UNIVERSITÄT
BIELEFELD

On some flexible extensions of hidden Markov models

Timo Adam

Dissertation

presented for the degree *Doctor rerum politicarum* (Dr. rer. pol.)
at the Faculty of Business Administration and Economics,
Bielefeld University

Bielefeld, May 2020

First examiner: Prof. Dr. Roland Langrock
Second examiner: Prof. Dr. Thomas Kneib
Third examiner: Prof. Dr. Christiane Fuchs

Date of the thesis defense: August 28, 2020

Printed on non-aging paper (in compliance with DIN-ISO 9706).

Preface

The successful completion of this thesis could not have been accomplished without the generous support of many people, whom I would like to thank in the following.

First and foremost, I wish to express my sincere gratitude to my principal advisor, Prof. Roland Langrock, who provided invaluable advice, continuous encouragement, and excellent support throughout the last four years. I am very grateful for the opportunity to be introduced to the academic world by such an enthusiastic and supportive supervisor, who taught me far more than the basics of hidden Markov models. I much enjoyed the last four years and can sincerely state that I could not have imagined a better place to begin my academic career. Thank you, Roland!

I would like to extend my thanks to Prof. Thomas Kneib, whom I am very grateful for his generous help, especially for sharing his outstanding expertise on non-parametric modeling techniques. Furthermore, I much appreciate his commitment to act as my second examiner. In addition, I want to thank Prof. Christiane Fuchs for her willingness to act as my third examiner.

I also wish to thank Prof. Andreas Mayr, Prof. Christian H. Weiß, Dr. Christopher A. Griffiths, Dr. Vianey Leos-Barajas, Emily N. Meese, Prof. Christopher G. Lowe, Dr. David Righton, Prof. Paul G. Blackwell, and Len- nart Oelschläger for their valuable contributions to the different projects this thesis is based on. Their feedback, as well as many insightful discussions, considerably improved this work. It was a great pleasure not only to collaborate with but also to learn from such helpful co-authors.

Financial support was gratefully received by the Deutsche Forschungsgemeinschaft (DFG), Projektnummer 316099922-TRR 212.

Special thanks deserve my fellow doctoral students and friends Jennifer Pohle, Marius Ötting, and Sina Mews, whom I am very grateful not only for their generous help and many insightful discussions but also for sharing my complaints whenever something initially did not work out, and for making the last four years a great time in many respects.

On a more personal note, I wish to express my sincere appreciation to my parents, who unconditionally supported me throughout my entire life. Their continuous encouragement means a lot to me and greatly helped to successfully complete this thesis.

Lastly, I would like to thank Silvana, for being such a wonderful partner, as well as for the countless evenings spent together in the library. Her endless patience, incredible support, and unconditional love are invaluable to me. I am so grateful to have you!

Bielefeld, May 2020

Timo Adam

Table of contents

Chapter 1

Introduction

Chapter 1

Introduction

> "Signals always come with noise: it is trying to separate out the two that makes the subject interesting."
>
> — *D. Spiegelhalter*

1.1 Introduction to hidden Markov models

Hidden Markov models (HMMs) constitute a versatile class of statistical models for time series where the observed variables are driven by latent states (ZUCCHINI *et al.*, 2016). Over the last decades, they have been successfully applied across a variety of scientific disciplines, including, *inter alia*, ecology, economics, medicine, meteorology, marketing, and sports. The basic HMM, however, often lacks the flexibility to adequately model complex types of data and, as a consequence, to address certain research questions of interest. In this thesis, we discuss three such problems related to HMMs and propose corresponding extensions of the basic model. Thereby, we aim at providing a small contribution to the toolbox of statistical modeling techniques that can help to address the various challenges arising with the increasingly complex types of data that are likely being collected over the next decades. In this first chapter, we briefly introduce HMMs and their various applications, outline important historical developments, and provide an overview of the flexible extensions of the basic model that are subject of this work.

A basic HMM comprises two stochastic processes that are connected with each other: an observed state-dependent process, which is denoted by $\{Y_t\}_{t=1,\ldots,T}$, and a hidden state process, which is denoted by $\{S_t\}_{t=1,\ldots,T}$. Although the states are not directly observed, they still determine the outcome of the state-dependent process, where the correlation

between the two processes can be exploited to make inference also on the underlying state process (ZUCCHINI *et al.*, 2016). Typical examples for the processes that can be modeled include:

- *observed* distances traveled by an animal that are driven by the animal's *hidden* behavioral modes (which could e.g. be resting, foraging, or traveling; cf. LANGROCK *et al.*, 2012b; PATTERSON *et al.*, 2017);
- *observed* stock prices that are driven by the market agents' *hidden* expectations on a company's future profits (which could e.g. be high or low; cf. RYDÉN *et al.*, 1998; HASSAN AND NATH, 2005);
- *observed* epileptic seizure counts that are driven by *hidden* physiological states of an epilepsy patient's brain (which could e.g. be inter-ictal, pre-ictal, ictal, or post-ictal periods; cf. ALBERT, 1991; WANG AND PUTERMAN, 2001)[1];
- *observed* wind speeds or rainfall occurrences that are driven by *hidden* climate regimes (which could e.g. be low- or high-pressure periods; cf. ZUCCHINI AND GUTTORP, 1991; PINSON AND MADSEN, 2012);
- *observed* product choices that are driven by the *hidden* state of a costumer's relationship to a brand (which could e.g. be weak or strong; cf. CHING *et al.*, 2004; NETZER *et al.*, 2008);
- *observed* performances of a professional baseball or darts player that are driven by the player's *hidden* "hot hand", meaning the experience of a period of exceptional success (cf. GREEN AND ZWIEBEL, 2018; ÖTTING *et al.*, 2020).

The research questions that can be addressed using HMMs are manifold, ranging from estimating the state-dependent distributions of the observed variables over state decoding to making inference on the drivers of the state-switching dynamics. In ecological applications, for instance, it is of particular interest to estimate an animal's step length's or turning angle's distribution conditional on the states, which can then often be linked to certain behaviors exhibited by the animal. Furthermore, it could also be of interest to decode the states, which can then be used to infer when an animal was likely to exhibit a certain behavior (MCCLINTOCK *et al.*, 2020; cf. also SECTION 4.4.1 for an example of such an

[1] An ictal period refers to the physiological state of an epilepsy patient's brain that is characterized by the presence of epileptic seizures.

application). In a clinical trial, to give another example, an interesting point to focus on could be to investigate how different covariates affect the probability of an epilepsy patient switching to an ictal state. In that regard, HMMs could also be used to quantify the extent to which this probability can be decreased by certain drugs (cf. WANG AND PUTERMAN, 2001). In these examples, but also in the several real-data applications that are being presented throughout this work, HMMs can be used to separate the signal from the noise and, ultimately, to extract information from data.

HMMs arguably constitute a rather specialized class of statistical models, as evidenced from the fact that they are not usually being taught in undergraduate programmes in statistics. However, they are very closely related to the much more widely known class of state-space models (SSMs; KIM AND NELSON, 1999; DURBIN AND KOOPMAN, 2012). SSMs were developed in the late 1960s and have been successfully applied e.g. in aerospace engineering, where they were used to filter spacecraft trajectories from inaccurate, noisy geo-positional data, which considerably contributed to the success of the NASA's various Apollo missions (GREWAL AND ANDREWS, 2010; AUGER-MÉTHÉ *et al.*, 2020). A general SSM is specified by two equations, one of which describes the relationship between the observed state-dependent variable and the hidden state variable, while the other one describes how the state variable evolves over time (ZENG AND WU, 2013). The serial dependence induced by these two equations can be modeled in various ways and either in discrete or in continuous time. Commonly used models for the state process of an SSM include simple autoregressive processes (cf. STATHOPOULOS AND KARLAFTIS, 2003), continuous-time correlated random walks (cf. JOHNSON *et al.*, 2008), and Ornstein-Uhlenbeck processes (cf. MICHELOT AND BLACKWELL, 2019), to name but a few examples.

