UNIVERSITÀ DEGLI STUDI DI PALERMO

Dottorato in Scienze Economiche e Statistiche

Dipartimento di Scienze Economiche, Aziendali e Statistiche

Clickstream Data Analysis: A Clustering Approach Based on Mixture Hidden Markov Models

CANDIDATO

Furio Urso

COORDINATORE

Andrea Consiglio

SUPERVISOR

Maria Francesca Cracolici

CO-SUPERVISOR

Antonino Abbruzzo

CICLO XXXV

ANNO DI CONSEGUIMENTO 2023

Clickstream Data Analysis: A Clustering Approach Based on Mixture Hidden Markov Models

Abstract

Nowadays, the availability of devices such as laptops and cell phones enables one to browse the web at any time and place. As a consequence, a company needs to have a website so as to maintain or increase customer loyalty and reach potential new customers. Besides, acting as a virtual *point-of-sale*, the company portal allows it to obtain insights on potential customers through clickstream data, web generated data that track users accesses and activities in websites. However, these data are not easy to handle as they are complex, unstructured and limited by lack of clear information about user intentions and goals. Clickstream data analysis is a suitable tool for managing the complexity of these datasets, obtaining a cleaned and processed sequential dataframe ready to identify and analyse patterns.

Analysing clickstream data is important for companies as it enables them to understand differences in web user behaviour while they explore websites, how they move from one page to another and what they select in order to define business strategies targeting specific types of potential costumers. To obtain this level of insight it is pivotal to understand how to exploit hidden information related to clickstream data.

This work presents the cleaning and pre-processing procedures for clickstream data which are needed to get a structured sequential dataset and analyses these sequences by the application of Mixture of discrete time Hidden Markov Models (MHMMs), a statistical tool suitable for clickstream data analysis and profile identification that has not been widely used in this context. Specifically, hidden Markov process accounts for a time-varying latent variable to handle uncertainty and groups together observed states based on unknown similarity and entails identifying both the number of mixture components relating to the subpopulations as well as the number of latent states for each latent Markov chain.

However, the application of MHMMs requires the identification of both the number of components and states. Information Criteria (IC) are generally used for model selec-

tion in mixture hidden Markov models and, although their performance has been widely studied for mixture models and hidden Markov models, they have received little attention in the MHMM context. The most widely used criterion is BIC even if its performance for these models depends on factors such as the number of components and sequence length. Another class of model selection criteria is the Classification Criteria (CC). They were defined specifically for clustering purposes and rely on an entropy measure to account for separability between groups. These criteria are clearly the best option for our purpose, but their application as model selection tools for MHMMs requires the definition of a suitable entropy measure.

In the light of these considerations, this work proposes a classification criterion based on an integrated classification likelihood approach for MHMMs that accounts for the two latent classes in the model: the subpopulations and the hidden states. This criterion is a modified ICL_BIC, a classification criterion that was originally defined in the mixture model context and used in hidden Markov models. ICL_BIC is a suitable score to identify the number of classes (components or states) and, thus, to extend it to MHMMs we defined a joint entropy accounting for both a component-related entropy and a state-related conditional entropy.

The thesis presents a Monte Carlo simulation study to compare selection criteria performance, the results of which point out the limitations of the most commonly used information criteria and demonstrate that the proposed criterion outperforms them in identifying components and states, especially in short length sequences which are quite common in website accesses. The proposed selection criterion was applied to real clickstream data collected from the website of a Sicilian company operating in the hospitality sector. Data was modelled by an MHMM identifying clusters related to the browsing behaviour of web users which provided essential indications for developing new business strategies. This thesis is structured as follows: after an introduction on the main topics in Chapter 1, we present the clickstream data and their cleaning and pre-processing steps in Chapter 2; Chapter 3 illustrates the structure and estimation algorithms of mixture hidden Markov models; Chapter 4 presents a review of model selection criteria and the definition of the proposed ICL_BIC for MHMMs; the real clickstream data analysis follows in Chapter 5.

Acknowledgements

This thesis work is the culmination of an important path to which many people have contributed enormously. I am deeply grateful to my supervisor, Professor Maria Francesca Cracolici, for all the support and advice she provided me in this years. Her support made me grow from an academic, professional and human point of view, making me stronger and more aware of my abilities, always reminging me the meaning of research: ask yourself the right questions, find the best path to reach the answers, understand what the results convey and if you fail don't stop. During this path she never stopped transmitting her passion for research, professionalism and humbleness that can hardly be found elsewhere. I thank my co-supervisor, Professor Antonino Abbruzzo, for his professionalism, his patience and always being present and available beyond what was necessary. His advice and his attention to details have been the backbone of my current preparation. Their help, their constructive criticisms led to this work and to the conclusion of an important path. I owe the beginning of my career to everything they passed on to me, voluntarily and not, in these years of research. I also want to thank my supervisor Reza Mohammadi who I worked with at Amsterdam Business School. He helped me define my research method, find new directions and opportunities. The period spent in Amsterdam was very important for my professional growth. I am grateful to all the colleagues I met during my PhD; above all Nico, Pri, Ale, Totò, Andre and Chiara, have made these years a family trip. We helped and supported each other every second, were happy with each other's highs and made the lows less deep. I thank all the colleagues, researchers and professors that helped me, sharing experiences and laughing together in Mineo classroom. They made the dSEAS department a wonderful place to work. A deep hug to my family. My mother Eliana, my father Franco, my sister Lucilla and my friends Giulia, Mik, Nik, Nadia and Ale for the strength they always give me in everything I do. You are the reason I achieve my goals and I won't stop making you proud.

Publication based on this Thesis

- Cracolici M.F., **Urso F.** (2020). Web Usage Mining and Website Effectiveness. Book of Short Papers, SIS2020, pp. 1129-1134. ISBN: 9788891910776

- Cracolici M.F., **Urso F.** (2020). Exploring User Behavior in Destination Websites: An Application of Web Mining Techniques. A Broad View of Regional Science: Essays in Honor of Peter Nijkamp. New Frontiers in Regional Science: Asian Perspectives. Suzuki, S., Patuelli, R., pp. 259-273. ISBN: 9789813340978

- **Urso F.**, Abbruzzo A., Cracolici M.F. (2021). Analysis of clickstream data with mixture hidden Markov models. Book of Short Papers, SIS2021, pp. 823-828. ISBN: 9788891910776

- **Urso F.**, Abbruzzo A., Cracolici M.F.(2021). Model selection procedure for mixture hidden Markov models. In CLADAG 2021. 13th Scientific Meeting of the Classification and Data Analysis Group 2021 (Vol. 128, pp. 243-246). Firenze University press. ISBN: 9788855183406

- **Urso F.**, Abbruzzo A., Cracolici M.F. , Chiodi M. (2022). Model selection for mixture Hidden Markov models: An application to Clickstream Data. *Statistical Papers* (Under review)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The interaction between business and consumer has changed dramatically over the past few decades as a result of new technology; in the present era, every company is online or exists virtually completely online, selling goods and services to customers directly through the Internet. The more established businesses understand the value of bolstering their marketing plans by setting up online sales channels, though. The digitalization of transactions, an online presence through web portals, social media advertising opportunities, and real-time product and person traceability give businesses access to new, highly-generated streams of data. If properly handled, this enormous volume of data can serve as a vital source of knowledge for developing timely and successful business strategies, particularly in industries where there is fierce competition, like retail and services. Web generated data such as clickstream data is specifically used to gather customer information from a company's website and can be used to explore browsing habits, track online purchases, enhance website functionality, and identify profiles or clusters so as to define strategic and operative marketing actions aimed at boosting business competitiveness and profitability as a result. Identifying user browsing behaviour profiles is the main focus of this thesis; to do that are both necessary techniques to manage web data and sequential analysis methods that enable to cluster user paths in the website. However, cluster identification for sequential data is not easy to perform and requires model-based approaches that takes into account web data limitations such as lack of information, handled via latent variables; moreover, it requires the definition of an adequate selection criterion.

As regards web data, web Usage Mining (WUM) is a branch of Web Mining (WM) and is essential for managing and analysing web related data in order to identify unknown

patterns and relationships. WM is generally divided into three branches: structure mining, content mining and usage mining. The aim of web structure mining is to obtain knowledge from the hypertexual structure representing the connections of the web. Web content mining information to be extracted directly from the content of the pages and to classify web pages according to similarity related to their purpose, key words, web media and text files. Web usage mining is the analysis of access patterns through website log files. Specifically, it focuses on the analysis of sequences of clicks made by users exploring websites. It requires long and time-consuming data pre-processing in order to obtain usable dataset.

Traces left by users interacting with the website, may allow managers to obtain information on purchase purposes, for example. Moreover, an important task concerns the classification of users according to a certain characteristic or different exploratory motivations. Some users, in the initial stages of browsing the website, focus on finding product information (Pavlou & Fygenson, 2006). In this phase users view product features and compare alternative products before making a purchase decision. After an information-gathering phase they switch to a buying phase. However, the collection of information is not always followed by the actual purchase of the product: it could be out of simple amusement or curiosity (Y.-H. Park & Fader, 2004). Sometimes users explore a website only for their information content in order to make comparisons with other portals that sell similar products. Users exploring websites behave in a similar way to consumers visiting retail stores; moving directly and quickly towards their goal with little time spent checking each product or moving slowly, spending more time analysing different brands and exploring every corner of the store (see Titus & Everett, 1995).

Understanding and classifying users by both movements and goals is one of the most important results that can be obtained from clickstream data analysis. However, although the amount web data is generally enormous, it is not simple to handle and does not report enough information to identify objectively the clusters or understand exploration behaviour. User' actions in website are recorded by *log-files*. They track in chronological order every web resource that is requested by web users while they access a web page. Specifically, these resources can be page HTML files, images, audio and video files, JAVA codes, etc, that are necessary to *download* the page.

Accurate analysis of clickstream data requires understanding how to clean and process them to obtain a sequential dataframe, then defining how to extract information from

them and identifying profiles. The cleaning process generally involves removing useless information such as page elements that are not the focus of the analysis. One of the main issues is identifying web users because IP addresses are insufficient, being assigned to users by the web server. Moreover, web programs known as *bots*, *spiders* and *crawlers* which scan websites to update recommendations in search engines, need to be removed. Unfortunately, their identification is not always simple to do.

Cleaned and processed sequential data is configured as a list of subjects to which one or more sets of ordered objects are assigned (the web pages). The selected object and the order of selection both represent the subjects' underlying motivations and desires. The only information available is generally related to the characteristics of the selected items/page (information about users may be limited or absent in *log-files*) and it is not clear which of these has the most significant influence on the selection and why.

That being said, as for the identification of consumer profiles, there is a lack of available information in clickstreams related to user characteristics and hidden goals and motivations. The only information available about consumers is their movements in the website as shown by the web sequences. Thus, the clustering approach is mostly based on the differences in sequence patterns. These differences may represent a heterogeneous population, with groups of subjects exhibiting similar behaviour. In order to identify user clusters we can rely on sequential clustering based on similarity measures between sequences. This approach is widely used to cluster static categorical sequences such as genetic sequences. However, its extension to a dynamic process is not simple, particularly as regards the definition of a similarity measure.

Another approach used to identify groups of sequences is model-based clustering, accounting for the evolution of sequences over time.

Given the discrete nature of web data, statistical models such as discrete- time Mixture Markov Models (MMMs) are becoming increasingly popular in analysing clickstreams as they can model sequences based on Markov processes and take into account categorical latent variables representing subgroups within the population.

Models based on mixture Markov models (MMMs) are used to identify clusters by assuming that users in the same group move similarly between web pages, i.e., that the sequences evolve according to the same Markovian process. This approach is an intermediate one between assuming that all sequences follow a single Markov model and assuming transition probabilities that vary for each sequence, which is rarely used due to

model complexity.

It is important to keep in mind that our goal is to identify browsing profiles to gain insight into user behavior. However, MMMs cannot capture similarities between sequences related to groups of pages responding to specific user needs, which may be crucial in understanding why certain pages are selected. For instance, web pages or thematic areas in the site may serve similar goals and needs during user website exploration. For example, different pages may be selected by users who are in a "mental state" of information search, or different pages or thematic areas in the site may be perceived as less important auxiliary pages and are selected in the advanced stages of exploration.

MMMs do not capture similarities between pages and instead perceive each page as an observed state completely unrelated to others. This results in high inter-sequence variability, leading MMMs to identify more behaviour profiles than necessary. While similarities between web pages can sometimes be deduced from the site itself, this is not always the case. The identification of behaviour profiles requires methods able to take into account the underlying reasons why a user accesses one page over another and this information is hidden and not present in the *log files*.

That being said, statistical models with hidden variables such as Mixture Hidden Markov Models (MHMMs, Vermunt et al., 2008) are particularly suitable for exploring web sequences and reveal unknown variables influencing browsing behaviour, even though they have received little attention from the empirical literature. MHMMs enable us to take two levels of uncertainty into account; the evolution of a latent process which explains why a subject moves from one item to another and a hidden variable related to the presence of clusters representing selection behaviour profiles. Specifically, such a model-based clustering approach enables us to objectively identify consumer groups accounting for dynamic sequence behaviour. However, using MHMMs to find clusters of sequences requires methods of model selection to identify the unknown number of clusters and hidden states.

Model selection criteria are generally based on scores derived from Information Criteria (IC), which, however, cannot select the correct number of components and states (see e.g. Celeux & Durand, 2008; Costa & Angelis, 2010; Dias, 2006; Helske et al., 2018). Biernacki et al. (2000) showed that IC, such as BIC, can identify the correct number of components in the context of mixture models if the latter show a high degree of sepa-

ration.[1] On the other hand, in identifying the number of latent states in hidden Markov models, Costa and Angelis (2010) showed that BIC tend to underestimate their number for short sequences. When we deal with mixture hidden Markov models, IC should be used carefully and their behaviour needs to be explored. It is worth noting that IC does 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.[2] Another approach to model selection is based on Classification Criteria (CC), which 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. However, this approach has received little attention in the literature on mixture hidden Markov models. Starting from the literature on model selection for mixture hidden Markov models, the main contribution of this thesis is the proposal of a classification criterion in the MHMM context based on a joint entropy measure that accounts for both mixture components and hidden states.

Finally, an application to real clickstream data is presented in the final chapter and aims to analyse web user accesses relating to the PalermoTravel[3] website in order to explore the browsing behaviour of its users. This is the website of a Sicilian company that offers hospitality services and information related to the province of Palermo. The portal is divided into thematic areas which can be accessed via a menu on the home page of the site. The Attraction area provides information on seaside and landscape areas, museums, monuments, gastronomy or general information on Sicily. The Accommodation area allows users to view pages of holiday apartments for rent and to carry out thematic research, based on the chosen city of destination, the period of interest, the number of guests, the preferred price range and other additional features such as the presence of certain services. The Service area is designed to meet the needs of tourists by offering the possibility of obtaining transport, special equipment or food products. The Experience area is for booking activities such as sports courses, food and wine experiences, guided tours or cultural activities. The Event area provides an overview of upcoming concerts,

[1] However, "if the correct model is not in the family of considered models, BIC criterion will tend to overestimate the correct size regardless of the separation of the clusters ", see Biernacki et al. (2000).

[2] A high-quality data partition is obtained when the class membership of each unit (probability of belonging to a class) is high in correspondence with a given class and low compared to the other classes. On the other hand, if the class memberships for a unit are close to each other, the classification quality is low.

[3] This name is a pseudonym

festivals, theatrical performances or other activities and is updated monthly. In addition, the site provides general information on the company and its staff, on partners and access to bloggers and user reviews. The web server log data is in *IIS-W3Cex* Extendend format and was collected over a period of four months: September, October, November and December 2017.

In summary, the goal of this thesis is to explore clickstream data used in business contexts to support the decision-making process of management by identifying browsing profiles based on both similarity between patterns in web sequences and between subject characteristics, understanding how similarities and differences affect their browsing behaviour. To identify behavioural profiles, we propose model-based sequential clustering via mixture Markov models. However, the lack of information related to *log files* makes it difficult to understand users' hidden goals and "mental states" behind their movements. For example, during their exploration, users may select different web pages to satisfy the same hidden need, such as informative or purchase-related, and mixture Markov models may miss this information, leading to a higher number of profiles to handle sequence variability. To address this issue, we apply mixture hidden Markov models that take into account unknown "mental states" affecting navigation and identify profiles based on what users want, rather than what they click. However, these models require selecting both clusters and hidden states, which we address by proposing to adapt existing likelihood- or entropy-based information criteria. Our contribution is an entropy-based criterion, which is a novelty in MHMMs for categorical series. We illustrate the behaviour of our criterion on simulated and real data.

The thesis is organized as follows. Chapter 2 illustrates clickstream data structure and their main issues in handling complexity. Specifically, the cleaning and pre-processing phases are presented in order to generate a sequential dataset. Sequence analysis techniques for data modelling and clustering are also introduced. Chapter 3 presents Markov models and their extension to mixture Markov models and hidden Markov models. Furthermore, the Estimation-Maximization algorithm for mixture hidden Markov models is illustrated. Chapter 4 focuses on the issue of model selection related to the identification of both number of mixture components and hidden states. Classic information criteria are presented in a literature review, and then a classification criterion based on entropy is proposed for MHMMs. The chapter ends with an assessment of the proposed criterion performance through a Monte Carlo simulation study. Chapter 5 presents an empiri-

cal study. Clickstream data from the PalermoTravel website are cleaned and processed. Profile identification is carried out by both mixture Markov models and mixture hidden Markov models. Finally, conclusions and future developments are presented.

Chapter 2

Clickstream data

The continuous global growth of eCommerce and online information services creates a continuous flow of massive amounts of data that represent user interactions with websites, such as how users navigate through a website while browsing, which pages they choose to access, what they purchase, the personal information they share in subscriptions, etc. By collecting and analysing this data, companies are able to obtain important information about their website users (and potential customers, if we consider eCommerce portals), classify them in behaviour categories, evaluate the effectiveness of the website and improve the structure and design of the portal, customize their products and services and adapt their business strategies. Such results can be obtained by means of Web Usage Mining (WUM), which is the identifying and analysing of the patterns in the flow of clickstreams (the sequential list of access requests to web resources) so as to model the browsing behaviour of web users identifying browsing profiles (Cooley et al., 1997). The web resources mentioned above refer to web pages, images, links and so on.

The web usage mining process can be divided into three phases (B. Liu, 2007); *i)* the data collection and pre-processing phase, *ii)* the pattern identification phase and the pattern analysis phase.

In the first phase, data is cleaned by eliminating the resources not relevant for the analysis, obtaining new information (e.g. by extracting the geographical positions of users from the IPs) and defining the sessions, i.e. the sequences of resources displayed by the same user in a defined period of time. In this preliminary phase it is possible to use other information, such as the contents and the structure of the site, to enrich the data.

In the second phase the data is processed using data mining techniques and the hidden

patterns reflecting user browsing behaviour are identified. Furthermore, indices are calculated that are representative of website users and their sessions.
In the final phase the obtained patterns are further processed, aggregated and analysed in order to identify hidden browsing motivations. WUM data sources are the server log files. In some cases it is also possible to have access to additional data obtained from external clickstreams or from demographic data sources collected from specific sites. Huang et al. (2009), Johnson et al. (2004) and Y.-H. Park and Fader (2004) have focused on navigation between websites while Bucklin and Sismeiro (2009), Moe (2003) and Olbrich and Holsing (2011) on that within a specific site. Resul et al. (2007) analysed secondary web data such as web server access logs, proxy server logs, browser logs, user profile registration data, user session data, cookies, user queries, bookmark data, mouse clicks and mouse scrolls.

The analysis of clickstream data is a precious resource for companies and can be used for a variety of purposes. For example, it is useful in a business context as a tool of purchasing prediction. Furthermore, it allows companies to evaluate how users react to the design of the site and how they interact with the structure, understood as a network of hypertext links. The second type of analysis is done through engagement measures that estimate the level of interest of users during navigation: examples are the duration of the visit in the website (Ghose et al., 2013), average number of pages viewed and average time spent in each page (Huang et al., 2009) and depth of search (Johnson et al., 2004). Olbrich and Holsing (2011) highlighted the positive correlation between these engagement measures and purchase intention, and that the sequence of pages viewed before a purchase is the best browsing behaviour predictor.
Looking at the sequences of pages that make up the navigation path, Canter et al. (1985) identified four types of behaviour/route-shape: *a)* the path, in which the same page is not visited more than once, *b)* the ring, a path that go back to the starting page, *c)* the loop, a path that revisits the same page several times and *d)* the spike, a path that at a certain point repeats the sequence backwards. The authors associated these types of behaviour with purchase intention.
Finally, Drott (1998) treated several methods of web server log mining in order to improve the design of the site and Sarukkai (2000) worked on link prediction and path analysis to improve user navigation.

In the light of these considerations, Section 1 of this chapter presents the data relating

to the accesses of a website i.e. log files and illustrates the cleaning and pre-processing procedures to recognize user accesses and obtain a structured sequential dataset representing user visits in the site. The next sections present web usage mining processes and techniques such as association analysis and Markov chains. Section 2 illustrates how the application of the association rules allows the possible links between different pages and sections of the website to be analysed. Section 3 presents Markov chains which allow the user path to be traced in terms of the probability of transitions from one page or section to another. Finally, Section 4 focuses on sequential clustering as a tool for identifying groups of sequences with similar patterns.

2.1 Data cleaning and pre-processing

Web usage data sources are *log-files* extracted from websites which are typically ASCII text files. Every time users have an interaction with a website, for example they click on a link to visualize a page, the browser sends the visualization request to the server and the server responds by sending the browser a large amount of different web resources needed to access the aforementioned page. Each of these web resources are cited in the individual lines of a log file.

These *log-files* register all the received requests in chronological order and the log lines are structured in fields containing information related to the requested web resource such as the IP address that sent the request, the time and day of the request, the URL of the requested resource (.html files, .jpeg files, .jv files etc.), the notification of the successful upload of a resource as a code, the number of bytes sent and received by the server, the user-agent field, the referrer field and the user ID if the site requires registration. The user-agent log field is one of the most important sources of information as it is a character string that contains codes related to the browser, software and device used by the IP to access the website,while the referrer field contains the URL of the web resource that was requested in the previous step of the same IP.

An example of a log file is shown in the box below and a scheme of the main log fields is presented in Table 2.2.

Log files are complex and unstructured and require cleaning and pre-processing operations before proceeding with a statistical analysis. Data cleaning involves the elimination of all the log lines not useful for the analysis to be carried out, for example by focusing

only on the pages viewed by the user, we exclude the information referring to images, structural and graphics elements of the site, elements that have loading errors and so on. Then, we remove the log fields that will not be used (for example bytes loaded).
Once the data has been cleaned, we proceed to the elimination of *bots*[1] as explained in the next sub-section.

```
Date:  2017-12-01 00:00:14 /Fields:  date time c-ip cs-username s-ip cs-method cs-uri-stem cs-uri-query sc-status sc-bytes cs-bytes
time-taken cs-version cs-host cs(User-Agent) cs(Cookie) cs(Referer)
2017-12-01 00:00:14 217.182.132.1 - 31.11.32.56 - /fr/appartements/appartamenti-
e-case-vacanza-in-sicilia/casa-design-a-palazzo-
merlo-245.html - 200 156155 261 947 - - Mozilla/5.0+(compatible;+AhrefsBot/5.2;
++http://ahrefs.com/robot/) - -
2017-12-01 00:00:38 207.46.13.1 - 31.11.32.56 - /vp-tour.asp - 500 878 298 522 - - Mozilla/5.0+(compatible;+bingbot/2.0;++
http://www.bing.com/bingbot.htm) - -
2017-12-01 00:00:55 40.77.167.1 - 31.11.32.56 - /public/Passaggio.jpg - 200 82970 294 579 - - Mozilla/5.0+(compatible;+bingbot/2.0;
++http://www.bing.com/bingbot.htm) - -
2017-12-01 00:01:31 216.244.66.1 - 31.11.32.56 - /casa-style.asp id-casa=434_hl=fr 301 1647 248 455 - -
Mozilla/5.0+(compatible;+DotBot/1.1;+
http://www.opensiteexplorer.org/dotbot,+help@moz.com) - -
2017-12-01 00:01:47 66.249.70.1 - 31.11.32.56 - / - 200 118273 296 1156 - - Mozilla/5.0+(compatible;+Googlebot/2.1;++http://
www.google.com/bot.html) - -
```

2.1.1 Bots and spiders filtering

A *web crawler*, *spider*, or search engine bot is a program that downloads and indexes content from every corner of the Internet, registering information about web pages and classifying them by content and keywords so that they can easily be retrieved. These bots are also called web crawlers or spiders as crawling is the technical term used to refer to accessing a website and retrieving data obtained through a computer program. These bots are almost always run by search engines. By applying a search algorithm to data collected by crawlers, search engines are able to provide useful links in response to user queries. So a search engine such as Google or Bing is able to generate a list of web pages every time a user send a request.

