

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

SCHOOL OF INFORMATION SCIENCES & TECHNOLOGY

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

Categorical time series

By

Vasiliki Daskalaki

M.Sc. Thesis

Submitted to the Department of Statistics

of the Athens University of Economics and Business

in partial fulfillment of the requirements for

the degree of Master of Science in Statistics

Athens, Greece

August 2023



ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ

Ανάλυση χρονολογικών σειρών για κατηγορικά δεδομένα

Βασιλική Δασκαλάκη

Διατριβή
Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα
Αύγουστος 2023

DEDICATION

To my friends and family
who were always there

ACKNOWLEDGEMENTS

I am deeply grateful to all those who have supported and guided me throughout my journey in completing this master's thesis on categorical time series. Their invaluable contributions have played a crucial role in shaping my research and enriching my academic experience.

I would like to express my sincere gratitude to my thesis advisor Dimitris Karlis, whose expertise, guidance, and support have been instrumental in every phase of this research.

I extend my appreciation to my friends and colleagues who have provided me with encouragement, fruitful discussions, and a supportive network. Their diverse perspectives have added an extra dimension to my research journey.

My heartfelt thanks also go to my family for their encouragement throughout my academic pursuits. Their constant belief in my abilities has been my driving force.

This research wouldn't have been possible without the collective efforts and contributions of everyone mentioned above. However, any errors or omissions are entirely my own.

Thank you all for being an integral part of this journey.

ABSTRACT

Categorical time series

by Vasiliki Daskalaki

August 2023

This thesis delves into the comprehensive exploration of categorical time series analysis, focusing on various tools and models that have been extensively researched in this domain. Categorical time series data, characterized by discrete states or categories, frequently arises in diverse fields. Understanding the underlying patterns and dependencies in such data is crucial for making informed decisions and predictions. The first part of the thesis presents an in-depth review of existing techniques and methodologies used for analyzing categorical time series. By examining the strengths and limitations of each tool, this research aims to provide a comprehensive overview of the state-of- the-art in categorical time series analysis. The second part of the thesis focuses on the investigation and application of Markov chains, Hidden Markov models, and their mixtures as powerful modeling techniques for categorical time series data. Markov chains are widely employed to model sequential dependencies between discrete states, while Hidden Markov models can capture latent or unobservable states underlying the observed categorical data. Additionally, mixtures of these models are explored, combining their strengths to create more flexible and expressive representations. The methodology is demonstrated by using real female-workers data.

ΠΕΡΙΛΗΨΗ

Ανάλυση χρονολογικών σειρών για κατηγορικά δεδομένα

Βασιλική Δασκαλάκη

Αύγουστος *2023*

Η παρούσα διατριβή εμβαθύνει στην ολοκληρωμένη διερεύνηση της ανάλυσης χρονοσειρών για κατηγορικά δεδομένα, εστιάζοντας σε ποικίλα εργαλεία και μοντέλα που έχουν ερευνηθεί εκτενώς σε αυτόν τον τομέα. Τα δεδομένα κατηγορικών χρονοσειρών, χαρακτηρίζονται από διακριτές καταστάσεις ή κατηγορίες και εμφανίζονται συχνά σε διάφορους τομείς. Η κατανόηση των υποκείμενων προτύπων και εξαρτήσεων σε τέτοια δεδομένα είναι κρίσιμης σημασίας για τη λήψη τεκμηριωμένων αποφάσεων και προβλέψεων. Στο πρώτο μέρος της διατριβής παρουσιάζεται μια εμπεριστατωμένη επισκόπηση των υφιστάμενων τεχνικών και μεθοδολογιών που χρησιμοποιούνται για την ανάλυση κατηγορικών χρονοσειρών. Με την εξέταση των πλεονεκτημάτων και των περιορισμών κάθε εργαλείου, η παρούσα έρευνα στοχεύει στην παροχή μιας ολοκληρωμένης επισκόπησης της κατάστασης προόδου στην ανάλυση κατηγορικών χρονοσειρών. Το δεύτερο μέρος της διατριβής επικεντρώνεται στην τη διερεύνηση και εφαρμογή των αλυσίδων **Markov**, των κρυφών **Markov** μοντέλων, και των μίξεών τους ως ισχυρές τεχνικές μοντελοποίησης για δεδομένα κατηγορικών χρονοσειρών. Οι αλυσίδες **Markov** χρησιμοποιούνται ευρέως για τη μοντελοποίηση διαδοχικών εξαρτήσεων μεταξύ διακριτών καταστάσεων, ενώ τα κρυφά **Markov** μπορούν να συλλάβουν λανθάνουσες ή μη παρατηρήσιμες καταστάσεις που υποκρύπτουν τα παρατηρούμενα κατηγορικά δεδομένα. Επιπλέον, διερευνώνται μίξεις αυτών των μοντέλων, συνδυάζοντας τα πλεονεκτήματά τους για τη δημιουργία πιο ευέλικτων και ευπερίστατων αναπαραστάσεων. Η μεθοδολογία αυτή καταδεικνύεται με τη χρήση πραγματικών δεδομένων γυναικών εργατριών.

Contents

1 Introduction **1**

1.1 Categorical time series 1

1.2 Software and packages 3

1.3 Exploratory data analysis for categorical time series 4

1.3.1 Descriptive statistics 5

1.3.2 Visual analysis 8

1.4 Outline of the thesis 13

2 Models for categorical time series **15**

2.1 Markov models 16

2.1.1 Markov Chain model 16

2.1.2 pth-order Markov model 17

2.1.3 Mixture Markov model 18

2.1.4 Variable length Markov model 20

2.1.5 Mixture transition distribution model 20

2.2 NDARMA 21

2.3 Hidden Markov Model 23

2.3.1 Extensions of Hidden Markov model 26

2.3.2 Mixture Hidden Markov model 26

2.4 Conditional random fields 27

3 Parameter estimation methods and algorithms **33**

3.1 Markov Chain model 33

3.1.1 Bayesian information criterion 35

3.2 Hidden Markov model 36

3.3 Expectation-maximization algorithm 40
3.3.1 EM algorithm for Hidden Maerkov model 42
3.4 Viterbi algorithm 45
3.5 Forward-backward algorithm 46
3.6 Mixture Hidden Markov model 50
3.7 Mixture Markov model 52

4 Applied analysis and results **53**
4.1 The dataset of the female workers 54
4.2 Exploratory data analysis 54
4.3 Model performance 57
4.3.1 Markov Chain model 58
4.3.2 Mixture Markov Chain models 59
4.3.3 Hidden Markov models 62
4.3.4 Mixture Hidden Markov models 65
4.4 Model selection 69

5 Conclusion and future directions **73**

Bibliography **77**

List of Tables

1.1 Table with Cohen's κ values from lag1 to lag 30 and critical value for user 350 of the msn dataset. 13

4.1 Frequency table of female workers dataset. 55
4.2 Frequency table and Shanon entropy measure of female workers dataset by age. 55
4.3 Markov model, order 1, estimated parameters for female workers dataset. 58
4.4 mixture Markov chain model, k=2 components, estimated parameters for female workers dataset. 61
4.5 Posterior counts and probabilities of 2-part mixture Markov model. 61
4.6 Classification table of 2-part mixture Markov chain model. 62
4.7 5-state Hidden Markov model, estimated parameters for female workers dataset. 63
4.8 Mixture 3-state/4-state Hidden Markov model, k=2 components, estimated parameters for female workers dataset. 67
4.9 Posterior counts and probabilities of the mixture 3-state/4-state Hidden Markov model. 68
4.10 Classification table of the mixture 3-state/4-state Hidden Markov model. 68
4.11 Log-likelihood, BIC and number of free parameters for the various models fitted on the female workers dataset. 70

List of Figures

1.1 Time series plot for record 4 of the female workers dataset (476 females, 6 categories). Non-informative for categorical time series. 9
1.2 Rate evolution plot for user 350 of the msn dataset (989818 users, 17 categories). 11
1.3 Cohen's κ plot for user 350 of the msn dataset (989818 users, 17 categories). 12

2.1 Directed graph of the basic Hidden Markov model. 23
2.2 Graph of the relationship between naive Bayes, logistic regression, HMMs, linearchain CRFs, generative models, and general CRFs, depicted in the book (Sutton, McCallum, et al., 2012). . . . 30

4.1 Barplots of each category by age for female workers dataset. . . . 56
4.2 Sequence of modal states for female workers. Category "Employed-PermJob" dominant in all life stages. 57
4.3 Ten most frequent sequences plot for female workers dataset. . . 58
4.4 Complete sequence plot for female workers dataset. 59
4.5 Markov chain plot for female workers dataset. 60
4.6 2-part mixture Markov chain plot for the female workers dataset. 62
4.7 5-state Hidden Markov model plot for female workers dataset. . . 64
4.8 5-state HMM plot for the most probable hidden state paths for the female workers dataset. 65
4.9 5-state HMM barplots for the most probable hidden states for ages 19 and 30 for the female workers dataset. 66

4.10 Visual representation of the mixture 3-state/4-state Hidden Markov model. 66
4.11 Stacked plots of most probable hidden states and observations of the mixture 3-state/4-state Hidden Markov model. 69
4.12 Plot BIC values vs the number of parameters for the various models fitted on the female workers dataset. 71
4.13 Stacked plot of most probable hidden states and observations of the 6-state Hidden Markov model. 72

Chapter 1

Introduction

1.1 Categorical time series

Time series models are widely applied in numerous scientific fields, including finance, computer science, epidemiology, and meteorology. However, the majority of research conducted on time series pertains to continuous data (Brockwell and Davis, 2002), and there is limited information available in scientific literature regarding time series for count data (McKenzie, 2003; Karlis, 2016), and even less on categorical data. This dearth of literature is not unexpected, given the numerous challenges associated with analyzing categorical time series data. Nevertheless, the growing demand for the utilization of such models, which has been brought about by advances in technology, has brought time series analysis on categorical data to the forefront of research.

Some very common applications of categorical time series are those of natural language processing, where categorical variables can be used to represent the parts of speech (Lawrence R Rabiner, 1989) or semantic categories of words in a text, on Web traffic where categorical variables can be used to represent the type of web traffic, such as different types of users, page views (Cadez et al., 2000), or geographic locations, or on EEG (electroencephalogram) data (Stoffer, Tyler, and Wendt, 2000). EEG data measures the electrical activity of the brain over time and is often used in neuroscience research. Categorical time series analysis can be used to identify patterns in the EEG signal, such as changes in frequency or amplitude, that may be associated with different cognitive or

neurological processes. Moreover, categorical time series can be applied in sociology, for understanding patterns of social inequality, such as changes in the gender wage gap or racial disparities in educational attainment. Finally, an extensive look at how categorical time series analysis even relates to biology and biological sequence analysis has been given by (Durbin et al., 1998).

Before starting getting a good grasp of the particularities that come along with time series analysis of this kind, it is imperative to undertake a thorough examination of the characteristics of categorical data and its potential impact.

A categorical time series stochastic process $\{X_t\}$ is formally characterized by the sequence of categorical values it generates: $X_t \in S$, where t indexes time and S is a finite set comprising d distinct categories, with $d \in \mathbb{N}$. In this context, the possible outcomes are arranged in a specific order, often either natural or lexicographical. For the scope of this thesis, our focus will be exclusively on discrete-time categorical time series, where t takes values in the set of natural numbers, $\mathbb{N}$.

A foundational assumption often made is the concept of **stationarity** within the investigation of categorical time series processes. Specifically, the sequence $(X_t)_{\mathbb{N}}$ is deemed stationary if the joint distributions of any finite selection of random variables $(X_{t_1}, ..., X_{t_n})$ remain invariant over time. More formally, this means that the joint distributions of $(X_{t_1}, ..., X_{t_n})$ with $t_1 < t_2 < \ldots < t_n$ and $(X_{t_1+k}, ..., X_{t_n+k})$ remain constant for all $k \in \mathbb{N_0}$.

In the context of weak stationarity, if $(X_t)_{\mathbb{N}}$ is considered weakly stationary, the marginal probabilities $P(X_t = s)$ remain constant for all t and for every category s within the set $S = s_1, s_2, \ldots, s_d$.

It is important to emphasize, as this introductory paragraph is concluded, that sequences produced by a nominal time series process exhibit a distinctive characteristic when compared to continuous data. Unlike continuous data, the outcomes of such sequences possess a qualitative nature that precludes meaningful ordering. While ordinal categorical data can be arranged based on a basic

or inherent order, as demonstrated by (Agresti, 2012), it is crucial to note that, in the context of this dissertation, no such inherent order will be presupposed for the variables.

1.2 Software and packages

The application of descriptive statistics, the generation of exploratory plots, and the fitting of statistical models for categorical time series analysis will be facilitated through the utilization of the versatile computing environment `R` (R Core Team, 2022). `R` serves as the cornerstone of all practical applications within this thesis, providing a robust framework to address the complexities inherent in categorical time series.

`R`, being an open-source language, is an ideal choice for this study due to its widespread accessibility and collaborative nature. Its open-source characteristic encourages a global community of users to contribute, fostering a continuous cycle of development and refinement. This makes `R` an attractive platform for cutting-edge research, where accessibility and adaptability are paramount.

In the context of this thesis, essential analytical tools are harnessed through specialized packages. Notably, the package `TraMineR` (**TraMineR**) plays a pivotal role in implementing fundamental exploratory analyses for categorical time series and sequences. With its comprehensive set of functions, `TraMineR` empowers researchers to explore intricate patterns, transitions, and trends within categorical data.

Additionally, the `ctsfeatures` package (Oriona and Fernández, 2023) stands as an indispensable asset in the toolkit, providing an array of analytical tools tailored to the nuances of categorical time series data. This package equips researchers with the capability to uncover latent features and nuanced behaviors hidden within the temporal sequences.

For the critical task of model fit and selection, the `seqHMM` package (J. Helske and S. Helske, 2023) is an ideal option, since the dataset of female workers that is to be analyzed, is consisted of a set of sequences. Through extensive utilization of the `seqHMM` package, this thesis explores Markov, Hidden Markov models, as well as mixtures thereof. The package's diverse features facilitate rigorous model assessment, enabling researchers to effectively capture the underlying dynamics of categorical time series data.

In essence, the combination of `R` 's computational power and the specialized packages —`TraMineR`, `ctsfeatures`, and `seqHMM`— forms the spine of this thesis's analytical journey. By harnessing these tools, a comprehensive exploration of categorical time series is embarked, unveiling intricate insights that pave the way for deeper understanding and knowledge advancement.

1.3 Exploratory data analysis for categorical time series

Exploratory data analysis (EDA) stands as a fundamental cornerstone in the realm of data analysis, serving as a critical process that unveils the hidden facets of datasets, enabling a comprehensive understanding of the underlying patterns, trends, and relationships. This significance is further amplified when delving into the realm of categorical time series data, where uncovering intricate dynamics and transitions is paramount for comprehending the complexities that drive systems and phenomena over time.

In the context of categorical time series, where observations are organized into discrete categories or classes, the exploration of data goes beyond mere numeric evaluation; it involves deciphering the evolution of states and transitions over temporal sequences. EDA offers a powerful toolkit to dissect such data, illuminating the behavior of discrete states and the shifts between them across

time steps. While the categorical nature of the data presents unique challenges, EDA serves as a guiding light, allowing us to navigate these intricacies and unlock meaningful insights.

1.3.1 Descriptive statistics

The role of descriptive statistics in EDA is undeniable. Measures such as transition probabilities, mode duration, and categorical frequency distributions serve as quantitative indicators that lend precision to the narrative. These statistics provide a framework to quantitatively assess the dominance of specific states, the likelihood of transitions, and the stability of categorical patterns. By quantifying categorical dynamics, a deeper comprehension of the system's behavior is attained, enabling more informed decision-making and hypothesis formulation.

The absence of a clear order within the variables' categories renders most common time series descriptive statistics unsuitable for categorical time series analysis. Measures of central tendency such as median, and moments such as mean and auto-covariance, which are commonly employed to perform a preliminary exploration of real-valued data, are not applicable to nominal categorical data. While measures of central tendency like quantiles may hold meaning for ordinal data, the investigation of categorical variables in their most extreme form, nominal, requires a different approach. When dealing with unordered categorical random variables, their location can be quantified by using the mode of their marginal distribution. However, it must be noted of course that there are situations where the mode of the marginal distribution is not uniquely determined, such as with uniform distribution.

Regarding the dispersion of categorical data, measures of dispersion play a pivotal role in the analysis of categorical time series. These measures provide

valuable insights into the variability and diversity inherent within the categorical data, offering a deeper understanding of the underlying patterns and dynamics. Some proposed measures can be found in literature such as the normalized Entropy (Shannon, 1948), the normalized Gini index (Rao, 1982) or the normalized Chebysev dispersion (Weiß, 2019). These measures are based on quantifying the uncertainty associated with the outcomes of a categorical random variable and its marginal properties. Their maximum value is taken in case the vector of the marginal probabilities $\mathbf{p} = (P(X = s_1), .., P(X = s_d))^\top \in [0,1]^d$ follows a uniform distribution and their minimum values in case of one-point distribution (Kvålseth, 2011). They incorporate the values of marginal probabilities and they are normed to $[0,1]$ for ease of interpretation.

Taking a closer look at Gini index and Entropy measures, let $s_k(\mathbf{p}) = \sum_{j=1}^{d} p_j^k$ and $k \in \mathbb{N}$, then the Gini index is defined as

$$v_g = \frac{d}{d-1}(1 - s_2(\mathbf{p})), \tag{1.1}$$

the Entropy measure is defined as

$$v_{En} = \frac{-1}{\log(d)} \sum_{i=1}^{d} p_i \log p_i, \tag{1.2}$$

with the convention that $0 \cdot \log 0 := 0$,

while the Chebycheff dispersion is given by the equation:

$$v_C = \frac{d}{d-1}(1 - max_i p_i). \tag{1.3}$$

High entropy indicates a diverse and less predictable series, while low entropy signifies a more stable and predictable pattern. By quantifying the information content and uncertainty, entropy aids in identifying periods of significant change or stability in the time series. Applying the Gini index in this context allows you to detect shifts in the distribution of categories, highlighting

potential shifts in patterns or anomalies, while Chebyshev dispersion can help the identification of time points with extreme or unusual category distributions. This is particularly useful for detecting outliers or unexpected fluctuations in the categorical data that may hold valuable information or require further investigation.

