)	0.50 (0.050)	0.52 (0.050)	0.60 (0.049)

Approximate success rate of identifying the correct model or an approximation of the model. Standard errors are shown in parentheses. See Appendix B.3 for the definition of the approximate success rate. The best among the selection criteria are indicated in bold.

Table 16 Results of the Monte Carlo study for $n = 300$, $T \in \{40, 60\}$

$n = 300$		AIC	BIC	ssBIC	BIC_H		
$T = 40$	O_2	U	0.66 (0.047)	0.90 (0.030)	0.84 (0.037)	0.37 (0.048)	
		O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.05 (0.022)	
	O_3	U	0.48 (0.050)	0.68 (0.047)	0.59 (0.049)	0.22 (0.041)	
		O	0.02 (0.014)	0.00 (0.000)	0.00 (0.000)	0.21 (0.041)	
	O_4	U	0.51 (0.050)	0.52 (0.050)	0.49 (0.050)	0.25 (0.043)	
		O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.12 (0.032)	
	O_5	U	0.50 (0.050)	0.65 (0.048)	0.64 (0.048)	0.33 (0.047)	
		O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.17 (0.038)	
	O_6	U	0.43 (0.050)	0.45 (0.050)	0.44 (0.050)	0.23 (0.042)	
		O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.05 (0.022)	
	O_7	U	0.54 (0.050)	0.58 (0.049)	0.58 (0.049)	0.42 (0.049)	
		O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.02 (0.014)	
	$T = 60$	O_2	U	0.56 (0.050)	0.80 (0.040)	0.66 (0.047)	0.00 (0.000)
			O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.61 (0.049)
O_3		U	0.48 (0.050)	0.65 (0.048)	0.66 (0.047)	0.36 (0.048)	
		O	0.02 (0.014)	0.01 (0.0099)	0.02 (0.014)	0.06 (0.024)	
O_4		U	0.70 (0.046)	0.66 (0.047)	0.66 (0.047)	0.34 (0.047)	

Table 16 continued

$n = 300$		AIC	BIC	ssBIC	BIC _H
	O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.02 (0.014)
O_5	U	0.64 (0.048)	0.54 (0.050)	0.55 (0.050)	0.40 (0.049)
	O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.01 (0.010)
O_6	U	0.43 (0.050)	0.44 (0.050)	0.41 (0.049)	0.33 (0.047)
	O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.01 (0.010)
O_7	U	0.50 (0.050)	0.50 (0.050)	0.48 (0.050)	0.40 (0.049)
	O	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)	0.00 (0.000)

Failure rate related to underestimate (U) and overestimate (O) the number of components. Standard errors are shown in parentheses. See Appendix B.3 for the definition of the failure rate

Appendix C MHMM: parameters' estimates

In this section, we present results related to the selected model referred to in the main text. The selected model is that with $K = 3$ components and $S = (3, 2, 4)$ hidden states. Moreover, the prior probabilities are estimated through a multinomial regression model, which was presented in the main text in Eq. (2), considering the additive contribution of the three covariates: Nationality, Device and Month (excluding interactions).

$$\ln \frac{\omega_i^k}{\omega_i^1} = \ln \frac{P(M^k|X_i)}{P(M^1|X_i)} = \gamma_0 + \gamma_1 \text{Nationality}_i + \gamma_2 \text{Device}_i + \gamma_3 \text{Month}_i.$$

Tables 17, 18, 19, 20, 21, 22, 23, 24, 25 and 26 show the parameter estimates obtained through the EM algorithm.

Table 17 Multinomial regression model results

Covariate effects	Estimate	Sd. error
Cluster 2		
(Intercept)	- 1.5852	0.267
Device:PC	- 1.2951	0.085
Zone:Africa	- 1.3114	0.892
Zone:Latin America	- 0.0514	0.537
Zone:Oceania	2.8312	1.109
Zone:North Europe	- 0.9248	0.252
Zone:Asia	- 0.6925	0.292
Zone:East Europe	- 1.4603	0.276
Zone:North America	- 1.6719	0.255
Zone:Middle East	- 0.4519	0.644
Zone:Italy	- 0.6502	0.245
Zone:Russia	- 0.3587	0.399
Month:October	- 1.6045	0.131
Month:November	2.3577	0.132
Month:December	2.3665	0.134
Cluster 3		
(Intercept)	- 1.6141	0.1697
Device:PC	- 0.3258	0.0708
Zone:Africa	- 1.1951	0.7813
Zone:Latin America	- 0.2512	0.3423
Zone:Oceania	2.5224	1.0392
Zone:North Europe	- 0.0479	0.1688
Zone:Asia	- 2.8154	0.2147
Zone:East Europe	- 2.2718	0.1730
Zone:North America	- 0.1630	0.1747
Zone:Middle East	- 0.7601	0.4921
Zone:Italy	- 0.0225	0.1625
Zone:Russia	- 1.3035	0.2580
Month:October	- 0.1619	0.0724
Month:November	- 0.2997	0.0772
Month:December	- 0.4363	0.0795
Cluster 1, Mobile, South Europe and September as the baselines		