In some cases the server is able to recognize bots and spiders but their identification is not always simple and usually requires heuristics that are defined to take into account the website type and structure (Kohavi, 2001; Suchacka, 2014). If the server recognizes them a string “Bot” or “Spider” is present in the user-agent field. Sometime the bots start their activity in the website by accessing a txt file called “robot.txt”. This file specifies the rules that these programs have to follow to access the hosted website such as which pages they can crawl and which link to use. A “robot.txt” file is recorded in the log as the

[1] Programs that scan websites

Table 2.1: Log fields example

Users	A	B
date	2017-10-20	2017-10-20
time	00:00:23	00:38:37
c.ip	157.55.39.1	121.235.159.1
cs.username	-	-
s.ip	31.11.32.56	31.11.32.56
cs.method	-	-
cs.uri.stem	/robots.txt	/palermo_eng.asp
cs.uri.query	-	-
sc.status	200	200
sc.bytes	317	119197
cs.bytes	252	227
time.taken	175	5979
cs.version	-	-
cs.host	-	-
cs.User.Agent.	Mozilla/5.0+(compatible;+bingbot/2.0; ++http://www.bing.com/bingbot.htm)	Mozilla/5.0+(Windows+NT+10.0; +Win64;+x64)+AppleWebKit/537.36+(KHTML,+like+Gecko) +Chrome/51.0.2704.106+Safari/537.36
cs.Cookie.	-	-
cs.Referrer.	-	-

first page of bot exploration. If the bot is not easy to spot, we need to define identification criteria. For example, it is possible to define time-based heuristics that identify a bot if an IP address accesses pages too quickly by changing user-agent at high speed.

The bot-filtering step leads to a further issue: how to recognize users exploring the site based on the list of accessed resources in the *log-files*.

2.1.2 User identification

The next step of pre-processing is to identify the users who access the website. Although a user can make multiple visits to the same site, assigning multiple accesses to the same user is almost impossible. What can be done to identify multiple accesses from the same user? If users register and log-in before exploring the site, we will have a personal id in the log field and their activity can be tracked over time. However, many users do not log-in. An alternative is to request that a client-side cookie be accepted during the initial exploration of the website, which allows the website to recognize subsequent visits without requiring additional information. Unfortunately, these cases are limited compared to users who browse anonymously.

Commonly, different accesses are identified by the IP address alone, which are not sufficient to identify a user. Actually, since the number of IPs are limited, they are assigned in rotation by the ISP proxy server and the same address at different times will correspond to different users (B. Liu, 2007). Techniques to recognize users accessing in close proximity on the basis of the IP include the user-agent field which identifies characteristics of the device and the referrer field which records the pre vious action of an IP. By cross-referencing the information (see Table 2.2), it is possible to distinguish users during exploration, though still with some uncertainty (Cooley et al., 1999).

2.1.3 Sessionization

Once the users have been identified by using log fields, their activity is divided into a set of website interactions $\mathscr{S} = \{\mathscr{S}_1, \mathscr{S}_2, \ldots, \mathscr{S}_n\}$. As previously mentioned, if the user is registered or has accepted the cookies it is possible to define sessions as the n_i visits made by the same i-th user at different moments in time $\mathscr{S}_i = \{S_{i,1}, S_{i,2}, \ldots, S_{i,n_i}\}$ with $S_{i,j}$ the sequence of ordered web pages selected by i-th user at their j-th visit. If this is not possible, and therefore there is no certain correspondence between an IP address and

Table 2.2: User identification through IP, user-agent and referrer

Time	**IP**	URL	Referer	**User Agent**
13:00	1.100.3	Page 1	-	Agent1
13:01	1.100.3	Page 2	Page 1	Agent1
13:20	1.100.3	Page 3	Page 2	Agent1
13:29	1.100.3	Page 1	-	Agent2
13:33	1.100.3	Page 5	Page 3	Agent1
13:35	1.100.3	Page 3	Page 1	Agent2
13:38	1.100.3	Page 2	Page 3	Agent2
13:42	1.100.3	Page 4	Page 2	Agent2
13:45	1.100.3	Page 5	Page 4	Agent2
13:51	1.100.3	Page 1	-	Agent3
13:57	1.100.3	Page 3	Page 1	Agent3
14:07	1.100.3	Page 6	Page 3	Agent3
14:09	1.100.3	Page 2	Page 6	Agent3
14:13	1.100.3	Page 4	Page 2	Agent3

a real person, the sessions will be obtained by dividing the activity of the same IP on the basis of structure or time heuristic.

As far as a structure heuristic is concerned, it uses the referrer field to track the user's journey through the site from the first page logged-in, and it uses this field's missing values to separate sessions (if this information is available). Another option is to take into account the structure of the website and the links. If an IP moves from one page to another where there is no direct link then the second access will be a different user.

A time heuristic can be based on *session-duration* or *time-of-stay*. By using *session-duration*, we define a threshold τ as the maximum time for the same session S and the time of the first click in a session is denoted as t_0. We assign to this session every web paged requested by the same IP at time t such that $|t - t_0| \leq \tau$. When using *time-of-stay*, we define a threshold ζ as the maximum time to spend in a page and times t_1 and t_2 the ones related to consecutive pages requested by the same IP. We assign these two selected pages to the same session S if $|t_2 - t_1| \leq \zeta$.

While the *time-of-stay* approach is the most common, it does have limitations that when the IP associated at a user remains on the same page for long time, i.e. a period greater than a conventional fixed time, the browsing is considered ended and the possible next page, which is explored by the same IP, starts a new session. This procedure is known

as *sessionization*. Although it is not possible to identify users, sessions obtained splitting the same IP address's activity cannot be considered related and, conventionally, they are treated as separate sequences. Thus, we obtain a sequential dataset $\mathscr{S} = \{S_1, S_2, \ldots, S_n\}$ and each element is assumed to represent a singular user visit and interaction in the website.
Table 2.3 shows an example of sessionization of user activity in a website via a Path Matrix.

The Path matrix can be analysed by applying well known techniques of data mining such as association rules to understand relationships between different web pages, statistical models such as Markov models to estimate probabilities related on users' movements in the website and sequential clustering techniques that reflect similarity between web sequences and different browsing behaviour (Liu 2007).

Table 2.3: an example of Path matrix. Each line refers to an IP address and each column is a click

$\mathscr{S}$	Click 1	Click 2	Click 3	Click 4	Click 5	$\cdots$
S_1	Page 1					
S_2	Page 1	Page 2	Page 5			
S_3	Page 1	Page 4	Page 3			
S_4	Page 1	Page 2				
S_5	Page 4					
S_6	Page 1	Page 3	Page 2	Page 5	Page 4	
S_7	Page 1					
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$		

2.2 Association rules

Association rules analysis is one of the fundamental techniques of data mining. Introduced by Agrawal et al. (1993), it allows relationships between items of a dataset to be identified, and in the context of clickstream analysis it is used to identify groups of web resources/pages where access has been requested by the same users. Understanding relationships between web pages enables a company to have information about the tastes of users and also to monitor the efficiency of the site structure and to consider on the basis

of the browsing paths whether to introduce shortcuts (links) between products and pages associated with each each other.

Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of items and $S = \{s_1, s_2, \ldots, s_n\}$ a set of transactions where the element $t_j = \{i_{j_1}, i_{j_2}, \ldots, i_{j_s}\}$ is a subset of items such that $t_j \subseteq I$. For example, let us consider the set of 5 items I and a set of 4 transactions T represented in Table 2.4.

Table 2.4: Set of transactions in binary matrix

T	i_1	i_2	i_3	i_4	i_5
t_1	1	1	0	1	0
t_2	0	0	1	1	0
t_3	1	1	1	1	0
t_4	1	0	1	0	1

An association rule is an implication of the form:

$$X \rightarrow Y, \tag{2.1}$$

where

$$X \subset I, Y \subset I, X \cap Y = \emptyset$$

Sets X and Y are known as the *antecedent* and *consequent* of the rule. It is possible to measure the effectiveness of the rule by computing its support and confidence.
The support of a rule is the percentage of transactions in T that contains $X \cup Y$, and can be seen as an estimate of the probability $P(X \cup Y)$. So, it is a measure of the rule reliability in the transaction set T. Let n be the number of transactions in T. The support of the rule $X \rightarrow Y$ is computed as follows:

$$support(X \rightarrow Y) = \frac{(X \cup Y)_{count}}{n} \tag{2.2}$$

The confidence of a rule is a percentage value that measures how many of the transactions in T, that contain X, also contain Y. It can be seen as an estimate of the conditional probability $P(Y|X)$, and it is computed as follows:

$$confidence(X \rightarrow Y) = \frac{(X \cup Y)_{count}}{X_{count}} \tag{2.3}$$

So, confidence determines the predictability of the rule and if the confidence value of a rule is low it is not possible to predict Y from X with enough reliability.

Among the data mining algorithms developed for association rules the best known is the *apriori* algorithm (Agarwal, Srikant, et al., 1994). This method determines the rules by firstly identifying the sets of items satisfying a minimum threshold of *support*, then identifying rules that satisfy a minimum threshold of *confidence*.

In the context of the association rules, the calculation of lift values is interesting. These are indices used to check whether the association rules are effective in predicting user behaviour. Given a sequence, the lift measure is calculated as the ratio between the conditional probability of an event and the probability of the event occurring in the absence of the sequence. The index provides an estimate of the improvement of the predictive capacity of the rule. Thus, if we consider as items the web pages belonging to different categories, the lift is the ratio between the probability of accessing a page in category Y conditional on the display of other category pages, and the probability that a page in category Y is accessed; that is

$$lift(X \rightarrow Y) = \frac{confidence(X \rightarrow Y)}{P(Y)} \tag{2.4}$$

Lift values greater than 1 indicate that when the probability of displaying the goal page increases so does the number of the user's display of other page categories. When lift values are lower than one, this indicates that the display of the other categories decreases the probability of visualization.

2.3 The statistical Markov models

The association rules presented in the previous section are simple tools to understand how certain items/pages can increase or decrease the visualization of others. However, a statistical approach would allow us to exploit additional information as covariates of a regression model and to predict certain outcomes. Generally, in clickstream data analysis regression models have been used to predict a purchase event in eCommerce websites through generalized linear models (see e.g. Moe et al., 2002; Moe & Fader, 2004; Sismeiro & Bucklin, 2004) and also to estimate session duration (Bucklin & Sismeiro, 2003). Association rules consider the set of pages as a static object, ignoring the se- lection order as users explore the website whereas Markov processes allow patterns to be analysed, taking into account the probabilistic nature of users' movements. Markov models are designed to catch the evolution of a sequence of random variables corresponding to the

states of a system, where the probability of a state depends only on some of the previous states in the sequence. Markov chains can be used for the analysis of clickstreams and browsing behaviour considering the different pages of the website as the system states and predicting the probability of being on a particular page given the previous one (or given a set of previous pages). Sarukkai (2000) used first-order Markov models to clickstream data adapting a different model for each web user. Eirinaki et al. (2005) proposed adding information about the structure of the website by considering the application of Markov model on a graph structure accounting for web pages links. Montgomery et al. (2004) considered a multinomial probit model accounting for the dynamic nature of clickstream through a hidden Markov process.

To be precise, a discrete-time Markov chain is a stochastic process Y_t satisfying the Markov property that at each time t takes state y_t from a countable set R with probability that depends only on the previous states. If the probability of being in a state does not depend on time t then it is a time-homogeneous Markov chain. Moreover, a Markov chain is said to be "of order g" if the transition probability depends on the previous g states. If the state space is finite, the process can be described by transition matrices $A^{(g)}$, where the generic element $a_{hj}^{(g)}$ is the probability of transitioning from state h at time $t-g$ to state j at time t.

In the WUM context, the probability of users moving to a web page depends on their navigation behaviour, represented by the previous g visited pages (Moe, 2003). The parameters of a time-homogeneous Markov chain of the order g can be estimated using the model proposed by Raftery (1985) based on an approximation of the distribution of state probabilities obtained as a weighted sum of g previous state probabilities.

$$A = [a_{hj}] = \begin{array}{c} \begin{array}{cccccc} \text{Page 1} & \dots & \text{Page j} & \dots & \text{Page R} & \end{array} \\ \begin{pmatrix} a_{11} & \dots & a_{1,R} \\ \vdots & \ddots & \vdots \\ a_{R,1} & \dots & a_{R,R} \end{pmatrix} \begin{array}{l} \text{Page 1} \\ \vdots \\ \text{Page h} \\ \vdots \\ \text{Page R} \end{array} \end{array}$$

Where A is an $R \times R$ transition probability matrix. Markov chains allow one to estimate transition probabilities that represent the whole dataset and, thus, they assume that all

users behave similarly when they access the site. However, this is clearly far from reality, and we are interested in understanding the differences in users' browsing attitudes. In the next section, we focus on sequential clustering as a tool to partition clickstream, identifying different patterns of behaviour.

2.4 Clustering web sequences

Clustering is a technique for grouping a set of elements based on similar characteristics and in the context of web usage mining it is about identifying groups of users based on behavioural similarities. Identifying behavioural clusters is particularly useful for management in deciding marketing strategies for different website user profiles.

As user behaviour is represented by categorical sequences, identifying such groups relies on sequential data analysis techniques to discover statistically relevant temporal structures and common patterns in sequences.

Identifying these patterns in sequential data is generally based on measures of similarity used to define distance matrices between sequences (Abbott & Forrest, 1986; Abbott & Tsay, 2000) or on model-based sequential clustering techniques such as mixture Markov models and their extensions (Vermunt et al., 1999).

As regards the former approach, defining what one means by similarity between categorical sequences is vital in sequence data analysis. Once a similarity measure or criterion is identified the sequential data can be partitioned in groups and further explored with respect to information from additional sequences.

Generally, similarities between sequences are measured by counting the matching attributes. Some examples of these are the Longest Common Prefix or LCP (Elzinga, 2006) based on the number of common elements in the same position and the Hamming measure (Hamming, 1950), which is based on a number of different elements. These two measures require that each sequence is of the same length.

Another way to define similarity accounts for a cost function, used to measure the minimal cost of transforming one sequence into another, e.g the Dynamic Hamming Distance or DHD (Lesnard, 2006). The previous measures, however, do not account for shifting, i.e., considering two sequences to be similar if they have the same elements but shifted by one or more positions. Measures such as the Optimal Matching distance OM (Levenshtein et al., 1966) allow for sequences of different lengths as when they transform one sequence

into another, they account for both the cost of changing elements and the cost of adding or deleting elements shifting their position.
Similarity measures between sequences allow groups of web sequences to be identified by defining distance matrices and applying classic clustering techniques (see e.g. Banerjee & Ghosh, 2000; Wei et al., 2012). A common clustering technique for categorical data is K-medoids (H.-S. Park & Jun, 2009), which is based on the most centrally located elements. However this approach is computationally intensive and requires a known number of clusters. A more efficient algorithm is the leader clustering algorithm which randomly selects medoids and assigns the other elements by similarity (Yu & Luo, 2011). Additional information can also be used in identifying clusters of web sequences. Banerjee and Ghosh (2001), for example, clustered web users by defining an algorithm based on the longest Common sequences that also account for time spent inside pages.

However, clustering methods based on distance and similarity cannot always be extended to time series as the definition of such a distance measure is not unimportant, unless we ignore the dynamic nature of the phenomenon by treating sequences as static. This approach, when applied to clickstream data, does not give any insight into web user behaviour.
It is clear, therefore, that the model-based sequential clustering approach is to be preferred and by adopting Mixture Markov Models (MMMs), it is possible to catch different subpopulations in the data. Each subpopulation has its Markov model so that they may differ for the initial probabilities, transition matrices or both.
In the context of clickstream data, these differences mean that sequences belonging to the same cluster describe different browsing behaviour, such as starting the web path from a specific page or the web pages they select while exploring a website.

Markov models have been used on clusters of web sequences Sarukkai (2000) used first-order Markov models to clickstream data adapting a different model for each web user. However, since the number of users in a real application is extremely high this approach is not feasible. Cadez, Heckerman, et al. (2000) focused on the problem of sequence classification for web data, applying a mixture of first-order Markov models to identify clusters of users by considering the web page categories accessed as states of the Markov chain (see also Cadez, Gaffney, et al. (2000)). Dias and Vermunt (2007) highlighted the usefulness of mixture Markov models for the analysis of web models as they adapt both to heterogeneity between sequences and to serial dependencies. Mel-

nykov (2016) implemented a bi-clustering approach with a reduced number of categories to avoid the issue of increasing the number of parameters.

However, it is worth noting that classic Markov models do not take into account the unobserved heterogeneity related to lack of information about navigation purposes. They consider users' movements from one page to another, but do not clarify the reason for these movements. It is legitimate to assume that the pages clicked on by users exploring the site are just the expression of "mental states "that change while browsing the site. For example, the user could start from a preliminary phase in which he selects generic pages, then moves on to an exploratory phase in which he clicks on information pages and finally moves on to a purchase phase and selects product pages. This assumes that the transitions actually happen between these hidden "mental states "and the probabilities of accessing the pages depend on them. Thus, if we extend to hidden Markov models, we suppose that sequences evolve according to a hidden Markov process.

The difference between mixture hidden Markov models and mixture Markov models is that MMMs cluster based on observed sequence similarity, whereas MHMMs cluster based on the hidden sequences representing attitude changes. Furthermore, the use of a HMM allows for the clustering of web pages as well as sequences using the latent states of the Markov process. Specifically, if a subset of observed states (i.e. clicks on web pages) has high probabilities only conditioned to a specific latent state and low probabilities to other latent states, that state reveals the similarity of intention and "value "attributed by users to that group of pages. This leads to clusters of sequences of similar hidden user motivations. Furthermore, hidden states will also account for part of sequence variability, allowing a smaller number of clusters to represent user behaviour.

Nevertheless, there are very 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 was provided by Ypma and Heskes (2002)), who classified users assuming prior information on the hidden states.

An advantage in analysing and clustering web sequences by using mixture hidden Markov models instead of mixture Markov models is that a latent Markov structure can

take into account sequence uncertainty and identify less components while the MMMs would need a higher number of components to represent homogeneous browsing behaviour.

2.5 Discussion

Over the past few years the speed with which it is possible to carry out commercial transactions on the world wide web has led to an exponential growth of online business. We are able to follow the browsing behaviour of users by identifying the paths taken in the form of sequences of pages viewed. Analysis of user behaviour allows website owners to gain in-depth knowledge about their potential online shoppers and to customize product messages. The tool for obtaining this level of information is web usage mining, defined as the process of applying data mining techniques in order to discover the browsing patterns of users from web data.

This chapter presented clickstream data and its source of information, i.e. *log-files*. We illustrated how to clean this data by recognizing *bots*, identifying web users and their activity in the website. Then we explained how to process clickstream data to obtain a structured sequential dataset that can be used to identify patterns, applying data mining techniques such as association rules.
Furthermore, we illustrated sequential modelling by using Markov models that allow us to understand users' movements by estimating transition probabilities between web pages.
Finally, we presented sequential clustering techniques based on similarity measures and model-based clustering. In particular, we focused on the model-based approach presenting mixture Markov models.

However, since classic Markov models do not take into account the unobserved heterogeneity related to clickstream data, statistical models with hidden variables are to be preferred in analysing web sequences. Among them, Hidden Markov Models (HMMs) are the most suitable to explore sequences and reveal the unknown variables influencing browsing behaviour.
HMMs take into account how the sequence of clicks that represent the user's path depends on a latent Markov chain in which latent states allow us to classify sequence observations (web pages) into page groups. These groups reflect similarities between pages that may be related to unknown "states of mind "that influence navigation choices and reveal users'

goals as they navigate the site. That being said, sequential clustering by means of mixture hidden Markov models would enable us to identify profiles that better reflect user behaviour since sequences are clustered by hidden goals rather than just accessed pages.

In the next chapter we will focus on the modelling aspect of sequence analysis and we will present the Markov models and their extension to mixture Markov models and mixture hidden Markov models. MHMMs are the modelling approach we selected for clustering clickstream and we will present parameter estimation algorithms such as the Estimation-Maximization and forward-backward.

Chapter 3

Mixture hidden Markov models

3.1 Introduction

As pointed out in previous chapters, clickstream data is an essential source of information that businesses exploit to explore user behaviour on the web and determine prompt and effective business strategies (Cooley & Srivastava, 2000; Das & Turkoglu, 2009; H. Liu & Kešelj, 2007). This data, collected in log files, enables user activity inside websites to be tracked as sequences of pages. Each sequence element corresponds to a web page visited by a user (identified by an IP address) at time t (a click).

A severe limitation in using clickstream data is related to the lack of information about browsing behaviour. This sequential data shows which pages users access but does not explain why they navigate from page to page or exactly what they are looking for. Sequences that are similar in their order can actually represent different exploratory behaviour and goals on the website. There is a need to understand the hidden motivations of users in navigating the site as they can indicate the presence of sub-populations related to different patterns of behaviour. Extracting this type of information from the data would allow for the identification of differences in browsing behavior and sub-populations and this would be extremely useful for companies in choosing suitable web marketing strategies.

To this end, Mixture Hidden Markov Models (MHMMs) can be used to detect underlying structures and cluster data into homogeneous subsets representing similar browsing behaviour and identify different profiles of users.

Starting from Markov models and hidden Markov models for categorical sequences, in this chapter Mixture Hidden Markov models will be presented. Specifically, we focus

on mixture of first-order hidden Markov models with 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 reflect different unknown subpopulations called latent classes, which, in turn, are obtained through sequences assigned to them with specific probabilities (Van de Pol & Langeheine, 1990).

The chapter is structured as follows. Section 2 introduces Markov models and their extension to mixture Markov models. Then it presents hidden Markov models and mixture hidden Markov models. Section 3 focuses on parameter estimation for MHMMs presenting the Estimation-Maximization and forward-backward algorithms.

3.2 Markov Models and extensions

Markov models are statistical tools for randomly changing systems and are commonly used for modelling sequential data. The main assumption of these models is that the dynamic behaviour of time series depends on a process that evolves according to a Markov chain (Ching & Ng, 2006).

Let $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iT_i})$ be the generic i-th sequence of length T_i with card$|Y_i| = R$ and assume n independent sequences. Let $\pi = \{\pi_r\}_{r \in R}$ be the probability of initial state and $A = \{a_{j,h}\}_{j \in R, h \in R}$ is the $R \times R$ transition probability matrix. The element $a_{j,h}$ indicates the probability of making a transition from state j at time $t-1$ to state h at time t. The log-likelihood for a Markov model is

$$\ell(\Theta; y) = \sum_{i=1}^{n} \log \left(\pi_{y_{i,1}} \prod_{t=2}^{T_i} a_{y_{i,t-1}, y_{it}} \right),$$

Where $\Theta = \{\pi, A\}$ is the set of parameters. In this chapter, we consider time-discrete homogeneous Markov models in which the state at time t only depends on the state at previous time $(t-1)$ and does not depend on the time. Markov models assume that the population is homogeneous and that the model parameters are the same for all sequences. However, we can account for a heterogeneous population through a time-constant latent variable, defining the mixture Markov model.