As the state space of an SSM is generally continuous, the state process of such a model can take on infinitely many values. On the one hand, this renders SSMs very flexible in terms of the variety of stochastic processes that can be modeled, but on the other hand it makes them less accessible in practice, which is mainly due to the fairly complex calculations that are required to evaluate the likelihood of the model (PATTERSON *et al.*, 2017). In many applications, it is in fact reasonable to assume a discrete, finite state space rather than a state process whose states gradually change over time, which directs us to HMMs[2]. Mathematically, an HMM is a special case of an SSM where the state process is modeled

[2]In contrast to SSMs, where the likelihood typically involves multiple, high-dimensional integrals, the likelihood of an HMM can be written as a matrix product, which substantially facilitates statistical inference (ZUCCHINI *et al.*, 2016).

by a discrete-time, N-state Markov chain. The state process of an HMM is assumed to satisfy the Markov property, $\Pr(S_{t+1} = s_{t+1}|S_t = s_t, \ldots, S_1 = s_1) = \Pr(S_{t+1} = s_{t+1}|S_t = s_t)$, i.e. the state at time $t+1$, S_{t+1}, is assumed to be conditionally independent of all previous states, $S_1, \ldots, S_{t-1}$, given the state at time t, S_t. Furthermore, it is typically assumed that the observations are conditionally independent of each other, given the states. The discrete nature of the states, along with the simplifying dependence assumption made above, renders HMMs relatively easy to deal with from a mathematical perspective and thereby offers various opportunities for statistical inference (ZUCCHINI *et al.*, 2016; PATTERSON *et al.*, 2017).

1.2 A brief history of hidden Markov models

The theoretical framework of HMMs is based on studies of serially correlated random variables conducted by the Russian mathematician A.A. Markov (1856–1922) a little more than a century ago (SCHUSTER-BÖCKLER AND BATEMAN, 2007). However, due to the computational complexity of the calculations that were required to render his findings feasible in practice, his studies were hardly recognized by the broader scientific community for several decades (RABINER AND JUANG, 1986). It was not before the late 1960s that HMMs started to become popular. This increasing popularity is mainly due to the development of the expectation-maximization (EM) algorithm (BAUM AND PETRIE, 1966; BAUM *et al.*, 1970; DEMPSTER *et al.*, 1977; WELCH, 2003), which provided an efficient technique to estimate the parameters of an HMM. About at the same time, the Viterbi algorithm (VITERBI, 1967) was developed for state decoding, which constitutes another important problem in many applications. These two notable developments, along with the increasing availability of computing capacity, provided the foundation that was required to finally render HMMs feasible in practice.

The first scientific discipline discovering the potential of HMMs for solving real-world problems was supervised machine learning: from the late 1960s onwards, HMMs have been successfully applied to speech recognition problems (cf. BAKER, 1975; JELINEK, 1969; BAHL AND JELINEK, 1975; JELINEK *et al.*, 1975; JELINEK, 1976). In these applications, the observations are typically noisy voice records (or, more precisely, Fourier transformations of the observed speech signals), where the states usually correspond to the actually spoken syllables, words, or sentences (RABINER, 1989). In the following decades, the applications of HMMs in supervised machine learning were expanded to various other pattern recognition problems, ranging from image- or video-based face recognition (cf.

TABLE 1.1: *Development of the popularity of HMMs over the last six decades. The table displays the number of results obtained for the search request "hidden"+"Markov" +"model" on Google Scholar*[3] *per decade.*

1960–1969	1970–1979	1980–1989	1990–1999	2000–2009	2010–2019
153	641	2,450	17,500	148,000	338,000

NEFIAN AND HAYES, 1998; LIU AND CHENG, 2003) over gesture recognition (cf. WILSON AND BOBICK, 1999; CHEN *et al.*, 2003) to handwriting recognition (cf. CHEN *et al.*, 1994; HU *et al.*, 1996; PLÖTZ AND FINK, 2009).

In the late 1980s, HMMs also became increasingly popular in computational biology, where their potential for biological sequence analyses was discovered (cf. KROGH *et al.*, 1994a; KROGH *et al.*, 1994b; HUGHEY AND KROGH, 1996). The observations in these applications are typically genomic sequences, where the states can e.g. be associated with genes, transcription factor binding sites, or members of a protein family from a set of unknown proteins (EDDY, 1996; SCHUSTER-BÖCKLER AND BATEMAN, 2007). In the following years, the applications of HMMs in computational biology were expanded to more complex problems such as gene prediction (cf. MUNCH AND KROGH, 2006; STANKE *et al.*, 2006), pairwise and multiple sequence alignment (cf. DURBIN *et al.*, 1998; PACHTER *et al.*, 2002), and protein secondary structure prediction (cf. ASAI *et al.*, 1993). For a comprehensive overview of the various applications of HMMs in computational biology, we refer to EDDY (1996) and YOON (2009).

While HMMs still play a key role in supervised machine learning and computational biology (where they are primarily used to solve classification problems), in the early 1990s HMMs also started to expand to a wide range of other fields. A number of seminal articles published across different disciplines demonstrated that HMMs prove useful also as a statistical modeling technique: ZUCCHINI AND GUTTORP (1991), for instance, used HMMs to model binary time series of precipitation occurrences, while ALBERT (1991) successfully applied HMMs to time series of epileptic seizure occurrences. MACDONALD AND RAUBENHEIMER (1995) demonstrated how HMMs can be used to model binary time series of animals' feeding behaviors, while MORALES *et al.* (2004) introduced HMMs for modeling animal telemetry data, which provided the foundation for many succeeding articles where HMMs were used to reveal the nature of animal movement (LANGROCK *et al.*, 2012b; MCCLINTOCK *et al.*, 2020). Influential articles in economics include HAMILTON

[3]https://scholar.google.de/scholar?hl=en. The numbers were downloaded on January 31, 2020, where results for patents and citations were excluded.

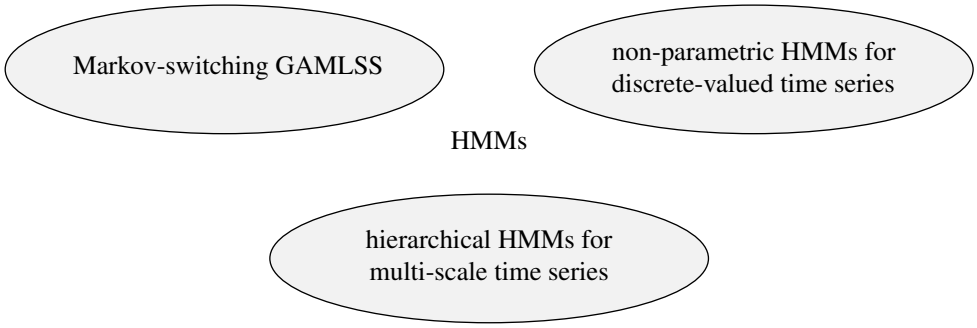

FIGURE 1.2: *Overview of the main topics of the thesis.*

(1989), where HMMs were used to model business cycles, and RYDÉN *et al.* (1998), who applied HMMs to derive stylized facts of stock returns. Finally, VISSER *et al.* (2002) proposed HMMs for implicit learning and concept identification problems, which provided the foundation for the application of HMMs in psychology.