The asymptotic behavior of the sample Gini index and Entropy measure is portrayed in detail again in (Weiß, 2019).

What is more, (Weiß and Göb, 2008) presented the concept of serial dependence as the categorical counterpart to the signed autocorrelation in a real-valued process, and they proposed several "absolute" measures for serial dependence. The evaluation of serial dependence measures plays a crucial role in determining a suitable model for a stationary categorical process and accurately estimating its model parameters. For example, to measure the degree of signed serial dependence, a type of Cohen's κ may be used:

$$\kappa(h) = \sum_{i=0}^{d} \frac{p_{ii(h)} - p_i^2}{1 - s_2(\mathbf{p})} \in \left[\frac{-s_2(\mathbf{p})}{1 - s_2(\mathbf{p})}; 1\right] \tag{1.4}$$

for lags $h \in \mathbb{N}$ where $p_{ij} = P(X_t = i, X_{t-h} = j), i, j = 1, .., d$. When κ equals to zero indicates serial independence for lag h while the strongest degree of positive (negative) dependence is indicated if the event $X_{t-h} = i$ is necessarily followed by $X_t = i (X_t \neq i)$. Equivalently, with the use of conditional probabilities $p_{i|j}(h) = P(X_t = i | X_{t-h} = j) = \frac{p_{ij}}{p_j}$, X_t and X_{t-h} are perfectly dependent if $p_{i|i}(h) = 1$ and perfectly independent if $p_{i|i}(h) = 0$.

To measure unsigned serial dependence, Cramer's v as well as Goodman and Kruskal's τ are considered with range $[0, 1]$ and serial independence and dependence is indicated by the boundaries. Cramer's v is given by the formula:

$$v(k) = \sqrt{\frac{1}{d-1} \sum_{i,j=0}^{d} \frac{(p_{ij}(k) - p_i p_j)^2}{p_i p_j}} \tag{1.5}$$

while Goodman and Kruskal's τ:

$$\tau(k) = \sqrt{\sum_{i,j=0}^{d} \frac{(p_{ij}(k) - p_i p_j)^2}{p_j(1 - s_2(\mathbf{p}))}} \tag{1.6}$$

Incorporating these measures of serial dependence into the analysis of categorical time series data, a deeper understanding of how categorical variables interact and influence each other across successive time points is gained. These measures provide a solid foundation for unraveling complex temporal patterns and dependencies.

1.3.2 Visual analysis

Visual analysis emerges as an indispensable aspect of EDA for categorical time series data. Visualizations not only provide an immediate impression of the data structure, but they also present a platform to unearth patterns, trends, and anomalies that might otherwise remain concealed. There have been developed a few techniques that offer an elegant means to depict the temporal evolution of categories and their interconnections. By translating raw data into visual narratives, we are empowered to discern the story encoded within, comprehend the sequences of events, and appreciate the dynamics at play.

When dealing with real-valued time series, certain diagrams, such as the time series plot, play a significant role in distinguishing patterns within the data, while other diagrams, such as the auto-correlation plot, aid in studying serial dependence. However, nominal data lack order and do not possess moments such as mean or auto-covariance, the auto-correlation function cannot be computed, rendering these types of plots meaningless or unattainable, as it becomes obvious in Figure 1.1. The vertical placement of categories on a time series plot can be misleading, as it might imply a continuous progression that doesn't exist. While trends might exist within categorical data, they are often

not as easily represented through continuous lines. For ordinal data, the time series plot is still possible to be created and interpreted somehow, but as it has been restated, for this thesis, only unordered data will be investigated.

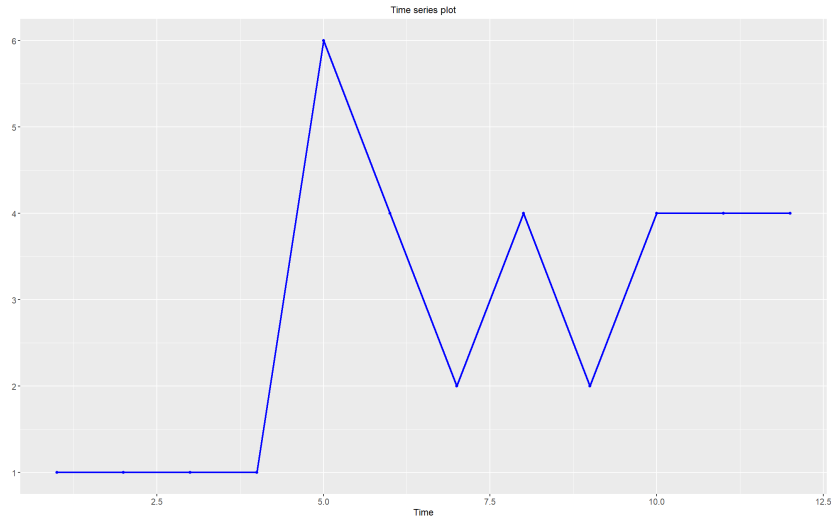


FIGURE 1.1: Time series plot for record 4 of the female workers dataset (476 females, 6 categories). Non-informative for categorical time series.

When dealing with categorical random variables that have a nominal range, alternative plots have been explored. For example stacked bar charts are often utilized to compare the relative frequency of each category at different points in time. Each of the time series is represented by stacked boxes that are colored according to the state at the successive positions. Another useful visual technique that can be conducted, is the symbol-by-symbol representation of the data. The resulting bars are put above each other to vertically align the positions. For each case, the individual longitudinal succession of states is depicted, as well as, through the length of each color segment, the duration spent in each successive state may be investigated. Unfortunately, for a large number of sequences or large length time series, this type of plots are difficult to interpret as one can hardly distinguish trends and patterns other than maybe the run of the symbols, if it is long or short (Weiß, 2008).The so called sequence plots, as well as

the stacked bar charts are going to be utilized extensively in Chapter 4 of this thesis.

Following the same line of reasoning, heatmaps are used to visualize the frequency of each category over time using a color- coded matrix. This can be useful for identifying outliers and unusual values. Heatmaps constitute an integral visualization approach especially for analyzing DNA sequences. They help researchers identify patterns, motifs, and relationships within the sequences and can be integrated with other genomics data to gain a deeper understanding of the functional aspects of DNA. In heatmaps the data is displayed in a grid where each row represents a gene and each column represents a sample. The colour and intensity of the boxes is used to represent changes (not absolute values) of gene expression.

Another popular visual tool are the state transition diagrams, also known as state diagrams or state transition diagrams. They are graphical representations that illustrate the transitions between different states or categories in a system over time. These diagrams are particularly useful for visualizing the sequence of events or transitions in a categorical time series data. This type of graphs can be also very assisting for the description of Markovian models and will be in effect fully in Chapter 4.

Another interesting tool that has been suggested throughout the bibliography is the rate evolution plot, which may provide information about the stationarity of a time series. Firstly, data need to be binarized. When working with categorical time series data, **binarization** is a common technique used to convert the categorical values into binary representations. Binarization essentially creates $y_{t,i} = \mathbb{1}_{\{X_t = s_i\}}$ binary variables or indicators for each category i, indicating the presence or absence of that category at each time point t. By binarizing the categorical time series, you can transform the data into a format that is compatible with various modeling techniques. After the binarization process, by plotting the cumulated sums, Equation 1.7, of binarized categorical

data

$$c_{t,i} = \sum_{t=1}^{T} y_{t,i} \tag{1.7}$$

where $i = 1, .., d$ the nominal variable categories, one may investigate through the linearity of the graph slopes the existence of stationarity for the marginal distributions (Ribler, 1997). In the case of a stationary and moderately serially dependent sequence, the graphs exhibit an approximately linear relationship with respect to t. Figure 1.2 is an example of non-stationarity for the marginal distributions.

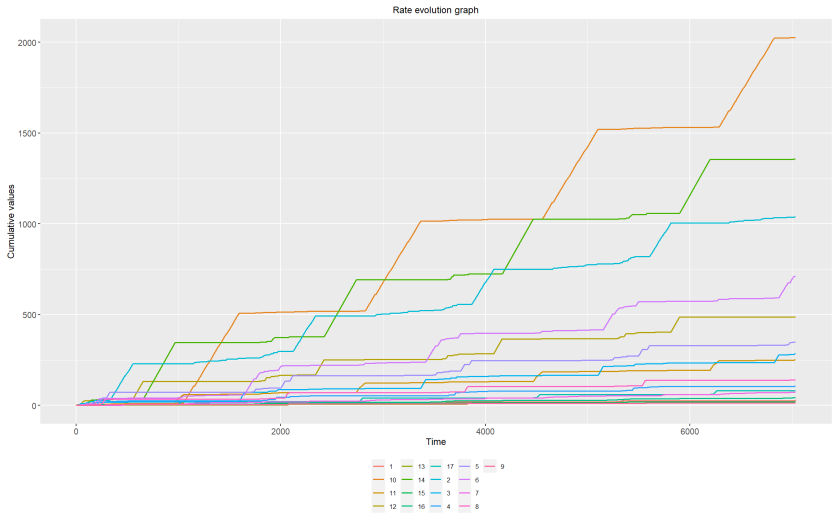


FIGURE 1.2: Rate evolution plot for user 350 of the msn dataset (989818 users, 17 categories).

In analogy to the case of the autocorrelation plot which is used to control serial dependence for continuous data, the need to use signed serial dependence measures for categorical data arises in order to produce a similar plot. Measures of serial dependence play a crucial role in both determining a suitable model for a stationary categorical process and in the estimation of its model parameters. Cohen's κ and its variates constitute serial dependency measures for different lags with extended range $[-1, 1]$. When these measures receive a value of zero, it is a sign of independence, when they receive a value of one, it indicates positive dependence and minus one, negative dependence. In Figure

1.3 and Table 1.1, a Cohen's κ serial dependence plot and its values for user 350 of the msn dataset is displayed, where positive serial dependence, which means the long residence of the process in a particular state, is obvious even for very high lags.

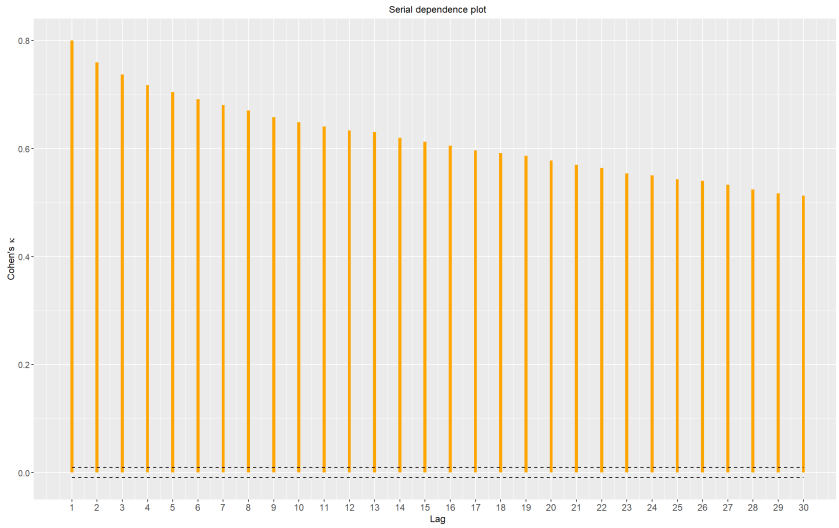


FIGURE 1.3: Cohen's κ plot for user 350 of the msn dataset (989818 users, 17 categories).

The dashed lines in the plot are calculated asymptotically for the i.i.d case with the approximation:

$$\sqrt{\frac{T}{V(\mathbf{p})}}(\kappa(h) + \frac{1}{T}) \sim N(0,1), \tag{1.8}$$

where T is the series length, $\hat{p}$ is the vector of estimated marginal probabilities of the series and $V(\mathbf{p}) = 1 - \frac{1+2\sum_{i=1}^{s} p_i^3 - 3\sum_{i=1}^{s} p_i^2}{(1-\sum_{i=1}^{s} p_i)^2}$.

A few other type of plots have been studied in the literature, such as IFS circle graphs, spectral envelope (Stoffer, Tyler, and McDougall, 1993), pattern histograms or the more delicate categorical control charts but they are either too intricate to build or the information they provide heavily depends on the data. Unfortunately to the knowledge of the author of this thesis, there is not yet a type of plot for categorical time series that would be as informative as

0.80	0.76	0.74	0.72	0.70
0.69	0.68	0.67	0.66	0.65
0.64	0.63	0.63	0.62	0.61
0.60	0.60	0.59	0.59	0.58
0.57	0.56	0.55	0.55	0.54
0.54	0.53	0.52	0.52	0.51
criticalvalu*e*:	0.01			

TABLE 1.1: Table with Cohen's κ values from lag1 to lag 30 and critical value for user 350 of the msn dataset.

the classic time series plot for real-valued time series which displays multiple insight at once.

1.4 Outline of the thesis

In Chapter 1, an introduction to the categorical time series was given. A comprehensive exploration was undertaken to understand the impact of the categorical nature of observations on exploratory data analysis. This exploration delved into the diverse spectrum of measures and graphic tools that have garnered research attention, illuminating their role in unraveling the intricate dynamics inherent in categorical time series.

Moving forward, Chapter 2 begins a browsing across the models landscape that are tailored for categorical time series analysis. Starting with the fundamental Markov models, such as Markov Chains and pth-order Markov models, it extends to embrace the nuances of more intricate models like Mixture Markov, Variable Length Markov, and Mixture Transition Distribution models. The chapter also also clarifies the capabilities of the NDARMA model and explores the realm of Hidden Markov models (HMMs), including their variants like Mixture Hidden Markov models. The discussion culminates with a glimpse into Conditional Random Fields and their relevance in modeling complex categorical dependencies.

Within Chapter 3, the focus narrows down to the methodology underlying parameter estimation. It unravels a set of algorithms designed to tease out the latent characteristics of the data. This includes an exploration of the Expectation- Maximization (EM) algorithm and its application in Hidden Markov models, as well as the Viterbi and Forward-Backward algorithms that straddle the ground of likelihood estimation and latent state inference. The chapter further unpacks the intricacies of parameter estimation for Mixture Hidden Markov and Mixture Markov models and how Markov Chain models can be perceived as special cases of Hidden Markov models.

The core of the thesis lies in Chapter 4, where the theoretical frameworks and methodologies converge with practical application. It commences with an in-depth examination of a dataset portraying categorical time series of female workers, setting the stage for comprehensive exploratory data analysis. The chapter further extends to scrutinize the performance of diverse models, ranging from Markov Chain to Hidden Markov and Mixture Hidden Markov models. Through model comparisons and evaluations, it seeks to uncover the models' effectiveness in capturing the essence of the data and the dynamics it embodies.

In Chapter 5, the insights garnered from the analysis are condensed and highlights their implications. It offers a concise summary of the findings, underscoring the contributions made to the field of categorical time series analysis. Additionally, the chapter paves the way for future research endeavors, suggesting potential avenues for expanding the horizons of knowledge in this realm.

Chapter 2

Models for categorical time series

Models for categorical time series differ in their assumptions and approaches due to the nature of the data they handle in comparison with the more common handled real-valued time series. The discrete and unordered categories that categorical time series models deal with, affect the mathematical representations and assumptions made by the models. Categorical time series models often assume a discrete probability distribution for each categorical state, while real-valued time series models typically assume a continuous distribution, such as Gaussian or exponential. The assumptions made in each case guide the model's behavior and parameter estimation techniques. Also these often focus on capturing dependencies between categorical states, such as transitions between different classes. In contrast, real-valued time series models typically emphasize temporal dependencies and trends in the continuous numerical values. Finally, forecasting in categorical time series often involves predicting the next categorical state, while real-valued time series forecasting usually aims to estimate the numerical value at the next time point. Interpretation of results also differs, as categorical models may provide insights into class transitions or trends, while real-valued models may offer insights into magnitude, trends, and continuous dynamics.

Here are a few prominent models commonly used for analyzing categorical time series data including Markov models, NDARMA, Hidden Markov models and Conditionl random fields.

2.1 Markov models

2.1.1 Markov Chain model

Markov chains were initially employed by (Andrey Andreyevich Markov, 1913; Andrey Andreevich Markov, 2006) for the purpose of predicting the forthcoming letter's nature, whether it would be a vowel or a consonant, in Pushkin's work "Eugene Onegin". A Markov chain model for categorical time series is a statistical framework that assumes the future state of a categorical variable depends only on its current state. If $(X_t)_{\mathbb{Z}}$ is a categorical process with state space $S = s_1, .., s_d$, this can be expressed as:

$$P(X_t = x | X_{t-1} = x_{-1}, X_{t-2} = x_{-2}, ...) = P(X_t = x | X_{t-1} = x_{-1}) \tag{2.1}$$

It provides a simple and intuitive approach to analyze and predict the behavior of categorical data over time. In the context of categorical time series, the system's states represent different categories, such as weather conditions (sunny, rainy, cloudy) or market trends (bullish, bearish, stagnant).