Table 18 Mixture hidden Markov model

State 1	State 2	State 3
0.7737	0.1370	0.0893
Initial probabilities in Cluster 1		

Table 19 Mixture hidden Markov model

State 1	State 2
0.655	0.345

Initial probabilities in Cluster 2

Table 20 Mixture hidden Markov model

State 1	State 2	State 3	State 4
0.561	0.115 0	0.214	0.110

Initial probabilities in Cluster 3

Table 21 Mixture hidden Markov model

	State 1	State 2	State 3
State 1	0.73217	0.1496	0.1182
State 2	0.00833	0.8958	0.0959
State 3	0.05551	0.0951	0.8494

Transition probabilities in Cluster 1

Table 22 Mixture hidden Markov model

	State 1	State 2
State 1	0.9947	0.00533
State 2	0.0347	0.96528

Transition probabilities in Cluster 2

Table 23 Mixture hidden Markov model

	State 1	State 2	State 3	State 4
State 1	0.2283	0.4275	0.2487	0.0955
State 2	0.0231	0.8494	0.0261	0.0231
State 3	0.0115	0.0390	0.9382	0.0113
State 4	0.0339	0.0389	0.0365	0.8907

Transition probabilities in Cluster 3

Table 24 Mixture hidden Markov model

	Accommodation	Attractions	Experiences	Events	Homepage	Info	Services
State 1	0.000	0.0115	0.000816	0.00000	0.9877	0.000	0.00000
State 2	0.188	0.2168	0.087138	0.05482	0.1113	0.000	0.34144
State 3	0.000	0.1327	0.001514	0.00419	0.0733	0.781	0.00717

Emission probabilities in Cluster 1

Table 25 Mixture hidden Markov model

	Accommodation	Attractions	Experiences	Events	Homepage	Info	Services
State 1	0.0194	0.908	0.0205	0.00105	0.0333	0.00750	0.01059
State 2	0.1657	0.109	0.0166	0.61265	0.0882	0.00423	0.00356

Emission probabilities cluster 2

Table 26 Mixture hidden Markov model

	Accommodation	Attractions	Experiences	Events	Homepage	Info	Services
State 1	0.000	0.0514	0.00929	0.028945	0.8973	0.00851	0.004511
State 2	0.141	0.7399	0.00000	0.000240	0.0407	0.05631	0.022153
State 3	0.912	0.0623	0.00500	0.000423	0.0172	0.00254	0.000904
State 4	0.210	0.1295	0.54329	0.016053	0.0475	0.00884	0.044802

Emission probabilities in Cluster 3

Acknowledgements The authors wish to thank two anonymous reviewers for their invaluable comments on a previous version of this paper. We are specially grateful to Christophe Biernacki and Giuseppe Sanfilippo for the useful suggestions on the derivation of the joint entropy measure.

Author contributions Furio Urso, Antonino Abbruzzo and Maria Francesca Cracolici conceived of the presented idea. Furio Urso developed the proposed measure and performed the computations and simulations. Antonino Abbruzzo verified the analytical method and simulations. Maria Francesca Cracolici encouraged Furio Urso to investigate modelling selection of mixture models considering clusters and hidden states, simultaneously; and supervised, jointly with Antonino Abbruzzo the findings of this work. Furio Urso, Antonino Abbruzzo and Marcello Chiodi contributed to the interpretation of Montecarlo simulations. Furio Urso and Maria Francesca Cracolici contributed to the interpretation of the results of the empirical analysis. All authors discussed the results and contributed to the final manuscript.

Funding Open access funding provided by Università degli Studi di Palermo within the CRUI-CARE Agreement. This work was carried out within GRINS project (Growing Resilient, INclusive and Sustainable) and received funding from the European Union Next-GenerationEU (*National Recovery and Resilience Plan*, NRRP, Mission 4, Component 2, Investment 1.3—D.D. 1558 11/10/2022, PE00000018, Spoke 7 Territorial sustainability). This manuscript reflects only the authors' views and opinions, neither the European Union nor the European Commission can be considered responsible for them.