In conclusion, it can be summarized that HMMs developed from a special-purpose classification technique primarily used in supervised machine learning towards general-purpose time series models that are now routinely applied in many different fields. Notably, the increasing popularity is also reflected by the numbers of articles that appeared over the last decades, which are listed in TABLE 1.1: while between 1960 and 1989 as few as 3,244 articles were published, this number increased to a total of 503,500 articles that appeared between 1990 and 2019. This development can likely be attributed to the versatility of the available HMM toolbox: by modifying the model formulation, a wide range of applications can be addressed, while the algorithms for parameter estimation, state decoding, and related problems essentially remain the same (SCHUSTER-BÖCKLER AND BATEMAN, 2007). This important property of the HMM framework is also exploited in this work, as will be outlined in the following section.

1.3 Outline of the thesis

While various special-purpose HMMs — such as mixed-effects HMMs for longitudinal data (cf. ALTMAN, 2007; MARUOTTI, 2011), HMMs with arbitrary state dwell-time distributions (cf. BULLA AND BULLA, 2006; LANGROCK AND ZUCCHINI, 2011), or non-parametric HMMs for continuous-valued time series (cf. LANGROCK *et al.*, 2015; LANGROCK *et al.*, 2018) — are now available, the ever-increasing complexity of the data being collected yields major challenges for statistical modeling in the 21st century. While the early applications of HMMs to statistical modeling problems involved relatively short, of-

ten binary time series, the complexity of the available data has considerably increased since the early 1990s. To adequately address the challenges arising therewith, new modeling techniques are required, which constitutes the main motivation for this work. In this thesis, we discuss three particular modeling challenges related to HMMs and propose corresponding flexible extensions of the basic model. Specifically, we propose i) Markov-switching generalized additive models for location, scale, and shape (GAMLSS), ii) non-parametric HMMs for discrete-valued time series, and iii) hierarchical HMMs for multi-scale time series (cf. FIGURE 1.2 for an overview of the main topics of the thesis).

These three flexible extensions of the basic HMM also constitute the main chapters of this thesis, which can be read independently and are briefly outlined in the following:

- In CHAPTER 2, we propose a novel class of flexible latent-state time series regression models, which we call Markov-switching GAMLSS. In contrast to conventional Markov-switching regression models, the presented methodology allows us to model different state-dependent parameters of the response distribution — not only the mean, but also variance, skewness, and kurtosis parameters — as potentially smooth functions of a given set of explanatory variables. In addition, the set of possible distributions that can be specified for the response is not limited to the exponential family but additionally includes, for instance, a variety of Box-Cox-transformed, zero-inflated, and mixture distributions. We propose an estimation approach that is based on the EM algorithm, where we exploit the gradient boosting framework to prevent overfitting while simultaneously performing variable selection. The feasibility of the suggested approach is assessed in simulation experiments and illustrated in a real-data application, where we model the conditional distribution of the daily average price of energy in Spain over time.

- In CHAPTER 3, we propose an effectively non-parametric approach to fitting HMMs to discrete-valued time series. While specifically for time series of counts, the Poisson distribution — or more flexible alternatives such as the negative binomial, zero-inflated, and mixture distributions — is often chosen for the state-dependent distributions, choosing an adequate class of parametric distributions remains difficult in practice, where an inadequate choice can have severe negative consequences. To overcome this problem, we estimate the state-dependent distributions in a completely data-driven way without the need to specify a parametric family of distributions, where a penalty based on higher-order differences between adjacent count probabilities is proposed to prevent overfitting. The suggested approach is assessed in simulation experiments and illustrated in a real-data application, where we model the

distribution of the annual number of earthquakes over time. The proposed methodology is implemented in the R package countHMM.

- In CHAPTER 4, we propose hierarchical HMMs as a versatile class of statistical models for multi-scale time series. While conventional HMMs are restricted to modeling single-scale data, in practice variables are often observed at different temporal resolutions. An economy's gross domestic product, for instance, is typically observed on a yearly, quarterly, or monthly basis, whereas stock prices are available daily or at even finer resolutions. Step lengths performed by an animal, to give another example, are often observed on a daily or hourly basis, whereas accelerations obtained from accelerometers are available at much higher frequencies, with observations typically made several times per second. To incorporate such multi-scale data into a joint HMM, we regard the observations as stemming from multiple, connected state processes, each of which operates at the time scale at which the corresponding variables were observed. The suggested approach is illustrated in two real-data applications, where we jointly model the distribution of i) daily horizontal movements and ten-minute vertical displacements of an Atlantic cod and ii) monthly trade volumes and daily log-returns of the Goldman Sachs stock, respectively.

Lastly, in CHAPTER 5, we conclude with a brief outlook on potential avenues for future research related to the different methods, including an outline of possible links between the three main chapters of this thesis, and provide some final remarks.

1.4 Statement of contribution and related work

This thesis is based on a number of collaborative projects, which involve contributions from many authors. The following research articles and conference proceedings papers were fully or partially integrated in this thesis, where my personal and the other authors' contributions are detailed in parentheses:

ADAM, T., MAYR, A., AND KNEIB, T. (2017a): Gradient boosting in Markov-switching generalized additive models for location, scale, and shape. *arXiv*, 1710.02385 (submitted to *Econometrics and Statistics, Part B: Statistics*).

(I conceived the idea, developed the research question, implemented the method, designed the simulation experiments, analyzed the data, and drafted the manuscript, A. Mayr and T. Kneib improved the manuscript.)

ADAM, T., MAYR, A., KNEIB, T., AND LANGROCK, R. (2018): Statistical boosting for Markov-switching distributional regression models. *Proceedings of the 33rd International Workshop on Statistical Modelling*, **1**, 30–35.

(The authors' contributions were as above, where R. Langrock further improved the manuscript.)

ADAM, T., LANGROCK, R., AND WEIß, C.H. (2019c): Penalized estimation of flexible hidden Markov models for time series of counts. *METRON*, **77**(2), 87–104.

(R. Langrock and C.H. Weiß conceived the idea and developed the research question, I implemented the method, designed the simulation experiments, analyzed the data, and drafted the manuscript, R. Langrock and C.H. Weiß improved the manuscript.)

ADAM, T., LANGROCK, R., AND WEIß, C.H. (2019d): Nonparametric inference in hidden Markov models for time series of counts. *Proceedings of the 34th International Workshop on Statistical Modelling*, **1**, 135–140.

(The authors' contributions were as above.)

ADAM, T., GRIFFITHS, C.A., LEOS-BARAJAS, V., MEESE, E.N., LOWE, C.G., BLACKWELL, P.G., RIGHTON, D., AND LANGROCK, R. (2019a): Joint modelling of multi-scale animal movement data using hierarchical hidden Markov models. *Methods in Ecology and Evolution*, **10**(9), 1536–1550.

(V. Leos-Barajas, R. Langrock, and I conceived the idea and developed the research question, V. Leos-Barajas and I implemented the method, E.N. Meese and D. Righton collected the data, V. Leos-Barajas and I analyzed the data, I drafted the manuscript, C.A. Griffith, V. Leos-Barajas, E.N. Meese, C.G. Lowe, P.G. Blackwell, and R. Langrock improved the manuscript.)