Markov chain models provide a straightforward representation of the probabilistic dependencies between categorical states over time, making them accessible to a wide range of users. The transition probabilities between states are typically represented by a transition matrix, where each element represents the probability of transitioning from one state to another. By analyzing the transition matrix, one can gain insights into the dynamics and patterns of state transitions in the time series. One key application of this model is generating future predictions or forecasts. Given the current state, the model can calculate the probabilities of transitioning to different states in the next time step. This allows for the generation of likely future scenarios for the categorical time series. Some basic restrictions encountered in relation with Markov Chain models are

the following. Precisely because Markov chain models assume that the future state depends only on the current state, that they disregard any long term dependencies or patterns in the time series. This limitation makes it challenging to capture complex relationships or trends that span multiple time steps. In addition, they often assume that the transition probabilities between states remain constant over time. That is, for any $i, j \in S$, the transition probabilities $P(X_t = i|x_{t-1} = j) = p_{i|j}$ do not vary with time for all $t \in \mathbb{N}$, called the **homogeneity** assumption. However, in real world scenarios, these probabilities may change due to external factors or evolving conditions. Failing to account for such variations can lead to inaccurate predictions and modeling assumptions.

Finally, the predictions and behavior of a Markov chain model can be highly sensitive to the initial conditions or starting state. Small changes in the initial state can have a significant impact on subsequent state transitions, which can lead to divergent outcomes or unstable predictions.

Overall, while Markov chain models offer simplicity and forecasting capabilities for categorical time series, they have limitations in capturing long-term dependencies, handling evolving transition probabilities, accommodating large state spaces, and being sensitive to initial conditions. These factors should be considered when applying the model and interpreting its results.

2.1.2 pth-order Markov model

A pth-order Markov process, where the process has memory of the previous p states:

$$P(X_t = x|X_{t-1} = x_{-1}, X_{t-2} = x_{-2}, ...) = P(X_t = x|X_{t-1} = x_{-1}, X_{t-2} = x_{-2}, .., X_{t-p} = x_{-p}), \tag{2.2}$$

is able to capture long-term dependencies and compensate for the analogous weakness of the basic Markov chain because the transition probabilities are

based on a sequence of previous states, rather than just the immediate preceding state. Let it be remarked here, that a p_{th}-order Markov process can always be transformed into a first- order one, into a Markov chain. By considering a longer history of states, higher order Markov models can potentially improve the accuracy of predictions compared to simpler models. While certain properties may present greater challenges in establishing, analyzing a higher-order Markov chain does not require fundamentally new theories compared to analyzing a first-order chain. But a pth-order Markov process comes with a high cost of exponentially increasing parameters $(d-1)d^p$, that rapidly become too large to be estimated efficiently, or even identifiably, with datasets of the sizes typically encountered in practice. Estimating transition probabilities in higher order models can be computationally intensive, especially when dealing with a large number of states and high-order dependencies and the memory requirements for storing transition probabilities and other model parameters also increase.

2.1.3 Mixture Markov model

A mixture Markov model (MMM) is a statistical model commonly used for analyzing categorical time series data (Cadez et al., 2000). It combines the principles of Markov models and mixture models to capture the complex dynamics and heterogeneity present in the data. An MMM assumes that the observed categorical states are governed by a finite mixture of underlying Markov chains. As pointed out in (Frühwirth-Schnatter, 2006) the prominence of finite mixture modelling is greater than ever. Many important statistical topics like clustering data, outlier treatment, or dealing with unobserved heterogeneity involve finite mixture models in some way or other. The MMM assumes that the observed categorical states are generated from a mixture of these Markov chains. Each component in the mixture model corresponds to a different Markov chain with

its own set of transition probabilities. It involves estimating several parameters. These include the mixing proportions, which represent the probabilities of belonging to each component in the mixture model, and the transition probabilities, which describe the probabilities of transitioning between states within each Markov chain. The mixture Markov model is particularly useful when there is heterogeneity or variability in the categorical time series data. It allows for the identification and characterization of different subgroups or classes within the data, each following their own Markov chain dynamics. These type of models pose some challenges though. MMMs can be computationally demanding and require estimation of numerous parameters. As the number k of mixture components and the complexity of the underlying Markov chains increase, the computational burden and the risk of overfitting also rise. Estimating $kd(d-1)+(k-1)$ independent parameters accurately becomes challenging, especially with limited data or large- scale datasets. Also, a critical aspect of MMMs is determining the appropriate number of mixture components to include in the model. Selecting the optimal number of components is often subjective and relies on statistical criteria, such as information criteria or model selection techniques. An incorrect choice of the number of components can lead to underfitting or overfitting the data, impacting the model's performance. An interesting extension of the basic MMM has been applied by (Frühwirth-Schnatter and Pamminger, 2010). Their paper covers two methods for model-based clustering of categorical time series using time-homogeneous first-order Markov chains. The first approach, known as Markov chain clustering, involves setting the individual transition probabilities to a group-specific transition matrix. The second approach, referred to as Dirichlet multinomial clustering, introduces a new method where the rows of individual transition matrices differ from the group mean and follow a Dirichlet distribution with unknown group-specific hyperparameters.

2.1.4 Variable length Markov model

In order to restrain the number of parameters that need to be estimated, in contrast to the aforementioned models, more parsimonious Markov models have been researched. Such a case is the variable length Markov model (VLMM) (Bühlmann and Wyner, 1999). Unlike Markov models with fixed order p, where each random variable in a sequence is dependent on a fixed number of preceding random variables, VLMM models allow for the number of conditioning random variables to vary based on the observed sequence realization. The principle behind this category of models is to investigate all possible paths of pth order Markov chain and combine branches that exhibit similar probability distributions. The maximal length of memory required in the model is determined by the order p, but the model can adapt to shorter memory lengths based on the observed past. This flexibility is why it is referred to as a "variable length" Markov model. VLMMs are particularly useful when modeling sequences with long-range dependencies or when the optimal order of the Markov model is unknown or variable. By allowing variable-length dependencies, VLMMs can capture complex patterns and dependencies in the data without making assumptions about a fixed order. A recent presentation of an R-based package for the computation of this type of models can be found at (Mächler and Bühlmann, 2004).

2.1.5 Mixture transition distribution model

Another type of model used alternatively to provide more parsimonious variants of the full Markov models is the pth- order mixture transition distribution (MTD(p)) model of (A. E. Raftery, 1985). Since then it has been generalized and successfully applied to a range of situations (Berchtold and A. Raftery, 2002). An extensive research and enrichment of the MTD model has been more recently conducted by (W. Ching, E. Fung, and M. Ng, 2003; Wai Ki Ching, Eric S

Fung, and Michael K Ng, 2004; W.-K. Ching, Michael K Ng, and Eric S Fung, 2008).

In the basic MTD(p) model, the transition probabilities between states are represented as a mixture of transition matrices:

$$P(X_t = x|X_{t-1} = x_{-1}, X_{t-2} = x_{-2}.., X_{t-p} = x_{-p}) = \Sigma_{k=1}^{p} \lambda_k q_{x|x_{-k}} \tag{2.3}$$

where $q_{x|x_{-k}}$ are the probabilities of an d*d transition matrix Q and λ_k is the weight parameter associated with lag k. This model has only $d(d-1)+(p-1)$ independent parameters and each additional lag adds only one additional parameter. Even if this is similar to an autoregressive (AR) model, the modeling is based on a transition probability matrix. The mixture weights determine the contribution of each context or history to the overall transition probability. The MTD model allows for a more flexible and nuanced representation of the underlying dynamics in categorical time series data compared to fixed-order Markov models. Some limitations that should be considered regarding this model is model complexity, since depending on the number of states and mixture components, the MTD model can become complex and computationally intensive. Like any statistical model, the MTD model requires a sufficient amount of data to accurately estimate its parameters and capture meaningful patterns. Finally, as with many complex models, interpreting the learned parameters and patterns in the MTD model might require domain knowledge and additional analysis.

2.2 NDARMA

(Jacobs and Lewis, 1983) introduced the concept of "new" discrete ARMA (NDARMA) models. These models incorporate a random mixture to create a dependence

structure similar to ARMA models. The NDARMA model can be seen as the categorical equivalent of the traditional autoregressive moving average (ARMA) model. On one hand, the NDARMA model is particularly suitable for scenarios involving an unordered range of categories. On the other hand, it shares certain characteristics with the conventional ARMA model. Specifically, the serial dependence structure in the NDARMA model shows similarities to the autocorrelation structure observed in ARMA processes. Various formulations exist for constructing these models, offering different approaches to capture the desired characteristics. NDARMA model for categorical time series data provides with a nice interpretability which, in addition, is parsimonious. However, this model only can capture positive autocorrelation as well as positive parameters.

Let $(X_t)_{\mathbb{Z}}$ be a stationary process taking values in S. Let $(\epsilon_t)_{\mathbb{Z}}$ be an i.i.d. discrete valued innovation process, such that ϵ_t is independent of $(X_s)_{,s<t}$ with mean $\mu_e = E(\epsilon_t)$ and variance $\sigma_\epsilon^2 > 0$. Let the i.i.d. multinomial random vectors

$$(\alpha_{t,1}, ..., \alpha_{t,p}, \beta_{t,1}, ..., \beta_{t,q}) \sim MULT(1; \Phi), t \in \mathbb{Z}$$

then $(X_t)_{\mathbb{Z}}$ is said to be a NDARMA(p, q) process if it follows the recursion

$$X_t = \Sigma_{i=1}^{p} \alpha_{t,i} X_{t-i} + \Sigma_{j=0}^{q} \beta_{t,j} e_{t-j}.$$

The cases $q = 0$ and $p = 0$ are referred to as a $DAR(p)$ process and DMA(q) process respectively. It should be noted that only one variable among $(\alpha_{t,1}, ..., \alpha_{t,p}, \beta_{t,1}, ..., \beta_{t,q})$ can take the value 1, while all others are set to 0. As a result, the NDARMA recursion indicates that each observation X_t selects either a past observation from $X_{t-1}, ..., X_{t-p}$ or one of the previous innovations $\epsilon_t, ..., \epsilon_{t-q}$. This mechanism leads to the stationary marginal distribution of X_t, $\pi_{i \forall i \in S}$ being identical to that of ϵ_t. NDARMA(p, q) model has $d + p + q$ parameters.

(Jacobs and Lewis, 1983) derived a few first properties of the model as e.g. Yule Walker equations and autocorrelation structure. Due to its nice structure,

(Weiß, 2013) extends the analysis of the model to further stochastic properties and dependency measures, e.g. signed and unsigned serial dependence measures as Cohen's κ and Cramer's v.

2.3 Hidden Markov Model

A hidden Markov model (HMM) is a particular kind of dependent mixture. It can be actually be examined as an extension of an independent mixture, overcoming the independence limitation by allowing for serial dependence in the parameter process through a Markov chain. (Zucchini and MacDonald, 2009) present an extensive examination of HMMs in their book and two rich chapters for HMMs for discrete and categorical data are offered in the book of (Weiß, 2018). Even earlier (Juang and Laurence R Rabiner, 1991) had published an analytical description of them, an inference map around them and parameter estimation techniques for them. In (Frühwirth-Schnatter, Celeux, and Robert, 2019) can be found an extensive look on HMMs applied in economics. Hidden Markov modelling have been applied in various longitudinal settings; for accounting for measurement error and unobserved heterogeneity, for finding latent sub-populations and for detecting true unobservable states. A graphical representation of a HMM can be seen in Figure 2.1.

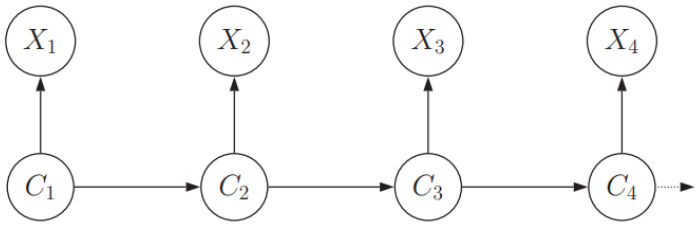


FIGURE 2.1: Directed graph of the basic Hidden Markov model.

The structure displayed in the graph for $(X_t)_{\mathbb{N}}$ is described by two basic parts. The unobserved part of the recurring states C_t which represent a Markov

chain and they therefore follow the Markov dependence structure and the observations X_t which are only affected from the current state the process is in. More specifically in terms of probabilities:

$$P(C_t|C_{t-1}, C_{t-2}, ..) = P(C_t|C_{t-1}),, t \in \mathbb{N} \tag{2.4}$$

$$P(X_t|X_{t-1}, X_{t-2}, ...C_t, C_{t-1}, ..) = P(X_t|C_t), t \in \mathbb{N}. \tag{2.5}$$

The t_{th} conditional probability of the bivariate discrete valued process $(X_t, C_t)_{\mathbb{N}}$ accounted for HMMs splits into:

$$P(X_t, C_t|X_{t-1}, .., C_{t-1}, ..) = P(X_t|X_{t-1}, .., C_t, ..)P(C_t|X_{t-1}, .., C_{t-1}, ..) = P(X_t|C_t)P(C_t|C_{t-1}) \tag{2.6}$$

So a discrete first order hidden Markov model for a single time series or sequence is characterized by the following:

- observed state sequence $\mathbf{x} = (x_1, x_2, ..., x_T)$ with state space $S = \{s_1, s_2, ..s_d\}$.
- hidden state sequence $\mathbf{c} = (c_1, c_2, ..., c_T)$ with state space $C = \{1, 2, .., d_c\}$.

Only homogeneous HMMs are going to be considered for this thesis, meaning that the the transition probabilities between the hidden states remain unchanged through different time-points. Matrix A of dimension $d_c * d_c$ includes the transition probabilities between states:

$$a_{rc} = P(C_t = c|C_{t-1} = r) \forall c, r \in C. \tag{2.7}$$

The observation process in a categorical HMM which specifies the probabilities of observing each category given the corresponding hidden state are represented by an emission matrix B of size $d * d_c$ and the emission probabilities:

$$p(s|c) = P(X_t = s|C_t = c), s \in S, c \in C \tag{2.8}$$

The state-dependent distributions $p(\cdot|c)$, which are invariant over time, where $P(X_t = s|C_t = c) = p(s|c)\forall t$, are represented as categorical distributions with d parameters. Therefore, in total, we have $d_c(d_c - 1)$ parameters for the hidden states, along with $d_c(d - 1)$ parameters associated with the observations. In relation with the choice of the initial distribution for C_1 and its probabilities, it either introduces extra model parameters or relies on a stationarity assumption.

In Rabiner's seminal work (Lawrence R Rabiner, 1989), the concept of hidden Markov models was introduced with a focus on three fundamental challenges:

Likelihood Estimation: In the context of an HMM represented as $\theta = (\pi, A, B)$ and an observation sequence denoted as X, the task is to compute the likelihood, $L(X|\theta)$.
Decoding: Given an observation sequence X and an HMM characterized by $\theta = (\pi, A, B)$, the objective is to determine the optimal sequence of hidden states, C, that best explains the observations.
Parameter Learning: When provided with an observation sequence X and the set of states within the HMM, the goal is to learn the parameters π, A and B that define the HMM.

These problems, as outlined by Rabiner, lay at the core of hidden Markov models and continue to be pivotal in their analysis and application and will be further discussed in Chapter 3.

It should be noted that the same HMM can be also fit for multiple subjects. Instead of one observed sequence x we have N sequences as $X = (\mathbf{x_1}, ..., \mathbf{x_N})^T$, where the observations $\mathbf{x_i} = (x_{i1}, ..., x_{iT})$ of each subject i take values in the observed state space. Observed sequences are assumed to be mutually independent given the hidden states. The observations are assumed to be generated by the same model, but each subject has its own hidden state sequence

$\mathbf{c_i} = (c_{i1}, ..., c_{iT})$.

If there is a (non-zero) probability of transitioning between any two latent states of an HMM, it is called a fully connected or ergodic HMM (Jurafsky and Martin, 2008). Sometimes, however, we have HMMs in which many of the transitions between states have zero probability. For example, in left-to-right (also called Bakis) HMMs, the state transitions proceed from left to right, no transitions go from a higher-numbered state to a lower-numbered state (or, more accurately, any transitions from a higher-numbered state to a lower numbered state have zero probability). Bakis HMMs are generally used to model temporal processes like speech.

2.3.1 Extensions of Hidden Markov model

There are various ways to extend the fundamental HMM (Chapter 8 in Zucchini and MacDonald, 2009). For instance, one possible approach involves allowing the state process to adhere to a higher-order Markov model. This extension may incorporate additional parametric assumptions, similar to the the MTD(p) models. Alternatively, one can consider augmenting the model by including covariate information or introducing additional dependencies at the observation level. Regression type models for categorical time series have been discussed by (Fahrmeir and Kaufmann, 1987), (Kaufmann, 1987), (Fokianos and Kedem, 2003).

2.3.2 Mixture Hidden Markov model

Another possible extension of the HMM is the mixture of HMMs (MHMMs) (S. Helske and J. Helske, 2017). The mixture of HMMs naturally enables clustering of categorical time series data. Each component in the mixture corresponds to a cluster, allowing the identification of different patterns or states within the data. These models allow the consideration for the time-constant unobserved

heterogeneity in the population and based on the assumption that the population is composed of unobservable groups. It provides flexibility in modeling diverse data sources. Different HMMs in the mixture can be tailored to handle specific patterns or behaviors in the data, making the model more adaptable to various scenarios. It is important to underline though that model selection should be carried on carefully since selecting the appropriate number of components (HMMs) in the mixture is not always straightforward (McLachlan, Lee, and Rathnayake, 2019). Determining the optimal number of hidden states can be challenging, and using too many components may lead to overfitting, while using too few may result in oversimplification and loss of important patterns. If the number of parameters of a HMM mixture grows quite a lot, the idea of parsimonious models for categorical time series is fading away. Closing, it should be mentioned that fitting a mixture of HMMs can be computationally demanding, especially when dealing with a large number of time series or complex data patterns. It may require powerful computational resources for training and inference.