3.2.1 Mixture Markov Models

Let $M = \{M^1, M^2, \ldots, M^K\}$ be a set of Markov models, $C = \{C_1, C_2, \ldots, C_n\}$ is a vector such that the variable $C_i = k$ is a cluster membership indicator for unit $i = 1, 2, \ldots, n$, $\Theta^k = \{\pi^k, A^k\}$ the set of parameters for each sub-model M^k related to each subpopulation, for $k = 1, \ldots, K$. For each sequence y_i, we can define the prior cluster probabilities as ω^k. Specifically, $P(C_i = k|\Theta) = \omega_k$ is the probability that the model parameters are those related to the k-th Markov model M^k in the k-th component.

Let $X_i = (X_{i1}, \ldots, X_{iQ})$ be a vector of Q time-constant covariates for the i-th sequence and γ_k the vector of regression coefficients corresponding to cluster k. This set of covariates can be used to estimate cluster memberships ω_i^k of each sequence according to the following multinomial logistic regression model:

$$\omega_i^k = P(C_i = k|X_i) = \frac{e^{\gamma_k \cdot X_i}}{1 + \Sigma_{j=2}^K e^{\gamma_j \cdot X_i}}, \tag{3.1}$$

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

The log-likelihood for mixture Markov models with covariates is

$$\begin{aligned}
\ell(\Theta; y, X) &= \sum_{i=1}^n \log P(y_i|\Theta, X_i) = \sum_{i=1}^n \log \left(\sum_{k=1}^K P(y_i, C_i = k, |\Theta^k, X_i) \right) \\
&= \sum_{i=1}^n \log \left(\sum_{k=1}^K \omega_i^k P(y_i|\Theta^k) \right) \\
&= \sum_{i=1}^n \log \left(\sum_{k=1}^K \omega_i^k \sum_u \pi_{y_{i1}}^k \prod_{t=2}^{T_i} a_{y_{i,t-1}, y_{i1}}^k \right).
\end{aligned} \tag{3.2}$$

Finally, the cluster posterior probabilities $P(C_i = k|y_i, X_i)$ are obtained as

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

where $P(C_i = k|X_i)$ are the cluster memberships defined in Eq. and $P(y_i|C_i = k, X_i)$ are conditional probabilities of the observed sequences in cluster k. $P(y_i|\Theta, X_i)$ is the likelihood of a MMM for a sequence i and the product $P(y_i|C_i = k, X_i) P(C_i = k|X_i)$ is the likelihood for sequence i given that we are in cluster k. These two quantities are

computed by using the forward-backward algorithm. Figure 3.1 schematizes mixture Markov models with time-constant covariates.

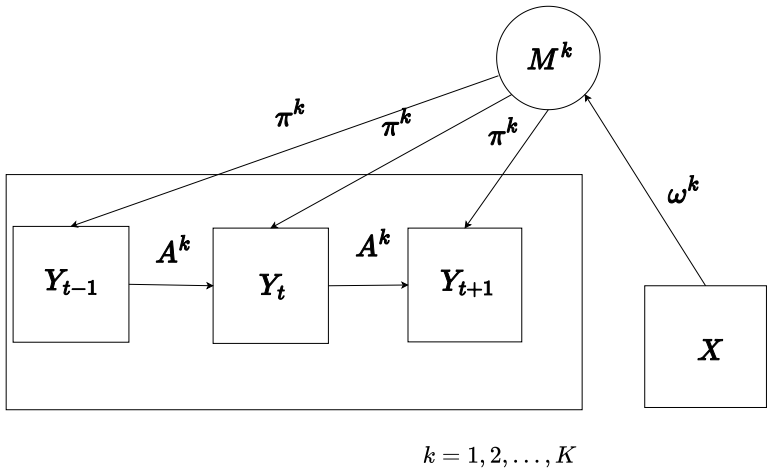

Figure 3.1: Structure of a mixture Markov model with covariates. The variables Y_t and X time-constant covariates are observable, M^k identify the k-th Markov model with $k = 1, 2, \ldots, K$ and K indicates the number of mixture components, ω^k are the mixture coefficients, $\Theta^k = \{\pi^k, A^k\}$ are the model parameters representing initial probabilities and transition matrix, respectively, for each component k

3.2.2 Hidden Markov Models

Let $Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iT_i})$ be a discrete random vector representing a sequence of length T_i with card$|Y_{ij}| = R$, $U_i = (U_{i1}, U_{i2}, \ldots, U_{iT_i})$ the i-th hidden random sequence, with card$|U_{ij}| = S$. Let $\pi = \{\pi_s\}_{s \in S}$ be the probability vector of initial hidden states and $A = \{a_{hj}\}_{h \in S, j \in S}$ a $S \times S$ transition probability matrix. An element a_{hj} indicates the probability of making a transition from state h at time $t-1$ to state j at time t. Finally, a $S \times R$ emission matrix $B = \{b_{sr}\}_{s \in S, r \in R}$ connects the hidden states and the observed states, where $b_{sr} = b_s(r)$ is a probability of hidden state s emitting observed state r. Let us assume that we collect n independent sequences $y = (y_1, y_2, \ldots, y_n)$ generated by the same hidden Markov model. The model parameter $\Theta = \{\pi, A, B\}$ is estimated by maximising the log-likelihood

$$\ell(\Theta;y) = \sum_{i=1}^{n} \log P(y_i|\Theta) = \sum_{i=1}^{n} \log\Big(\sum_{u} P(y_i|u,\Theta)P(u|\Theta)\Big)$$

$$= \sum_{i=1}^{n} \log\left(\sum_{u} P(u_{i1}|\Theta)P(y_{i1}|u_{i1},\Theta)\prod_{t=2}^{T_i} P(u_{it}|u_{i,t-1},\Theta)P(y_{it}|u_{it},\Theta)\right)$$

$$= \sum_{i=1}^{n} \log\left(\sum_{u} \pi_{u_{i1}} b_{u_{i1}}(y_{i1})\prod_{t=2}^{T_i} a_{u_{i,t-1},u_{it}} b_{u_{it}}(y_{it})\right),$$

where the hidden state sequences $u_i = (u_{i1}, u_{i2}, \ldots, u_{iT_i})$ take all possible combinations of values in the hidden state space S, and where y_{it} are the observations of unit i at time t, $\pi_{i1} = P(U_{i1} = s) \quad \forall i \in \{1,2,\ldots,n\}$ and $s \in \{1,\ldots,S\}$ is the initial probability of the hidden state at time $t = 1$ in sequence u_i; $a_{u_{i,t-1},u_{it}} = P(U_{it} = j|U_{i,t-1} = h) \quad \forall i \in \{1,2,\ldots,n\}$ and $h, j \in \{1,\ldots,S\}$ is the transition probability from the hidden state at time $t-1$ to the hidden state at t; and $b_{u_{it}}(y_{it}) = P(Y_{it} = r|U_{it} = s)$ with $s \in \{1,\ldots,S\}$ and $r \in \{1,\ldots,R\}$ is the probability that the hidden state of subject i at time t emits the observed state at t, in other words the state-dependent conditional probabilities.

3.2.3 Mixture Hidden Markov Models

Let $M = \{M^1, M^2, \ldots, M^K\}$ be a set of HMMs, $C = \{C_1, C_2, \ldots, C_n\}$ is a vector such that the variable $C_i = k$ is a cluster membership indicator for unit $i = 1,2,\ldots,n$, $\Theta^k = \{\pi^k, A^k, B^k\}$ is the set of parameters for each sub-model M^k, related to each subpopulation $k = 1,\ldots,K$. For each observed sequence y_i, we can define the prior cluster probabilities that the model parameters are those related to the k-th sub-model M^k as $P(C_i = k) = \omega^k$. Let $X_i = (X_{i1},\ldots,X_{iQ})$ be a vector of Q time-constant covariates for the i-th sequence and γ_k the vector of regression coefficients corresponding to cluster k. This set of covariates can be used to estimate cluster memberships ω_i^k as seen for mixture Markov models in equation 3.1.

The log-likelihood for MHMMs is

$$\ell(\Theta; y, X) = \sum_{i=1}^{n} \log P(y_i|\Theta, X_i) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} P(y_i, C_i = k|\Theta^k, X_i) \right)$$

$$= \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \omega_i^k P(y_i|\Theta^k) \right) \tag{3.3}$$

$$= \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \omega_i^k \sum_{u} \pi_{u_{i1}}^k b_{u_{i1}}^k (y_{i1}) \prod_{t=2}^{T_i} a_{u_{i,t-1},u_{i1}}^k b_{u_{it}}^k (y_{it}) \right).$$

Cluster posterior probabilities $P(C_i = k|y_i, X_i)$ are obtained as in subsection 3.2.1, equation 3.1.

Figure 3.2 schematizes mixture hidden Markov models with time-constant covariates.

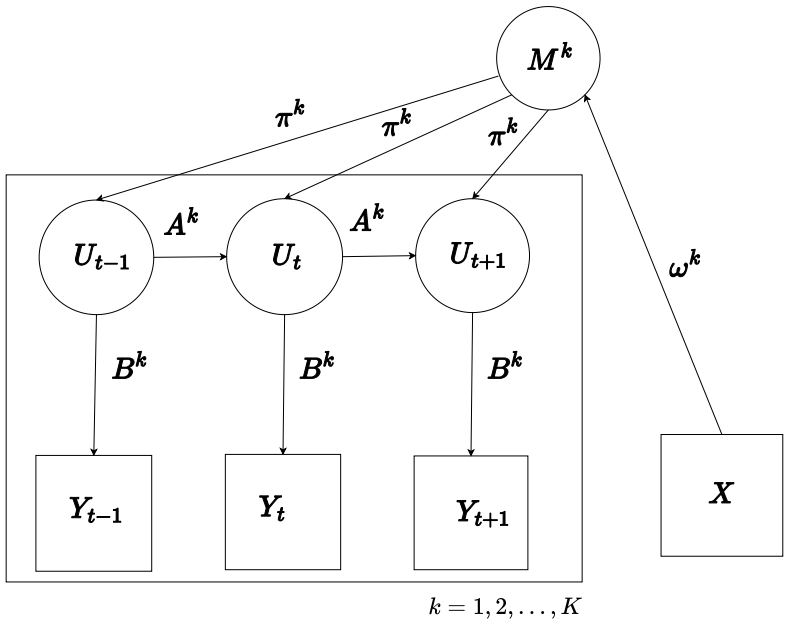

Figure 3.2: Structure of a mixture hidden Markov model with covariates. The variables Y_t and X time-constant covariates are observable, the U_t are hidden variables, M^k identify the k-th hidden Markov model with $k = 1, 2, \ldots, K$ and K indicates the number of mixture components, ω^k are the mixture coefficients, $\Theta^k = \{\pi^k, A^k, B^k\}$ are the model parameters representing initial probabilities, transition matrix and emission matrix, respectively, for each component k

3.3 Inference

The log-likelihood of MHMM complexity does not allow us to obtain an estimator in closed form. To overcome this problem, the estimation of parameters can be carried out through the expectation-maximization algorithm (Baum et al., 1970) jointly with the forward-backward algorithm (Paas et al., 2007; Rabiner, 1989).

Firstly, we need to understand how to maximize the log-likelihood of MHMMs. In these models the cluster label variable C_i and hidden sequence u_i are latent. Therefore, to obtain the maximization we will consider the complete-data log-likelihhod $\ell(\Theta, u, C; y)$ instead, assuming that the clusters and hidden states are known. We define the indicator functions as follows

$$w^{i,k} = \begin{cases} 1 & \text{if } C_i = k \\ 0 & \text{if } C_i \neq k, \end{cases} \tag{3.4}$$

$$p_t^{i,k}(h) = \begin{cases} 1 & \text{if } u_{i,t} = h \text{ and } C_i = k \\ 0 & \text{if } u_{i,t} \neq h \ , \end{cases} \tag{3.5}$$

$$z_t^{i,k}(h,j) = \begin{cases} 1 & \text{if } u_{i,t-1} = j \text{ and } u_{i,t} = h \text{ and } C_i = k \\ 0 & \text{if } u_{i,t-1} \neq j \text{ or } u_{i,t} \neq h, \text{ and } C_i = k \ . \end{cases} \tag{3.6}$$

Thus, $w^{i,k}$ takes value equal to one if the i-th hidden sequence is generated by the k-th component, $p_t^{i,k}(h)$ takes value equal to one if the i-th hidden sequence at the t-th click from the k-th component takes hidden state h, and $z_t^{i,k}(h,j)$ takes value equal to one if in the i-th hidden sequence from k-th component there is a movement from state j at the $(t-1)$-th click to state h at the t-th click. The complete-data log-likelihood is defined by using these indicator functions for the hidden variables. Specifically, the complete log-likelihood is obtained as the sum of three elements related to initial, transition and emission probabilities as follows

$$\ell(\Theta,u,C;y) = \sum_{i=1}^{n} w^{i,k} \sum_{h} p_1^{i,k}(h) \log \pi^k_{u_{i1}=h}$$

$$+ \sum_{i=1}^{n} w^{i,k} \sum_{h,j} \Big(\sum_{t=2}^{T_i} z_t^{i,k}(h,j) \Big) \log a^k_{u_{i,t-1}=j,u_{it}=h} \tag{3.7}$$

$$+ \sum_{i=1}^{n} w^{i,k} \sum_{h} \sum_{t=1}^{T_i} p_t^{i,k}(h) \log b^k_{u_{it}=h}(y_{it}).$$

It is possible to maximize the aforementioned complete log-likelihood by maximizing these three expressions.

3.3.1 Expectation-Maximization algorithm

The expectation-maximization algorithm is used to compute maximum likelihood estimates of parameters where there are unknown latent variables. This algorithm is a two-step procedure. Firstly, since the hidden sequence is not observed, the algorithm considers it as missing data and estimates the expected value of the complete-data log-likelihood $\ell(\Theta,u,C;y)$ seen in equation (3.7), given the observed sequences and parameters obtained in a previous step. Secondly, in the maximization step, the expected complete-data log-likelihood is maximized, and latent Markov model parameters $\Theta^k = \{\pi^k, A^k, B^k\}$ are estimated to be used in a new expectation step (Muthen & Shedden, 1999). Thus, the estimation consists of two steps, i.e. E-step and M-step.

As far as the first step is concerned (i.e. E-step), we select initial values for Θ^k in each cluster to start the iterations. We estimate $w^{i,k}$, $p_t^{i,k}(h)$ and $z_t^{i,k}(h,j)$ (defined in equations 3.4, 3.5 and 3.6) as conditional expectations given current estimates of parameter Θ^k obtained in the M-step

$$\begin{aligned}\hat{w}^{i,k} &= \mathbb{E}\left[w^{i,k}\right] \\ &= \frac{\omega^k P(y_i|C_i=k,\Theta^k)}{\sum_{k=1}^{K}\omega^k P(y_i|C_i=k,\Theta^k)}\end{aligned} \tag{3.8}$$

$$\begin{aligned}\hat{p}_t^{i,k}(h) &= \mathbb{E}\left[s_t(h)|\hat{\Theta}^k\right] \\ &= P(U_{i,t}=h|Y_i=y_i,C_i=k,\hat{\Theta}^k) \\ &= \frac{\alpha_{it}^k(h)\beta_{it}^k(h)}{\mathscr{L}(\hat{\Theta}^k,u_i;y_i)}\end{aligned} \tag{3.9}$$

$$\begin{aligned}\hat{z}_t^{i,k}(h,j) &= \mathbb{E}\left[z_t^{i,k}(h,j)|\hat{\Theta}^k\right] \\ &= P(U_{i,t-1}=j,U_{i,t}=h|Y_i=y_i,C_i=k,\hat{\Theta}^k) \\ &= \frac{\hat{a}^k_{u_{i,t-1}=j,u_{it}=h}\alpha^k_{i,t-1}(h)\hat{b}^k_{u_{it}=h}(y_{it})\beta^k_{it}(h)}{\mathscr{L}(\hat{\Theta}^k,u_i;y_i)}\end{aligned} \tag{3.10}$$

where estimates $\hat{w}^{i,k}$, $\hat{p}_t^{i,k}(h)$ and $\hat{z}_t^{i,k}(h,j)$ take values between 0 and 1.

$\mathscr{L}(\hat{\Theta}^k,u_i;y_i)$ is the complete likelihood evaluated at a sequence i with parameter estimate $\hat{\Theta}^k$ as

$$\mathscr{L}(\hat{\Theta}^k,u_i;y_i) = \hat{\pi}^k_{u_{i1}}\prod_{t=2}^{T_i}\hat{a}^k_{u_{i,t-1},u_{it}}\prod_{t=1}^{T_i}\hat{b}^k_{u_{it}}(y_{it})$$

and α_{it}^k and β_{it}^k are the forward and backward probabilities presented in the next subsection.

Regarding the M-step, we substitute previous values in equation (3.7) to maximize the complete log-likelihood (Harte, 2006; Jamalzadeh, 2011). We obtain new parameter estimate for each $\hat{\Theta}^k$ with $k = 1,2,\ldots,K$.

In the next paragraphs related to parameter estimation, we will use the simplified notations π_h^k instead of $\pi^k_{u_{i1}=h}$ to identify initial probabilities, $a^k_{j,h}$ instead of $a^k_{u_{i,t-1}=j,u_{it}=h}$ to identify transition probabilities and $b_h^k(r)$ instead of $b^k_{u_{it}=h}(y_{it}=r)$ to identify emission probabilities.

1. The first element to maximize in equation (3.7) is

$$\sum_{i=1}^{n} w^{i,k} \sum_{h} p_1^{i,k}(h) \log \pi_h^k.$$

under the constraint $\sum_h \pi_h^k = 1$. Let define the function f_1

$$f_1 = \sum_{i=1}^{n} w^{i,k} \sum_{h} p_1^{i,k}(h) \log \pi_h^k + \lambda_1 \Big[1 - \sum_{h} \pi_h^k\Big]$$

with λ_1 the Lagrange multiplier. Then the derivative is

$$\frac{\partial f_1}{\partial \pi_h^k} = \frac{\sum_{i=1}^{n} w^{i,k} p_1^{i,k}(h)}{\pi_h^k} - \lambda_1,$$

for $h = 1, 2, \ldots, S^k$. Finally, we obtain

$$\hat{\pi}_h^k = \frac{\sum_{i=1}^{n} \hat{w}^{i,k} \hat{p}_1^{i,k}(h)}{\sum_{i=1}^{n} \hat{w}^{i,k} \sum_{h'} \hat{p}_1^{i,k}(h')}, \tag{3.11}$$

2. The second element to maximize in equation (3.7) is

$$\sum_{i=1}^{n} w^{i,k} \sum_{h,j} \big(\sum_{t=2}^{T_i} z_t^{i,k}(h,j) \big) \log a_{j,h}^k$$

under S^k constraints $\sum_j a_{j,h}^k = 1$. Let define the function f_2

$$f_2 = \sum_{i=1}^{n} w^{i,k} \sum_{h,j} \big(\sum_{t=2}^{T_i} z_t^{i,k}(h,j) \big) \log a_{j,h}^k + \sum_{h} \lambda_{h,2} \Big[1 - \sum_{j} a_{j,h}^k\Big]$$

with $\lambda_{h,2}$ such that $h = 1, 2, \ldots, S^k$ the Lagrange multipliers. Then we compute the derivatives by each $a_{j,h}^k$ with $h, j = 1, 2, \ldots, S^k$

$$\frac{\partial f_2}{\partial a_{j,h}^k} = \frac{\sum_{i=1}^{n} w^{i,k} \sum_{t=2}^{T_i} z_t^{i,k}(h,j)}{a_{j,h}^k} - \lambda_{h,2},$$

setting the derivatives equal to 0 and knowing that

$$\sum_j \Big[\sum_{i=1}^{n} w^{i,k} \sum_{t=2}^{T_i} z_t^{i,k}(h,j) - \lambda_{h,2} a_{j,h}^k \Big] = 0,$$

Thus, $\lambda_{h,2} = \sum_j \sum_{i=1}^{n} w^{i,k} \sum_{t=2}^{T_i} z_t^{i,k}(h,j)$ because $\sum_j a_{j,h}^k = 1$ as it is the sum of elements of a transition matrix row.

Finally, the estimates are

$$\hat{a}_{j,h}^k = \frac{\sum_{i=1}^{n} \hat{w}^{i,k} \sum_{t=2}^{T_i} \hat{z}_t^{i,k}(h,j)}{\sum_j \sum_{i=1}^{n} \hat{w}^{i,k} \sum_{t=2}^{T_i} \hat{z}_t^{i,k}(h,j)}, \tag{3.12}$$

3. The third element to maximize in equation (3.7) is

$$\sum_{i=1}^{n} w^{i,k} \sum_h \sum_{t=1}^{T_i} p_t^{i,k}(h) \log b_h^k(y) = \sum_{i=1}^{n} w^{i,k} \sum_h \sum_{t=1}^{T_i} \sum_r o_t^{i,k}(h,r) \log b_h^k(r)$$

Where the formula is modified by using the new indicator function

$$o_t^{i,k}(h,r) = \begin{cases} 1 & \text{if } u_{i,t} = h \text{ and } y_{i,t} = r \text{ and } C_i = k \\ 0 & \text{if } u_{i,t} \neq h \text{ or } y_{i,t} \neq r, \text{ and } C_i = k \ , \end{cases}$$

such that $o_t^{i,k}(h,r)$ takes value equal to one if in the i-th observed sequence from the k-th component at the t-th click, the hidden state h emits the observed state r. Let define the function f_3 as

$$f_3 = \sum_{i=1}^{n} w^{i,k} \sum_h \sum_{t=1}^{T_i} \sum_r o_t^{i,k}(h,r) \log b_h^k(r) + \sum_{h=1}^{S^k} \lambda_{h,3} \Big[1 - \sum_{r=1}^{R} b_h^k(r) \Big]$$

with $\lambda_{h,3}$ such that $h = 1,2,\ldots,S^k$ the Lagrange multipliers. Then, we compute the derivatives for each $b_h^k(r)$ with $h = 1,2,\ldots,S^k$ and $r = 1,2,\ldots,R$ as

$$\frac{\partial f_3}{\partial b_h^k(r)} = \sum_{i=1}^{n} w^{i,k} \sum_{t=2}^{T_i} \frac{o_t^{i,k}(h,r)}{b_h^k(r)} - \lambda_{h,3},$$

setting the derivatives equal to 0 and knowing that

$$\sum_{r=1}^{R} \Big[\sum_{i=1}^{n} w^{i,k} \sum_{t=2}^{T_i} o_t^{i,k}(h,r) - \lambda_{h,3} b_h^k(r) \Big] = 0,$$

Thus, $\lambda_{h,3} = \sum_{r=1}^{R} \sum_{i=1}^{n} w^{i,k} \sum_{t=2}^{T_i} o_t^{i,k}(h,r)$ because $\sum_{r=1}^{R} b_h^k(r) = 1$ as it is the sum of elements of an emission matrix row. Finally, noticing that the count $\sum_{r=1}^{R} o_t^{i,k}(h,r) = p_1^{i,k}(h)$, the parameter estimate is

$$\hat{b}_h^k(r) = \frac{\sum_{i=1}^{n} \hat{w}^{i,k} \sum_{t=1 \Lambda y_t = r}^{T_i} \hat{p}_1^{i,k}(h)}{\sum_{i=1}^{n} \hat{w}^{i,k} \sum_{t=2}^{T_i} \hat{p}_t^{i,k}(h)}. \tag{3.13}$$

4. Moreover, the new estimates for mixture coefficients $\hat{\omega}^k$ are computed as

$$\hat{\omega}^k = \Big[\sum_{i=1}^{n} \hat{w}^{i,k} \Big] / n$$

These new parameter estimates are substituted in the complete log-likelihood for the following E-step. Expectation and maximization steps are iterated till convergence.