Data availability The authors declare that the dataset belongs to LovePanormus and can not be shared.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Consent for publication The authors declare that all listed authors have approved the manuscript before submission, including the names and order of authors and gave their consent to publication.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, 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 licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If

material is not included in the article's Creative Commons licence 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. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Akaike H (1974) A new look at the statistical model identification. *IEEE Trans Autom Control* 19(6):716–723
- Altman RM (2007) Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting. *J Am Stat Assoc* 102(477):201–210
- Baudry JP, Raftery AE, Celeux G et al (2010) Combining mixture components for clustering. *J Comput Graph Stat* 19(2):332–353
- Biernacki C, Govaert G (1997) Using the classification likelihood to choose the number of clusters. *Comput Sci Stat* 451–457
- Biernacki C, Govaert G (1999) Choosing models in model-based clustering and discriminant analysis. *J Stat Comput Simul* 64(1):49–71
- Biernacki C, Celeux G, Govaert G (2000) Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans Pattern Anal Mach Intell* 22(7):719–725
- Boucheron S, Gassiat E (2007) An information-theoretic perspective on order estimation. In: Cappé, Olivier and Moulines, Eric and Rydén, Tobias, (eds) *Inference in hidden Markov models*, pp 565–601
- Celeux G, Durand JB (2008) Selecting hidden Markov model state number with cross-validated likelihood. *Comput Stat* 23(4):541–564
- Cooley RW, Srivastava J (2000) Web usage mining: discovery and application of interesting patterns from web data. Citeseer
- Das R, Turkoglu I (2009) Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method. *Expert Syst Appl* 36(3):6635–6644
- Dias JG (2006) Model selection for the binary latent class model: a Monte Carlo simulation. In: *Data science and classification*. Springer, New York, pp 91–99
- Dias JG (2007) Model selection criteria for model-based clustering of categorical time series data: a Monte Carlo study. In: *Advances in data analysis*. Springer, New York, pp 23–30
- Dias JG, Vermunt JK, Ramos S (2009) Mixture hidden Markov models in finance research. In: *Advances in data analysis, data handling and business intelligence*. Springer, New York, pp 451–459
- Dias JG, Vermunt JK, Ramos S (2015) Clustering financial time series: New insights from an extended hidden markov model. *Eur J Oper Res* 243(3):852–864
- Du J, Hu Y, Jiang H (2011) Boosted mixture learning of gaussian mixture hidden markov models based on maximum likelihood for speech recognition. *IEEE Trans Audio Speech Lang Process* 19(7):2091–2100
- Durand JB, Guédon Y (2016) Localizing the latent structure canonical uncertainty: Entropy profiles for hidden Markov models. *Stat Comput* 26(1–2):549–567
- Helske S, Helske J (2019) Mixture hidden Markov models for sequence data: the seqHMM package in R. *J Stat Softw* 88(3):1–32
- Helske S, Helske J, Eerola M (2018) Combining sequence analysis and hidden Markov models in the analysis of complex life sequence data. *Sequence analysis and related approaches*. Springer, Cham, pp 185–200
- Hernando D, Crespi V, Cybenko G (2005) Efficient computation of the hidden Markov model entropy for a given observation sequence. *IEEE Trans Inf Theory* 51(7):2681–2685
- Humphreys K (1998) The latent Markov chain with multivariate random effects: an evaluation of instruments measuring labor market status in the british household panel study. *Sociol Methods Res* 26(3):269–299
- Liu H, Kešelj V (2007) Combined mining of web server logs and web contents for classifying user navigation patterns and predicting users' future requests. *Data Knowl Eng* 61(2):304–330
- Marino MF, Alfó M (2020) Finite mixtures of hidden markov models for longitudinal responses subject to drop out. *Multivar Behav Res* 55(5):647–663
- McLachlan GJ, Peel D (2004) *Finite mixture models*. Wiley, New York
- Rissanen J (1978) Modeling by shortest data description. *Automatica* 14(5):465–471
- Schwarz G et al (1978) Estimating the dimension of a model. *Ann Stat* 6(2):461–464

-
- Scott SL, Hann IH (2006) A nested hidden Markov model for internet browsing behavior. *Marshall School of Business*, pp 1–26
- Smyth P (1997) Clustering sequences with hidden Markov models. In: *Advances in neural information processing systems*, pp 648–654
- Smyth P (1999) Probabilistic model-based clustering of multivariate and sequential data. In: *Proceedings of the seventh international workshop on AI and statistics*, Citeseer, pp 299–304
- Van de Pol F, Langeheine R (1990) Mixed Markov latent class models. *Sociol Methodol* 213–247
- Vermunt JK, Tran B, Magidson J (2008) Latent class models in longitudinal research. *Design, measurement, and analysis, Handbook of longitudinal research*, pp 373–385
- Volant S, Bérard C, Martin-Magniette ML et al (2014) Hidden Markov models with mixtures as emission distributions. *Stat Comput* 24(4):493–504
- Ypma A, Heskes T (2002) Automatic categorization of web pages and user clustering with mixtures of hidden Markov models. In: *International workshop on mining web data for discovering usage patterns and profiles*. Springer, New York, pp 35–49
- Zucchini W, MacDonald IL, Langrock R (2017) *Hidden Markov models for time series: an introduction using R*. CRC Press, Boca Raton

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.