2.4 Conditional random fields

Conditional Random Fields (CRFs) are a popular probabilistic graphical model that have found extensive applications in sequence labeling and structured prediction tasks. In this section, we explore the application of CRFs to model categorical time series, which involve sequential data with discrete categorical values. We discuss the formulation of CRFs for categorical time series, the choice of features, and the inference techniques involved.

A Conditional Random Field is a probabilistic model that captures the conditional distribution of a sequence of labels given an input sequence. For categorical time series, let $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T\}$ be the observed input sequence of

length T, where $\mathbf{x}_t$ represents the categorical value at time step t. The corresponding label sequence is denoted as $\mathbf{Y} = \{y_1, y_2, \ldots, y_T\}$.

The goal is to model the conditional distribution $p(\mathbf{Y}|\mathbf{X})$ using a CRF.

CRFs model the conditional distribution using energy functions, which quantify the compatibility between labels and observations. These energy functions capture both the relationships between adjacent labels (transition potentials) and the relationships between labels and observations (emission potentials).

For a pairwise CRF, the energy function can be defined as:

$$E(y_{t-1}, y_t, x_t) = \theta_{y_{t-1}, y_t, x_t} \tag{2.9}$$

Here, y_{t-1} and y_t are adjacent labels, and x_t is the observed categorical value. $\theta_{y_{t-1}, y_t, x_t}$ represents the parameter associated with the specific label configuration and observation.

The transition probability between adjacent labels is defined using the energy function:

$$P(y_t|y_{t-1}, x_t) = \frac{\exp(-\theta_{y_{t-1}, y_t, x_t})}{\sum_{y'} \exp(-\theta_{y_{t-1}, y', x_t})} \tag{2.10}$$

Where y_t and y_{t-1} are adjacent labels, x_t is the observation, and the sum in the denominator ranges over all possible labels y'. The energy function in a CRF is so named because it represents a notion of "energy" or "cost" associated with a particular configuration of labels in the context of the given input sequence. The concept of an energy function is derived from the field of statistical physics and relates to the idea of finding the most probable or optimal configuration of labels in a way that minimizes this energy. The emission probability of an observation given a label is defined using the energy function:

$$P(x_t|y_t) = \frac{\exp(-\theta_{y_t, x_t})}{\sum_{x'} \exp(-\theta_{y_t, x'})} \tag{2.11}$$

Where x_t is the observation, y_t is the label, and the sum in the denominator ranges over all possible observations x'. The probabilities calculated using energy functions need to be properly normalized to obtain a valid probability distribution.

CRF parameters θ can be learned from labeled training data using techniques like maximum likelihood estimation.

The effectiveness of CRFs relies on the choice of features that capture relevant patterns in the data. In the context of categorical time series, features can include:

- **Observation Features**: These capture the local relationship between the categorical value and the input features. For instance, for categorical variables representing weather conditions, observation features might include temperature, humidity, and wind speed.
- **Transition Features**: These encode the transition patterns between neighboring categorical values. Transition features could account for dependencies such as temporal trends or seasonal patterns.
- **Higher-Order Features**: To capture more complex dependencies, higher-order features consider interactions between multiple consecutive time steps. These features can capture longer-term patterns in the categorical time series.

Inference in CRFs involves finding the most likely sequence of labels given the observed input sequence. This is typically done using techniques such as the Viterbi algorithm or belief propagation. The Viterbi algorithm efficiently computes the most probable label sequence by exploiting the graph structure of the CRF.

CRFs have been successfully applied to various sequential data tasks, including part-of-speech tagging, named entity recognition, and speech recognition. In the context of categorical time series, they have been employed for

applications like activity recognition, speech emotion analysis, and financial market modeling. It is rather interesting the comparison of this type of models with Hidden Markov models (HMMs). CRFs (Discriminative models) offer relaxation on independence assumptions and their efficiency in conjunction with HMMs (Generative models) is explicitly researched in (Lafferty, McCallum, and Pereira, 2001). A generative model, exemplified by the naive Bayes classifier and the Hidden Markov Model (HMM), constitutes a collection of joint distributions that factorize as $p(y, x) = p(y)p(x|y)$. In other words, it outlines the procedure for generating values for features, given the label. Conversely, a discriminative model, like the logistic regression model, encompasses a range of conditional distributions $p(y|x)$. In this case, the primary focus is directly modeling the classification rule. While a discriminative model theoretically has the potential to derive a joint distribution $p(y, x)$ by incorporating a marginal distribution $p(x)$ for inputs, this is seldom necessary in practice (Sutton, McCallum, et al., 2012). A visual representation, Figure 2.2, which is contained at this same book makes this distinction idea more handy.

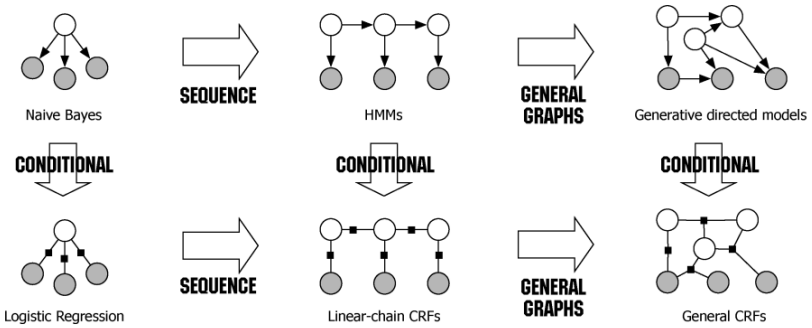


FIGURE 2.2: Graph of the relationship between naive Bayes, logistic regression, HMMs, linearchain CRFs, generative models, and general CRFs, depicted in the book (Sutton, McCallum, et al., 2012).

Concluding this section, let it be noted that CRFs tend to be more parsimonious models compared to HMMs, with their number of parameters constituted just from transition and emission probabilities and not the initial state

probabilities. However, it's important to note that the choice between CRFs and HMMs depends on the specific problem at hand, the nature of the data, and the modeling assumptions that are appropriate for the task.

Chapter 3

Parameter estimation methods and algorithms

The accurate determination of model parameters is essential for building reliable statistical models. In this research, it is aimed to explore a range of sophisticated techniques designed to achieve precise parameter estimates for the Markov chain and Hidden markov models. Through a thorough evaluation of these methods, their strengths, weaknesses, and potential applications are identified.

3.1 Markov Chain model

For the estimation of parameters θ of a Markov chain, the maximum likelihood (ML) approach is going to be applied. Parameter values are selected such as the observed time series become more reasonable. Likelihood function is defined as the probability of observing what has already been observed, as a function of the unknown θ:

$$L(\theta) = P(X_T = x_T, .., X_1 = x_1|\theta) \tag{3.1}$$

The ML estimate of θ is defined to be a value $\hat{\theta}_{ML}$ maximizing L(θ). In view of computational issues, it is common practice to determine $\hat{\theta}_{ML}$, not by

maximizing $L(\theta)$, but by maximizing the logarithmic likelihood function (log-likelihood) instead; that is, $l(\theta) = \log L(\theta)$. The value of the maximized log-likelihood will be denoted as l_{max}. For a Markov chain using the definition of conditional probability and the Markov assumption of dependence the likelihood function becomes equal to:

$$L(\theta) = P(X_1 = x_1) \cdot \prod_{t=2}^{T} p_{x_t|x_{t-1}} \tag{3.2}$$

and the log-likelihood function:

$$l(\theta) = \log P(X_1 = x_1) + \sum_{t=2}^{T} \log p_{x_t|x_{t-1}} \tag{3.3}$$

where $P(X_1 = x_1)$ is the initial probability. If we want to avoid the calculation of the initial distribution we estimate the conditional likelihood (CL) function, we condition on X_1. The CL and log-CL are then taking the form:

$$L(\theta|X_1) = \prod_{t=2}^{T} p_{x_t|x_{t-1}} = \prod_{i=1}^{d}\prod_{j=1}^{d} p_{ij}^{n_{ij}} \tag{3.4}$$

$$l(\theta|X_1) = \sum_{t=2}^{T} \log p_{x_t|x_{t-1}} = \sum_{i=1}^{d}\sum_{j=1}^{d} n_{ij} \log p_{ij} \tag{3.5}$$

where $n_{ij} = \sum_{k=1}^{n-1} I(X_k = i, X_{k+1} = j)$ is the number of transitions from state i to state j.

The stationarity assumption, which is adopted in many applications, is as follows for a Markov process. A Markov chain with transition probability matrix $\boldsymbol{P}$ is said to have stationary distribution $\boldsymbol{\pi}$ (a row vector with non-negative elements) if $\boldsymbol{\pi P} = \boldsymbol{\pi}$ and $\boldsymbol{\pi 1'} = \mathbf{1}$. The initial requirement pertains to stationarity, while the second requirement ensures that $\boldsymbol{\pi}$ conforms to the characteristics of a probability distribution. If a Markov chain is initiated from its stationary distribution, it will maintain that distribution at all future time

points. Such process is commonly referred to as a stationary Markov chain. $(X_t)_{\mathbb{N}}$ is stationary if the initial distribution is an invariant vector. That said, the invariant distributions are the possible stationary marginal distributions of $(X_t)_{\mathbb{N}}$. Let's remark that an irreducible (homogeneous, discrete-time, finite state-space) Markov chain has a unique, strictly positive, stationary distribution. In an irreducible Markov chain it is possible to get to any state form any state.

Finally, it is important regarding the parameter estimation procedures to explain the connection between Markov models and Hidden Markov models. Considering a scenario where the hidden states in the HMM correspond one-to-one with the observable states in the basic Markov model, you can think of the HMM as essentially having the hidden states "revealed" as observable states. In this case, the "hidden" states in the HMM are essentially just the same as the observable states in the basic Markov model. In essence, a Markov model can be seen as a specific instance of a hidden Markov model where the hidden states are treated as directly observable. The emission probability $p(x_{it}|c_t) = 1$ if $c_t = x_{it}$ and 0 otherwise, meaning that the emission matrices are identity matrices. The hidden states are identical to the observed states and all emission probabilities are deterministic. For this thesis the parameter estimation for the Markov chain models will be carried out, by treating them as a special case of HMMs and taking advantage of the core algorithms inside `seqHMM` package originally designed for HMM application.

3.1.1 Bayesian information criterion

The maximum likelihood approach can be utilized in conjunction with information criteria for model selection when multiple candidate models are available for the time series. One aim is to choose a model with the maximum achievable l_{max}. However, it is also important to consider the interpretability of smaller

models and prevent overfitting. Information criteria address this by striking a balance between the goodness of fit and the model's size, achieved by introducing a suitable penalty term to log- likelihood l. Bayesian information criterion (BIC) (Schwarz, 1978) which will be accounted for this thesis is given by:

$$BIC = pl \log T - 2l \tag{3.6}$$

where p is the number of model parameters and T is the number of observations and l is the log-likelihood of the fitted model. BIC strongly penalizes the model size by incorporating a penalty term for model complexity.

In the context of model selection, the goal is to compare BIC scores across different candidate models and select the model with the lowest BIC score. The model with the lowest BIC strikes a balance between fitting the data well and having a reasonable number of parameters. This process inherently performs a trade-off between model complexity and goodness of fit, preventing both overfitting and underfitting.

BIC has become quite popular for model selection in sociology, particularly in log-linear and other models for categorical data (A. E. Raftery, 1995). For a more comprehensive look at practical applications of BIC one can consult (Anderson and Burnham, 2004).

3.2 Hidden Markov model

For the parameter estimation process, maximum likelihood approach is going to be applied, for the estimation of the transition probabilities matrix, the emission probabilities matrix and the initial state probabilities.

It is reminded here that likelihood function is defined as the probability of observing what has already been observed, as a function of the unknown $\theta = \{\boldsymbol{A}, \boldsymbol{p}(\cdot|c), \boldsymbol{\pi}\}$. For the discrete case, which has initial distribution $\boldsymbol{\pi}$, transition

probability matrix A for the Markov chain and emission probability matrix B, the likelihood function turns as:

$$L(X|\theta) = P(X_T = x_T, .., X_1 = x_1|\theta) = \sum_c P(X_t, C_t) \tag{3.7}$$

For the bivariate probability of $(X_t, C_t)_{\mathbb{N}}$ it holds that:

$$P(X_t, C_t) = P(C_1)\prod_{t=2}^{T} P(C_t|C_{t-1})\prod_{t=1}^{T} P(X_t|C_t) = \pi_1 p(x_1|c_1)\prod_{t=2}^{T}\alpha_{t-1,t}p(x_t|c_t), \tag{3.8}$$

leading to likelihood function:

$$L(X|\theta) = \sum_c \pi_1 p(x_1|c_1)\prod_{t=2}^{T}\alpha_{t-1,t}p(x_t|c_t) = \sum_c(\pi_1\alpha_{12}\alpha_{23}\cdots\alpha_{T-1,T})(p(x_1|c_1)p(x_2|c_2)\cdots p(x_T|c_T)) \tag{3.9}$$

The interpretation of the computation in the equation above, is the following. Initially at time t=1 the system is in state c_1 with probability π_1 and generate observation x_1 with probability $p(x_1|c_1)$. We then make a transition to state c_2 with probability α_{12} and generate observation x_2 with probability $p(x_2|c_2)$. This process continues until we make the last transition from state c_{T-1} to state c_T with probability $\alpha_{T-1,T}$ and generate observation x_T with probability $p(x_T|c_T)$.

Next, in order to switch into convenient matrix notation, the diagonal matrix $\mathbf{P}(x)$ is going to be introduced, with i_{th} diagonal element $p(x|c_i)$. Now the likelihood function can be written as:

$$L(\theta) = \boldsymbol{\pi P}(x_1)A\boldsymbol{P}(x_2)\cdots\boldsymbol{AP}(x_T)\mathbf{1}^{\boldsymbol{T}} \tag{3.10}$$

A recursive scheme is easily noticed and based on it, the likelihood function can be described as:

$$\boldsymbol{\alpha_1} = \boldsymbol{\pi P}(x_1),$$

$$\boldsymbol{\alpha_t} = \boldsymbol{\alpha_{t-1}AP}(x_t)_{t\in(2,3,..)} = \boldsymbol{\pi P}(x_1)\prod_{t=2}^{T}\boldsymbol{AP}(x_t),$$

$$L(\theta) = \boldsymbol{\alpha_T}\mathbf{1^T} \tag{3.11}$$

Here α_t can be denoted as the vector of forward probabilities at time t:

$$\boldsymbol{\alpha_t(c)} = P(C_t = c, X_1 = x_1, .., X_t = x_t) \forall c \in C \tag{3.12}$$

These probabilities are not only useful in view of likelihood computation, but also for forecasting future observations. The computation of the h-step-ahead forecasting distribution for the observations, given the sequence of observations $x_1, ..., x_t$, is performed as follows:

$$P(X_{t+h} = x | x_1, .., x_t) = \frac{P(X_1 = x_1, .., X_t = x_t, X_{t+h} = x)}{P(X_1 = x_1, .., X_t = x_t)} = \frac{\boldsymbol{\alpha_t A^h P}(x)\mathbf{1^T}}{\boldsymbol{\alpha_t}\mathbf{1^T}} \tag{3.13}$$

Moreover state prediction at time t is possible to be carried out through the calculation of the conditional probability:

$$P(C_{t+h} = c | x_1, .., x_t) = \frac{\boldsymbol{\alpha_t A^h e_c^T}}{L(\theta)} \tag{3.14}$$

for $t > T$, where $e_c = (0, ..., 0, 1, 0, ..., 0)$ has a one in the c_{th} position only. At this point let also the vector of backward probabilities to be denoted as:

$$\boldsymbol{\beta_t(c)} = P(X_T = x_T, ..., X_{t+1} = x_{t+1} | C_t = c) \forall c \in C \tag{3.15}$$

It can be additionally expressed as:

$$\boldsymbol{\beta_t^T} = \boldsymbol{\beta_{t+1}^T P}(x_{t+1})\boldsymbol{A} = (\prod_{t=t+1}^{T} \boldsymbol{AP}(x_t))\mathbf{1^T} \tag{3.16}$$

and $\boldsymbol{\beta_T} = \mathbf{1}$.

Having defined the forward and backward probabilities, the concept of decoding in HMM can be now investigated. *Local decoding* of the state at time t refers to the determination of that state which is most likely at that time given

the observations, the posterior probabilities. Posterior probabilities can be used to find the locally most probable hidden state at each time point, but the resulting sequence is not necessarily globally optimal (check global decoding Chapter 3.6).

$$P(C_t = c|X_1 = x_1, .., X_T = x_T) = \frac{P(C_t = c, X_1 = x_1, ..X_T = x_T)}{P(X_1 = x_1, ..., X_T = x_T)} = \frac{\boldsymbol{\alpha_t(c)\beta_t(c)}}{L(\theta)} \tag{3.17}$$

Scaling is necessary in order to prevent numerical underflow in the evaluation of the product $\alpha_t(c)\beta_t(c)$. Similarly, $L(\theta)$ can be calculated by applying a scaling method. (see Chapter 3.5 for details)

As mentioned earlier, stationarity is often assumed for the data generating process. This means that the initial distribution, which is defined as π, satisfies the invariance equation $A\pi = \pi$. For the purposes of this thesis, the focus will be mainly on stationary HMMs, where $P(C_t = c) = \pi_c$ holds true for all $c \in \mathbb{N}$ and all $t \in \mathbb{N}$. If $\boldsymbol{\pi}$ is the stationary distribution of the Markov chain, we have $\boldsymbol{\pi P}(x_1) = \boldsymbol{\pi AP}(x_1)$, hence likelihood equation turns as:

$$L(\theta) = \boldsymbol{\pi AP}(x_1)\boldsymbol{AP}(x_2)\cdots\boldsymbol{AP}(x_T)\mathbf{1}^{\boldsymbol{T}} \tag{3.18}$$

which involves an extra factor of A but may be slightly simpler to code.