ADAM, T. AND OELSCHLÄGER, L. (2020): Hidden Markov models for multi-scale time series: an application to stock market data. Available on request (submitted to the *Proceedings of the 35th International Workshop on Statistical Modelling*).

(I conceived the idea and developed the research question, implemented the method, analyzed the data, and drafted the manuscript, L. Oelschläger improved the manuscript.)

The following research articles and conference proceedings papers are closely related to the work presented in this thesis but were neither fully nor partially integrated:

LEOS-BARAJAS, V., GANGLOFF, E.J., ADAM, T., LANGROCK, R., VAN BEEST, F.M., NABE-NIELSEN, J., AND MORALES, J.M. (2017b): Multi-scale modeling of animal movement and general behavior data using hidden Markov models with hierarchical structures. *Journal of Agricultural, Biological and Environmental Statistics*, **22**(3), 232–248.

ADAM, T., LEOS-BARAJAS, V., LANGROCK, R., AND VAN BEEST, F.M. (2017b): Using hierarchical hidden Markov models for joint inference at multiple temporal scales. *Proceedings of the 32nd International Workshop on Statistical Modelling*, **2**, 181–184.

LANGROCK, R., ADAM, T., LEOS-BARAJAS, V., MEWS, S., MILLER, D.L., AND PAPASTAMATIOU, Y.P. (2018): Spline-based nonparametric inference in general state-switching models. *Statistica Neerlandica*, **72**(3), 179–200.

ADAM, T., LANGROCK, R., AND KNEIB, T. (2019e): Model-based clustering of time series data: a flexible approach using nonparametric state-switching quantile regression models. *Book of Short Papers of the 12th Scientific Meeting on Classification and Data Analysis*, 8–11.

Chapter 2

Gradient boosting in Markov-switching generalized additive models for location, scale, and shape

Chapter 2

Gradient boosting in Markov-switching generalized additive models for location, scale, and shape[1]

> "An economic model conditioned on the notion that nothing major will change is a useless one."
>
> — *N. Silver*

Summary

In this chapter, we propose a novel class of flexible latent-state time series regression models, which we call Markov-switching GAMLSS. In contrast to conventional Markov-switching regression models, the presented methodology allows us to model different state-dependent parameters of the response distribution — not only the mean, but also variance, skewness, and kurtosis parameters — as potentially smooth functions of a given set of explanatory variables. In addition, the set of possible distributions that can be specified for the response is not limited to the exponential family but additionally includes, for instance, a variety of Box-Cox-transformed, zero-inflated, and mixture distributions. We propose an estimation approach that is based on the EM algorithm, where we exploit the gradient boosting framework to prevent overfitting while simultaneously performing variable selection. The feasibility of the suggested approach is assessed in simulation experiments and illustrated in a real-data application, where we model the conditional distribution of the daily average price of energy in Spain over time.

[1] This chapter is based on ADAM *et al.* (2017a) and ADAM *et al.* (2018).

2.1 Introduction

In recent years, latent-state models — particularly HMMs — have become increasingly popular tools for time series analyses. In many applications, the data at hand follow some patterns within some periods of time but reveal different stochastic properties during other periods (ZUCCHINI *et al.*, 2016). Typical examples are economic time series, e.g. share returns, oil prices, or bond yields, where the functional relationship between response and explanatory variables differs in periods of high and low economic growth, inflation, or unemployment, respectively (HAMILTON, 1989). Since their introduction by GOLDFELD AND QUANDT (1973) nearly half a century ago, Markov-switching regression models, i.e. time series regression models where the functional relationship between response and explanatory variables is subject to state-switching controlled by an unobservable Markov chain, have emerged as a versatile method to account for the dynamic patterns described above (KIM *et al.*, 2008; DE SOUZA AND HECKMAN, 2014; DE SOUZA *et al.*, 2017; LANGROCK *et al.*, 2017).

While Markov-switching regression models are typically restricted to modeling the mean of the response (treating the remaining parameters as nuisance and constant across observations), it often appears that other parameters — including variance, skewness, and kurtosis parameters — depend on explanatory variables as well rather than being constant (RIGBY AND STASINOPOULOS, 2005). A motivating example to bear in mind is the daily average price of energy, which we present in detail in SECTION 2.5 for Spain as a specific case study. When the energy market is in a calm state, which implies relatively low prices alongside a moderate volatility, then the oil price exhibits positive correlation with the mean of the conditional energy price distribution, but the variance is usually constant across observations. In contrast, when the energy market is nervous, which implies relatively high and volatile prices, then also the variance of energy prices is strongly affected by the oil price. This latter possible pattern cannot be addressed by existing Markov-switching regression models. As a consequence, by neglecting the strong heteroskedasticity in the process, price forecasts may severely under- or overestimate the associated uncertainty. This is problematic in scenarios where the interest lies not only in the expected prices, but also in quantiles, e.g. when the costs of forecast errors are asymmetric.

Since their introduction in the seminal work of RIGBY AND STASINOPOULOS (2005) a little more than a decade ago, GAMLSS have emerged as the standard framework for distributional regression models, where not only the mean, but also other parameters of the response distribution are modeled as potentially smooth functions of a given set of explanatory variables. Over the last decade, GAMLSS have been applied in a variety of

different fields, ranging from the analysis of insurance (HELLER *et al.*, 2007) and long-term rainfall data (VILLARINI *et al.*, 2010) over phenological research (HUDSON, 2010) and energy studies (VOUDOURIS *et al.*, 2011) to clinical applications, including long-term survival models (DE CASTRO *et al.*, 2010), childhood obesity (BEYERLEIN *et al.*, 2008), and measurement errors (MAYR *et al.*, 2017).

GAMLSS are applied primarily to data where it is reasonable to assume that the given observations are independent of each other. This is rarely the case when the data have a time series structure. In fact, when the data are collected over time, as e.g. daily energy prices, then the functional relationship between response and explanatory variables may actually change over time. This results in serially correlated residuals due to an under- or overestimation of the true functional relationship. To exploit the flexibility of GAMLSS also within time series settings, we propose a novel class of flexible latent-state time series regression models, which we call Markov-switching GAMLSS. In contrast to conventional Markov-switching regression models, the presented methodology allows us to model different state-dependent parameters of the response distribution as potentially smooth functions of a given set of explanatory variables.

A practical challenge that emerges with the flexibility of Markov-switching GAMLSS is the potentially high dimension of the set of possible model specifications. Each of the parameters of the response distribution varies across two or more states, and each of the associated predictors may involve several explanatory variables, the effect of which may even need to be estimated non-parametrically. Thus, a grid-search approach for model selection, e.g. based on information criteria, is usually practically infeasible. We therefore propose the MS-gamboostLSS algorithm for model fitting, which incorporates the gradient boosting framework into Markov-switching GAMLSS. Gradient boosting emerged from the field of machine learning, but was later adapted to estimate statistical models (cf. MAYR *et al.*, 2014). The basic idea is to iteratively apply simple regression functions (which are denoted as base-learners) for each potential explanatory variable one-by-one and to select in every iteration only the best performing one. The final solution is then an ensemble of the selected base-learner fits including only the most important variables. The design of the algorithm thus leads to automated variable selection and is even feasible for high-dimensional data settings, where the number of variables exceeds the number of observations.