3.3.2 Forward-backward algorithm

The EM algorithm requires computing the model likelihood, for example, at each iteration to obtain estimates in equations 3.9 and 3.10. Since this can be computationally intensive, recursive procedures are used. The forward-backward algorithm allows the likelihood and the posterior probabilities of latent states to be calculated.

Let us focus on the algorithm for HMM (Vermunt et al., 2008), its extension to MHMM being obtained by combining the K models in an HMM having $S = \sum_{k=1}^{K} S^k$ hidden states, initial probabilities $\pi_i = (\omega_{i1}\pi^1, \omega_{i2}\pi^2, \ldots, \omega_{iK}\pi^K)$ and transition and emission matrices as block diagonal matrices where the blocks are given by A^k, B^k for $k = 1, 2, \ldots, K$.

Let $\alpha_{it}(s)$ be the forward probability defined as

$$\alpha_{it}(s) = P(y_{i1}, y_{i2}, \ldots, y_{it}, u_{it} = s|\Theta),$$

i.e., the joint probability of the observed sequence i-th up to time t and the hidden state at time t given model parameter Θ. Forward probabilities are solved recursively for $i = 1, 2, \ldots, n$ and $s = 1, 2, \ldots, S$ starting as

$$\begin{aligned}\alpha_{i1}(s) &= P(y_{i1}, u_{i1} = s|\Theta) \\ &= P(y_{i1}|u_{i1} = s, \Theta)P(u_{i1} = s|\Theta) \\ &= b_s(y_{i1})\pi(s),\end{aligned} \tag{3.14}$$

where $\pi(s)$ are initial probabilities and $b_s(y_{i1})$ are emission probabilities.

$$\begin{aligned}\alpha_{i2}(s) &= P(y_{i1}, y_{i2}, u_{i2} = s|\Theta) \\ &= \sum_{z=1}^{S} P(y_{i1}, y_{i2}, u_{i1} = z, u_{i2} = s|\Theta) \\ &= \sum_{z=1}^{S} P(y_{i1}|u_{i1} = z, \Theta)P(y_{i2}|u_{i2} = s, \Theta)\times \\ &\qquad P(u_{i2} = s|u_{i1} = z, \Theta)P(u_{i1} = z|\Theta) \\ &= \sum_{z=1}^{S} \alpha_{i1}(z)a_{zs}b_s(y_{i2}),\end{aligned}$$

Then, if we generalize for $t = 1, 2, \ldots, T-1$

$$\begin{aligned}\alpha_{i,t+1}(s) &= P(y_{i1}, y_{i2}, \ldots, y_{i,t+1}, u_{i,t+1} = s|\Theta) \\ &= \left[\sum_{z=1}^{S} \alpha_{it}(z)a_{zs}\right] b_s(y_{i,t+1}),\end{aligned}$$

where $s = 1, 2, \ldots, S$
where a_{zs} are transition probabilities from state z to state s.
The likelihood function can then be computed in the last step as

$$\mathscr{L}(\Theta; y) = \sum_{i=1}^{n}\sum_{z=1}^{S} \alpha_{i,T}(z).$$

The backward probabilities $\beta_{it}(s)$ are defined as

$$\beta_{it}(s) = P(y_{i,t+1}, y_{i,t+2}, \ldots, y_{iT}|u_{it} = s, \Theta),$$

the probability of the observed sequence from $t+1$ to T given the hidden state at time t and the model parameter Θ.

Backward recursion is initialized starting with $\beta_{iT}(s) = 1$ for $i = 1, 2, \ldots, n$, and $s = 1, 2, \ldots, S$.

Then, for $t = T-1, T-2, \ldots, 1$

$$\beta_{it}(s) = \sum_{z=1}^{S} \left[a_{sz} b_z(y_{i,t+1}) \beta_{i,t+1}(z) \right],$$

The likelihood function can also be computed by using both the forward and backward probabilities as

$$\mathscr{L}(\Theta; y) = \sum_{i=1}^{n} \sum_{z=1}^{S} \alpha_{i,t}(z) \beta_{i,t}(z).$$

for any $t = 1, 2, \ldots, T$.

Furthermore, forward and backward probabilities are used to compute the posterior state probabilities as

$$L_{it}(s) = P(u_{it} = s | y_i, \Theta) = \frac{\alpha_{i,t}(s) \beta_{i,t}(s)}{\sum_{z=1}^{S} \alpha_{i,t}(z) \beta_{i,t}(z)}, \tag{3.15}$$

Forward and backward probabilities are often scaled to avoid numerical instability. Thus, forward probabilities become $\hat{\alpha}_{it}(s) = N_{it} \alpha_{it}(s)$ where $N_{it} = 1 / \sum_{z=1}^{S} \alpha_{it}(z)$.

3.4 Discussion

This chapter presented mixtures of Markov models and how they are used to identify clusters of sequences based on the assumption that the sequences in a cluster follow the same Markov process. We then moved on to hidden Markov models characterized by the presence of latent Markov processes in which states emit the observations, i.e. the observed states on the basis of emission probabilities. The forward-backward algorithm combined with the EM were presented as a parameter estimation method. However, these models and estimation techniques assume that the number of mixture components and the number of hidden states is known *a priori*. In reality, these numbers are unknown, so we need model selection techniques to identify them. The next chapter illustrates model

selection criteria for MHMMs and points out their limitations. Then we present a proposal of a classification criterion based on a joint entropy measure defined to account for both component and hidden states. The performance of this criterion is compared with the most widely used information criteria.

Chapter 4

Model selection for mixture hidden Markov models

As argued in the previous chapter, MHMM is a combination of mixture and hidden Markov models; hence, estimating parameters is constrained to identify the number of hidden states and subpopulations (i.e. the components) of the models. MMs and HMMs usually identify the number of components and hidden states, respectively, assuming prior information. If there is no such prior knowledge, identifying the components and hidden states becomes more difficult.

In this chapter, after providing a brief overview of the main studies on model selection for MMs and HMMs, in Section 1 we point out the main gaps in the literature on model selection for MHMMs. In Section 2, an entropy-based model selection criterion defined as an ICL_BIC with a joint entropy measure is proposed. Finally, Section 3 presents a Monte Carlo simulation study to assess the performance of the criterion.

4.1 Profile identification: Components and states

One of the main issues in applying MHMMs actually concerns selecting the unknown number of mixture components and states in grouping sequences. Generally, the selection of the number of components and states is based on *a priori* information about the problem under analysis. In such cases, researchers can perform the model selection procedure under the assumption that the number of components or the number of states is fixed. However, a priori information is not always known.

Classic Information Criteria (IC) have been widely employed for mixture models or Markov models, and their use has been extended to MHMMs even if the limits of its application have received little attention in the literature. Commonly, the BIC is used under the assumption that either the number of hidden states or the number of clusters is known. Being simple to use, BIC (Bayesian Information Criterion) is also chosen in cases where both these numbers are unknown, but the performance of this criterion is not satisfactory and depends on the length of the sequence and the number of clusters. In the light of these considerations, selection measures taking into account the latent levels of MHMMs should be defined. Thus, the aim of this chapter is to enrich this stream of literature and to propose a new selection measure for MHMMs based on an integrated classification likelihood approach that outperforms traditional criteria in identifying components and states simultaneously.

This criterion enables us to group similar sequences (i.e., web sessions) in order and time, under the assumption that they have been generated by a hidden Markov model (HMM) having a specific number of states. Therefore, we will assume that sequences related to different HMMs (i.e. the mixture components) define well-separated clusters of sequences with a high level of dissimilarities among them.

As regards the model used, we refer to time-constant covariates, single-channel and single response MHMMs (see Vermunt et al., 2008) because they are particularly appropriate in managing clickstream data, and we consider a single web-page sequence for each user with the only available covariates being features related to IP addresses. In the next subsections we present model selection procedures. We start with mixture models identifying the number of components, then move on to HMMs and the identification of the number of hidden states. In the final subsection, we summarize what we presented for MMs and HMMs and focus on MHMMs, where both components and hidden states must be identified, and we emphasize how identifying component and state numbers is a difficult task that requires a suitable criterion.

4.1.1 Selecting the number of components in mixture models

As mentioned in the previous chapter, mixture models are a class of latent variable modelling which allows us to deal with situations where some of the variables are unobserved. Depending on the domain of the random variables, MMs are called latent class (discrete

random variables) or latent profile models (continuous random variables). The main concern is selecting the number of components or subpopulations (McLachlan & Basford, 1988). Model selection for MMs is commonly performed by using information and classification criteria. The former tends to estimate the best approximating mixture model in terms of log-likelihood, whereas the latter should be the best way of determining the number of clusters (Biernacki & Govaert, 1997; McLachlan & Peel, 2004).

As far as the information criterion approach is concerned, Dias (2006) compared its performance in identifying the number of classes in a latent class model, using a modified AIC (Akaike Information Criterion) [1] which performed better than other IC but had a tendency to overfit. Nylund et al. (2007) examined the performance of information criteria in identifying the number of components in different MMs, varying the sample size. Their study highlighted that, in both latent class and profile models, BIC performs better than IC indices such as AIC , CAIC (the consistent AIC, (Bozdogan, 1987)) and ssBIC (sample size adjusted BIC, (Rissanen, 1978)).

Fonseca (2008) showed that BIC behaves similarly to AIC3, and AICu (i.e. a bias correction of AIC proposed by McQuarrie et al. (1997)) and performs very well with several sample sizes. Lukociene and Vermunt (2009) argued that AIC worked better with small sample sizes and that CAIC performs almost as well as BIC. Biernacki et al. (2000) argued that the goal of clustering is not the same as that of estimating the best approximating mixture model. Therefore, IC-based measures may not be the best way of determining the number of clusters, even though they can perform well in selecting the number of components. The authors pointed out that there is generally no actual correspondence between the number of components of a mixture and the number of clusters and also that more than one mixture component could represent a cluster, e.g. with a mixture of normal distributed data two components could have close means and standard deviations and identify a singular cluster.

In the light of that, classification criteria, which consider the quality of the data partition in selecting the optimal number of groups, have been introduced to perform model-based clustering. These classification criteria measure the degree of class separation by entropy-based approaches (Ramaswamy et al., 1993). The most widely used classification methods for mixture models are the classification likelihood criteria (CLC) (Biernacki &

[1] Specifically, the author used the AIC3, defined by Bozdogan (1994) as an AIC with 3 as the penalising factor.

Govaert, 1997), and an approximation of the integrated completed likelihood based on BIC, the ICL_BIC (Biernacki et al., 2000).

Baudry et al. (2010) provided a further approach to identify clusters from the components of a mixture model. In their two-step procedure, they initially identify the components of a Gaussian mixture using BIC, but they then used a classification criterion (the entropy criterion to be precise) to combine pairs of components hierarchically by comparing the degree of separation before and after the aggregation. Summing up, the choice of a suitable selection approach depends on whether the aim is to identify mixture components or clusters and on the sample size or on other model features. However, the literature shows that the BIC index is generally to be preferred to other information criteria since it performs better and produces similar results to the AIC index in small samples. Nevertheless, the CC is to be preferred if the aim is to recover the number of clusters rather than the number of mixture components.

4.1.2 Selecting the number of states in hidden Markov models

In the previous subsection, we presented model selection in mixture models. MMs are defined by assuming a time-constant latent variable in which classes represent the mixture components, so the target of model selection is to identify the number of components. Now we will focus on latent class models, those with a latent variable that depends over time, that is the hidden Markov models. In HMM framework, model selection plays a prominent role since it corresponds to the choice of the number of latent states of the unobserved Markov chain underlying the observed data. The number of states should be chosen to enable the model to account for the dynamic pattern and covariance structure of the observed time series (Costa & Angelis, 2010).

Identifying a suitable number of hidden states is usually carried out through IC measures for HMMs, although none could be considered preferable. Costa and Angelis (2010) applied information criteria to HMMs showing that AIC and AICc (a bias correction of AIC proposed by Hurvich and Tsai (1989)) outperform BIC in identifying the correct number of hidden states for shorter time series. Moreover, they show that BIC and CAIC seriously underestimate the number of hidden states when the number of observations is small. By performing a simulation study, Celeux and Durand (2008) showed that AIC, BIC and the Integrated Classification Likelihood (ICL) behave similarly to MMs

(McLachlan & Peel, 2004) and that ICL performs better than BIC when it comes to identifying the number of hidden states. They also proposed a new criterion based on cross-validated likelihood, which, although accurate, tends to overestimate the number of states and is more time consuming than traditional criteria. Furthermore, Boucheron and Gassiat (2007) proved the consistency of BIC in retrieving the number of hidden states in HMMs, and for this reason it is often used to identify the number of states when the sequence length is large (Costa & Angelis, 2010; Paas et al., 2007).

As far as classification criteria are concerned, their application in HMMs has received little attention. An interesting attempt is the study of Bacci et al. (2014), which proposed a modified Normalized Entropy Criterion (NEC; Biernacki and Govaert (1999)) to HMMs defining a simplified entropy measure, obtained as an approximation of the measure proposed by Hernando et al. (2005). Their approximation is based on the assumption that the hidden variables are independent given the observed ones, and thus they obtained a normalized entropy measure considering marginal entropy profiles for each state. However, their results indicate that NEC is outperformed by other classification or information criteria. More recently, Pohle et al. (2017) suggested a stepwise procedure to select the correct number of states. However, since their empirical criterion is based on *a priori* information from the theoretical framework underlying the analysed problem, their proposal cannot be used in cases where there is no such information about the hidden states. In a nutshell, the literature shows that classification criteria and AIC are to be preferred to BIC for multiple short sequences, while BIC is the preferable criterion for long series.

4.1.3 Model selection in mixture hidden Markov models

Mixture hidden Markov models combine mixture and hidden Markov models by assuming several HMMs generated by different subpopulations related to a number of covariates (Bartolucci & Pandolfi, 2015; Helske & Helske, 2017; Van de Pol & Langeheine, 1990; Vermunt et al., 2008). MHMMs have been used to model longitudinal sequences relaxing single population assumptions of classic HMMs (Vermunt et al., 2008). We point out that unlike mixture models, where there is no correspondence between clusters and components, in the context of MHMMs we assume that each component of the mixture identifies a cluster of sequences generated by the same HMM.

Model selection for MHMMs involves finding both the number of clusters and the

number of latent states. To the best of our knowledge, model selection criteria related to these MHMMs are only based on information criteria (Du et al., 2011; Marino & 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 analysing short-length categorical sequences, but the information criterion "*kept suggesting models with more and more states*" (Helske et al., 2018, p.192). Dias et al. (2009) applied an MHMM to analyse 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). Crayen et al. (2012) analysed categorical sequences and noted that using BIC or AIC3 leads to similar results as regards the number of clusters and states. Unlike MMs and HMMs, as far as we know, no attempt has been made in the literature to apply classification criteria based on entropy in the context of MHMMs, although they seem particularly suitable given the structure of these models characterized by latent classes on two levels. Thus, it would be very useful to define a model selection procedure that focuses on the identification of the number of states and clusters taking into account two levels of entropy, the first related to hidden states classifying sequence elements; and the second related to clusters classifying the sequences.

An interesting proposal along these lines is the study presented by Volant et al. (2014). The authors referred to an HMM presenting mixture components in the emission probabilities of the chain, which are added to a classic HMM for a continuous-valued sequence. They identified groups of observations by combining the components in the emission probabilities and, to determine the number of groups, and proposed an ICL_BIC that involves only a measure of entropy related to the hidden states.[2]

Following Volant et al. (2014), this thesis aims to enrich the literature on model selection for MHMMs by proposing an approach that uses an integrated classification likelihood obtained by defining a joint entropy measure taking into account the latent classes of clusters and hidden states. Unlike the authors, however, the model used in this thesis presents mixtures of HMMs each having a different number of latent states and we define an ICL_BIC to identify both clusters and states (not only the number of states).

[2]Their proposal is based on 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

Previous subsections are schematized in Table 4.1. Summing up, the literature suggests that CC are the best performers when the aim is to identify clusters in MMs or hidden states in HMMs for short time series, while IC such as BIC outperform other criteria in identifying hidden states in HMMs for long time series. The use of ICL_BIC in the context of MHMMs does seem to be the best choice as the aim is to identify both hidden states and clusters of sequences and this criterion, being a CC computed as an entropy-penalized BIC, takes into account both tasks. In the next section the hidden Markov models will be presented in order to introduce the issue of model selection in the context of MHMMs.

Table 4.1: Model selection criteria for MMs, HMMs and MHMMs; key papers and best criteria

	key papers	criteria
MM	Dias (2006)	AIC3
	Nylund et al. (2007)	BIC
	Fonseca (2008)	BIC, AIC3, AICu
	Lukociene and Vermunt (2009)	BIC and CAIC (AIC for small sample size)
	Biernacki et al. (2000)	ICL
	Baudry et al. (2010)	Entropy
HMM	Costa and Angelis (2010)	AIC and AICc (short sequence)
	Celeux and Durand (2008)	ICL
	Boucheron and Gassiat (2007)	BIC for long sequence
	Bacci et al. (2014)	modified NEC
MHMM	Helske et al. (2018)	BIC
	Dias et al. (2009)	BIC (states known)
	Crayen et al. (2012)	BIC and AIC3
	Volant et al. (2014)	ICL_BIC

4.1.4 Model selection: Information criteria for MHMMs

As highlighted in the previous sections, MHMMs can detect underlying latent structures and enable clustering sequences to be arranged into homogeneous subpopulations. The model selection process in MHMMs can be performed by fitting models to the data varying the number of subgroups and states and comparing them using information criteria. The most widely used information criteria for MHMMs are

$$\begin{aligned}
\text{AIC} &= -2\hat{\ell} + 2\text{df}, & \text{AIC}_3 &= -2\hat{\ell} + 3\text{df}, \\
\text{CAIC} &= -2\hat{\ell} + \big(1 + \log N\big)\text{df}, & \text{BIC} &= -2\hat{\ell} + \big(\log N\big)\text{df}, \\
\text{ssBIC} &= -2\hat{\ell} + \log\left(\frac{2+N}{24}\right)\text{df}, & & \qquad (4.1)
\end{aligned}$$

where $\hat{\ell}$ denotes the maximum of the log-likelihood related to the MHMM with Q categorical time-constant covariates in equation (3.3), df denotes the number of free parameters computed as

$$\text{df} = \sum_{k=1}^{K} [S^k(R-1) + S^k - 1 + S^k(S^k-1)] + (K-1)\prod_{q=1}^{Q} x_q, \qquad (4.2)$$

where K is the number of components, S^k is the number of hidden states in component k, R is the number of observed states and x_q is the number of values of the categorical covariates. $N = \sum_{i=1}^{n}\sum_{t=1}^{T_i} I(y_{it})$, where I is the indicator function equal to 1 if y_{it} is observed. This summation is equal to $n \times T$ if all sequences have the same length.

4.2 Entropy-based model selection

As highlighted in the previous chapter, MHMMs can detect underlying latent structures and enable clustering sequences to be arranged into homogeneous subpopulations whose evolution follows the same hidden Markov process having specific parameters and number of states. Therefore, a model selection criterion based on a measure of separability (i.e. an entropy measure) is to be preferred in order to identify the number of components and hidden states simultaneously.

To use an entropy-based criterion enables us to identify latent states so that the distributions of the observations (i.e. the elements in the sequences) conditional to the latent states, will have a high degree of separability and consequently each latent state will be identified in groups of observations seizing the unknown similarities among the observations. In the next subsection, the proposed entropy measure (i.e. ICL_BIC) for MHMMs is presented.

4.2.1 Proposed model selection criterion: ICL_BIC for MHMMs

As mentioned above, from a model-based clustering perspective, the separability among possible clusters should be taken into account when selecting the correct model. In this regard, Biernacki et al. (2000) proposed an Integrate Complete Likelihood (ICL), approximated by means of a Bayesian information criterion, to assess 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 complete likelihood (ICL) instead of the observed one so as to account for the degree of separation between latent classes. More precisely, the ICL criterion can be approximated as an entropy-penalized BIC[3] obtaining the so-called ICL_BIC (McLachlan & Peel, 2000).

To define the ICL_BIC for MHMMs, we have to consider two latent variables, the time-varying variable U of the latent Markov process; and the time-constant latent variable C related to components.[4]

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

$$P(m|y) = \frac{P(y|m)P(m)}{P(y)}$$

where y are the data and $P(m)$ the model prior distribution. Under the assumption of non-informative prior for m, we can identify the model by directly maximizing the integrated likelihood $P(y|m)$.

As pointed out by Biernacki, when the target is to identify an unknown number of latent classes, we should consider the Integrated Complete Likelihood (ICL) instead, related to complete data that in the MHMM context are (y, U, C) .

The ICL for MHMMs is defined as

$$P(y, C, U|m) = \int P(y, C, U|m, \Theta)\pi(\Theta|m)d\Theta. \tag{4.3}$$

[3]This derivation was obtained approximating the complete log-likelihood as the sum of two elements, the observed log-likelihood and the data entropy; see Biernacki et al. (2000) for its derivation.

[4]A 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.