For direct numerical maximization (DNM) of the log-likelihood, any general-purpose optimization routines such as BFGS or Nelder–Mead can be used (with suitable reparameterizations). Another commonly employed estimation technique is the expectation-maximization (EM) algorithm, often referred to as the Baum-Welch algorithm when applied to Hidden Markov models. The EM algorithm exhibits rapid convergence towards a local optimum, although its convergence speed near the optimum is often slower compared to DNM (S. Helske and J. Helske, 2017). Interestingly and depending on each problem, DNM

methods and EM may be combined for higher efficiency. For the purposes of this thesis, the EM algorithm is going to be applied for parameter estimation.

An additional important point that is worth to be made, is that of dealing with the missing observations. In the context of Hidden Markov models it is a straightforward process. When a missing observation is encountered, denoted as x_{it}, it does not provide any new information about the hidden states. In such situations, a fixed emission probability $p(x_{it}|c) = 1$ is assigned for all possible states c, ranging from 1 to d_c. To accommodate sequences of varying lengths, missing values are handled by incorporating them either before and/or after the observed states.

Closing this section, model selection has to be commented on. In the fundamental HMM with d_c states, enhancing the number of states (d_c) consistently enhances the model's fit, as determined by the likelihood. However, this improvement is accompanied by a quadratic growth in the number of parameters, necessitating a trade-off between the enhanced fit and parameter increase. Consequently, a model selection criterion is necessary to make informed decisions. As per with the case of MM models, BIC criterion is going to be used.

$$BIC = p \log(\sum_{i=1}^{N} \sum_{t=1}^{T} I(x_{it})) - 2l \tag{3.19}$$

where p is the number of estimated model parameters, N is the number of subjects and l is the log-likelihood of the fitted model.

3.3 Expectation-maximization algorithm

Moving forward, regarding finding the maximum-likelihood estimate of the parameters for HMMs, one commonly employed method is the Baum-Welch algorithm (Baum et al., 1970), which falls under the umbrella of the expectation-maximization (EM) algorithm. The EM algorithm was introduced for general

latent variable models in the seminal paper by (Dempster, Laird, and Rubin, 1977), who also mentioned applications to finite mixture models. This is the method that will be applied in this thesis to obtain the estimates of HMMs and it will be accomplished through the utilization of forward and backward probabilities, which as mentioned above, also serve as essential tools for state prediction and as it will be mentioned below, for decoding.

In general, the EM algorithm is an iterative approach used for maximum likelihood estimation when there are or there assumed to be missing data points (Bilmes et al., 1998). It takes advantage of the fact that maximizing the complete log- likelihood, which incorporates both observed and missing data, can be simpler than maximizing the likelihood of the observed data alone.

Let the observed data be represented as X and the unobserved variables as Z. The parameters of the model are denoted as θ. The complete log-likelihood function is $Q(\theta, \theta^{(t)})$, where $\theta_{(t)}$ represents the current parameter estimates at iteration t. The general form of the EM is as follows:

Expectation Step (E-step): Calculate the expectation of the complete log-likelihood function $Q(\theta, \theta^{(t)})$ with respect to the conditional distribution of the unobserved variables Z given the observed data X and the current parameter estimates $\theta^{(t)}$:

$$Q(\theta, \theta^{(t)}) = \mathbb{E}_{Z|X,\theta^{(t)}} \left[\log p(X, Z|\theta)\right] \tag{3.20}$$

Maximization Step (M-step): Update the parameter estimates θ by maximizing the expectation obtained from the E-step:

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta, \theta^{(t)}) \tag{3.21}$$

Convergence Check: Evaluate the change in the log-likelihood (or other convergence criteria) between consecutive iterations. If the change is below a predefined threshold or after a maximum number of iterations, terminate the algorithm, otherwise, go back to the E-step.

The specific form of the complete log-likelihood function $Q(\theta, \theta^{(t)})$ and the model-specific expressions for the E-step and M-step depend on the particular model being used. Under fairly mild regularity conditions, the EM algorithm converges to a local maximum of the likelihood function.

It is interesting to note that (Roderick, Little, and Rubin, 2002) emphasize that the key idea of EM is not necessarily replacing the missing data itself with their conditional expectations in the complete log-likelihood, but rather replacing the missing data functions that appear in the complete log-likelihood.

3.3.1 EM algorithm for Hidden Maerkov model

More specific, for a Hidden Markov model, hidden states are naturally going to be accounted as the "missing data". In order to treat the model states as observed data for the complete likelihood, two simple variables are going to be introduced:

$$z_j(t) = 1 \text{ if } c_t = j \ _{(t=1,..,T)}$$

$$z_{jk}(t) = 1 \text{ if } c_{t-1} = j \text{ and } c_t = k \ _{(t=2,..,T)}$$

It is reminded that the log-bivariate probability of (X_t, C_t) is configured as:

$$\log P(X_t, C_t) = \log P(C_1) \sum_{t=2}^{T} \log P(C_t|C_{t-1}) \sum_{t=1}^{T} logP(X_t|C_t) \tag{3.22}$$

The complete log-likelihood is then configured as:

$$CLL = \sum_{j=1}^{c} z_j(1) \log \pi_j + \sum_{j=1}^{c} \sum_{k=1}^{c} (\sum_{t=2}^{T} z_{jk}(t)) \log \alpha_{jk} + \sum_{j=1}^{c} \sum_{t=1}^{T} z_j(t) \log p(x_t|j) \tag{3.23}$$

consisted of three terms.

A step-by-step description of the EM algorithm is given below:

Initialization: Set initial values for the model parameters, including the initial state probabilities, transition probabilities, and emission probabilities. These initial values can be randomly assigned or based on prior knowledge.

E-step: Compute the expected sufficient statistics for the latent (hidden) variables. In the categorical time series HMM context, this involves estimating the conditional probabilities of being in each hidden state at each time step, given the observed data and the current parameter estimates.

$$\hat{z}_j(t) = P(C_t = j|x_1, ..x_T) = \frac{\alpha_t(j)\beta_t(j)}{L(\theta)},$$

$$\hat{z}_{jk}(t) = P(C_{t-1} = j, C_t = k|x_1, ..x_T) = \frac{\alpha_{t-1}(j)\alpha_{jk}p(x_t|k)\beta_t(k)}{L(\theta)}$$

This step utilizes the forward-backward algorithm or other similar methods to calculate the posterior probabilities of the hidden states.

M-step: Update the model parameters using the expected sufficient statistics obtained in the E-step. Specifically, re-estimate the initial state probabilities, transition probabilities, and emission probabilities based on the computed posterior probabilities of the hidden states. This step involves maximizing the complete log-likelihood of the observed data under the current parameter estimates, while taking into account the expectations obtained from the E-step. Therefore the three parts of the CLL, Equation 3.23, that were described above,

will be maximized separately with respect to the initial distribution, the transition probabilities and the state-dependent parameters. To update the initial state probabilities, the complete log-likelihood is maximized by taking the derivative with respect to π and setting it to zero. A Lagrange multiplier is used to satisfy the constraint that the probabilities sum to 1.

$$\hat{\pi_j} = \frac{\hat{z_j}(1)}{\sum_{j=1}^{c} \hat{z_j}(1)} = \hat{z_j}(1),$$

displaying the expected frequency (number of times) in state $C_1 = j$ at time $(t = 1)$.

The same procedure is followed for the transition probabilities to be updted:

$$\hat{\alpha_{jk}} = \frac{\sum_{t=2}^{T} \hat{z_{jk}}(t)}{\sum_{k=1}^{c} \sum_{t=2}^{T} \hat{z_{jk}}(t)} = \frac{\sum_{t=2}^{T} \hat{z_{jk}}(t)}{\sum_{t=2}^{T} \hat{z_j}(t)},$$

which describes the fraction of expected number of transitions from state $C_{t-1} = j$ to state $C_t = k$ to the expected number of transitions from state $C_t = k$.

For the emission probabilities:

$$p(\hat{x_t}|j) = \frac{\sum_{t=1}^{T} \hat{z_j}(t) \cdot \mathbb{I}(x = x_t)}{\sum_{t=1}^{T} \hat{z_j}(t)}$$

Where $\mathbb{I}(x = x_t)$ is the indicator function that takes the value 1 when $x_i = x_t$ and 0 otherwise. $p(\hat{x_t}|j)$ is the fraction of the expected number of times in state $C_t = j$ and observing symbol x_t to the expected number of times in state $C_t = j$.

Iterate E-step and M-step: Repeat the E-step and M-step until convergence is reached. In each iteration, update the expected sufficient statistics using the current parameter estimates and then update the parameters based on these statistics. Convergence is typically determined by checking the change in the log-likelihood or the parameter values between iterations. If the change is below a predetermined threshold, the algorithm is considered to have converged.

Final parameter estimates: Once the algorithm has converged, the final parameter estimates represent the maximum likelihood estimates (MLE) of the model parameters given the observed data.

In case the underlying Markov chain is stationary rather than just homogeneous, the parameter vector ß becomes completely determined by the transition probabilities **A**, rendering the estimation of ß irrelevant. However, the M-step introduces an optimization problem. The objective instead of maximizing the three CLL parts separately, Equation 3.23, is to maximize the sum of the first two terms together. Usually numerical optimization is utilized because analytical solution is hard to come up. (Bulla and Berzel, 2008) suggest an alternative to EM algorithm for stationary Hidden Markov models.

Before concluding the EM topic, it is important to highlight that the calculation of the forward and backward probabilities can be prone to underflow or overflow errors. This case shall be mostly dealt with the application of scaling the forward and backward probabilities. If scaling still produces underflow in backward algorithm, the forward and backward variables can be computed in log-space. These methods are applied by the package `seqHMM` (J. Helske, 2023) which will be used for model estimations for this thesis.

3.4 Viterbi algorithm

Determining the entire series of hidden states $c_1, ..., c_T$ is called *global decoding*. In order to find the state sequence that maximizes the probability

$$P(C_1 = c_1, ..., C_T = c_T | x_1, ..., x_T),$$

the Viterbi algorithm can be employed (Lawrence R Rabiner, 1989). The Viterbi algorithm was first applied to speech and language processing in the context of

speech recognition by (Vintsyuk, 1968).

For the initial step $(t = 1)$ of the algorithm the following probabilities are calculated:

$$m_{1,c} = P(C_1 = c, X_1 = x_1) = p(x_1|c)\pi_c \forall c \in C, \tag{3.24}$$

while during the next steps and for $t >= 1$, the probabilities that are recursively computed are:

$$m_{t,c} = max_{c_1,..,c_{t-1}} P(C_t = c, C_{t-1} = c_{t-1}, .., C_1 = c_1, X_t = x_t, ..., X_1 = x_1) = p(x_t|c) max_r \{m_{t-1,r} a_{rc}\}, \tag{3.25}$$

and with path backtracking, starting from the hidden state with the maximum probability in the last step, the previous states are traced back by following the recorded previous state c at each time step:

$$\hat{c}_T = argmax_c \{m_{T,c}\}$$

$$\hat{c}_t = argmax_c \{m_{t,c} a_{t,t+1}\}_{t=T-1,...,1}$$

In order to prevent numerical underflow that can occur when multiplying numerous small probabilities, the Viterbi algorithm can be easily implemented in log space. This is because the objective of maximizing the quantity in global decoding involves a simple product of probabilities.

(... ΜΠΟΡΩ ΝΑ ΠΡΟΣΘΕΣΩ ΛΕΠΤΟΜΕΡΕΙΕΣ ΕΔΩ..)

3.5 Forward-backward algorithm

The forward algorithm is a fundamental algorithm used in Hidden Markov models (HMMs) to compute the probability of observing a particular sequence

of observations at time t and hidden state c at time t, given the model parameters:

$$\boldsymbol{\alpha_t(c)} = P(C_t = c, X_1 = x_1, .., X_t = x_t) \forall c \in C. \tag{3.26}$$

It is also known as the forward procedure of the forward- backward algorithm.

The algorithm is initialized as follows:

$$\alpha_1(c) = \pi_c p(x_1|c) \forall c \in C \text{ (initialization for the first time step)}$$

The recursive procedure that is then executed for $t = 2, .., T$ is:

$$\alpha_t(c) = (\sum_c \alpha_{t-1}(c) a_{t-1,t}) p(x_t|c)$$

This equation calculates the probability of being in state c_t at time t by summing over all possible previous states c. The probability of transitioning from state c_{t-1} to state c_t is given by $a_{t-1,t}$, and the probability of observing x_t from state c_t is given by $p(x_t|c)$. The forward procedure is terminated with the calculation of the observed sequence given the model parameters θ:

$$L(X|\theta) = P(X_T = x_T, .., X_1 = x_1|\theta) = \sum_c \alpha_T(c) \tag{3.27}$$

The forward algorithm computes the probability of the observed sequence efficiently by utilizing dynamic programming and the Markov property of the HMM. It breaks down the calculation into smaller steps, making it feasible to handle longer sequences and larger state spaces.

The probabilities calculated by the forward algorithm are crucial for various tasks involving HMMs, such as decoding the most likely state sequence (Viterbi algorithm), estimating the model parameters (Baum-Welch algorithm), or performing inference and prediction tasks.

The backward algorithm is another fundamental algorithm used in Hidden

Markov models (HMMs) to compute the probability of observing a particular sequence of observations given the model parameters. It is also known as the backward procedure or the forward-backward algorithm. The backward variable

$$\boldsymbol{\beta_t(c)} = P(X_T = x_T, ..., X_{t+1} = x_{t+1} | C_t = c) \forall c \in C \tag{3.28}$$

represents the probability of starting in state c at time t and generating the remaining observations $x_{t+1}, x_{t+2}, ..., x_T$.

The algorithm is initialized as follows:

$$\beta_1(c) = \pi_c p(x_1|c) \forall c \in C \text{ (initialization for the first time step)}$$

The recursive procedure that is then executed for $t = T - 1, .., 1$ is:

$$\beta_t(c) = \sum_c (a_{t,t+1} p(x_{t+1}|c) \beta_{t+1}(c))$$

This equation calculates the probability of being in state c_t at time t by summing over all possible next states c. The probability of transitioning from state c_t to state c_{t+1} is given by $a_{t,t+1}$, the probability of observing x_{t+1} from state c is given by $p(x_{t+1}|c)$, and the probability of being in state j at time t+1 is given by $\beta_{t+1}(c)$. The backward procedure is terminated with the calculation of the observed sequence:

$$P(X_T = x_T, .., X_1 = x_1 | \theta) = \sum_c (\pi_c p(x_1|c) \beta_1(c)) \tag{3.29}$$

This equation sums over the initial state distribution π, the probability of observing the first observation from each state c, and the backward variable $\beta_1(\cdot)$ for the first time step.

The backward algorithm computes the probability of the observed sequence by calculating the backward variable $\beta_t(\cdot)$ in a similar manner to the forward algorithm. It propagates probabilities backward in time, starting from the last

time step and moving towards the first time step.

The probabilities calculated by the backward algorithm are useful for various tasks involving HMMs, such as estimating the model parameters (Baum-Welch algorithm), performing smoothing, or computing the probability of being in a particular state at a specific time given the observed sequence (the forward-backward algorithm).

It needs to be underlined though, that in practice the forward-backward algorithm is prone to numerical instabilities and they are typically scaled. When working with probabilities, especially when multiplying many small probabilities together, numerical precision can become a problem. This is particularly common in sequences with many time steps. Scaling is introduced to mitigate these precision issues.

The idea is to normalize the probabilities at each time step by dividing them by a scaling factor. This maintains the relative probabilities but prevents them from becoming too small and losing precision. The scaled forward variable $\hat{\alpha}_t(c)$ is defined as:

$$\hat{\alpha}_t(c) = \frac{\alpha_t(c)}{\gamma_t}$$

Where γ_t is the scaling factor at time step t. The scaling factor is chosen so that the scaled probabilities still sum to 1. This prevents numerical underflow and maintains accuracy.

Similar to the forward algorithm, scaling can also be applied to the backward algorithm to maintain numerical stability. The scaled backward variable $\hat{\beta}_t(c)$ is defined as:

$$\hat{\beta}_t(c) = \frac{\beta_t(c)}{\gamma_t}$$

The equations for the scaling factors in the forward and backward algorithms are the following:

For the Forward Algorithm:

$$\gamma_t = \frac{1}{\sum_c \alpha_t(c)}$$

For the Backward Algorithm:

$$\gamma_t = \frac{1}{\sum_c \beta_t(c)}$$

Where:
- c is the total number of hidden states in the HMM.
- $\alpha(t, i)$ is the scaled forward probability at time step t and state c.
- $\beta_t(c)$ is the scaled backward probability at time step t and state c.

These scaling factors γ_t are used to normalize the scaled probabilities, maintaining numerical stability during calculations in the forward and backward algorithms.

By using scaling in both the forward and backward algorithms, you can ensure that the calculations involving small probabilities are performed with greater precision, leading to more accurate estimations of the likelihoods and model parameters in the HMM. This is crucial for applications like parameter estimation, decoding, and training of HMMs with categorical time series data.

3.6 Mixture Hidden Markov model

In comparison with the conventional hidden Markov model likelihood, Equation 3.9, an additional summation is introduced when considering the clusters.