This chapter is structured as follows: in SECTION 2.2, we introduce the different components of Markov-switching GAMLSS and discuss the underlying dependence assumptions. In SECTION 2.3, we derive the MS-gamboostLSS algorithm and give a brief overview of related topics, including model selection. The synergy of HMMs and GAMLSS,

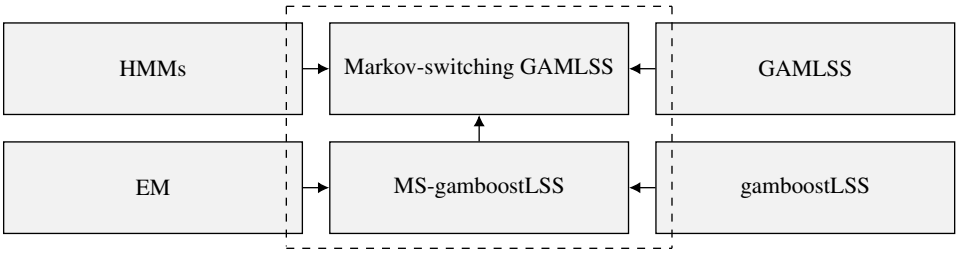

FIGURE 2.1: *Overview of the Markov-switching GAMLSS framework. In* SECTION *2.2, we introduce the components of Markov-switching GAMLSS and discuss the underlying dependence assumptions, which involve features from both HMMs and GAMLSS. In* SECTION *2.3, we present the MS-gamboostLSS algorithm, which incorporates gradient boosting into Markov-switching GAMLSS.*

which lies at the core of this work, is illustrated in FIGURE 2.1. In SECTION 2.4, we assess the suggested approach in simulation experiments, where we consider both linear and non-linear base-learners. In SECTION 2.5, we illustrate the proposed methodology in a real-data application, where we model the conditional distribution of the daily average price of energy in Spain over time.

2.2 Model formulation and dependence structure

In this section, we introduce the model formulation and dependence structure of Markov-switching GAMLSS, which constitute an extension of the closely related but less flexible and in fact nested class of Markov-switching generalized additive models (Markov-switching GAMs[2]; LANGROCK *et al.*, 2017).

2.2.1 The state process

Markov-switching GAMLSS comprise two stochastic processes, one of which is hidden and the other one is observed. The hidden process, which is denoted by $\{S_t\}_{t=1,\dots,T}$ and referred to as the state process, is modeled by a discrete-time, N-state Markov chain. Assuming the Markov chain to be time-homogeneous, we summarize the state transition probabilities, i.e. the probabilities of switching from state i at time t to state j at time $t+1$,

[2]Markov-switching GAMs as proposed in LANGROCK *et al.* (2017) extend conventional Markov-switching regression models to potentially smooth covariate effects and general response distributions from the exponential family.

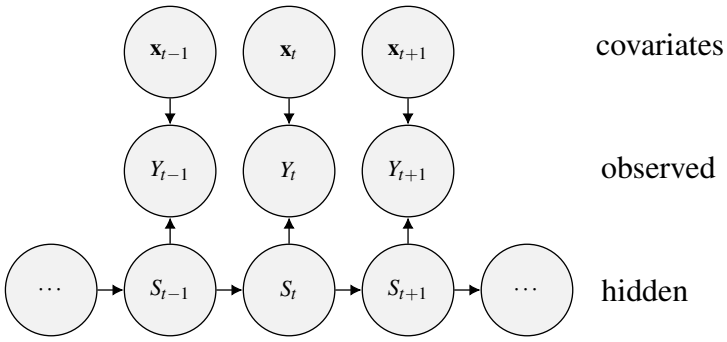

FIGURE 2.2: *Dependence structure of a Markov-switching GAMLSS. In contrast to the basic HMM, here the state-dependent process does not only depend on the underlying states but also on covariates.*

in the $N \times N$ transition probability matrix (t.p.m.) $\Gamma = (\gamma_{i,j})$, with elements

$$\gamma_{i,j} = \Pr(S_{t+1} = j | S_t = i), \tag{2.1}$$

$i, j = 1, \ldots, N$. The initial state probabilities, i.e. the probabilities of the process being in the different states at time 1, are summarized in the row vector $\boldsymbol{\delta} = (\delta_i)$, with elements

$$\delta_i = \Pr(S_1 = i), \tag{2.2}$$

$i = 1, \ldots, N$. If the Markov chain is assumed to be stationary, which is adequate in many applications, then the initial distribution is the stationary distribution, i.e. the solution to the equation system $\boldsymbol{\delta}\Gamma = \boldsymbol{\delta}$ subject to $\sum_{i=1}^{N} \delta_i = 1$ (ZUCCHINI *et al.*, 2016). However, if the Markov chain is not assumed to be stationary, then the initial state probabilities are additional parameters that need to be estimated. The state process is completely specified by the initial state and the state transition probabilities as given by EQUATIONS (2.1) and (2.2), respectively.

Throughout this chapter, we consider first-order Markov chains, i.e. we assume that the state process satisfies the Markov property, $\Pr(S_{t+1} = s_{t+1} | S_1 = s_1, \ldots, S_t = s_t) = \Pr(S_{t+1} = s_{t+1} | S_t = s_t)$, $t = 1, \ldots, T-1$. This simplifying dependence assumption is exploited in the likelihood calculations provided in SECTION 2.3.1. While certainly being a strong assumption, in practice it is often a good proxy for the actual dependence structure, and could in fact be relaxed to higher-order Markov chains if deemed necessary (ZUCCHINI *et al.*, 2016; cf. also SECTION 2.6 for an overview of possible model extensions).

2.2.2 The state-dependent process

The observed process, which is denoted by $\{Y_t\}_{t=1,\dots,T}$ and referred to as the state-dependent process, can take on either discrete or continuous values. We denote the conditional probability density function (p.d.f.) or, in the discrete case, probability mass function (p.m.f.), of Y_t by

$$f_Y\left(y_t; \theta_t^{(s_t)}\right) = f_Y\left(y_t; \theta_{1,t}^{(s_t)}, \dots, \theta_{K,t}^{(s_t)}\right). \tag{2.3}$$

In EQUATION (2.3), $\theta_t^{(s_t)} = (\theta_{1,t}^{(s_t)}, \dots, \theta_{K,t}^{(s_t)})$ is the parameter vector associated with the distribution assumed for Y_t. It may depend both on the current state, s_t, and on the explanatory variables at time t, $\mathbf{x}_t = (x_{1,t}, \dots, x_{P,t})$, with P denoting the number of variables included in the model (cf. FIGURE 2.2 for an illustration of the dependence structure). The first parameter of the response distribution, $\theta_{1,t}^{(s_t)}$, often denotes the conditional mean of Y_t. Depending on the distribution family assumed, the other parameters may relate to the conditional variance, the conditional skewness, and the conditional kurtosis, respectively, though other parameters are also possible. The set of possible distributions that can be specified for the response is not limited to the exponential family; in fact, any parametric distribution (including Box-Cox-transformed, zero-inflated, and mixture distributions; cf. RIGBY AND STASINOPOULOS, 2006; RIGBY *et al.*, 2019) can be considered. Notably, not all parameters contained in EQUATION (2.3) need to depend on the states or covariates. Assuming $\theta_{1,t}^{(s_t)}$ to be the conditional mean of Y_t and treating all other parameters as nuisance parameters (which depend on the states but not on covariates), for instance, leads to the nested special case of Markov-switching GAMs (cf. DE SOUZA AND HECKMAN, 2014; LANGROCK *et al.*, 2017).