The quantity $2\log P(y,C,U|m)$ can be approximated as

$$\log P(y,\hat{C},\hat{U}|m,\hat{\Theta}) - (\log N)\frac{\mathrm{df}}{2},$$

where $\hat{C}$ and $\hat{U}$ are posterior modes. Then, $\hat{C}$ and $\hat{U}$ are replaced by considering the conditional expectation given the observation y of the previous approximated quantity (see McLachlan and Peel (2000)), leading to

$$\begin{aligned}
\mathrm{ICL} &= \mathbb{E}_Y[\log P(y,C,U|m,\hat{\Theta})] - (\log N)\frac{\mathrm{df}}{2} \\
&= \mathbb{E}_Y[\log\{P(U|C,y,m,\hat{\Theta})P(C|y,m,\hat{\Theta})P(y|m,\hat{\Theta})\}] - (\log N)\frac{\mathrm{df}}{2} \\
&= \mathbb{E}_Y[logP(y|m,\hat{\Theta})] + \mathbb{E}_Y[\log P(C|y,m,\hat{\Theta})] + \mathbb{E}_Y[\log P(U|C,y,m,\hat{\Theta})] - (\log N)\frac{\mathrm{df}}{2} \\
&= \log P(y|m,\hat{\Theta}) - \mathrm{H}(C|y) - \mathrm{H}(U|C,y) - (\log N)\frac{\mathrm{df}}{2}.
\end{aligned}$$

Here, $\mathrm{H}(C|y)$ is the entropy measure related to the number of components, given by

$$\mathrm{H}(C|y) = -\sum_{i=1}^{n}\sum_{k=1}^{K} P(C_i = k|y_i)\log P(C_i = k|y_i), \tag{4.4}$$

where $P(C_i = k|y_i)$ are component posterior probabilities, i.e. the probability that given the i-th observed sequence, the latter was generated by the k-th model. Furthermore, $\mathrm{H}(U|C,y)$ is the entropy related to the sequence of hidden states given by

$$\mathrm{H}(U|C,y) = -\sum_{i=1}^{n}\sum_{k=1}^{K} P(C_i = k|y_i)\left[\mathrm{H}(U_{i1}|y_i,C_i = k) + \sum_{t=2}^{T_i}\mathrm{H}(U_{it}|U_{i,t-1},y_i,C_i = k)\right], \tag{4.5}$$

where the $t = 1$ elements $\mathrm{H}(U_{i1}|y_i,C_i = k)$ and the $\mathrm{H}(U_{it}|U_{i,t-1},y_i,C_i = k)$, i.e. the conditional entropy profiles, are computed as proposed by (Durand & Guédon, 2016) based on the Hernando et al. (2005) conditional entropy definition for HMM. Specifically,

$$\begin{aligned}
&\mathrm{H}(U_{it}|U_{i,t-1},y_i,C_i = k) = \\
&\qquad = \sum_{s,z\in S_k} P(U_{it} = s, U_{i,t-1} = z|y_i,C_i = k)\log P(U_{it} = s|U_{i,t-1} = z,y_i,C_i = k),
\end{aligned}$$

where

$$P(U_{it}=s,U_{i,t-1}=z|y_i,C_i=k)=L_{it}^k(s)a_{zs}^k\hat{\alpha}_{i,t-1}^k(z)/G_{it}^k(s),$$

and

$$P(U_{it}=s|U_{i,t-1}=z,y_i,C_i=k)=L_{it}^k(s)a_{zs}^k\hat{\alpha}_{i,t-1}^k(z)/\{G_{it}^k(s)L_{i,t-1}^k(z)\}.$$

Here, a_{zs}^k are transition probabilities from state z to state s in cluster k, $\alpha_{it}^k(s)$ are the forward probabilities, the posterior state probabilities $L_{it}^k(s)=P(U_{it}=s|y_i,C_i=k)$ being obtained in the forward-backward algorithm. The $G_{it}^k(s)$ are called predicted probabilities and are computed from the forward probabilities as

$$G_{i,t+1}^k(s)=P(U_{i,t+1}=s|y_{i1},y_{i2},\ldots,y_{it},C_i=k)=\sum_{z=1}^{S_k}a_{zs}^k\hat{\alpha}_{it}^k(z).$$

Finally, the ICL_BIC for MHMMs is obtained by multiplying ICL by a constant of minus two

$$\text{ICL\_BIC}=-2\hat{\ell}+2\text{H}(C,U|y)+(\log N)\text{df}, \tag{4.6}$$

where $\hat{\ell}$ denotes the maximum of the log-likelihood (see equation (3.3)), df is the number of free parameters in equation (4.2), the joint entropy is $\text{H}(C,U|y)=\text{H}(C|y)+\text{H}(U|C,y)$, $N=\sum_{i=1}^n\sum_{t=1}^{T_i}I(y_{it})$, where I is the indicator function equal to 1 if y_{it} is observed. This summation is equal to $n\times T$ if all sequences have the same length.

4.3 Simulation Study

To assess the quality of this proposed model selection criterion (ICL_BIC in equation (4.6)), a Monte Carlo study, with different model features, has been performed. First, we simulate longitudinal datasets from six different MHMMs with a known number of components K and latent states $(S^1,S^2,\ldots,S^K)$. We generate the datasets by considering the sample size $n\in\{300,500\}$ (n refers to the number of sequences), creating sequences of length $T\in\{10,20\}$ the elements of which can take one of four observed states ($card|Y|=4$). Table 4.2 reports the number of components K and hidden states

$(S^1, S^2, \ldots, S^K)$ for each MHMM O_h with $h \in \{2,3,4,5,6,7\}$, used to generate the sequences (h refers to the number of components).

Table 4.2: MHMMs component and hidden state numbers

Model	K	$(S^1, S^2, \ldots, S^K)$
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)

Varying the length of the sequences $T$and the number of components and states, we investigate 14 different scenarios to assess the effect of different design features. We generate 70 datasets for each scenario and evaluate which information and classification criteria are able to identify the correct numbers of states and components.

Parameters $\Theta^k = \{\pi^k, A^k, B^k\}$ are used to generate the sequences, varying the number of components and hidden states.

1. The initial probabilities of hidden states for each component:

$$\pi^1 = (0.30, 0.70); \qquad \pi^2 = (0.40, 0.20, 0.40); \qquad \pi^3 = (0.30, 0.30, 0.20, 0.20);$$
$$\pi^4 = (0.40, 0.60); \qquad \pi^5 = (0.20, 0.30, 0.15, 0.35); \qquad \pi^6 = (0.10, 0.40, 0.50)$$
$$\pi^7 = (0.40, 0.10, 0.20, 0.30).$$

2. Hidden states transition matrices for each component:

$$A^1 = \begin{bmatrix} 0.80 & 0.20 \\ 0.20 & 0.80 \end{bmatrix}; \qquad A^2 = \begin{bmatrix} 0.10 & 0.80 & 0.10 \\ 0.70 & 0.10 & 0.20 \\ 0.30 & 0.10 & 0.60 \end{bmatrix}; \qquad A^3 = \begin{bmatrix} 0.50 & 0.20 & 0.10 & 0.20 \\ 0.00 & 0.60 & 0.20 & 0.20 \\ 0.10 & 0.30 & 0.00 & 0.60 \\ 0.20 & 0.00 & 0.20 & 0.60 \end{bmatrix}$$

$$A^4 = \begin{bmatrix} 0.20 & 0.80 \\ 0.80 & 0.20 \end{bmatrix}; \qquad A^5 = \begin{bmatrix} 0.10 & 0.20 & 0.10 & 0.60 \\ 0.20 & 0.10 & 0.50 & 0.20 \\ 0.10 & 0.10 & 0.60 & 0.20 \\ 0.70 & 0.00 & 0.10 & 0.20 \end{bmatrix} \qquad A^6 = \begin{bmatrix} 0.35 & 0.50 & 0.15 \\ 0.40 & 0.30 & 0.30 \\ 0.50 & 0.20 & 0.30 \end{bmatrix}$$

$$A^7 = \begin{bmatrix} 0.30 & 0.35 & 0.05 & 0.30 \\ 0.30 & 0.20 & 0.00 & 0.50 \\ 0.00 & 0.30 & 0.60 & 0.10 \\ 0.10 & 0.50 & 0.30 & 0.10 \end{bmatrix}.$$

3. Observed States emission matrices for each component:

$$B^1 = \begin{bmatrix} 0.10 & 0.20 & 0.10 & 0.60 \\ 0.50 & 0.05 & 0.20 & 0.25 \end{bmatrix}; \qquad B^2 = \begin{bmatrix} 0.10 & 0.10 & 0.60 & 0.20 \\ 0.20 & 0.50 & 0.20 & 0.10 \\ 0.15 & 0.15 & 0.10 & 0.60 \end{bmatrix}$$

$$B^3 = \begin{bmatrix} 0.00 & 0.10 & 0.30 & 0.60 \\ 0.20 & 0.60 & 0.20 & 0.00 \\ 0.55 & 0.00 & 0.10 & 0.35 \\ 0.25 & 0.00 & 0.60 & 0.15 \end{bmatrix}; \qquad B^4 = \begin{bmatrix} 0.10 & 0.60 & 0.10 & 0.20 \\ 0.20 & 0.10 & 0.45 & 0.25 \end{bmatrix}$$

$$B^5 = \begin{bmatrix} 0.20 & 0.65 & 0.15 & 0.00 \\ 0.50 & 0.00 & 0.25 & 0.25 \\ 0.05 & 0.00 & 0.35 & 0.60 \\ 0.15 & 0.00 & 0.70 & 0.15 \end{bmatrix}; \qquad B^6 = \begin{bmatrix} 0.30 & 0.40 & 0.10 & 0.20 \\ 0.00 & 0.40 & 0.30 & 0.30 \\ 0.35 & 0.00 & 0.25 & 0.40 \end{bmatrix}$$

$$B^7 = \begin{bmatrix} 0.00 & 0.35 & 0.35 & 0.30 \\ 0.10 & 0.30 & 0.60 & 0.00 \\ 0.30 & 0.00 & 0.30 & 0.40 \\ 0.20 & 0.40 & 0.10 & 0.30 \end{bmatrix}$$

Results presented are success rates and are defined as the identification of correct numbers of components and states or an approximation (defined by using Manhattan distance, see equation (4.7)). Once the sequences are generated for a specific scenario, we fitted

to this data a model having the correct number of components and states and 29 other models having these numbers randomly generated:

- The number of components K from a discrete uniform distribution taking values between 1 (no mixture) and 7 i.e. $K \sim \mathscr{U}(1,7)$
- The number of hidden states S^k for each components from a discrete uniform distribution taking values between 2 and 6 i.e. $S^k \sim \mathscr{U}(2,6)$

This restriction is due to the number of models that should be considered, e.g. if $K \in \{1,2,\ldots,7\}$ and $S^k \in \{2,3,\ldots,6\}$ for each component k, the number of models can be computed as a sum of K combinations with repetition:

$$\sum_{k=1}^{K} \frac{(5+k-1)!}{k!(n-1)!},$$

where 5 represents the S^k alternatives. Thus, there are 791 combinations of clusters and states for our setting.

Simulation study was carried out by using R package "seqHMM" (Helske & Helske, 2017). To assess criteria performance, we consider the success rate (correct model identification) of BIC, AIC, ssBIC and the ICL_BIC for MHMM. The simulations demonstrate how difficult it is to identify the correct number of components and hidden states in MHMMs, with all indices at zero or meagre success rates. So, we investigate whether the criteria are able to identify the correct number of components K and numbers of hidden states equal or *close enough* to the correct ones, where *close enough* is defined as the minimal Manhattan distance between the correct vector of states and any permutation of the selected vector of states is less than 3. Thus, let $\mathrm{s}_c = (S_c^1, S_c^2, \ldots, S_c^{K_c})$ be the correct hidden states and $\mathrm{s}_{sl} = (S_{sl}^1, S_{sl}^2, \ldots, S_{sl}^{K_c})$ the selected ones, the selection is a success if

$$\min_{\eta[\mathrm{s}_{sl}(k)]} \quad \sum_{k=1}^{K_c} |\mathrm{s}_c(k) - \mathrm{s}_{sl}(k)| < 3, \tag{4.7}$$

where $\eta[\mathrm{s}_{sl}(k)]$ is a permutation of the selected numbers of states.

Furthermore, due to the number of possible models that could be investigated, we restrict the simulation study to a scenario in which a certain amount of prior information about the problem under analysis is available. Specifically, we compared criteria performance results related to ten models with a number of components close to the correct

one. ICL_BIC performance is compared with the most widely used information criteria for MHMMs such as AIC, BIC and ssBIC.

The results are displayed in Tables 4.3 and 4.5; the best performances are in bold. In Tables 4.4 and 4.6, we highlight criteria proportions related to overestimation and underestimation of the number of components.

In Tables 4.3 and 4.4 we fixed the number of sequences at $n = 300$. Focusing on Table 4.3, when $T = 10$ the proposed ICL_BIC outperforms the classic information criteria which struggle to identify even an approximation of the model. The IC seem to systematically underestimate the number of components and select models with no mixtures if the option is available. When the number of components is high ($K \in \{6,7\}$) all criteria struggle. Overall, the ICL_BIC performs best. As the sequence length is increased to $T = 20$, we note that the previous results are confirmed and ICL_BIC have a higher success rate. Comparing AIC and ICL_BIC, the two criteria have similar results when $T = 20$ even though the latter seems to outperform the former. Increasing the number of sequences to $n = 500$ does not seem to change the results, confirming the limitations of classic IC and ICL_BIC performance.

Table 4.3: Results of the Monte Carlo study for $n = 300$, $T \in \{10,20\}$. **Success rate** of identifying the correct model or an approximation of the model; best results are in bold; standard errors are shown in parentheses

			AIC	BIC	ssBIC	ICL_BIC
$n = 300$	$T = 10$	O_2	0.51 (0.060)	0.19 (0.047)	0.22 (0.049)	**0.62** (0.058)
		O_3	0.46 (0.059)	0.31 (0.055)	0.36 (0.057)	**0.55** (0.059)
		O_4	0.58 (0.060)	0.41 (0.059)	0.46 (0.059)	**0.73** (0.053)
		O_5	0.38 (0.058)	0.34 (0.057)	0.36 (0.057)	**0.60** (0.058)
		O_6	0.20 (0.048)	0.14 (0.041)	0.16 (0.044)	**0.48** (0.060)
		O_7	0.12 (0.039)	0.08 (0.032)	0.08 (0.032)	**0.40** (0.058)
	$T = 20$	O_2	0.60 (0.058)	0.24 (0.051)	0.29 (0.054)	**0.69** (0.055)
		O_3	0.49 (0.060)	0.32 (0.056)	0.34 (0.057)	**0.53** (0.060)
		O_4	0.59 (0.058)	0.42 (0.058)	0.46 (0.059)	**0.73** (0.052)
		O_5	0.32 (0.056)	0.28 (0.054)	0.28 (0.054)	**0.44** (0.059)
		O_6	0.35 (0.057)	0.10 (0.037)	0.15 (0.043)	**0.41** (0.059)
		O_7	0.24 (0.051)	0.20 (0.048)	0.20 (0.048)	**0.58** (0.059)

Table 4.4: Results of the Monte Carlo study for $n = 300$, $T \in \{10, 20\}$. Rates related to an underestimation (U) or an overestimation (O) of the number of components

				AIC	BIC	ssBIC	ICL_BIC
$n = 300$	$T = 10$	O_2	U	0.49	0.81	0.78	0.00
			O	0.00	0.00	0.00	0.38
		O_3	U	0.44	0.66	0.61	0.21
			O	0.10	0.03	0.03	0.24
		O_4	U	0.42	0.59	0.54	0.25
			O	0.00	0.00	0.00	0.02
		O_5	U	0.62	0.66	0.64	0.38
			O	0.00	0.00	0.00	0.02
		O_6	U	0.80	0.86	0.84	0.46
			O	0.00	0.00	0.00	0.06
		O_7	U	0.88	0.92	0.92	0.55
			O	0.00	0.00	0.00	0.05
	$T = 20$	O_2	U	0.36	0.76	0.71	0.00
			O	0.04	0.00	0.00	0.31
		O_3	U	0.48	0.76	0.74	0.42
			O	0.03	0.02	0.02	0.05
		O_4	U	0.41	0.58	0.54	0.25
			O	0.00	0.00	0.00	0.2
		O_5	U	0.66	0.72	0.72	0.40
			O	0.02	0.00	0.00	0.16
		O_6	U	0.65	0.90	0.85	0.50
			O	0.00	0.00	0.00	0.09
		O_7	U	0.76	0.80	0.80	0.32
			O	0.00	0.00	0.00	0.10

Table 4.5: Results of the Monte Carlo study for $n = 500$, $T = 10$. **Success rate** of identifying the correct model or an approximation of the model; best results are in bold; standard errors are shown in parentheses

			AIC	BIC	ssBIC	ICL_BIC
$n = 500$	$T = 10$	O_2	0.47 (0.059)	0.17 (0.042)	0.19 (0.046)	**0.67** (0.053)
		O_3	0.68 (0.056)	0.64 (0.057)	0.62 (0.058)	**0.70** (0.055)
		O_4	0.48 (0.060)	0.35 (0.057)	0.35 (0.057)	**0.61** (0.058)
		O_5	0.49 (0.060)	0.35 (0.057)	0.35 (0.057)	**0.65** (0.057)
		O_6	0.26 (0.052)	0.10 (0.036)	0.12 (0.039)	**0.62** (0.058)
		O_7	0.20 (0.048)	0.14 (0.041)	0.14 (0.041)	**0.46** (0.059)
	$T = 20$	O_2	0.68 (0.055)	0.54 (0.059)	0.54 (0.059)	**0.75** (0.051)
		O_3	0.64 (0.057)	0.60 (0.058)	0.66 (0.057)	**0.72** (0.054)
		O_4	0.50 (0.060)	0.38 (0.058)	0.34 (0.057)	**0.66** (0.057)
		O_5	**0.62** (0.058)	0.46 (0.059)	0.48 (0.060)	**0.62** (0.058)
		O_6	0.42 (0.059)	0.06 (0.028)	0.06 (0.028)	**0.48** (0.060)
		O_7	0.32 (0.056)	0.20 (0.048)	0.20 (0.048)	**0.56** (0.059)

Table 4.6: Results of the Monte Carlo study for $n = 500$, $T = 10$. Rates related to an underestimation (U) or an overestimation (O) of the number of components

				AIC	BIC	ssBIC	ICL_BIC
$n = 500$	$T = 10$	O_2	U	0.46	0.83	0.81	0.00
			O	0.07	0.00	0.00	0.33
		O_3	U	0.24	0.36	0.38	0.10
			O	0.08	0.00	0.00	0.20
		O_4	U	0.52	0.65	0.65	0.23
			O	0.00	0.00	0.00	0.16
		O_5	U	0.51	0.65	0.65	0.21
			O	0.00	0.00	0.00	0.14
		O_6	U	0.74	0.90	0.88	0.32
			O	0.00	0.00	0.00	0.06
		O_7	U	0.80	0.86	0.86	0.46
			O	0.00	0.00	0.00	0.08
	$T = 20$	O_2	U	0.32	0.46	0.46	0.00
			O	0.00	0.00	0.00	0.25
		O_3	U	0.16	0.36	0.28	0.08
			O	0.20	0.04	0.06	0.20
		O_4	U	0.50	0.62	0.66	0.12
			O	0.00	0.00	0.00	0.22
		O_5	U	0.38	0.54	0.52	0.28
			O	0.00	0.00	0.00	0.10
		O_6	U	0.58	0.94	0.94	0.40
			O	0.00	0.00	0.00	0.12
		O_7	U	0.68	0.80	0.80	0.35
			O	0.00	0.00	0.00	0.09

4.4 Discussion

Following the literature on model selection in mixture models and hidden Markov models, which focuses on entropy-based measures, we propose a score based on classification criteria, i.e. an ICL_BIC. The most suitable model is selected, from a set of candidates, by minimizing the ICL_BIC. By employing the proposed criterion, we 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 ICL_BIC by varying sample size and sequence length. Simulations demonstrate 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 ICL_BIC seems to outperform the other criteria. However, the ICL_BIC requirement of computing entropy is one of its drawbacks when compared to using the BIC for MHMM. In particular, the conditional entropy associated with the HMMs is computationally intensive and requires more processing as the sequence length increases.

However, as we will point out in the next chapter, since clickstream data has, on average, a short sequence length ICL_BIC would be preferable, performing better than IC and not requiring too much time for entropy computation. We considered, in our simulation study, scenarios related to different numbers of components and states. A broader range of scenarios should be considered in future studies, such as varying model state-dependent conditional probabilities $b_{u_t}^k(y_{it})$. These probabilities represent the uncertainty in hidden state classifications of observations. A situation of low uncertainty, for example, would be when a hidden state has a probability of emitting an observed state greater than 0.9, while the sum of the probabilities of emitting the other states is less than 0.1. On the other hand, if a latent state emits the observed states with probabilities that are close to each other, we would be in a high uncertainty scenario. The effect of different levels of uncertainty on the performance of the selection criteria should be investigated.

Finally, further research is needed to improve the measure of entropy and the procedure of modelling selection of MHMMs because of their increasing use in the analysis of clickstream data from websites.

Chapter 5

Case study: PalermoTravel website

To demonstrate the effectiveness of MHMMs in managing clickstream to identify customer profiles, we propose an application to real data. We have analysed the website data of a Sicilian company operating in the tourism sector.

eTourism, or the online tourism market, is a prime example of how consumers have changed the way they make purchases by going from tourism organized by tour operators and travel agencies to do-it-yourself tourism. The travelers' need for informa tion on their destination and the ability to compare different offers has pushed tourism companies to improve websites to meet the needs of potential buyers. The tourism sector is an interesting case study because consumers in Tourism 2.0 (Sparks et al., 2013) obtain information anytime and anywhere to better plan their trips; hence, holiday packages must be customizable and the site structure should be attractive and constantly improved. Providing a well-structured and user- friendly website is obviously a strategic factor of marketing, and could be achieved by exploiting the information on the browsing behaviour of users obtained through web usage mining techniques.

Here, we consider the case study of PalermoTravel, a company that provides short-term house rental. Recently, the company has diversified its business offering services related to experiential holidays. Hence, it re-designed its website to offer tourists both apartments for rent and a booking service for experiential holidays. To increase the website's appeal, the company decided to add general information on Sicilian tourist attractions and activities, assuming that enriching the informative content would attract potential new customers.

By using web usage mining descriptive measures and an MHMM, we aim here to investi-

gate whether the management's decision to use the website as a showcase for information prompts users to explore the accommodation pages or if it actually distracts customers and works against the goals of the business. We also apply our proposal of entropy based selection criterion to identify the most suitable model.
The chapter is structured as follows. In Section 1 the data from the PalermoTravel website is presented. Section 2 illustrates an exploratory analysis where association rules and simple transition probabilities were obtained. In Section 3, behavioural profiles were identified by means of MMM (selected via BIC) and MHMM where the number of components and states are selected by using our proposed ICL_BIC.

5.1 PalermoTravel dataset

Data from the website of PalermoTravel was collected from September to December 2017.

Original rough data consists of $2,487,802$ observations arranged in chronological order. This dataset was cleaned and pre-processed in order to obtain a functional dataset for the application of statistical data mining techniques. After eliminating all the irrelevant log lines and the likely bots, the new dataset contains only the resources with *html* extension (i.e. the pages viewed) and consists of 95,201 lines by IP address. We also extracted information about user browsers, software and devices by using *uaparserjs* R package (Rudis et al., 2021). After this the dataset was ready for the session identification procedure. The procedure chosen was time-oriented and differentiated the navigation path of the individual user into sessions according to the time spent on the individual pages. Taking into consideration the structure of the site and the content of the individual pages, the time threshold of 10 minutes was chosen. Thus, if a user remains on the same page for more than 600 seconds the current session is considered concluded and the pages visited by the user in subsequent clicks are attributed to a new session. Using this methodology, we obtained $43,182$ user sessions, which allows us to follow the path of users within the site in the form of a sequence of pages viewed. Figure 5.1 shows the empirical distribution of the number of clicks, with an average number of clicks of 2.13 and standard deviation of 4.14. Most of the accesses are actually accidental visits related to users who leave after just one click.

Sessions having $T = 1$ clicks (i.e. short length sessions) have been removed and

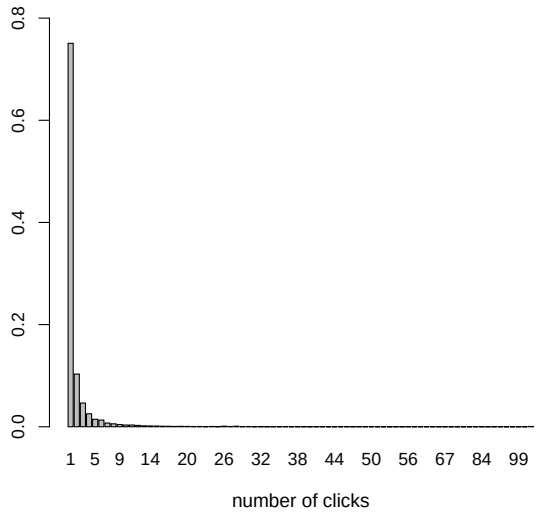

Figure 5.1: Number of clicks distribution; $n = 43,182$ sequences

the reduced dataset consists of $n = 10,252$ user sessions. Figure 5.2 shows the length distribution.

Instead of considering web pages as observed states of the sequence we decided to consider page categories. We made this decision both because the company wanted to compare pages of different types (purchase-oriented and information-oriented) and because of the large number of pages on the site. In fact, during the reference period pages with the same URL reported different information (replacing an apartment with another located in the same neighbourhood). Reliable information in the four months was in fact the only subject area of the pages concerned. These categories correspond to the thematic areas on the site and are presented in Table 5.1.