The likelihood of a mixture of Hidden Markov models (HMMs) can be expressed as:

$$L(X|\theta) = \sum_{k=1}^{K} w_k \sum_{c} P(X_t, C_t|\theta_k) = \sum_{k=1}^{K} w_k \sum_{c} \pi_1^k p^k(x_1|c_1) \prod_{t=2}^{T} \alpha_{t-1,t}^k p^k(x_t|c_t), \tag{3.30}$$

and the log-likelihood as:

$$l(\theta) = \log L(\theta) = \log \left(\sum_{k=1}^{K} w_k \sum_{c} P(X_t, C_t|\theta_k) \right) = \log \left(\sum_{k=1}^{K} w_k \sum_{c} \pi_1^k p^k(x_1|c_1) \prod_{t=2}^{T} \alpha_{t-1,t}^k p^k(x_t|c_t) \right), \tag{3.31}$$

where

w_k is the mixing coefficient for mixture component k

$\theta_k = (\pi^k, \alpha^k, p^k(x_t|c_t))$ represents the parameters of the mixture model

Having already defined the algorithms for calculating the parameters of the basic Hidden Markov models, it should be underlined here that the mixtures of hidden Markov models can be defined in the context of the aforementioned. The K submodels are combined into one large Hidden Markov model consisting of K states, where the initial state vector contains elements of the form $w_k \pi_k$. The transition matrix would be block diagonal where the diagonal blocks A_k, are square matrices containing the transition probabilities of one cluster. The off-diagonal blocks are zero matrices, so transitions between clusters are not allowed. Similarly, the emission matrices contain stacked emission matrices B_k. The EM algorithm for MHMMs might require more iterations to converge due to the added complexity of updating parameters for multiple mixture components. .

Let it be reminded here, that in case multiple subjects, multiple sequencies are being studied then the equations of likelihood and log-likelihood are represented by a summation over all subjects. That being said, above are given the equations for a single time series.

3.7 Mixture Markov model

The parameters of the Mixture Markov Model (MMM) are typically estimated using the expectation-maximization (EM) algorithm and a maximum likelihood approach. The log-likelihood function of the MMM takes the following form:

$$l(\theta) = \log \sum_k w_k P(X_T = x_T, \ldots, X_1 = x_1 | \theta) \tag{3.32}$$

where w_k represents the mixing proportion for component k, indicating the probability of selecting component k. The MMM can be regarded as a special case of the Mixture Hidden Markov model (MHMM) (Chapter 3.6), where there is no hidden structure. Through the use of `seqHMM` package it is regarded as such and thus takes advantage of the already existing algorithms for MHMMs.

Chapter 4

Applied analysis and results

In this chapter, the focus shifts from the theoretical and methodological aspects discussed in previous sections to the empirical evidence and findings derived from the data analysis. It provides an opportunity to showcase the research's contribution to the field. It aims to provide an in-depth interpretation and analysis of a dataset of sequences of female workers, allowing for meaningful insights and implications to be derived.

A sequence refers to an ordered collection of categorical states. It can represent various types of data, such as a time series (e.g., career trajectory or residential history), DNA sequences, or story structures. Sequence data often involve multiple independent subjects, making it multivariate, and sometimes there are multiple interdependent sequences per subject, known as multichannel or multidimensional sequence data.

The structure of this chapter is designed to facilitate a systematic and coherent presentation of the results. It begins with a brief overview of the data collection and analysis process, including the research instruments employed and any specific considerations taken into account. This serves to establish a context for the subsequent presentation of the results. The subsequent sections will delve into the core findings, organized according to the research questions or themes explored in the study. Each finding will be presented concisely, supported by appropriate visual aids such as tables, charts, or graphs, to enhance clarity and comprehension. In addition to the presentation of the raw results, this chapter will provide a comprehensive analysis and interpretation of the

findings. The interpretation will be guided by the relevant theoretical framework and supported by the existing literature. It is worth noting that while this chapter presents the results of the research, it also acknowledges the limitations and potential sources of error inherent in the study. Recognizing these limitations helps to provide a balanced perspective and offers avenues for future research and further refinement of the findings.

4.1 The dataset of the female workers

The dataset that this thesis is going to be focused on, is consisted of 472 female workers, for which, 12 observations were recorded for each of them from the age of 18 to the age of 30. The data are coming from the British Labor Force survey in 2014. The observations for each female worker are relevant to her occupational stage. Each observation falls into one of six categories: Student, Temporary worker, Self-employed, Permanent job, Family care, Unemployed. The sequential and categorical nature of the observations makes them an ideal candidate for the application of many of the aforementioned analysis approaches related to categorical time series. The main scope of the investigation of this type of data is to extract meaningful insight regarding the working life of the modern woman, while providing valuable insights into the unique dynamics of the female workforce, ultimately contributing to a more equitable and inclusive understanding of labor markets.

4.2 Exploratory data analysis

Starting with the exploratory data analysis (EDA), it is going to be conducted by evaluating specific descriptive measures and by calculating different metrics for the female worker dataset. It becomes obvious from Table 4.1 which includes the category frequencies described in the dataset, that most of the observations

collectively appear to belong to permanent job category. On the other hand, the category observed the least, is this of the self employment.

	Frequency	Percent
Student	569	9.96
Temporary employed	309	5.41
Self employed	257	4.50
Permanently employed	3689	64.58
Family care	538	9.42
Unemployed	350	6.13

TABLE 4.1: Frequency table of female workers dataset.

Since the exploratory process regards sequential data, it is important to check the category frequencies split by the age of the female workers, as well as the entropy values for each one of them. By analyzing Table 4.2, it can be derived that females are students more often at the early stages of their life. The same occurs, though in a lesser magnitude, for women which are temporary employed or unemployed. On the contrary, females with a permanent job are increased as they get older and also the dominance of this category throughout all the different age groups is prominent. Self employment and family care categories are more frequent in older age groups too. As complementary information for each age group, the Entropy measure appears to have higher values at early ages 19-22 and lower values for the rest of the age groups. This fact indicates a higher uncertainty and variety during young women's professional stages than during latter ones.

	age19	age20	age21	age22	age23	age24	age25	age26	age27	age28	age29	age30
Student	0.345	0.294	0.229	0.124	0.071	0.032	0.025	0.025	0.021	0.011	0.011	0.008
Temporary employed	0.080	0.082	0.080	0.082	0.069	0.044	0.059	0.034	0.027	0.029	0.036	0.027
Self employed	0.011	0.021	0.021	0.029	0.036	0.044	0.044	0.057	0.048	0.071	0.071	0.086
Permanently employed	0.414	0.452	0.513	0.607	0.664	0.725	0.716	0.744	0.733	0.731	0.729	0.723
Family care	0.046	0.046	0.065	0.071	0.101	0.105	0.111	0.103	0.118	0.122	0.120	0.122
Unemployed	0.105	0.105	0.092	0.086	0.059	0.050	0.044	0.038	0.053	0.036	0.034	0.034
Entropy	0.76	0.77	0.76	0.71	0.65	0.56	0.57	0.53	0.54	0.53	0.53	0.53

TABLE 4.2: Frequency table and Shanon entropy measure of female workers dataset by age.

The horizontal reading of the above table is best illustrated in the barplots Figure 4.1 where, keeping the same scale in each category, the observations are confirmed both for the magnitude of their frequencies and for the increasing or decreasing trend observed by age in each of the six cases.

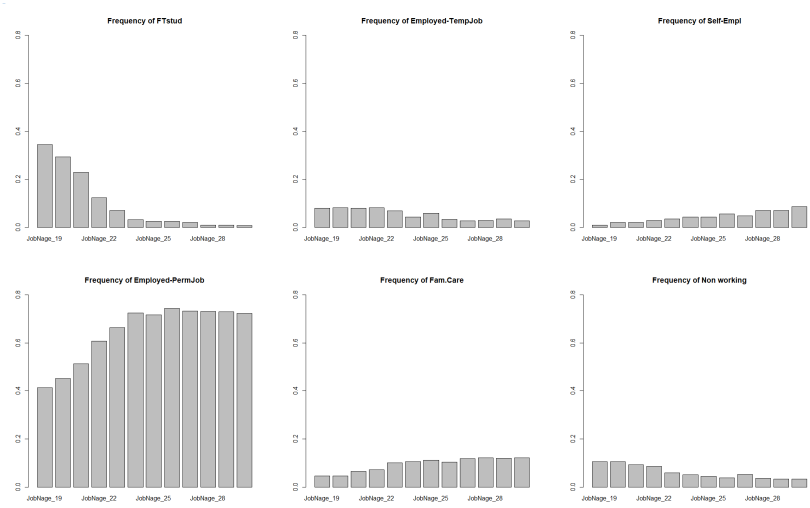


FIGURE 4.1: Barplots of each category by age for female workers dataset.

An interesting summary that can be derived from the state distributions is the sequence made of the most frequent state at each position. For the female workers dataset this sequence is represented at each age stage by the category of the employed women as Figure 4.2 clearly displays. This is an anticipated visual result this is the mode category already realised. In that same plot it is displayed that the shown sequence of modal states belongs to 71 women in the whole dataset.

In Figure 4.3 where the ten most frequent sequences are depicted, it can been once more confirmed, the sequence of modal states. It is followed by nine other frequent sequences where eight out of nine times women at early life stages seem to be engaged with university studies and later on obtain a steady job. There is also at some level the case where a woman with a permanent job as she gets older, she gets directed to the care of the family.

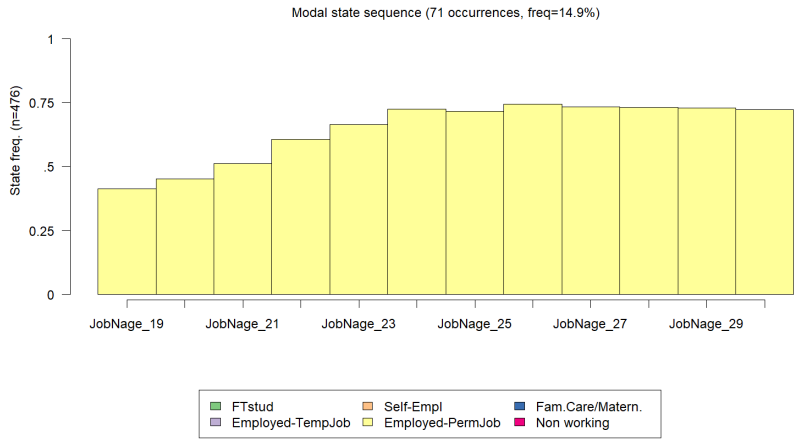


FIGURE 4.2: Sequence of modal states for female workers. Category "Employed-PermJob" dominant in all life stages.

Of course the complete sequence collection can be plotted, but as it was highlighted in previous chapters, when the amount of records is large, it becomes difficult to draw conclusions. In Figure 4.4, two facts that can be vaguely noticed is the dominance of permanent employed women and the existence of student categories at early life stages at most.

Overall, EDA provided a crucial step in the analysis of these categorical time series data, allowing the gain of insights into the inherent dynamics, patterns, and relationships within the data. By employing various EDA techniques, the understanding of the data is now enhanced and the decisions regarding subsequent modeling, forecasting, and analysis approaches will be more informed.

4.3 Model performance

The exhibition of dynamic categorical patterns over time, points towards the analysis and understanding of these temporal patterns through modelling. It is an essential part for decision-making, predictions, and uncovering valuable insights.

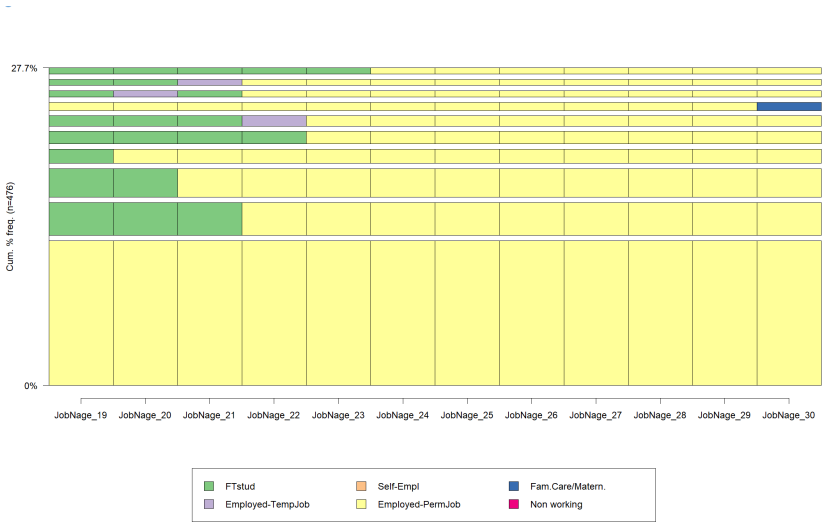


FIGURE 4.3: Ten most frequent sequences plot for female workers dataset.

4.3.1 Markov Chain model

The Markov chain model offers a typical framework for capturing the temporal dependencies and transitions within categorical time series and it is the first and simplest model which will be applied on the data. For this type of model the estimates of the initial probabilities and the transition matrix were based on observations. The values of the estimated initial probabilities and the transition probability matrix is presented in Table 4.3.

Transition probabilities						
from/to	Student	Temp.job	Self.emp.	Perm.emp.	Fam.care	Unemp.
Student	0.588	0.099	0.011	0.214	0.014	0.074
Temp.job	0.077	0.277	0.040	0.466	0.027	0.111
Self.emp.	0.005	0.051	0.745	0.157	0.023	0.018
Perm.emp.	0.008	0.022	0.016	0.901	0.033	0.019
Fam.care	0.010	0.016	0.012	0.192	0.727	0.042
Unemp.	0.054	0.117	0.036	0.275	0.105	0.413
Initial probabilities	0.344	0.080	0.010	0.414	0.046	0.105

TABLE 4.3: Markov model, order 1, estimated parameters for female workers dataset.

The initial probabilities display a high probability for this process to start

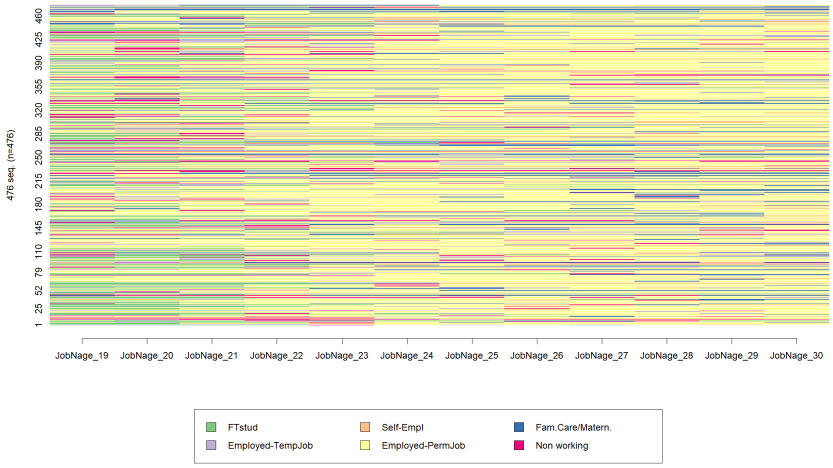


FIGURE 4.4: Complete sequence plot for female workers dataset.

from the category permanent job or from the category student. From the transition matrix it can be spotted that for all the categories other than temporary job category, the highest probabilities lie on the diagonal. This means that when a female starts from one category it is more possible to stay at it for the next year. Except if she finds herself with a temporary job. Then it is more possible to move on next year and get a permanent job. In Figure 4.5 a visual representation of this Markov chain is depicted. Each vertex represents a different category, the initial probabilities of each category are illustrated and finally the transition probability paths between the categories are represented by edges with width analogous to their size. For example as regards the case where a female is temporary employed and next year it is highly expected for her to get a more steady one, this can be spotted at the graph as the purple vertex connected with a bold edge with the yellow vertex.

4.3.2 Mixture Markov Chain models

The female workers dataset is consisted of multiple objects sequence observations, so it is only natural to try modelling them with mixtures of Markov chain

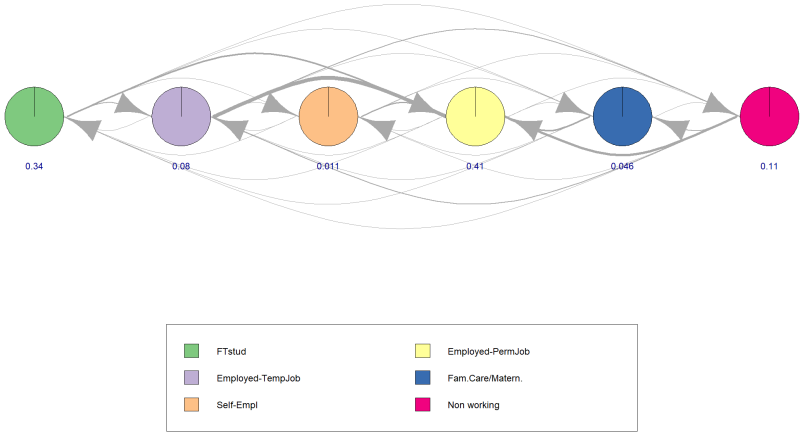


FIGURE 4.5: Markov chain plot for female workers dataset.

models as an attempt to distinguish specific clusters in the data, if there are any, and explain the observed information in a more sufficient way, since mixture models aim to add flexibility to the basic model, making it advantageous when dealing with complex data that exhibits multiple behavioral patterns or modes.

Two, three and four cluster mixtures were tested as extensions to the basic Markov chain model. More clusters were avoided since the number of parameters was growing large in comparison with the total number of observations. Regarding the EM algorithm, in order to reduce the risk of being trapped in a poor local optimum, a large number of initial values was tested for the EM repetitions.

In this section, two cluster mixture model will be inferenced, the selection of which was guided by BIC results. Details about model selection criteria values in regards with all models that were tested, will be described in model selection subsection.

For a two-part Markov chain mixture for the six-category observed variables, 71 parameters were estimated, representing the two prior mixture probabilities, the two sets of initial state probabilities and the two transition matrices Table 4.4.