As the parameters are possibly constrained (the conditional variance, for instance, typically needs to be strictly positive), we introduce a monotonic link function, which is denoted by $g_k(\theta_{k,t}^{(s_t)})$, for each parameter $\theta_{k,t}^{(s_t)}$, $k = 1, \dots, K$. The link function maps the constrained parameter onto some real-valued predictor function, which is denoted by $\eta_k^{(s_t)}(\mathbf{x}_t)$, the choice of which is determined by the respective parameter constraints. For instance, the log link function, $g_k(\theta_{k,t}^{(s_t)}) = \log(\eta_k^{(s_t)}(\mathbf{x}_t))$, is typically chosen for the conditional variance, such that the inverse function, $\theta_{k,t}^{(s_t)} = \exp(\eta_k^{(s_t)}(\mathbf{x}_t))$, is strictly positive. The form of the predictor function is determined by the specification of the base-learners, the discussion of which is subject of SECTION 2.3.2.

The variables $Y_1, \dots, Y_T$ are assumed to be conditionally independent of each other, given the states, as illustrated in the graphical model depicted in FIGURE 2.2. While

certainly being an adequate assumption in many applications, serial correlation that is not sufficiently captured by the basic model could potentially be modeled using autoregressive terms in the state-dependent process if deemed necessary (ZUCCHINI *et al.*, 2016; cf. also SECTION 2.6 for an overview of possible model extensions).

2.3 Model fitting

In this section, we derive the MS-gamboostLSS algorithm to estimate the state transition probabilities as given by EQUATION (2.1), the initial state probabilities as given by EQUATION (2.2), and the state-dependent parameters of the response distribution contained in EQUATION (2.3).

2.3.1 The MS-gamboostLSS algorithm

The MS-gamboostLSS algorithm comprises an outer and an inner cycle, which incorporate two different model fitting procedures into a joint algorithm. The outer cycle is the EM algorithm (BAUM AND PETRIE, 1966; BAUM *et al.*, 1970; DEMPSTER *et al.*, 1977; WELCH, 2003), which constitutes a popular method for iteratively maximizing the likelihood of a statistical model in the presence of missing data, and has become one of the standard procedures for model fitting in HMMs. The inner cycle is a weighted version of the gamboostLSS algorithm (MAYR *et al.*, 2012; HOFNER *et al.*, 2016), which is exploited to carry out one part of the maximization (M-) step of the EM algorithm, namely the estimation of the state-dependent parameters of the response distribution contained in EQUATION (2.3).

The missing data — or, more precisely, functions of the missing data — can be estimated, which is referred to as the expectation (E-) step. Based on the obtained estimates, the complete-data log-likelihood (CDLL; i.e. the joint log-likelihood of the observations and the states) is then maximized with respect to the state transition probabilities as given by EQUATION (2.1), the initial state probabilities as given by EQUATION (2.2), and the state-dependent parameters of the response distribution contained in EQUATION (2.3), which is referred to as the M-step.

To derive the CDLL, we represent the state sequence $\{S_t\}_{t=1,\dots,T}$ (i.e. the missing data) by the binary random variables $u_i(t) = \mathbb{1}_{\{S_t=i\}}$ and $v_{i,j}(t) = \mathbb{1}_{\{S_{t-1}=i,S_t=j\}}$, $i,j = 1,\dots N$, $t = 1,\dots,T$ (i.e. functions of the missing data). Assuming the $u_i(t)$'s and $v_{i,j}(t)$'s to be

observed, the CDLL can be written as

$$\begin{aligned} l^{\text{CDLL}}(\theta|y_1,\ldots,y_T) &= \log\left(\delta_{s_1}\prod_{t=2}^{T}\gamma_{s_{t-1},s_t}\prod_{t=1}^{T}f_Y\left(y_t;\theta_t^{(s_t)}\right)\right) \\ &= \log(\delta_{s_1})+\sum_{t=2}^{T}\log(\gamma_{s_{t-1},s_t})+\sum_{t=1}^{T}\log\left(f_Y\left(y_t;\theta_t^{(s_t)}\right)\right) \\ &= \underbrace{\sum_{i=1}^{N}u_i(1)\log(\delta_i)}_{\text{dependent on } \delta_i,\ i=1,\ldots,N}+\underbrace{\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{t=2}^{T}v_{i,j}(t)\log(\gamma_{i,j})}_{\text{dependent on } \gamma_{i,j},\ i,j=1,\ldots,N} \\ &\quad+\underbrace{\sum_{i=1}^{N}\sum_{t=1}^{T}u_i(t)\log\left(f_Y\left(y_t;\theta_t^{(i)}\right)\right)}_{\text{dependent on } \eta_k^{(i)}(\mathbf{x}_t),\ k=1,\ldots,4}. \end{aligned} \tag{2.4}$$

Note that the CDLL as given by EQUATION (2.4) consists of three separate summands, each of which only depends on i) $\delta=(\delta_i)$, ii) $\Gamma=(\gamma_{i,j})$, and iii) $\theta_t^{(i)}=(g_k^{-1}(\eta_k^{(i)}(\mathbf{x}_t))), i=1,\ldots,N$, which considerably simplifies the maximization in the M-step.

The E-step consists of the computation of the conditional expectations of the $u_i(t)$'s and $v_{i,j}(t)$'s, namely the $\hat{u}_i(t)$'s and $\hat{v}_{i,j}(t)$'s, respectively. To compute these conditional expectations, we require the forward and backward probabilities. The forward probabilities, which are denoted as $\alpha_t(i)=f(y_1,\ldots,y_t,\ S_t=i|\mathbf{x}_1,\ldots,\mathbf{x}_t)$, are summarized in the row vectors $\alpha_t=(\alpha_t(1),\ldots,\ \alpha_t(N))$, which can be evaluated via the forward algorithm by applying the recursion

$$\begin{aligned} \alpha_1 &= \delta\mathbf{P}(y_1); \\ \alpha_t &= \alpha_{t-1}\Gamma\mathbf{P}(y_t), \end{aligned} \tag{2.5}$$

$t=2,\ldots,T$, with $N\times N$ diagonal matrix

$$\mathbf{P}(y_t)=\begin{pmatrix} f_Y\left(y_t;\theta_t^{(1)}\right) & & 0 \\ & \ddots & \\ 0 & & f_Y\left(y_t;\theta_t^{(N)}\right) \end{pmatrix}. \tag{2.6}$$

The backward probabilities, which are denoted as $\beta_t(j)=f(y_{t+1},\ldots,y_T|S_t=j,\mathbf{x}_{t+1},\ldots,\mathbf{x}_T)$, are summarized in the row vectors $\beta_t=(\beta_t(1),\ldots,\beta_t(N))$, which can be evaluated via the backward algorithm by applying the recursion

$$\begin{aligned} \beta_T &= 1; \\ \beta_t^\top &= \Gamma\mathbf{P}(y_{t+1})\beta_{t+1}^\top, \end{aligned}$$

$t = T-1,\ldots,1$, with $\mathbf{P}(y_{t+1})$ as defined in EQUATION (2.6) above. We let $\alpha_t^{[m]}(i)$ and $\beta_t^{[m]}(j)$ denote the forward and backward probabilities obtained in the m-th iteration, which are computed using the predictors obtained in the $(m-1)$-th iteration (or offset values in case of the first iteration).