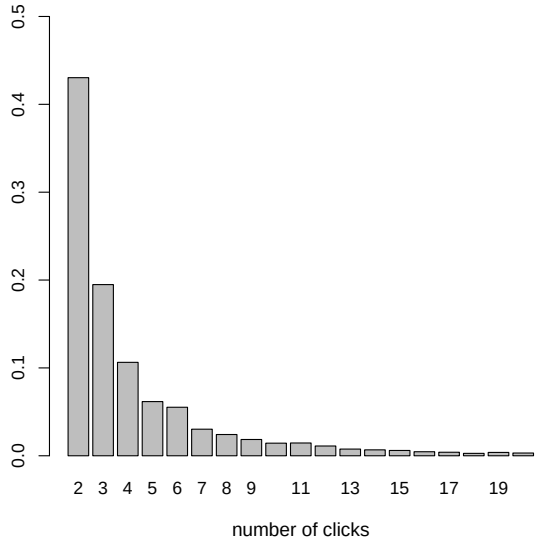

Figure 5.2: Number of clicks distribution; $n = 10,252$ sequences

Table 5.1: PalermoTravel website thematic areas

Area	
Homepage	The homepage contains several sheets. These pages contain suggestions as image links to access other areas
Attraction	This area contains information on tourist attractions or general information on Sicily
Accommodation	In this area, the user can access pages that offer apartments and rooms for rent and use a search engine to identify accommodation by time period, number of guests and price
Event	This area provides a calendar of the main seasonal concerts and festivals in Sicily
Experience	This area contains additional bookable tourist activities such as art and cuisine workshops, hiking, museum visits etc.
Service	An area for additional services to add to your package. It is possible to find bookable transport, food delivery services, childcare services etc.
Info	This area provides general information on the company and its staff and business partners. Moreover there are FAQ section and reviews

To summarise, the dataset contains sequences of pages visited by users, where a se-

quence derives from a combination of seven basic observed states: Homepage, Attraction, Accommodation, Event, Experience, Service and Info. This dataset can be considered as an example of single-channel sequence data with seven categorical states.[1] The maximum sequence length in the considered data was $T = 20$ (the few sequences exceeding this length were removed as outliers). Following Helske and Helske (2017), in the analysis the sequences are set at the same length $T = 20$ by adding missing values and emission probabilities $b_{u_{it}}(y_{it}) = P(Y_{it} = r|U_{it} = s)$ related to these clicks are set to 1 $\forall s \in \{1,2,\ldots,S\}$ and $\forall r \in \{1,2,\ldots,R\}$.

5.2 Exploratory analysis

To obtain preliminary information on browsing behaviour of users, an exploratory analysis was performed on post-processed *log-files*. The distribution of variable "Area" (Figure 5.3) shows that over the course of the 4 months, in 27.3% of the cases the users of the site viewed pages in the "Homepage" category, indicative of the simple access to the site and the use of the personalized search menu. The most viewed areas are those relating to "Attraction" in 33.5% of cases and to "Accommodation" in 22.3% of cases, indicating that users are more interested in viewing information relating to the city of Palermo than booking apartments. The other areas are considerably less popular; the "Experience" area was visited 5.3% of the times and the "Event" area just 3.2% of the time.

The association rules were obtained through an *apriori* algorithm implemented in the R *arules* (Hahsler et al., 2020) package. The threshold values for support and confidence are 5% and 30%.

Tables 5.2 and 5.3 show the results obtained by considering all 43,182 sessions and the "Attraction" and "Accommodation" areas as *consequent*.

It should be remembered that lift values greater than 1 mean that visualization of the *antecedent* areas increases the probability of display for "Attraction" and "Accommodation" pages.

It should be noted in Table 5.2 that "Attraction" and "Accommodation" pages were jointly accessed 20% of the time, the percentage of joint visualization of "Attraction"

[1] Multi-channel sequence data refers 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).

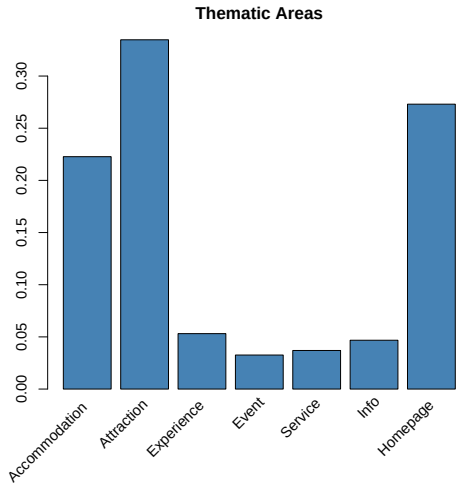

Figure 5.3: Thematic areas distribution

and “Homepage” is equal to 28%. Except for the joint display of “Accommodation”, “Homepage” and “Attraction” at 12%, all other rules have less than 10% support.

It is noteworthy that the confidence levels are higher than 50% highlighting the validity of the rules, the only exceptions being the rules relating to the “Accommodation-Attraction” and “Homepage-Attraction” pairings.

Turning our attention the lift values, we notice that only 6% of sessions jointly visualized “Attraction” and “Service”. However, the visualization of a “Service” tends to increase the probability of visualizing an “Attraction” with a lift value of 1,149. Similarly, viewing “Service” and “Homepage” or “Experience” and “Homepage” increases the probability of viewing “Attraction” with lifts equal to 1,199 and 1,111. The highest lift value is 1,206 when sessions accessing “Accommodation” and “Service”, increase the probability of selecting “Attraction”. However, these favorable cases correspond to less than 10% support, while rules with more than 20% support have list values less than 1. For example, the joint visualization of “Attraction” and “Accommodation” is present in 20% of sessions with a lift value of 0.889, i.e. users viewing an “Accommodation” have a lower probability of visualizing an “Attraction”.

Regarding Table 5.3, as in the previous case, “Attraction” and “Experience” visualizations, or those of “Experience” and “Homepage” have an increased probability of displaying “Accommodation” with lift values of 1.579 and 1.713, respectively. The rules with the highest support are those with a joint display of “Homepage” and “Accommo-

Table 5.2: Association rules, attraction consequent

rule	support	confidence	lift
$\{Event\} \rightarrow \{Attraction\}$	0.057	0.532	0.979
$\{Service\} \rightarrow \{Attraction\}$	0.066	0.625	1.149
$\{Info\} \rightarrow \{Attraction\}$	0.069	0.552	1.017
$\{Experience\} \rightarrow \{Attraction\}$	0.091	0.569	1.048
$\{Accommodation\} \rightarrow \{Attraction\}$	0.207	0.483	0.889
$\{Homepage\} \rightarrow \{Attraction\}$	0.289	0.499	0.918
$\{Homepage, Service\} \rightarrow \{Attraction\}$	0.062	0.655	1.206
$\{Homepage, Info\} \rightarrow \{Attraction\}$	0.051	0.563	1.036
$\{Accommodation, Experience\} \rightarrow \{Attraction\}$	0.062	0.655	1.206
$\{Experience, Homepage\} \rightarrow \{Attraction\}$	0.055	0.603	1.111
$\{Accommodation, Homepage\} \rightarrow \{Attraction\}$	0.129	0.546	1.011

Table 5.3: Association rules, accommodation consequent

rule	support	confidence	lift
$\{Event\} \rightarrow \{Accommodation\}$	0.058	0.544	1.269
$\{Service\} \rightarrow \{Accommodation\}$	0.057	0.532	1.241
$\{Info\} \rightarrow \{Accommodation\}$	0.052	0.419	0.977
$\{Experience\} \rightarrow \{Accommodation\}$	0.094	0.588	1.372
$\{Attraction\} \rightarrow \{Accommodation\}$	0.207	0.381	0.889
$\{Homepage\} \rightarrow \{Accommodation\}$	0.237	0.408	0.951
$\{Attraction, Experience\} \rightarrow \{Accommodation\}$	0.062	0.677	1.579
$\{Experience, Homepage\} \rightarrow \{Accommodation\}$	0.067	0.734	1.713
$\{Attraction, Homepage\} \rightarrow \{Accommodation\}$	0.129	0.447	1.042

dation" or "Attraction" and "Accommodation"; in both cases the probability of viewing "Accommodation" decreases. Therefore, it seems that users visualize mostly "Accommodation" or "Attraction" pages but accessing pages in one area decreases the probability of accessing the other.

Association rules highlight relationships between thematic areas, but do not account for the order of access in web sequences. Thus, to analyse the paths of the users exploring the portal we apply Markov chains, as they allow us to represent users' movements from one page to another in probabilistic terms. We can see the graphs related to the Markov chain transition probabilities of different orders in the succeeding graphs.

Figures 5.10, 5.11 and 5.12 show that users tend to stay in the same subject area in the first clicks, and this behaviour persists even to the seventh click for the "Attraction" and "Accommodation" areas. For example, let us focus on first order transition probabilities,

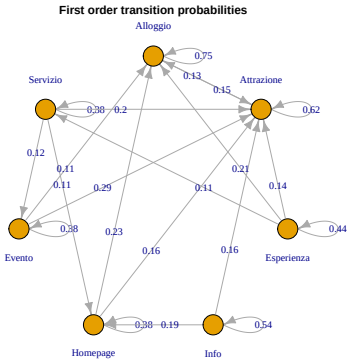

Figure 5.4: Transition probabilities orders 1

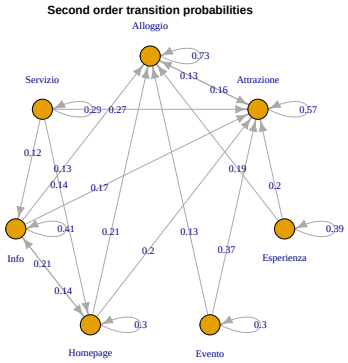

Figure 5.5: Transition probabilities orders 2

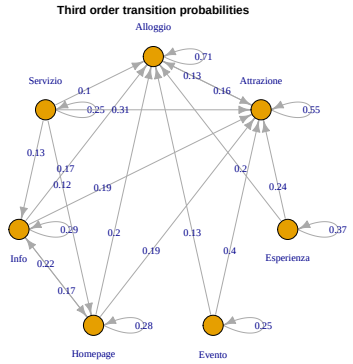

Figure 5.6: Transition probabilities orders 3

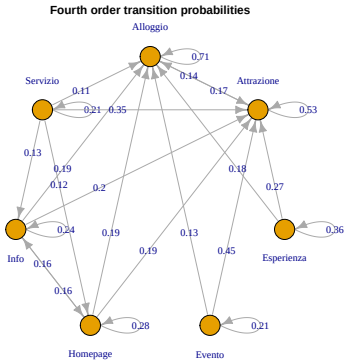

Figure 5.7: Transition probabilities orders 4

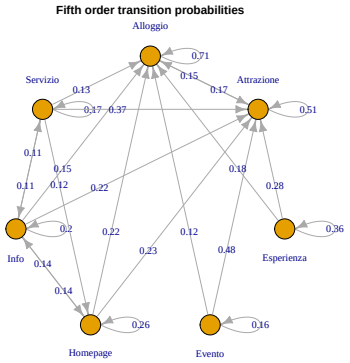

Figure 5.8: Transition probabilities orders 5

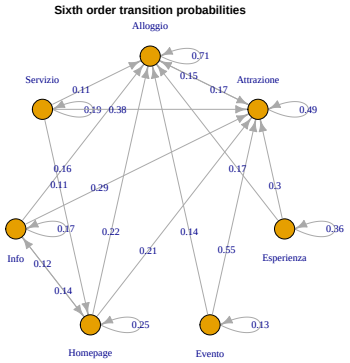

Figure 5.9: Transition probabilities orders 6

the intensity of which is visible in Figure 5.4 (transition probabilities with a value less than 0.1 has been removed). Starting from the "Accommodation" category, a user has a 75% chance of moving to a page from the same category. Similarly, a user on a page belonging to the "Attraction" category has a 62% chance of remaining in the same category. Other thematic areas have greater transition probabilities in case of a passage within the same category, but to a lesser extent compared to the two previous categories (values between 38% and 54%).

In the "Homepage" category, users who access the site have a 38% transition probability of remaining in the same category interacting with custom search engines and a 23% chance of moving to "Accommodation" and 16% of moving to the "Attraction". The latter are the most visited areas coming from each of the categories. Let us now consider the transition probability of going from an area to another. Users who are in the "Accommodation" area have a 12.9% probability of viewing an "Attraction", the transition probabilities to other areas being below 6%. Starting from an "Attraction" page, we have a 14.6% probability of viewing an "Accommodation" and a probability less than 9% of viewing another area. From the "Experience" area, there is a 21.3% probability of viewing an "Accommodation", a 14.4% of viewing "Attraction" and 10% of viewing the "Service" area. The other transition probabilities are less than 6%. The "Event" area has a greater probability of moving into an "Attraction" (28.7%) compared to "Accommodation" (10.7%). Similarly, the "Info" and "Service" areas are also more likely to make the transition into "Attraction" (16.4% and 20.4%) than "Accommodation" (4.7% and

5.1%). In addition, users who are on an “Info” page have a 20% probability of viewing the homepage.

If we take into consideration higher order transition probabilities, the intensity of which can be seen in Figures 5.5 - 5.9, we notice just a slight decrease in probabilities related to movements in the same area. After a third click the transition probabilities in the main diagonal (with the exception of “Accommodation” and “Attraction”) have values below 36%, in addition, users in the “Service” category are more likely (30.6%) to move to an “Attraction” than remain in the same category (24.6%). Both the “Service” category and the “Event” category as the order increases have very high transition probabilities towards the “Attraction” category while the probability of transitioning to “Accommodation” decrease dramatically.

In the next sections, web sequences will be classified by using model-based approaches (MMMs and MHMMs).

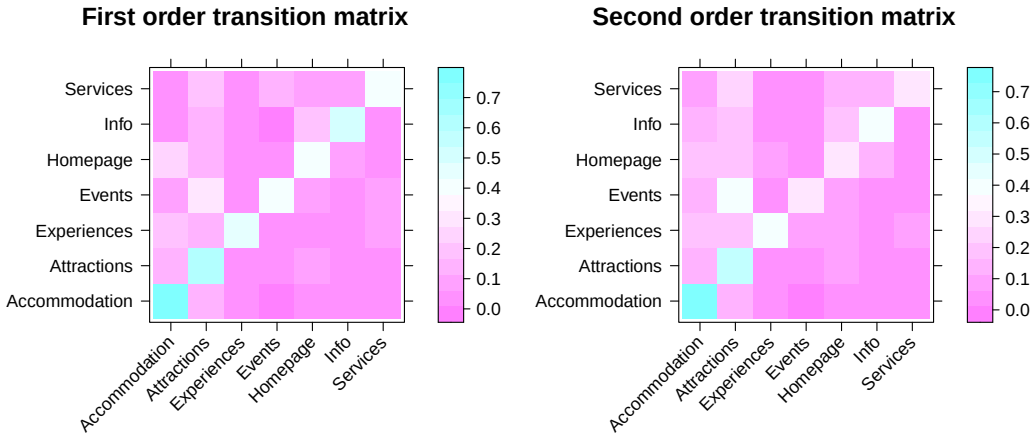

Figure 5.10: Transition Matrices order 1 and 2

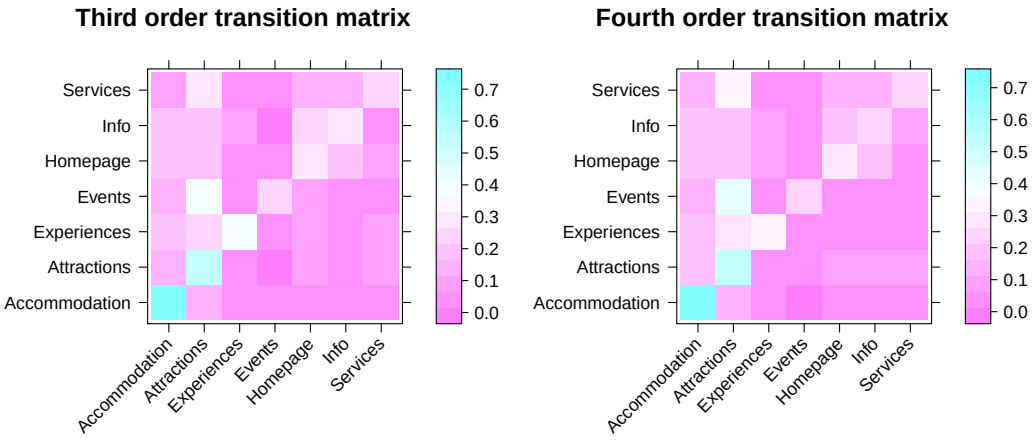

Figure 5.11: Transition Matrices order 3 and 4

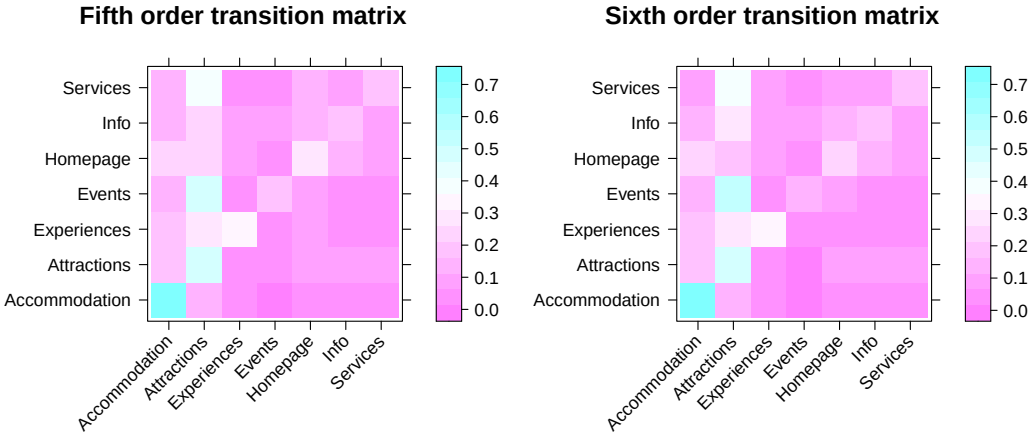

Figure 5.12: Transition Matrices order 5 and 6

5.3 Profiles identification

In the previous section, we estimated website users' transition probabilities assuming a homogeneous population and we obtained general indications on global navigation behaviour.
In this section, we will proceed to relax this assumption in order to identify clusters of sequences representing different profiles of navigational behaviour by using two methods of model-based clustering.
In the first case, we will apply an MMM, a method commonly used in the literature. The identification of clusters will be based directly on the observed web sequences and the method used to select the number of clusters is the BIC.
In the second case, we will use an MHMM, a method that takes into account the heterogeneity hidden in the clickstream data, assuming that users' movements evolve on the basis of a latent process. The states of this process identify latent "mental states" that influence users to select one page or another. To take into account two levels of uncertainty and identify clusters and states, we used the selection criterion proposed in chapter 4.

In both cases additional information obtainable from clickstream data will also be used as covariates in the mixture model (see Equation 3.1), i.e. the nationality extracted from the IP addresses, the access device and the month of access.

5.3.1 Mixture Markov model

We applied MMMs in order to identify behavioural profiles across users. Different MMMs are considered by varying the number of clusters and a $K = 4$ model was selected by comparing BIC scores.

The sizes of clusters are displayed in Figure 5.13 and are 23%, 46%, 24% and 7%.

The distribution of clusters according to geographic area, device and month of access can be seen in Figures 5.14 - 5.24. The first cluster is mostly composed of Italian accesses (62/%) followed by North American and North European (16% and 11%). The second cluster (the largest one) consists of 43% of Italians, 23% of North Europeans and 19% of North Americans. In the third cluster we find a more diverse composition, with 33% Italians, 16% of North Europeans and 25% East Europeans; moreover 8% of the sessions in this cluster are from Asians (indeed, 60% of the Asian accesses are classified in this cluster). The last cluster, as regards nationality, is quite similar to the second one, with

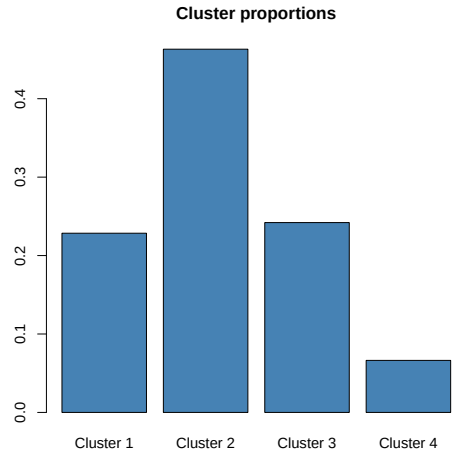

Figure 5.13: Mixture Markov model: Clusters sizes

40% of Italians, 29% of North Europeans and 16% of North Americans. Most website accesses are from a Desktop computer, with 74% of them in first cluster, 69% in the second, 83% in the third and 62% in the fourth. There are no clear preferences as regards month of access with the exceptions of users in first cluster that accessed in September in 30% of cases and the fourth cluster that rarely explored the website that month.

Figures 5.17-5.18 and 5.19-5.20 show the different clusters initial probability vectors and transition matrices representing different behaviour patterns while exploring the website.

64% of users in *Cluster 1* start from "Homepage" area and 16% from "Attraction" area. They move from "Homepage" to "Attraction" and "Accommodation" with similar probabilities (25% and 21%) or stay in the "Homepage" area with a 23% probability. If users accessed pages in "Attraction" or "Accommodation", they tend to stay in the same areas with 49% and 47% probabilities. We notice probabilities $> 40\%$ of remaining in "Service" and "Info" and a 38% probability of reaching "Accommodation" from "Experience".

Moving to *Cluster 2*, 57% of users start from "Homepage" area, 15% from "Attraction" and 14% from "Experience". There is a 75% probability they will stay on the homepage and if they do reach other areas, they will stay in "Attraction" with a probability of 57% and move from "Accommodation" to "Attraction" with a probability of 51%.

As regards *Cluster 3*, 44% of users start from the "Attraction" area and 84% stay there.