	Cluster 1	Cluster 2				
Means of prior cluster probabilities	0.792	0.208				
Cluster 1						
Transition probabilities						
from/to	Student	Temp.job	Self.emp.	Perm.emp.	Fam.care	Unemp.
Student	0.591	0.107	0.012	0.220	0.000	0.070
Temp.job	0.088	0.269	0.049	0.508	0.003	0.083
Self.emp.	0.005	0.039	0.771	0.168	0.009	0.008
Perm.emp.	0.008	0.023	0.017	0.919	0.018	0.015
Fam.care	0.016	0.017	0.000	0.868	0.099	0.000
Unemp.	0.059	0.186	0.068	0.444	0.011	0.232
Initial probabilities	0.400	0.074	0.009	0.443	0.008	0.066
Cluster 2						
Transition probabilities						
from/to	Student	Temp.job	Self.emp.	Perm.emp.	Fam.care	Unemp.
Student	0.563	0.0391	0.000	0.168	0.124	0.105
Temp.job	0.0255	0.317	0.000	0.259	0.145	0.253
Self.emp.	0.000	0.225	0.367	0.000	0.229	0.178
Perm.emp.	0.009	0.019	0.011	0.756	0.157	0.047
Fam.care	0.010	0.016	0.014	0.104	0.809	0.047
Unemp.	0.049	0.053	0.006	0.120	0.191	0.580
Initial probabilities	0.131	0.103	0.015	0.303	0.193	0.254

TABLE 4.4: mixture Markov chain model, k=2 components, estimated parameters for female workers dataset.

Some conclusions that can be drawn from such a partitioning of the model is that female workers can be divided in two groups. Those who will display paths where studies and permanent employment are gonna be dominant and those who will easier transition to life stages like family care, a temporary job or unemployment. In such a configuration female workers will be most probably come from the first cluster rather than the second. Table 4.5 depicts the relevant posterior cluster probabilities

Most probable clusters		
	Cluster 1	Cluster 2
count	386	90
proportion	0.811	0.189

TABLE 4.5: Posterior counts and probabilities of 2-part mixture Markov model.

and Table 4.6 is the classification table, defined from posterior cluster probabilities, where the values near 1 on the diagonal point out a good model.

Classification table		
Mean cluster probabilities (in columns) by the most probable cluster (rows)		
	Cluster 1	Cluster 2
Cluster 1	0.967	0.033
Cluster 2	0.044	0.956

TABLE 4.6: Classification table of 2-part mixture Markov chain model.

A visual representation of the two clustered configuration of the Markov chains is displayed in Figure 4.6.

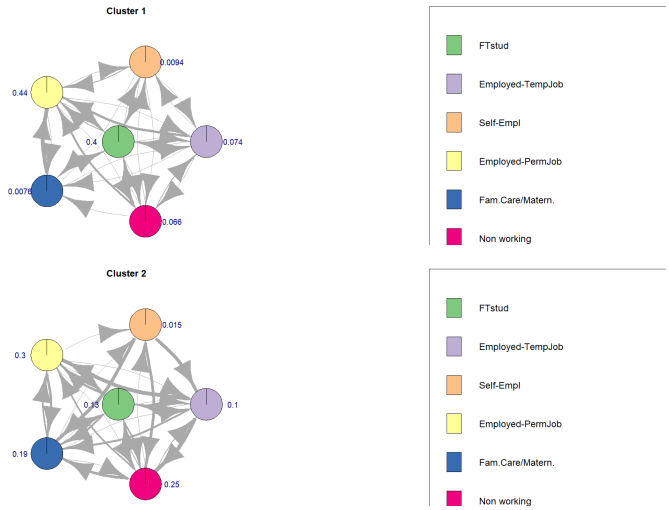


FIGURE 4.6: 2-part mixture Markov chain plot for the female workers dataset.

4.3.3 Hidden Markov models

The need of more parsimonious models lead to the testing of various Hidden Markov models (HMMs) with different number of states. The existence of latent states that affect the observed working stage of the females in the dataset is allowed. HMMs offer a level of interpretability by providing a pretty clear representation of the latent states and their transitions. This can aid in understanding the underlying process generating the time series data and in making informed decisions based on the model's results. Models with 2 to 7 hidden

states were tested. Randomly generated values were used each time for the initialization of the parameters for the EM algorithm. In order for the procedure not to get trapped at a local maximum or saddle, the EM algorithm was restarted several times with different starting points. The HMMs with 5 and 6 states are the most appropriate for inference based on BIC criterion. 6 state HMM had a slightly lower BIC but since the two models were so close, 5 state HMM is going to be used for interpretation because of lower number of parameters and for the sake of simplicity, in order not to have the same number of states and variable categories. In Table 4.7 the initial, the transition and the emission probabilities of the 5-state HMM are displayed.

Initial probabilities	0.009	0.382	0.393	0.176	0.04	
Transition probabilities						
	State 1	State 2	State 3	State 4	State 5	
State 1	0.890	0.000	0.074	0.036	0.000	
State 2	0.009	0.656	0.199	0.128	0.007	
State 3	0.010	0.006	0.959	0.012	0.013	
State 4	0.032	0.024	0.253	0.654	0.036	
State 5	0.004	0.009	0.045	0.011	0.931	
Emission probabilities						
	Student	Temp.job	Self.emp.	Perm.emp.	Fam.care	Unemp.
State 1	0.000	0.025	0.911	0.048	0.012	0.004
State 2	0.892	0.058	0.003	0.021	0.000	0.026
State 3	0.001	0.013	0.004	0.960	0.017	0.005
State 4	0.002	0.366	0.006	0.112	0.017	0.498
State 5	0.007	0.006	0.003	0.050	0.894	0.039

TABLE 4.7: 5-state Hidden Markov model, estimated parameters for female workers dataset.

It is noticed that the process generating the random variables which represent the working stages of the female workers, may most likely start from latent States 3 or 2. State 3 seems to be close enough to a self absorbing state. If a sequence is found in State 2 it is possible to move to State 3 or less likely to State 4 every 1-2 years out of 10. If a sequence is brought to State 4, it remains there or switch to State 3. State 3 is a latent dimension which produces mainly the observed job stage of the permanent employment. State 2 leads with higher

probability to female workers which are students. State 4 unveils most probably the category of unemployed women or those with a temporary job. Placing all the puzzle pieces together, the estimated 5-state HMM describes a situation where on the one hand a female worker could start from the age of 19 with a permanent job and remain permanently employed until her 30s. On the other hand, there may be females who will start by studying at an early stage in their lives and then after some years get a permanent job. That or after their studies some of them may remain unemployed for some time or with a temporary job before eventually get a permanent job. This process can be visually investigated in Figure 4.7. The initial probabilities are depicted under every vertex, the transition probabilities between states are displayed by the edges between the vertices and the emission probabilities are the pie slices. Emission probabilities with very low values are aggregated in the "others" slice.

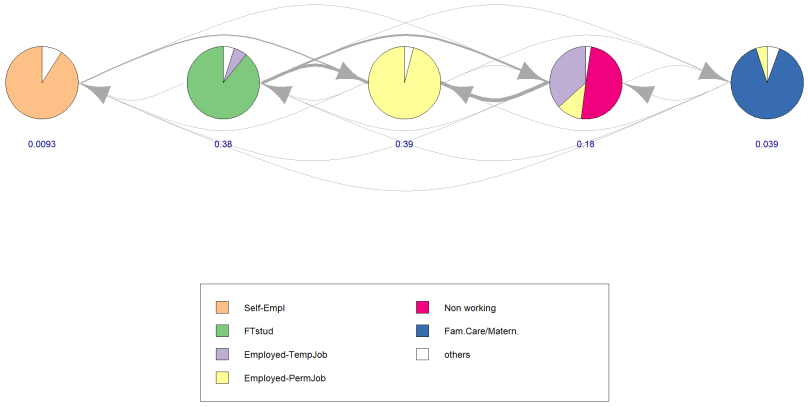


FIGURE 4.7: 5-state Hidden Markov model plot for female workers dataset.

Family care is one of the rarest categories to be visited in a sequence of this model. Taking a closer look at these cases, family care is a category mainly emissioned by latent State 5 of the model. It is highly unlikely for a working

woman to start a family at an early age and if at some point until her 30s she reaches this stage, she will most probably not go back to work.

Having estimated the 5-state HMM parameters and given the data, the most probable hidden paths for every woman can be estimated by utilizing the Viterbi algorithm which is described in detail in chapter 3.6. In Figure 4.8 the most probable hidden paths for all females are depicted at once.

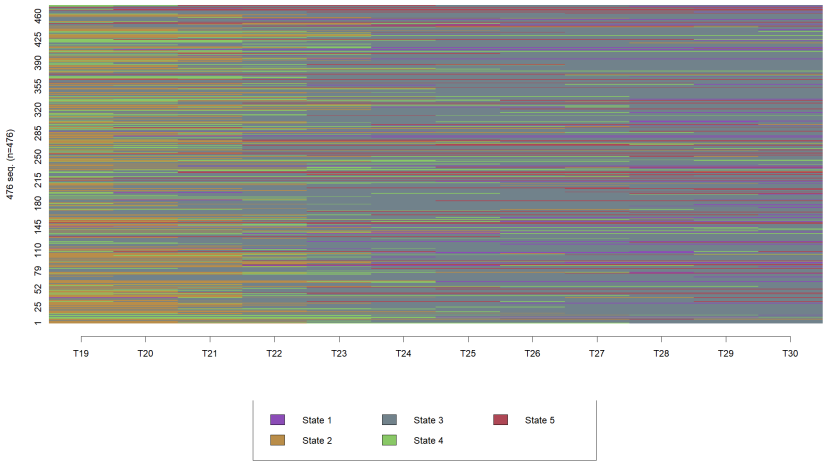


FIGURE 4.8: 5-state HMM plot for the most probable hidden state paths for the female workers dataset.

State 2 and State 4 are visited mostly at the early life stages of the working women, while State 3, is overall the most dominant state and even more after the age of 22-23 for these women. This fact is becoming clear by looking at Figure 4.9 which shows the barplots of most probable hidden states for ages 19 and 30. Of course every female's most probable path can be investigated more specifically on its own.

4.3.4 Mixture Hidden Markov models

The final type of models that are going to be fitted to the female-workers dataset are the mixtures of Hidden Markov models (MHMM). The procedure that was

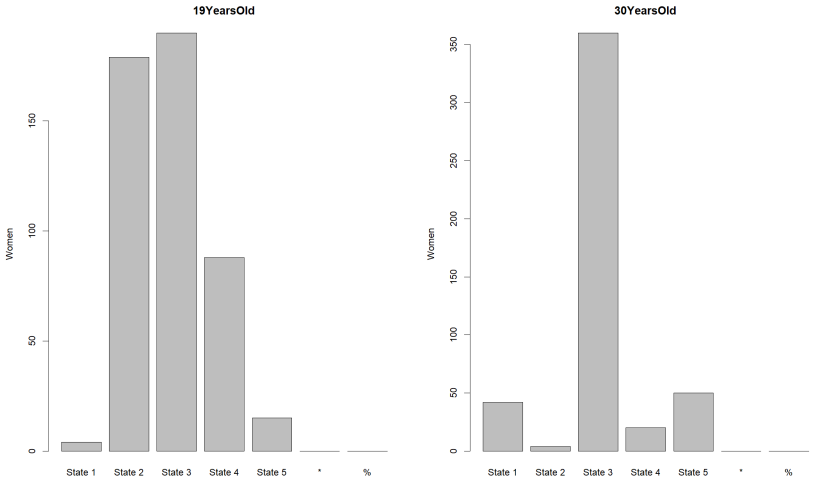
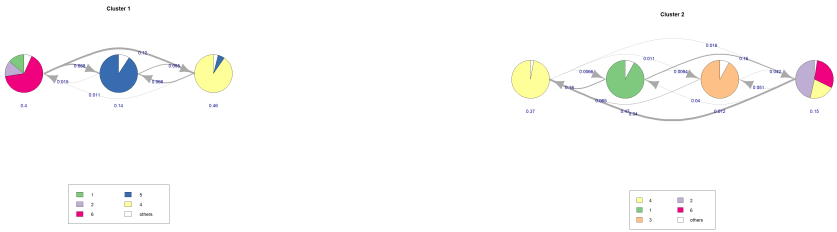


FIGURE 4.9: 5-state HMM barplots for the most probable hidden states for ages 19 and 30 for the female workers dataset.



(A) Cluster 1. (B) Cluster 2.

FIGURE 4.10: Visual representation of the mixture 3-state/4-state Hidden Markov model.

followed was the trial of different number of clusters which also included Hidden Markov models with various number of states for the whole data. Sequences are not in fixed clusters but get assigned to clusters with certain probabilities during the modeling process. The optimum combination BIC-wise was given by the formation of two clusters each represented by a 3-state and a 4-state HMM respectively. The inference related to these kind of mixture models is often intricate. Table 4.8 displays all the estimated model parameters (initial, transition, emission probabilities) and a visual representation of the model is given in Figure 4.10.

	Cluster 1	Cluster 2				
Means of prior cluster probabilities	0.634	0.366				
Cluster 1						
Transition probabilities						
from/to	State 1	State 2	State 3			
State 1	0.791	0.088	0.122			
State 2	0.015	0.930	0.055			
State 3	0.011	0.066	0.923			
Initial probabilities	0.404	0.136	0.460			
Emission probabilities						
	Student	Temp.job	Self.emp.	Perm.emp.	Fam.care	Unemp.
State 1	0.140	0.134	0.004	0.039	0.023	0.659
State 2	0.009	0.009	0.009	0.035	0.907	0.030
State 3	0.001	0.022	0.010	0.902	0.057	0.007
Cluster 2						
Transition probabilities						
from/to	State 1	State 2	State 3	State 4		
State 1	0.966	0.006	0.011	0.016		
State 2	0.176	0.649	0.009	0.165		
State 3	0.065	0.000	0.893	0.042		
State 4	0.337	0.040	0.051	0.572		
Initial probabilities	0.366	0.473	0.0124	0.148		
Emission probabilities						
	Student	Temp.job	Self.emp.	Perm.emp.	Fam.care	Unemp.
State 1	0.007	0.007	0.003	0.972	0.012	0.005
State 2	0.921	0.039	0.004	0.012	0.000	0.024
State 3	0.000	0.0164	0.919	0.049	0.009	0.005
State 4	0.001	0.464	0.004	0.213	0.020	0.299

TABLE 4.8: Mixture 3-state/4-state Hidden Markov model, k=2 components, estimated parameters for female workers dataset.

In Table 4.9 the most probable cluster distribution of the subjects is described. In Table 4.10 the classification table is pictured, where the values near 1 on the diagonal point out a good model.

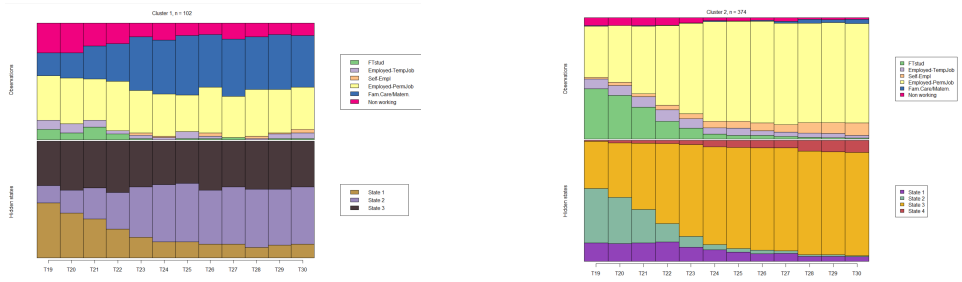
Most probable clusters		
	Cluster 1	Cluster 2
count	104	372
proportion	0.218	0.782

TABLE 4.9: Posterior counts and probabilities of the mixture 3-state/4-state Hidden Markov model.

Classification table		
Mean cluster probabilities (in columns) by the most probable cluster (rows)		
	Cluster 1	Cluster 2
Cluster 1	0.938	0.061
Cluster 2	0.066	0.934

TABLE 4.10: Classification table of the mixture 3-state/4-state Hidden Markov model.

Plotting the revealed most probable hidden state sequences along with the observations for each subject in each cluster in Figure 4.11, gives a general picture of the different characteristics existing in each group. More specific one may easily spot that the first and smaller cluster includes a bigger percentage of women that will turn to maternity and family care at some point in their lives and also unemployed ones. As opposed to cluster 2 in which, as it becomes obvious in Figure 4.11b, the dominant categories are those of employment and studies at the early years. This cluster represents most working women in the dataset. In both clusters temporal dependencies are taken into consideration by the different states included.



(A) Cluster 1. (B) Cluster 2.

FIGURE 4.11: Stacked plots of most probable hidden states and observations of the mixture 3-state/4-state Hidden Markov model.

4.4 Model selection

By employing a comprehensive range of candidate models and evaluating them through BIC, it is aimed to distinguish the most parsimonious and informative model that best characterizes the relationships within the female workforce dataset. In Table 4.11 the log-likelihood, the BIC-values and the number of free parameters for the various models fitted on the female workers dataset are investigated. It is easily noticed that, judging from the BIC numbers, the lower BIC values for the 5-7 state HMMs indicate that the HMM is favored as a simpler and more parsimonious model for the dataset in comparison with the mixtures of HMMs. This suggests that the complexity introduced by the mixture components in the MHMMs might not be justified by the data and that the dataset might be relatively homogeneous. The single HMMs might be capturing the underlying patterns and behaviors in the categorical time series effectively without the need for multiple components. the same pattern is found in the case of comparing the Markov Chain models with the mixtures of Markov Chain models. Eventually, the model that gives the maximum log-likelihood valued is the Hidden Markov model with seven latent states while the model with the lowest BIC value is the Hidden Markov model with six hidden states. It is the one that will be regarded as the best of the fitted models on the female

workers dataset.