The m-th E-step involves the computation of the conditional expectations of the $u_i(t)$'s and $v_{i,j}(t)$'s given the current parameter estimates, which leads to the following results:

- Since $\hat{u}_i(t) = \Pr(S_t = i|y_1,\ldots,y_T,\mathbf{x}_1,\ldots,\mathbf{x}_T) = f(y_1,\ldots,y_t,S_t = i|\mathbf{x}_1,\ldots,\mathbf{x}_T)$ $f(y_{t+1},\ldots,y_T|S_t = i,\mathbf{x}_1,\ldots,\mathbf{x}_T)/f(y_1,\ldots,y_T|\mathbf{x}_1,\ldots,\mathbf{x}_T)$ and $f(y_1,\ldots,y_T|\mathbf{x}_1,\ldots,$ $\mathbf{x}_T) = \sum_{i=1}^N f(y_1,\ldots,y_T,S_t = i|\mathbf{x}_1,\ldots,\mathbf{x}_T)$, it follows immediately from the definition of the forward and backward probabilities that

$$\hat{u}_i^{[m]}(t) = \frac{\alpha_t^{[m]}(i)\beta_t^{[m]}(i)}{\sum_{k=1}^N \alpha_T^{[m]}(k)}, \tag{2.7}$$

  $i = 1,\ldots,N, t = 1,\ldots,T$.

- Since $\hat{v}_{i,j}(t) = \Pr(S_{t-1} = i, S_t = j|y_1,\ldots,y_T,\mathbf{x}_1,\ldots,\mathbf{x}_T) = f(y_1,\ldots,y_{t-1},S_{t-1} = i|\mathbf{x}_1,$ $\ldots,\mathbf{x}_T)\Pr(S_t = j|S_{t-1} = i)f(y_t,\ldots,y_T|S_t = j,\mathbf{x}_1,\ldots,\mathbf{x}_T)/f(y_1,\ldots,y_T|\mathbf{x}_1,\ldots,\mathbf{x}_T)$, it follows immediately from the definition of the forward, backward, and state transition probabilities that

$$\hat{v}_{i,j}^{[m]}(t) = \frac{\alpha_{t-1}^{[m]}(i)\hat{\gamma}_{i,j}^{[m-1]} f_Y\left(y_t;\hat{\theta}_t^{(j)[m-1]}\right)\beta_t^{[m]}(j)}{\sum_{j=1}^N \alpha_T^{[m]}(j)},$$

  $i,j = 1,\ldots,N, t = 1,\ldots,T$.

The m-th M-step involves the maximization of the CDLL with the $u_i(t)$'s and $v_{i,j}(t)$'s replaced by their current conditional expectations with respect to the model parameters:

- As only the first term in the CDLL depends on δ_i, using a Lagrange multiplier to ensure $\sum_{i=1}^N \hat{\delta}_i^{[m]} = 1$ results in

$$\hat{\delta}_i^{[m]} = \frac{\hat{u}_i^{[m]}(1)}{\sum_{i=1}^N \hat{u}_i^{[m]}(1)} = \hat{u}_i^{[m]}(1),$$

  $i = 1,\ldots,N$.

- As only the second term in the CDLL depends on $\gamma_{i,j}$, using a Lagrange multiplier

to ensure $\sum_{j=1}^{N} \hat{\gamma}_{i,j}^{[m]} = 1$, $i = 1, \ldots, N$, results in

$$\hat{\gamma}_{i,j}^{[m]} = \frac{\sum_{t=2}^{T} \hat{v}_{i,j}^{[m]}(t)}{\sum_{k=1}^{N} \sum_{t=2}^{T} \hat{v}_{i,k}^{[m]}(t)},$$

$i, j = 1, \ldots, N$.

- As only the third term in the CDLL depends on the state-dependent parameters of the response distribution contained in EQUATION (2.3), the optimization problem effectively reduces to maximizing the weighted log-likelihood of a separate, conventional GAMLSS for each state, where the t-th observation is weighted by the $\hat{u}_i^{[m]}(t)$'s as given by EQUATION (2.7). We can therefore exploit the gamboostLSS algorithm (MAYR *et al.*, 2012; HOFNER *et al.*, 2016) to iteratively maximize this weighted log-likelihood. More specifically, we consider the computationally more efficient non-cyclical variant of the gamboostLSS algorithm (THOMAS *et al.*, 2018)[3]:

  - Initialize the additive predictors $\hat{\eta}_k^{(i)[0]}(\mathbf{x}_t), i = 1, \ldots, N, k = 1, \ldots, 4, t = 1, \ldots T$, with offset values. For each of the additive predictors, specify a set of base-learners $h_{k,1}^{(i)}(x_{1,t}), \ldots, h_{k,J_k^{(i)}}^{(i)}(x_{J_k^{(i)},t})$ (e.g. simple linear models or penalized B-splines, i.e. P-splines; EILERS AND MARX, 1996), where $J_k^{(i)}$ denotes the cardinality of the set of base-learners specified for $\eta_k^{(i)}(\mathbf{x}_t)$.
  - For $i = 1$ to N:
    - For $n = 1$ to $n_{\text{stop}}^{(i)}$:
      - For $k = 1$ to 4:
        - Compute the gradients of the CDLL with respect to $\eta_k^{(i)}(\mathbf{x}_t)$ (using the current estimates $\hat{u}_i^{[m]}(t)$ as given by EQUATION (2.7) and $\hat{\theta}_t^{(i)[n-1]} = g_k^{-1}(\hat{\eta}_k^{(i)[n-1]}(\mathbf{x}_t)), k = 1, \ldots, 4$),

$$\begin{aligned} \nabla_{k,t}^{(i)} &= \frac{\partial l^{\text{CDLL}}(\theta | y_1, \ldots, y_T)}{\partial \eta_k^{(i)}(\mathbf{x}_t)} \\ &= \frac{\partial \sum_{t=1}^{T} \hat{u}_i^{[m]}(t) \log\left(f_Y\left(y_t; \hat{\theta}_t^{(i)[n-1]}\right)\right)}{\partial \eta_k^{(i)}(\mathbf{x}_t)}, \end{aligned}$$

[3] While the cyclical gamboostLSS algorithm estimates the different distribution parameters separately, the non-cyclical variant proposed in THOMAS *et al.* (2018) incorporates an additional selection step of the best-fitting distribution parameter in each boosting iteration.

$t = 1, \ldots, T$, and fit each of the base-learners contained in the set of base-learners specified for $\eta_k^{(i)}(\mathbf{x}_t)$ to these gradients.

· Select the best-fitting base-learner $h_{k,j^*}^{(i)}(x_{j^*,t})$ by the residual sum of squares of the base-learner fit with respect to the gradients,

$$j^* = \underset{j \in (1,\ldots,J_k^{(i)})}{\operatorname{argmin}} \sum_{t=1}^{T} \left(\nabla_{k,t}^{(i)} - \hat{h}_{k,j}^{(i)}(x_{j,t}) \right)^2.$$

· Select, among the base-learners selected the previous loop, the best-fitting base-learner $\hat{h}_{k^*,j^*}^{(i)}(x_{j^*,t})$ by the weighted log-likelihood,

$$k^* = \underset{k \in (1,\ldots,4)}{\operatorname{argmax}} \sum_{t=1}^{T} \hat{u}_i^{[m]}(t) \log \left(f_Y \left(y_t; \hat{\theta}_t^{(i)[n-1]} \right) \right),$$

where $\hat{\theta}_{k,t}^{(i)[n-1]}$ is replaced by its potential update, $g_k^{-1}(\hat{\eta}_k^{(i)[n-1]}(\mathbf{x}_t) + \text{sl} \cdot \hat{h}_{k,j^*}^{(i)}(x_{j^*,t}))$, to update the corresponding predictor,

$$\hat{\eta}_{k^*}^{(i)[n]}(\mathbf{x}_t) = \hat{\eta}_{k^*}^{(i)[n-1]}(\mathbf{x}_t) + \text{sl} \cdot \hat{h}_{k^*,j^*}^{(i)}(x_{j^*,t}),$$

where $0 < \text{sl} < 1$ is some small step length (typically, $\text{sl} = 0.1$).