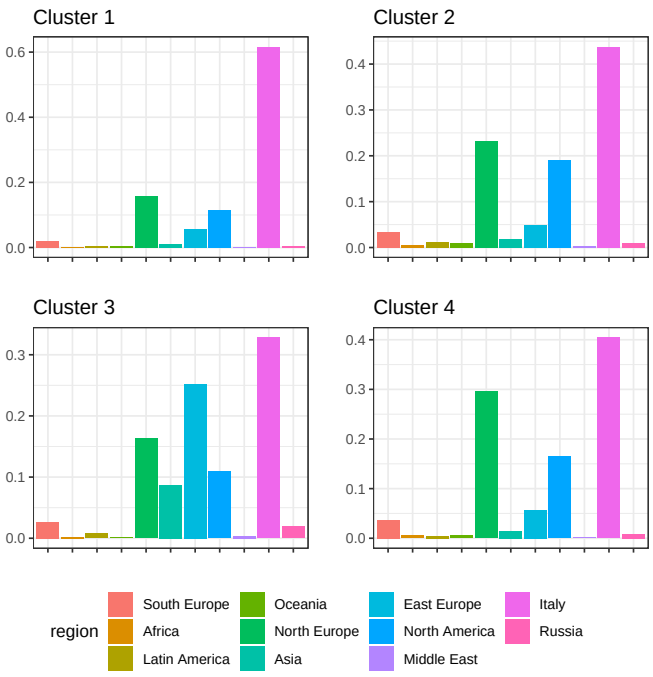

Figure 5.14: Mixture Markov model: IP geographic position distribution

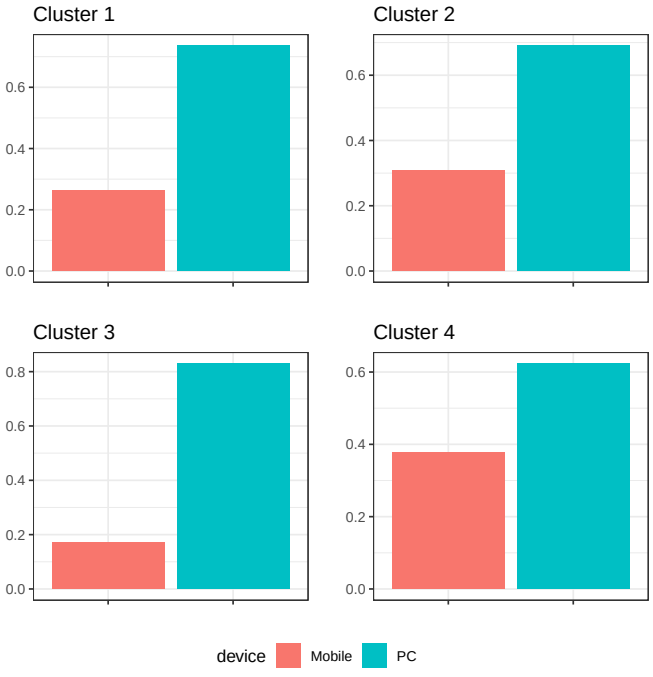

Figure 5.15: Mixture Markov model: IP access device distribution: Pc and mobile

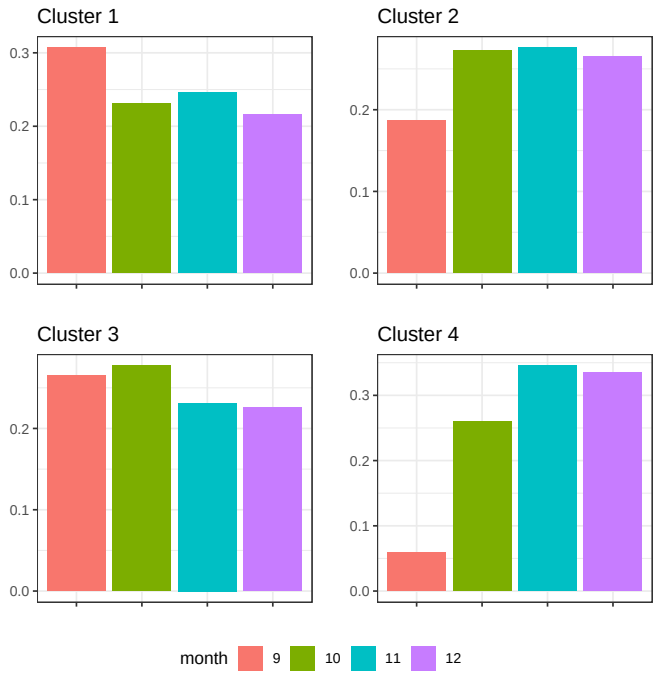

Figure 5.16: Mixture Markov model: IP access month distribution

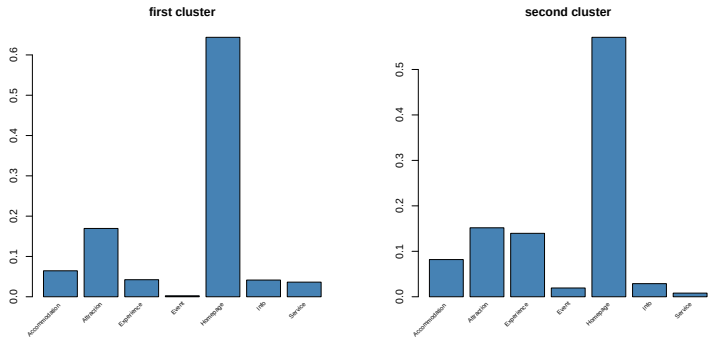

Figure 5.17: Initial probability vector π^k for clusters 1 and 2

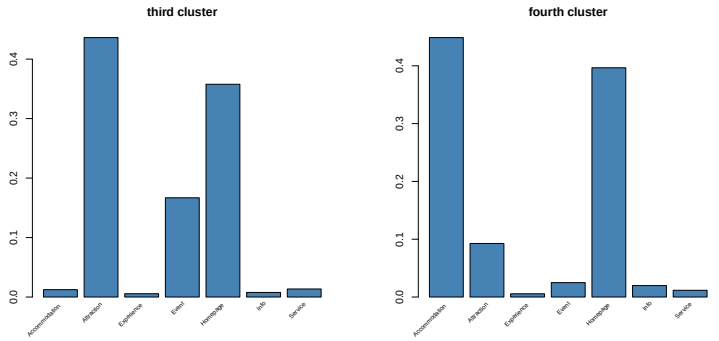

Figure 5.18: Initial probability vector π^k for clusters 3 and 4

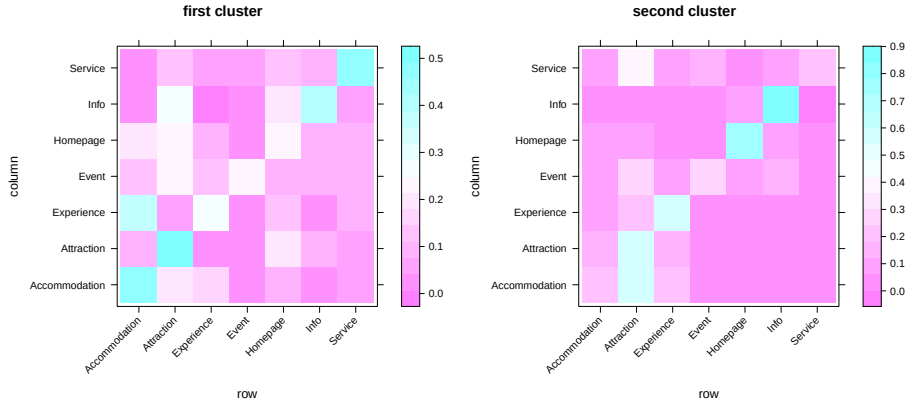

Figure 5.19: Transition matrices A^k for clusters 1 and 2

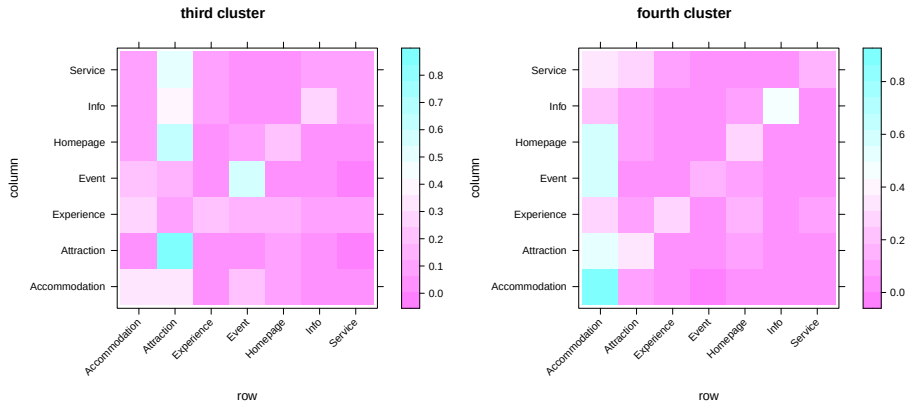

Figure 5.20: Transition matrices A^k for clusters 3 and 4

36% start from “Homepage”, with an 18.8% probability of staying in the same area, and a probability of 63% of moving to the “Attraction” area.

In *Cluster 4* 45% start from “Accommodation” and 40% from “Homepage”. There is a 58% probability they will move from “Homepage” to “Accommodation” and an 87% probability they will stay there.

5.3.2 Mixture hidden Markov model

We consider MHMMs to explore if and how browsing behaviour differs across users accounting for an additional level of complexity. Specifically, let us suppose the sequences follow a latent Markov process. Prior cluster membership was estimated through a multinomial logistic model (see equation (3.1)) having 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).[2] Different MHMMs are considered by varying the number of components $K \in \{2,3,4,5\}$, and hidden states for each component $S^k \in \{2,3,4,5\}$.

To identify the number of components and states the proposed model selection score ICL_BIC was used. We selected the MHMM consisting of three components and hidden states (3,2,4) (i.e. identifying three clusters/browsing profiles). The sizes of clusters are displayed in Figure 5.21 and are 22%, 19%, and 59%.

Figures 5.22-5.24 show, for each cluster, the distribution of geographical area, access device, and access month.

Figure 5.22 shows that users, from different geographical areas, differ as to their browsing behaviour. For example, 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%). They seem interested in exploring Sicily’s attractions and tourist products (i.e., cultural, beach, experiential holidays) as potential tourists. Asian and East European users, sparsely featured in the data, are assigned to cluster 1 (10% and 30% of this cluster), while their presence

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

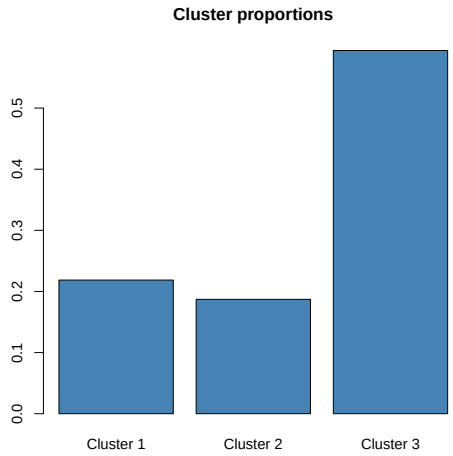

Figure 5.21: Mixture hidden Markov model: Clusters sizes

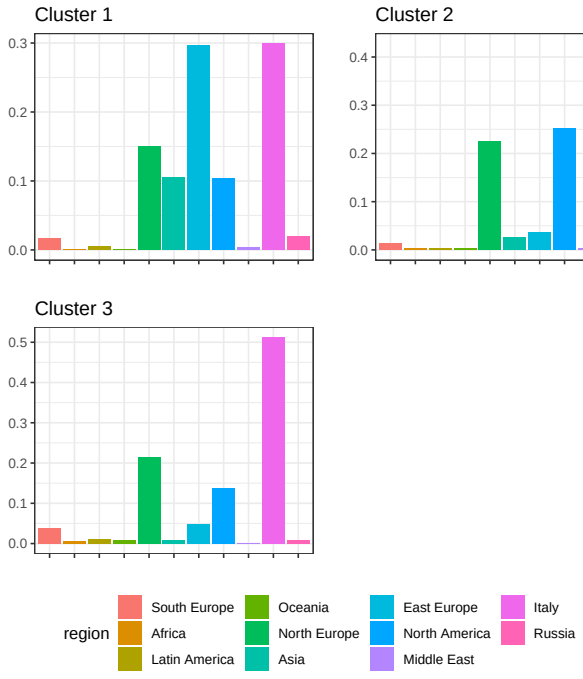

Figure 5.22: ixture hidden Markov model: IP geographic position distribution

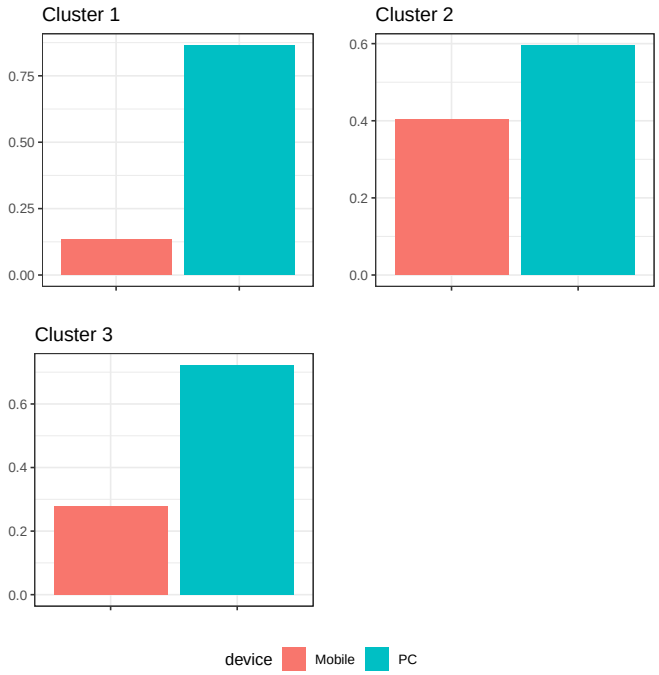

Figure 5.23: ixture hidden Markov model: IP access device distribution: Pc and mobile

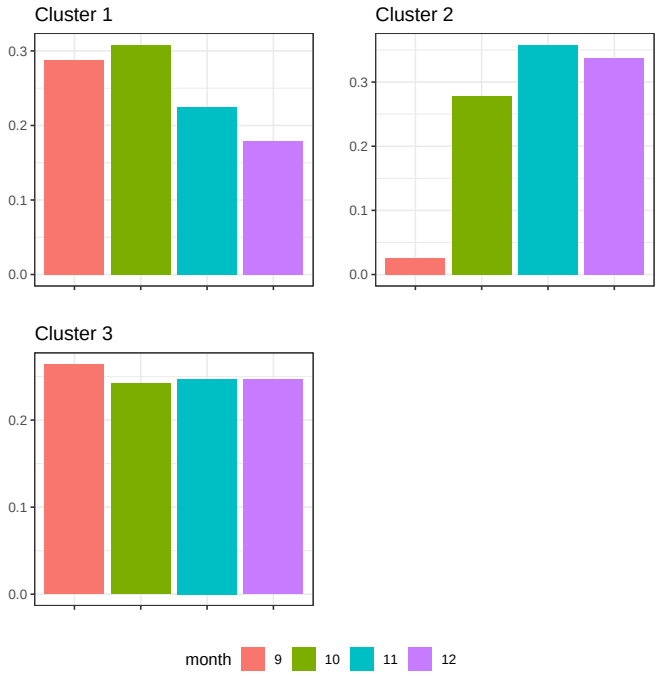

Figure 5.24: ixture hidden Markov model: IP access month distribution

in other clusters is below 3% and 5%, respectively. As concerns the access device, users show a preference for the PC, with users in cluster 1 preferring it in 86% of cases and those in cluster 2 in 59% of cases.
Finally, most users in cluster 1 visited the website in October (30.8%), and less during December (17.9%). Users in cluster 2 accessed in October, November and December, with just 2.6% of them accessing in September. Users in cluster 3 do not show any particular preference, with all access probabilities at around 25%.

Figure 5.25 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 charts represent the emission probabilities of the observed states (the thematic area of the page). Emission probabilities lower than 0.05 are classified as "others".

Focusing on the first cluster in Figure 5.25 top-left, we see that users start their web navigation from state 2 (that emits the observed state "Homepage"with a probability of 0.98). Then they keep accessing Homepage pages with a probability of 0.73 or transition to state 1 with a probability of 0.15 and state 3 with a probability of 0.12 and remain in these states (with probabilities of 0.89 and 0.84 respectively). Specifically, state 1 tends to emit "Service" with a probability of 0.34, "Attraction" with a probability of 0.22 and "Accommodation" with a probability of 0.19. State 3 emits "Info " with probability 0.78.

In cluster 2 (Figure 5.25 top-right), users have a higher probability of starting from state 2 (0.66) and a 0.99 probability of staying there. State 2 emits "Attraction" with a probability of 0.91. If they start from state 1 with a probability of 0.34 and remain there with a probability of 0.96, they have a high probability (0.61) of accessing "Event" pages.

Finally, moving on to cluster 3 (see Figure 5.25 bottom), users start from state 3 with a probability of 0.56 and move to state 1 with a probability of 0.43. Once users reach state 1, the probability of them staying there is 0.85. State 1 emits "Attraction" with a probability of 0.74. Another path is moving from state 3 to state 4 with a probability of 0.25 and staying in this state with a probability of 0.94. State 4 emits "Accommodation" with a probability of 0.91.

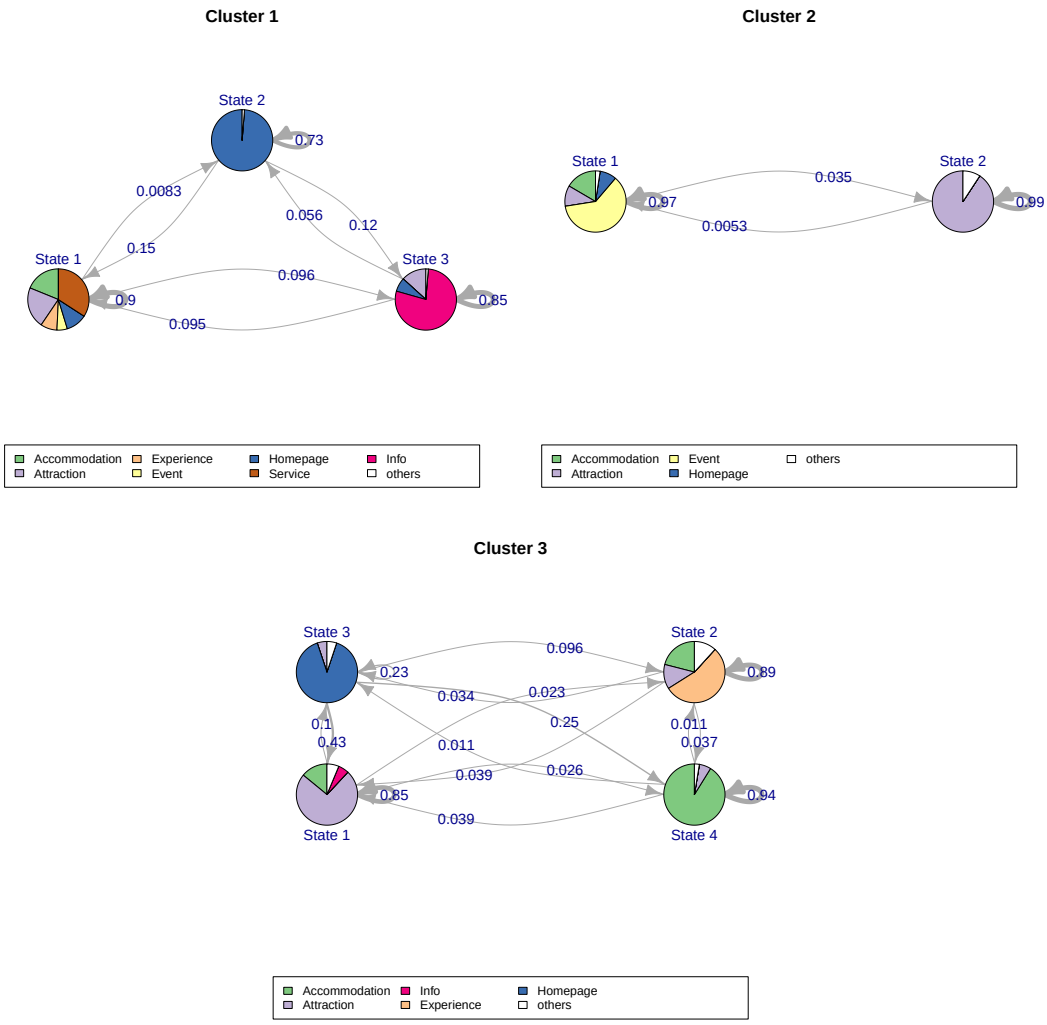

Figure 5.25: Hidden Markov process structures for clusters 1,2, and 3. Vertices represent hidden states. The slices show emission probabilities, and the edges show the transition probabilities

5.4 Discussion

In this chapter, web usage mining techniques have been used to handle and analyse click-stream data from a Sicilian company's website. Our empirical analysis concerned the tourism sector and the data extracted from the website was used to understand user behaviour.

Association rules analysis showed that users seem to be more interested in obtaining general information about the province of Palermo rather than viewing accommodation offers. In fact, the "Attraction" area , unlike the items related to other thematic areas, are shown to users directly in the homepage in the form of a link-image below the menu, with clear intent of encouraging the exploration of the site using cultural information.

Then, we analysed how users explored the website through Markov chains, identifying how they move between pages. As regards the order of the Markov chain, we computed transition matrices of different orders (first to sixth), displaying little differences that suggested a user behaviour of relatively homogeneous sessions in terms of the thematic areas visited.

Since the length of the sessions quite short on average (most of them below 5 clicks), we decided to consider a simpler model assumption. An application of Markov chains allowed us to estimate the exit probability from the different areas of the site. The analysis shows that users conclude navigation on a page of the "Attraction" category in 31.7% of cases and on a page of the "Accommodation" category in 33.5% of cases, confirming that these categories are the most attractive. The analysis of the Markov chains highlighted a "single track" navigation behaviour. Users tend to view pages in the same thematic area sequentially. As clicks increase, users who are viewing "Accommodation" or "Attraction" continue browsing without changing areas, while those in another area begin to vary their route. Taking into consideration what has been learned thanks to the rules of association, at this level of the analysis there seem to be three predominant user paths: a) the *consumer* has sessions browsing pages of apartments and accesses the site already as a potential traveler; b) the *missing consumer* accesses the site to view the "Attractions" simply for general cultural reasons and his failure to view "Services", "Events" or "Experiences" denotes the lack of interest in visiting the region; c) the *explorer* constitutes a minority among users and the wealth of information in the website (with particular emphasis on experiences) arouses their interest. They conclude their navigation by viewing one of the

two macro-areas “Attraction” and “Accommodation”.

The exploratory analysis carried out gave us insights on user behaviour and suggested the presence of different behavioural profiles. Therefore, we applied sequential clustering to explore user browsing behaviour and to identify clusters of sequences the evolution of which reflects differences in interest and aims among users. Specifically, we compared the profiles identified through a method commonly used in the literature, i.e. an MMM with identification of the number of clusters via BIC, with those identified through an MHMM and the proposed selection criterion.

Profile identification via MMM led to the selection of four clusters. The first cluster seems to identify a rather mixed behavioural pattern, with users viewing the “Homepage”, “Attraction” and “Accommodation” pages. The users of the second cluster on the other hand, seldom leave the “Homepage” area, with a 75% probability of always selecting a homepage sheet. The third cluster users tend to prefer to view the “Attraction” pages, while the fourth cluster prefer viewing sequences of “Accommodation” pages. The cluster composition as regards devices or nationalities does not provide particularly interesting information apart from the fact that the majority of Asian users are in the third cluster and that the first cluster is mainly made up of Italians. The model seems to have isolated the users most interested in selecting “Accommodation” in cluster 4 which is the smallest cluster with just 7% of users.

To account for hidden heterogeneity in clickstream we identified behavioural profiles by adopting an MHMM, selected through the proposed ICL_BIC. We identified three user profiles of website browsing: the casual explorer/potential partner, the information seeker and the potential tourist. The discriminant factors in these profiles seem to be the geographical location of the IP address, the access device and the access month. It is interesting that by using MHMMs we identified three components instead of four as in the MMMs example, due to hidden states accounting for sequences variability. This is because MMMs identify clusters having similar *observed* sequences and will need more clusters to obtain relatively homogeneous groups. On the other hand, MHMMs identify clusters based on *hidden* sequences. In this case a cluster will be identified by similar *hidden* sequences representing the same hidden behaviour and will group pages into categories which reflect user goals and hidden “mental-states”.

The *casual explorer or potential Partner (cluster 1):* hidden states in this cluster identify three different “mental-states” and three sub-paths in this group. The first one is

related to a lack of interest in the website or to an interest in general tourist information, as this state emits pages in the “Homepage” area. A second hidden state emits different areas and seem related to 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 seeker (cluster 2):* users who are looking for tourist information and primarily view “Attraction” and “Event”. This cluster has the highest percentage of mobile access (although PC access is still the preferred one) and makes up 19% of users.

The *potential tourist (cluster 3):* users who view the website to search for both tourist information and tourist products (59% of the sessions). Hidden states represent preliminary states related to “Homepage ” visualization, an exploratory state related to “Attraction” visualization and a buying state related to “Accommodation” visualization. There is also a fourth state that accounts for the other thematic areas that are rarely accessed in this cluster. Users in this cluster begin by selecting homepage sheets before moving on to tourist attractions pages or bookable apartments. It 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.