	Log-likelihood	BIC	df
MarkovChain	-4385.7	9074.162	35
2-MarkovChain	-4272.075	9158.322	71
3-MarkovChain	-4205.185	9335.954	107
4-MarkovChain	-4169.39	9575.776	143
2-state HMM	-5518.175	11148.8	13
3-state HMM	-4812.105	9823.168	23
4-state HMM	-4497.425	9297.612	35
5-state HMM	-4271.262	8966.39	49
6-state HMM	-4192.484	8947.24	65
7-state HMM	-4165.289	9048.563	83
2-HMM 3-state/3-state	-4374.75	9156.065	47
2-HMM 3-state/4-state	-4256.427	9023.223	59
2-HMM 4-state/4-state	-4214.738	9043.65	71
2-HMM 4-state/5-state	-4182.429	9100.135	85

TABLE 4.11: Log-likelihood, BIC and number of free parameters for the various models fitted on the female workers dataset.

In Figure 4.12 the models that were tested are plotted with coordinates the number of the parameters they include and the BIC they suggest. One may notice after inspection that the Hidden Markov models belong indeed to the more parsimonious modelling side, in comparison with the basic Markov models.

After selecting the HMM-6 as the best fitted model, let it be displayed here in Figure 4.13 a concluding image of the distribution of the observations and at the same time, the most probable hidden states conditional on the HMM-6 model parameters.

The findings that follow the selection of this model as the optimal one, are quite consistent with what was mentioned above in relation to the interpretation of the 5-state hidden Markov model. Which means, that a working woman is most likely to be studying when she is 19 years old or already working a permanent job. Sometimes, it can also happen to start as unemployed at this age. After studying, a female will continue her professional path quite possibly by finding a permanent or temporary job. If she is working and she is 19, she

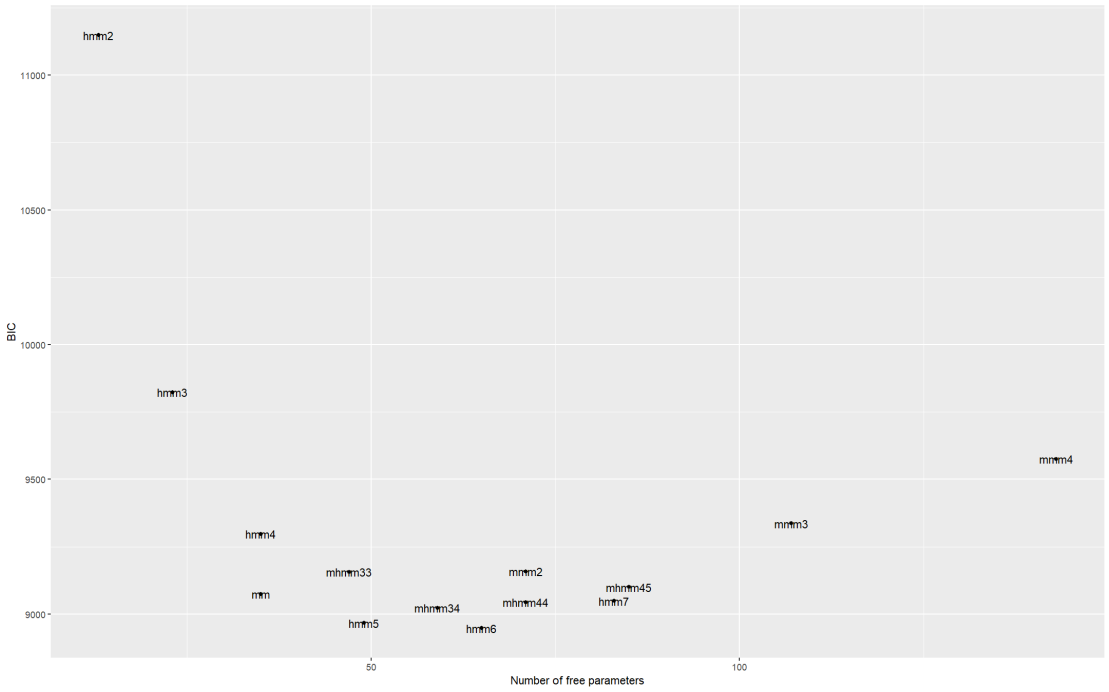


FIGURE 4.12: Plot BIC values vs the number of parameters for the various models fitted on the female workers dataset.

will continue having a permanent job or in more rare cases she will maybe start focusing on family care, but when she is closer at the age of 30. Let it be highlighted once more at this point that transition between working life stages and job categories is quite flexible due to the complex of transition and emission probabilities offered by a Hidden Markov model with six states.

If a more general observation were to be made, this would be that women joining the working force is highly unlikely to be turning into parenthood until their 30s and if they do, they will hardly ever be returning back to full working mode. This fact indicates that some legal or social action should be held to encourage a healthier combination of family care and employment mode for women. The more common recurrence of job phases for female workers is to first study at their early life and then continue working until their 30s. This pattern is quite dominant and it would be interesting for more patterns that would include for example self employment or family orientation to be more

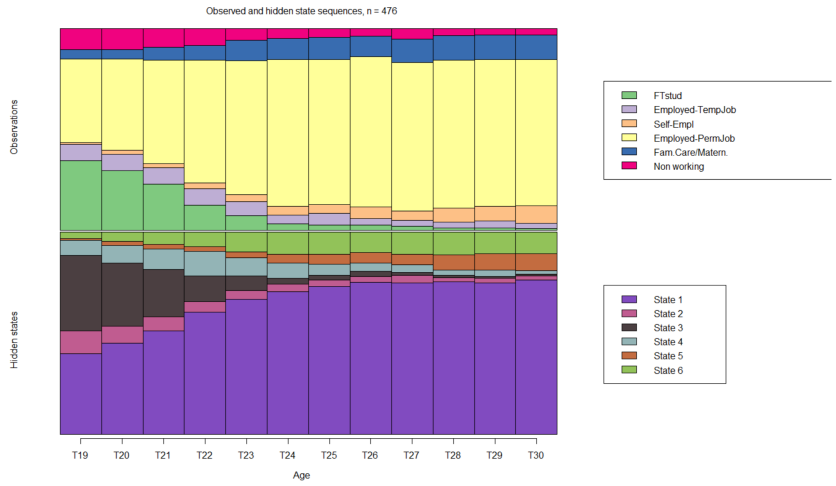


FIGURE 4.13: Stacked plot of most probable hidden states and observations of the 6-state Hidden Markov model.

easily adopted. For this to happen of course professional, social and legal issues should be regulated.

Chapter 5

Conclusion and future directions

In this master's thesis, an in-depth exploration of categorical time series was undertaken, encompassing a comprehensive review of existing literature, analytical techniques, and model frameworks. The overarching objective was to enhance our understanding of the intricate patterns embedded within categorical time series data. Through an exhaustive investigation, this study aimed to shed light on effective methodologies for analysis and modeling, ultimately leading to improved insights and predictive capabilities.

The preliminary phase of this research delved into a comprehensive survey of the current landscape of categorical time series, encompassing both visual and descriptive analysis techniques. By assimilating knowledge from diverse sources, a solid foundation was laid for the subsequent analytical journey. One distinct conclusion drawn was that in the field of categorical time series the various tools and descriptive measures should be combined with each other in each case or tailor made solutions should be created depending on the particular case. In this area of analysis there is much scope for further research and study.

The subsequent phase of this study focused on an extensive exploration of model frameworks suitable for categorical time series. From a multitude of options, Markov chain models, Hidden Markov models (HMMs), and their mixtures emerged as promising candidates. A meticulous application of these models was carried out on a real-world dataset pertaining to female workers, with the intention of unraveling underlying patterns and latent structures

within the data. Almost all intensive computations were done by leveraging C++ through the use of R software and by utilizing parallel computation.

The empirical analysis culminated in the selection of a Hidden Markov model with six distinct states, a choice substantiated by the Bayesian Information Criterion (BIC). This chosen model not only demonstrated superior performance in capturing the heterogeneity present within the dataset but also showcased the potential for unearthing valuable insights from categorical time series data.

Looking ahead, this research opens up intriguing avenues for further exploration and refinement. Future investigations may delve into advanced goodness-of-fit techniques, such as pseudo residuals (Zucchini and MacDonald, 2009) Chapter 6.2, to enhance the model validation process and ensure robustness of results. Moreover, the incorporation of covariates stands as a promising avenue for augmenting the explanatory power of the developed model, thereby enriching the cluster analysis with a contextual layer of interpretation (Vermunt, Tran, and Magidson, 2008). Personal characteristics and other relevant factors, time-constant as well as time-varying, may be used to predict transition probabilities between life stages. Naturally, an additional ambitious act should be the application of several other models referred in Chapter 2, such as for example VLMMs, MTDs, NDARMA or even CRFs on the same female workers dataset. Further investigation and comparison between the modelling results would be valuable and possibly lead to more parsimonious models than the six-state HMM. Last but not least, for the calculation of confidence intervals for the model parameters of the Hidden Markov model for categorical time series data and specifically for this dissertation the six-state HMM, parametric bootstrap (Efron and Tibshirani, 1994; Langeheine, Pannekoek, and Van de Pol, 1996; Visser, Raijmakers, and Molenaar, 2000) may be efficient. Despite the high number of parameters, it would be engaging to check the results of such a procedure and overcome the obstacle of label switching in each bootstrap repetition.

In conclusion, this thesis has contributed to the realm of categorical time series analysis by amalgamating existing knowledge, practically applying it in most of its part and showcasing the potential of advanced modeling techniques. Navigating the ever-evolving landscape of data analytics, hopefully the insights gained from this study will enhance in a way the future endeavors in the dynamic domain of categorical time series.

Bibliography

Agresti, Alan (2012). *Categorical data analysis*. Vol. 792. John Wiley & Sons.

Anderson, D and K Burnham (2004). "Model selection and multi-model inference". *Second. NY: Springer-Verlag* 63.2020, p. 10.

Baum, Leonard E et al. (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.

Berchtold, André and Adrian Raftery (Aug. 2002). "The Mixture Transition Distribution (MTD) Model for High-Order Markov Chains and Non-Gaussian Time". *Statistical Science* 17.

Bilmes, Jeff A et al. (1998). "A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models". *International computer science institute* 4.510, p. 126.

Brockwell, Peter J and Richard A Davis (2002). *Introduction to time series and forecasting*. Springer.

Bühlmann, Peter and Abraham J Wyner (1999). "Variable length Markov chains". *The Annals of Statistics* 27.2, pp. 480–513.

Bulla, Jan and Andreas Berzel (2008). "Computational issues in parameter estimation for stationary hidden Markov models". *Computational Statistics* 23, pp. 1–18.

Cadez, Igor et al. (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*, pp. 280–284.

Ching, Wai Ki, Eric S Fung, and Michael K Ng (2004). "Higher-order Markov chain models for categorical data sequences". *Naval Research Logistics (NRL)* 51.4, pp. 557–574.

Ching, Wai-Ki, Michael K Ng, and Eric S Fung (2008). "Higher-order multivariate Markov chains and their applications". *Linear Algebra and its Applications* 428.2-3, pp. 492–507.

Ching, WK, ES Fung, and MK Ng (2003). "A higher-order Markov model for the Newsboy's problem". *Journal of the Operational Research Society* 54.3, pp. 291–298.

Dempster, Arthur P, Nan M Laird, and Donald B Rubin (1977). "Maximum likelihood from incomplete data via the EM algorithm". *Journal of the Royal Statistical Society Series B: Statistical Methodology* 39.1, pp. 1–22.

Durbin, Richard et al. (1998). *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press.

Efron, Bradley and Robert J Tibshirani (1994). *An introduction to the bootstrap*. CRC press.

Fahrmeir, Ludwig and Heinz Kaufmann (1987). "Regression models for non-stationary categorical time series". *Journal of Time Series Analysis* 8.2, pp. 147–160.

Fokianos, Konstantinos and Benjamin Kedem (2003). "Regression theory for categorical time series". *Statistical Science* 18.3, pp. 357–376.

Frühwirth-Schnatter, Sylvia (2006). *Finite mixture and Markov switching models*. Springer.

Frühwirth-Schnatter, Sylvia, Gilles Celeux, and Christian P Robert (2019). *Handbook of mixture analysis*. CRC press.

Frühwirth-Schnatter, Sylvia and Christoph Pamminger (2010). "Model-based clustering of categorical time series". *Bayesian Analysis* 5.2, pp. 345–368.

Helske, Jouni (2023). "The main algorithms used in the `seqHMM` package".

Helske, Jouni and Satu Helske (2023). *seqHMM: Mixture hidden Markov models for social sequence data and other multivariate, multichannel categorical time series*. R package version 1.2.6. URL: https://cran.r-project.org/package=seqHMM.

Helske, Satu and Jouni Helske (2017). "Mixture hidden Markov models for sequence data: The seqHMM package in R". *arXiv preprint arXiv:1704.00543*.

Jacobs, Patricia A and Peter AW Lewis (1983). "Stationary discrete autoregressive-moving average time series generated by mixtures". *Journal of Time Series Analysis* 4.1, pp. 19–36.

Juang, Biing Hwang and Laurence R Rabiner (1991). "Hidden Markov models for speech recognition". *Technometrics* 33.3, pp. 251–272.

Jurafsky, Daniel and James H Martin (2008). "Speech and Language Processing: An introduction to speech recognition, computational linguistics and natural language processing". *Upper Saddle River, NJ: Prentice Hall*.

Karlis, Dimitris (2016). "Models for multivariate count time series". *Handbook of discrete-valued time series* 407, p. 424.

Kaufmann, Heinz (1987). "Regression models for nonstationary categorical time series: asymptotic estimation theory". *The Annals of Statistics*, pp. 79–98.

Kvålseth, Tarald O (2011). "Variation for categorical variables". *International Encyclopedia of Statistical Science*. Springer, pp. 1642–1645.

Lafferty, John, Andrew McCallum, and Fernando CN Pereira (2001). "Conditional random fields: Probabilistic models for segmenting and labeling sequence data".

Langeheine, Rolf, Jeroen Pannekoek, and Frank Van de Pol (1996). "Bootstrapping goodness-of-fit measures in categorical data analysis". *Sociological Methods & Research* 24.4, pp. 492–516.

Mächler, Martin and Peter Bühlmann (2004). "Variable length Markov chains: methodology, computing, and software". *Journal of Computational and Graphical Statistics* 13.2, pp. 435–455.

Markov, Andrey Andreevich (2006). "An example of statistical investigation of the text Eugene Onegin concerning the connection of samples in chains". *Science in Context* 19.4, pp. 591–600.

Markov, Andrey Andreyevich (1913). "An example of statistical investigation in the text of 'Eugene Onyegin'illustrating coupling of 'tests' in chains". *Proceedings of the Academy of Sciences of St. Petersburg* 7, pp. 153–162.

McKenzie, Eddie (2003). "Ch. 16. discrete variate time series". *Handbook of statistics* 21, pp. 573–606.

McLachlan, Geoffrey J, Sharon X Lee, and Suren I Rathnayake (2019). "Finite mixture models". *Annual review of statistics and its application* 6, pp. 355–378.

Oriona, Ángel López and José Antonio Vilar Fernández (2023). *Analyzing categorical time series with the R package ctsfeatures*. arXiv: 2304.12332 [stat.ML].

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/.

Rabiner, Lawrence R (1989). "A tutorial on hidden Markov models and selected applications in speech recognition". *Proceedings of the IEEE* 77.2, pp. 257–286.

Raftery, Adrian E (1985). "A model for high-order Markov chains". *Journal of the Royal Statistical Society Series B: Statistical Methodology* 47.3, pp. 528–539.

— (1995). "Bayesian model selection in social research". *Sociological methodology*, pp. 111–163.

Rao, C Radhakrishna (1982). "Diversity and dissimilarity coefficients: a unified approach". *Theoretical Population Biology* 21.1, pp. 24–43.

Ribler, Randy Louis (1997). "Visualizing categorical time series data with applications to computer and communications network traces". PhD thesis. Virginia Polytechnic Institute and State University.

Roderick, J, A Little, and Donald B Rubin (2002). *Statistical analysis with missing data*.

Schwarz, Gideon (1978). "Estimating the dimension of a model". *The Annals of Statistics*, pp. 461–464.

Shannon, Claude Elwood (1948). "A mathematical theory of communication". *The Bell System Technical Journal* 27.3, pp. 379–423.

Stoffer, David S, David E Tyler, and Andrew J McDougall (1993). "Spectral analysis for categorical time series: Scaling and the spectral envelope". *Biometrika* 80.3, pp. 611–622.

Stoffer, David S, David E Tyler, and David A Wendt (2000). "The spectral envelope and its applications". *Statistical Science*, pp. 224–253.

Sutton, Charles, Andrew McCallum, et al. (2012). "An introduction to conditional random fields". *Foundations and Trends® in Machine Learning* 4.4, pp. 267–373.

Vermunt, Jeroen K, Bac Tran, and Jay Magidson (2008). "Latent class models in longitudinal research". *Handbook of longitudinal research: Design, measurement, and analysis*, pp. 373–385.

Vintsyuk, Taras K (1968). "Speech discrimination by dynamic programming". *Cybernetics* 4.1, pp. 52–57.

Visser, Ingmar, Maartje EJ Raijmakers, and Peter CM Molenaar (2000). "Confidence intervals for hidden Markov model parameters". *British Journal of Mathematical and Statistical Psychology* 53.2, pp. 317–327.

Weiß, Christian H (2008). "Visual analysis of categorical time series". *Statistical Methodology* 5.1, pp. 56–71.

— (2013). "Serial dependence of NDARMA processes". *Computational Statistics & Data Analysis* 68, pp. 213–238.

— (2018). *An introduction to discrete-valued time series*. John Wiley & Sons.

— (2019). "Measures of dispersion and serial dependence in categorical time series". *Econometrics* 7.2, p. 17.

Weiß, Christian H and Rainer Göb (2008). "Measuring serial dependence in categorical time series". *AStA Advances in Statistical Analysis* 92, pp. 71–89.

Zucchini, Walter and Iain L MacDonald (2009). *Hidden Markov models for time series: an introduction using R*. Chapman and Hall/CRC.