Conclusions and future work

The study of browsing behaviour has become a fundamental component of online shopping and an understanding of it allows companies to implement effective strategies in order to increase customer relationships and sales. Browsing behaviour reveals the most frequently taken paths, the most purchased products or simply the most visited pages. The information obtained can be used to improve companies' websites in terms of structure and content, by improving the links preferred by users or by relocating the less viewed pages. In highly competitive environments, it is therefore crucial to be able to adapt websites to the needs of consumers, simplifying navigation and personalizing the service.

In this thesis we analysed real clickstream data from a Sicilian hospitality company to highlight the usefulness of mixture hidden Markov models as a tool for identifying and analysing behavioural profiles in a business context and to verify the effectiveness of business decisions. To identify these profiles we operated a model-based sequential clustering. We used MHMMs to analyse web sequences under assuming that there was a heterogeneous population and that each subpopulation relates to cluster of web sequences which evolve according to a specific hidden Markov model. In order to select a model and so the number of clusters and states for each hidden Markov model, we defined an entropy-based criterion that accounts for hidden heterogeneity in clickstream data.
The company had recently had the website modified by enriching the tourist information provided and requested an analysis of the clickstream data so as to better understand their users' browsing behaviour, how many users selected purchase-oriented pages, how they behaved before accessing the "Accommodation" pages, and the characteristics of potential buyers.

The dataset consists in *log-files* extracted from the company website and refers to the period September-December 2017. In Chapter 2, we presented clickstream data and their main limitations. Firstly, the data does not have an analysis-ready structure and must

undergo cleaning and restructuring procedures in order to obtain a sequential dataset. Cleaning involves removing unnecessary web resources such as audio and video data. A further issue is user identification. We have said that IP addresses are assigned in turn and do not allow for the identification of a single user; commonly the analysis involves assumptions on the identity of the users which is matched to those IP addresses that do not change browsers and devices in a defined period of time (extracted from the user-agent field) and that respect the path tracked in the referrer field. If previous information is not available, the activity of the same IP is divided into sessions on the basis of temporal rules such as the time spent on the same page i.e. if it exceeds 10 minutes the next page is part of a new session.
The sessions obtained are the statistical units and are used as a basis for data mining techniques such as association rules and sequential analysis.

After pre-processing, we analysed the clickstream data to identify browsing behavioural profiles by using mixture hidden Markov models. In order to select the number of profiles we had to choose the best criterion for MHMMs. However, a Monte Carlo simulation study pointed out that the most common criteria used in literature for mixture models and hidden Markov models do not perform satisfactorily if applied to MHMMs, where both the number of components and states have to be identified. 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 classification likelihood approach that accounts for the two latent classes in the model: subpopulations (i.e. mixture components) and hidden states. We used a modified ICL BIC by defining a new entropy penalization i.e. a joint entropy obtained as the sum of cluster-level entropy and state-level entropy, where the former is the mixture model entropy and the latter is based on the Hernando et al. (2005) conditional entropy definition.
The new criterion, although it requires a range of possibilities for the number of clusters and states to be defined, outperformed the other information criteria. The BIC, which is 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.

As regards the analysis of the company's clickstream data, the application of an MHMM and the proposed criterion identified three behavioural profiles. We also used a mixture of Markov models selecting the number of profiles based on BIC scores as is commonly done in the literature.

The MMM approach selected four profiles as it identifies similarities between observed se quences and needs more clusters to handle variability.
The MHMM approach allows us to identify, through the 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 common accessed areas into one state. The MHMMs' hidden states also provided a better understanding of user movements as they highlight sub-behaviours inside clusters.

Our results have given the company management pause for thought. The homepage offers the user a personalized search menu, links to different areas of the site and, at the bottom, displays a list of image links relating to the "Attraction" area with the clear intention of stimulating the user to explore the website further. Unfortunately, once the exploration of the attractions has begun, the user is no longer encouraged to change area, in fact the attraction pages only have links to other attractions, generating the aforementioned mono-thematic exploration. Furthermore, the links at the top of the page to other areas are small making exploration of these areas difficult. This problem is exacerbated even further on the small screen of a smartphone where the touch technology favours image links as a means of moving around the website accentuating the navigation to "separate tracks".
One way of overcoming these problems would be to diversify the image links provided in each area, thus improving site links. However, taking into consideration 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 "Attraction" page. The same method could also be applied to the "Experiences" area which has proved to be attractive to users.
Finally, it also seems that a change in the layout of the homepage menu would be useful, enlarging the entries of individual areas in order to make browsing easier on mobile phones. Profile identification results 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 is to be 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 potential Asian and Eastern European business partners who are interested in promoting outgoing tourism toward Italy. Since this

second group of users is not really catered for by PalermoTravel's B2C business model, the company will also need to focus on selling products and services to other companies (i.e. the B2B, business-to-business model).
This would involve not only sales to other companies but also creating business partnerships to promote Sicily as a tourist destination. Furthermore, our findings have demonstrated that MHMMs, although not commonly used for clickstream data, are very useful in identifying user profiles with similar browsing behaviour.

Summing up, sequential data is a valuable resource in modern society and it is therefore necessary to have computationally efficient and reliable statistical tools that can be applied to diversified phenomena. Such tools could be used to create routines that companies could deploy in a variety of different projects. The statistical analysis of sequential big data is extremely important in defining strategic market policies, targeted communication activities and marketing solutions.

It needs to be said , however, that this work does have its limitations. Representing human behaviour, by means of Markov models, is clearly an approximation of reality. It is common practice also to consider web sessions as time-discrete sequences of web pages accessed by users. Cadez, Heckerman, et al. (2000) focused on sequence classification for web data, applying a mixture of first-order Markov models to identify clusters of users by considering the web pages accessed as states of the Markov chain (see Cadez, Gaffney, et al., 2000). Dias and Vermunt (2007) highlighted the usefulness of mixture Markov models in the analysis of web data as they adapt both to heterogeneity between sequences and serial dependencies. However, Markov models do not consider the unobserved heterogeneity related to the lack of information about navigation purposes. Thus, statistical models with hidden variables are to be preferred in analysing clickstream data. Among them, discrete-time Hidden Markov Models (HMMs) are the most suitable to explore sequences and reveal the unknown variables that influence browsing behaviour.
Continuous-time HMMs can also be used to take into account how holding time affects transitions between states. Xu et al. (2013) proposed a single Hidden semi-Markov model to analyse clickstream data. However, due to the complexity of mixture hidden Markov models and the number of parameters to estimate, we approximated user behaviour to time-discrete sequences and the extension to continuous time sequences is the subject of further research.

The study on model selection for MHMMs, conducted in this thesis, provides the basis

for further developments. In our study, an Estimation-Maximization algorithm was used to estimate model parameters, although an alternative approach would be to use Bayesian approach based on MCMC sampler considering Dirichlet priors for transition matrices rows. This approach would overcome a limitation of the EM algorithm that may struggle during the M-step if there are no transitions between two states. Moreover, as we pointed out, although the proposed entropy criterion outperformed previous ones, its ability to identify clusters and states was still strongly influenced by different features such as the number of clusters and the length of sequences.

However, a different technique would be to estimate the number of clusters along with the other parameters considering a Bayesian framework and the reversible-jump Monte Carlo Markov chain algorithm (RJMCMC), allowing this number to be inferred via their posterior distribution. This approach was proposed for hidden Markov models by Robert et al. (2000) and is based on Gibbs sampler algorithm. Each iteration considers splitting a cluster into two or combining two clusters into one and is similar to the split/merge moves of Richardson and Green (1997). Having fixed the number of clusters, we can update the model parameters using standard MCMC moves.

The extension of the algorithm to MHMMs has recently been proposed by Luo and Stephens (2021) and requires the following additional steps; a split/combine step to select the number of clusters; then fixing these numbers and repeating the RJMCMC moves for the number of states in each cluster. Although this algorithm was able to identify both the number of clusters and states, it comes with several limitations: i) results were highly affected by conditional distribution variability, ii) the analysis was carried out considering continuous observations only. Thus, this estimation technique will need to be adapted to categorical data in order to be applied in the context of clickstream analysis.

Another issue in applying RJMCMC is label switching. This refers to unidentifiable models in data fitting if priors are invariant to relabelling. If K latent states/clusters are present there are $K!$ possible ways to label them and when we sample from posterior density during RJMCMC iterations, state/cluster permutations can alternate. Several approaches have been proposed to solve this issue which are generally divided into two categories. One is postprocessing (Marin et al., 2005)), which allows the algorithm to explore all state and cluster permutations and provides Bayesian estimates by minimizing posterior expectation of a suitable loss function. The other uses a data-driven identifiability constraint on model parameters (Green and Richardson (2002); Frühwirth-Schnatter

(2001); Robert et al. (2000)). In this case the algorithm fixes a single cluster/state permutation, rejecting the split/combine move if the constraint is not satisfied.

Bibliography

Abbott, A., & Forrest, J. (1986). Optimal matching methods for historical sequences. *The Journal of Interdisciplinary History*, *16*(3), 471–494.

Abbott, A., & Tsay, A. (2000). Sequence analysis and optimal matching methods in sociology: Review and prospect. *Sociological methods & research*, *29*(1), 3–33.

Agarwal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. *Proc. of the 20th VLDB Conference*, *487*, 499.

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, 207–216.

Bacci, S., Pandolfi, S., & Pennoni, F. (2014). A comparison of some criteria for states selection in the latent markov model for longitudinal data. *Advances in Data Analysis and Classification*, *8*(2), 125–145.

Banerjee, A., & Ghosh, J. (2000). Concept-based clustering of clickstream data.

Banerjee, A., & Ghosh, J. (2001). Clickstream clustering using weighted longest common subsequences. *Proceedings of the web mining workshop at the 1st SIAM conference on data mining*, *143*, 144.

Bartolucci, F., & Pandolfi, S. (2015). Lmest: Latent markov models with and without covariates. *R package version*, *2*.

Baudry, J.-P., Raftery, A. E., Celeux, G., Lo, K., & Gottardo, R. (2010). Combining mixture components for clustering. *Journal of computational and graphical statistics*, *19*(2), 332–353.

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *The annals of mathematical statistics*, *41*(1), pp 164–171.

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, *22*(7), pp 719–725.

Biernacki, C., & Govaert, G. (1997). Using the classification likelihood to choose the number of clusters. *Computing Science and Statistics*, 451–457.

Biernacki, C., & Govaert, G. (1999). Choosing models in model-based clustering and discriminant analysis. *Journal of Statistical Computation and Simulation*, *64*(1), 49–71.

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*, 565–601.

Bozdogan, H. (1987). Model selection and akaike's information criterion (aic): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370.

Bozdogan, H. (1994). Mixture-model cluster analysis using model selection criteria and a new informational measure of complexity. *Proceedings of the first US/Japan conference on the frontiers of statistical modeling: An informational approach*, 69–113.

Bucklin, R. E., & Sismeiro, C. (2003). A model of web site browsing behavior estimated on clickstream data. *Journal of marketing research*, *40*(3), 249–267.

Bucklin, R. E., & Sismeiro, C. (2009). Click here for internet insight: Advances in clickstream data analysis in marketing. *Journal of Interactive marketing*, *23*(1), 35–48.

Cadez, I., Gaffney, S., & Smyth, P. (2000). A general probabilistic framework for clustering individuals and objects. *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 140–149.

Cadez, I., Heckerman, D., Meek, C., Smyth, P., & White, S. (2000). Visualization of navigation patterns on a web site using model-based clustering. *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 280–284.

Canter, D., Rivers, R., & Storrs, G. (1985). Characterizing user navigation through complex data structures. *Behaviour & Information Technology*, *4*(2), 93–102.

Celeux, G., & Durand, J.-B. (2008). Selecting hidden markov model state number with cross-validated likelihood. *Computational Statistics*, *23*(4), pp 541–564.

Ching, W.-K., & Ng, M. K. (2006). Markov chains. *Models, algorithms and applications*.

Cooley, R., Mobasher, B., & Srivastava, J. (1997). Web mining: Information and pattern discovery on the world wide web. *Proceedings ninth IEEE international conference on tools with artificial intelligence*, 558–567.

Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data preparation for mining world wide web browsing patterns. *Knowledge and information systems*, *1*(1), 5–32.

Cooley, R., & Srivastava, J. (2000). *Web usage mining: Discovery and application of interesting patterns from web data*. Citeseer.

Costa, M., & Angelis, L. D. (2010). *Model selection in hidden markov models : A simulation study* (Working paper No. 2010/7). http://amsacta.unibo.it/2909/

Crayen, C., Eid, M., Lischetzke, T., Courvoisier, D. S., & Vermunt, J. K. (2012). Exploring dynamics in mood regulation—mixture latent markov modeling of ambulatory assessment data. *Psychosomatic medicine*, *74*(4), 366–376.

Das, R., & Turkoglu, I. (2009). Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method. *Expert Systems with Applications*, *36*(3), pp 6635–6644.

Dias, J. G. (2006). Model selection for the binary latent class model: A monte carlo simulation. In *Data science and classification* (pp. 91–99). Springer.

Dias, J. G., & Vermunt, J. K. (2007). Latent class modeling of website users' search patterns: Implications for online market segmentation. *Journal of Retailing and Consumer Services*, *14*(6), 359–368.

Dias, J. G., Vermunt, J. K., & Ramos, S. (2009). Mixture hidden markov models in finance research. In *Advances in data analysis, data handling and business intelligence* (pp. 451–459). Springer.

Dias, J. G., Vermunt, J. K., & Ramos, S. (2015). Clustering financial time series: New insights from an extended hidden markov model. *European Journal of Operational Research*, *243*(3), 852–864.

Drott, M. C. (1998). Using web server logs to improve site design. *Proceedings of the 16th annual international conference on Computer documentation*, 43–50.

Du, J., Hu, Y., & Jiang, H. (2011). Boosted mixture learning of gaussian mixture hidden markov models based on maximum likelihood for speech recognition. *IEEE transactions on audio, speech, and language processing*, *19*(7), 2091–2100.

Durand, J.-B., & Guédon, Y. (2016). Localizing the latent structure canonical uncertainty: Entropy profiles for hidden markov models. *Statistics and Computing*, *26*(1-2), pp 549–567.

Eirinaki, M., Vazirgiannis, M., & Kapogiannis, D. (2005). Web path recommendations based on page ranking and markov models. *Proceedings of the 7th annual ACM international workshop on Web information and data management*, 2–9.

Elzinga, C. H. (2006). Sequence analysis: Metric representations of categorical time series. *Sociological methods and research*.

Fonseca, J. R. (2008). The application of mixture modeling and information criteria for discovering patterns of coronary heart disease. *Journal of Applied Quantitative Methods*, *3*(4), 292–303.

Frühwirth-Schnatter, S. (2001). Markov chain monte carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, *96*(453), 194–209.

Ghose, A., Goldfarb, A., & Han, S. P. (2013). How is the mobile internet different? search costs and local activities. *Information Systems Research*, *24*(3), 613–631.

Green, P. J., & Richardson, S. (2002). Hidden markov models and disease mapping. *Journal of the American statistical association*, *97*(460), 1055–1070.

Hahsler, M., Buchta, C., Gruen, B., Hornik, K., Johnson, I., Borgelt, C., & Hahsler, M. M. (2020). Package 'arules'.

Hamming, R. W. (1950). Error detecting and error correcting codes. *The Bell system technical journal*, *29*(2), 147–160.

Harte, D. (2006). Mathematical background notes for package "hiddenmarkov". *Statistics Re*.

Helske, S., & Helske, J. (2017). Mixture hidden markov models for sequence data: The seqhmm package in r. *arXiv preprint arXiv:1704.00543*.

Helske, S., Helske, J., & Eerola, M. (2018). Combining sequence analysis and hidden markov models in the analysis of complex life sequence data. In *Sequence analysis and related approaches* (pp. 185–200). Springer, Cham.

Hernando, D., Crespi, V., & Cybenko, G. (2005). Efficient computation of the hidden markov model entropy for a given observation sequence. *IEEE transactions on information theory*, *51*(7), pp 2681–2685.

Huang, P., Lurie, N. H., & Mitra, S. (2009). Searching for experience on the web: An empirical examination of consumer behavior for search and experience goods. *Journal of marketing*, *73*(2), 55–69.

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, *76*(2), 297–307.

Jamalzadeh, M. (2011). *Analysis of clickstream data* (Doctoral dissertation). Durham University.

Johnson, E. J., Moe, W. W., Fader, P. S., Bellman, S., & Lohse, G. L. (2004). On the depth and dynamics of online search behavior. *Management science*, *50*(3), 299–308.

Kohavi, R. (2001). Mining e-commerce data: The good, the bad, and the ugly. *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 8–13.

Lesnard, L. (2006). Optimal matching and social sciences.

Levenshtein, V. I., et al. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, *10*(8), 707–710.

Liu, B. (2007). *Web data mining: Exploring hyperlinks, contents, and usage data*. Springer Science & Business Media.

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 & Knowledge Engineering*, *61*(2), pp 304–330.

Lukociene, O., & Vermunt, J. K. (2009). Determining the number of components in mixture models for hierarchical data. In *Advances in data analysis, data handling and business intelligence* (pp. 241–249). Springer.

Luo, Y., & Stephens, D. A. (2021). Bayesian inference for continuous-time hidden markov models with an unknown number of states. *Statistics and computing*, *31*(5), 1–15.

Marin, J.-M., Mengersen, K., & Robert, C. P. (2005). Bayesian modelling and inference on mixtures of distributions. *Handbook of statistics*, *25*, 459–507.

Marino, M. F., & Alfo, M. (2020). Finite mixtures of hidden markov models for longitudinal responses subject to drop out. *Multivariate behavioral research*, *55*(5), 647–663.

McLachlan, G. J., & Basford, K. E. (1988). *Mixture models: Inference and applications to clustering* (Vol. 38). M. Dekker New York.

McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. John Wiley & Sons.

McLachlan, G. J., & Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

McQuarrie, A., Shumway, R., & Tsai, C.-L. (1997). The model selection criterion aicu. *Statistics & probability letters*, *34*(3), 285–292.

Melnykov, V. (2016). Model-based biclustering of clickstream data. *Computational Statistics & Data Analysis*, *93*, 31–45.

Moe, W. W. (2003). Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of consumer psychology*, *13*(1-2), 29–39.

Moe, W. W., Chipman, H., George, E. I., & McCulloch, R. E. (2002). A bayesian treed model of online purchasing behavior using in-store navigational clickstream. *Revising for 2nd review at Journal of Marketing Research*.

Moe, W. W., & Fader, P. S. (2004). Dynamic conversion behavior at e-commerce sites. *Management Science*, *50*(3), 326–335.

Montgomery, A. L., Li, S., Srinivasan, K., & Liechty, J. C. (2004). Modeling online browsing and path analysis using clickstream data. *Marketing science*, *23*(4), 579–595.

Muthen, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the em algorithm. *Biometrics*, *55*(2), pp 463–469.

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural equation modeling: A multidisciplinary Journal*, *14*(4), 535–569.

Olbrich, R., & Holsing, C. (2011). Modeling consumer purchasing behavior in social shopping communities with clickstream data. *International Journal of Electronic Commerce*, *16*(2), 15–40.

Paas, L. J., Vermunt, J. K., & Bijmolt, T. H. (2007). Discrete time, discrete state latent markov modelling for assessing and predicting household acquisitions of financial products. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *170*(4), pp 955–974.

Park, H.-S., & Jun, C.-H. (2009). A simple and fast algorithm for k-medoids clustering. *Expert systems with applications*, *36*(2), 3336–3341.

Park, Y.-H., & Fader, P. S. (2004). Modeling browsing behavior at multiple websites. *Marketing Science*, *23*(3), 280–303.

Pavlou, P. A., & Fygenson, M. (2006). Understanding and predicting electronic commerce adoption: An extension of the theory of planned behavior. *MIS quarterly*, 115–143.

Pohle, J., Langrock, R., van Beest, F. M., & Schmidt, N. M. (2017). Selecting the number of states in hidden markov models: Pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological and Environmental Statistics*, *22*(3), 270–293.

Rabiner, L. R. (1989). A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), pp 257–286.

Raftery, A. E. (1985). A model for high-order markov chains. *Journal of the Royal Statistical Society: Series B (Methodological)*, *47*(3), 528–539.

Ramaswamy, V., DeSarbo, W. S., Reibstein, D. J., & Robinson, W. T. (1993). An empirical pooling approach for estimating marketing mix elasticities with pims data. *Marketing Science*, *12*(1), 103–124.

Resul, D., Turkoglu, I., & Poyraz, M. (2007). Analyzing of system errors for increasing a web server performance by using web usage mining. *Istanbul University-Journal of Electrical & Electronics Engineering*, *7*(2), 379–386.

Richardson, S., & Green, P. J. (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, *59*(4), 731–792.

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*(5), 465–471.

Robert, C. P., Ryden, T., & Titterington, D. M. (2000). Bayesian inference in hidden markov models through the reversible jump markov chain monte carlo method. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *62*(1), 57–75.

Rudis, B., Simon, L., & Langel, T. (2021). Uaparserjs: Parse browser'user-agent'strings into data frames. r package version 0.1. 0.

Sarukkai, R. R. (2000). Link prediction and path analysis using markov chains. *Computer Networks*, *33*(1-6), 377–386.

Scott, S. L., & Hann, I.-H. (2006). A nested hidden markov model for internet browsing behavior. *Marshall School of Business*, 1–26.

Sismeiro, C., & Bucklin, R. E. (2004). Modeling purchase behavior at an e-commerce web site: A task-completion approach. *Journal of marketing research*, *41*(3), 306–323.

Smyth, P. (1997). Clustering sequences with hidden markov models. *Advances in neural information processing systems*, 648–654.

Smyth, P. (1999). Probabilistic model-based clustering of multivariate and sequential data. *Proceedings of the Seventh International Workshop on AI and Statistics*, 299–304.

Sparks, B. A., Perkins, H. E., & Buckley, R. (2013). Online travel reviews as persuasive communication: The effects of content type, source, and certification logos on consumer behavior. *Tourism Management*, *39*, 1–9.

Suchacka, G. (2014). Analysis of aggregated bot and human traffic on e-commerce site. *2014 Federated Conference on Computer Science and Information Systems*, 1123–1130.

Titus, P. A., & Everett, P. B. (1995). The consumer retail search process: A conceptual model and research agenda. *Journal of the Academy of Marketing Science*, *23*(2), 106–119.

Van de Pol, F., & Langeheine, R. (1990). Mixed markov latent class models. *Sociological methodology*, 213–247.

Vermunt, J. K., Langeheine, R., & Bockenholt, U. (1999). Discrete-time discrete-state latent markov models with time-constant and time-varying covariates. *Journal of Educational and Behavioral Statistics*, *24*(2), 179–207.

Vermunt, J. K., Tran, B., & Magidson, J. (2008). Latent class models in longitudinal research. *Handbook of longitudinal research: Design, measurement, and analysis*, 373–385.

Volant, S., Bérard, C., Martin-Magniette, M.-L., & Robin, S. (2014). Hidden markov models with mixtures as emission distributions. *Statistics and Computing*, *24*(4), pp 493–504.

Wei, J., Shen, Z., Sundaresan, N., & Ma, K.-L. (2012). Visual cluster exploration of web clickstream data. *2012 IEEE conference on visual analytics science and technology (VAST)*, 3–12.

Xu, C., Du, C., Zhao, G., & Yu, S. (2013). A novel model for user clicks identification based on hidden semi-markov. *Journal of network and computer applications*, *36*(2), 791–798.

Ypma, A., & Heskes, T. (2002). Automatic categorization of web pages and user clustering with mixtures of hidden markov models. *International Workshop on Mining Web Data for Discovering Usage Patterns and Profiles*, 35–49.

Yu, H., & Luo, H. (2011). A novel possibilistic fuzzy leader clustering algorithm. *International Journal of Hybrid Intelligent Systems*, *8*(1), 31–40.