∗ Set $\hat{\eta}_k^{(i)[n]}(\mathbf{x}_t) = \hat{\eta}_k^{(i)[n-1]}(\mathbf{x}_t)$ for all $k \neq k^*$.

– Use the predictors obtained in the final iteration as estimates obtained in the m-th M-step, $\hat{\eta}_k^{(i)[m]}(\mathbf{x}_t) = \hat{\eta}_k^{(i)[n_{\text{stop}}^{(i)}]}(\mathbf{x}_t)$ for all i, k.

The MS-gamboostLSS algorithm alternates between the E- and the M-step, each of which involves $n_{\text{stop}}^{(i)}$ boosting iterations for each state, i, until some convergence threshold, e.g. based on the difference between the CDLLs (or, alternatively, based on the difference between the estimates) obtained in two consecutive iterations, is satisfied.

2.3.2 Specification of base-learners

The specification of the base-learners, $h_{k,j}^{(i)}(x_{j,t})$, which are used to fit the gradient vectors, is crucial, as they define the type of predictor effect. If the base-learners have a linear form, then the resulting fit is also linear, whereas if non-linear base-learners are chosen, then this fit may also be non-linear. Generally, base-learners can be any kind of prediction functions — in the classical machine learning context, gradient boosting is most often applied with trees or stumps as base-learners (RIDGEWAY, 1999). In the specific case of

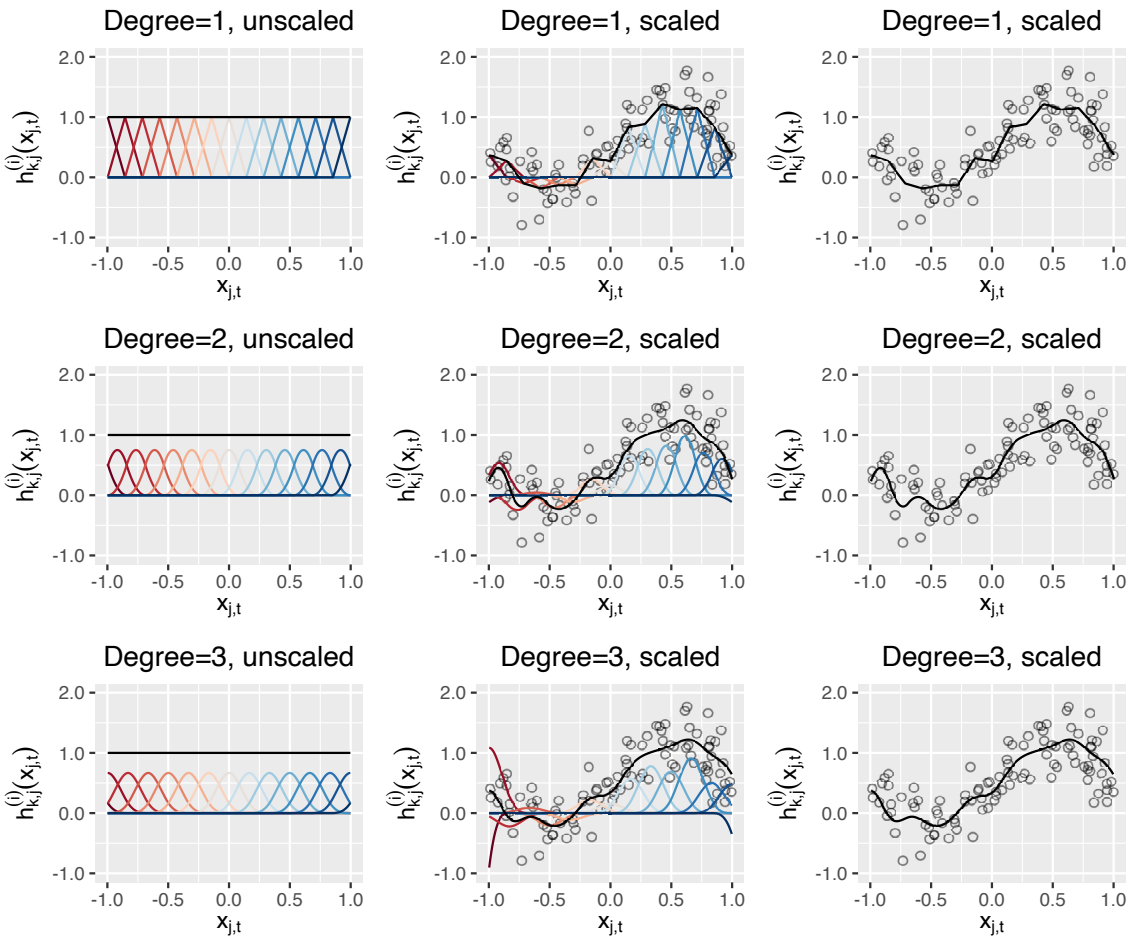

FIGURE 2.3: *Construction of B-spline bases of different degrees. Displayed are unscaled (left panel) and scaled (middle and right panel) B-spline bases of degrees $d = 1$ (upper panel), $d = 2$ (middle panel), and $d = 3$ (lower panel), respectively. Colors were used to indicate individual basis functions. Basis function coefficients were estimated via least squares (i.e. without penalization).*

boosting algorithms for statistical modeling, it is, however, reasonable to select regression-type functions that can be combined to additive models (MAYR *et al.*, 2014).

Due to their high flexibility, popular base-learners are P-splines (EILERS AND MARX, 1996) based on B-spline basis functions (DE BOOR, 1978). Following FAHRMEIR *et al.* (2013), B-spline basis functions of degree $d = 0$ can be constructed as follows:

$$B_q^0(x_{j,t}) = \mathbb{1}_{\{\kappa_q \leq x_{j,t} < \kappa_{q+1}\}} = \begin{cases} 1 & \kappa_q \leq x_{j,t} < \kappa_{q+1} \\ 0 & \text{otherwise,} \end{cases} \tag{2.8}$$

$q = 1, \ldots, Q-1$, where κ_q, denotes the location and Q is the number of the (typically equidistantly spaced) knots. Starting from basis functions of degree $d = 0$ as given by EQUATION (2.8), basis functions of general (higher) degree $d > 0$ can be constructed

recursively based on basis functions of degree $d-1$ via

$$B_q^d(x_{j,t}) = \frac{x_{j,t} - \kappa_{q-1}}{\kappa_q - \kappa_{q-1}} B_{q-1}^{d-1}(x_{j,t}) + \frac{\kappa_{q+1} - x_{j,t}}{\kappa_{q+1} - \kappa_q} B_q^{d-1}(x_{j,t}).$$

The base-learner is then a linear combination of the B-spline basis functions scaled by some basis function coefficients, which are denoted by $\phi_{j,k,q}^{(i)}$, such that

$$h_{k,j}^{(i)}(x_{j,t}) = \sum_{q=1}^{Q} \phi_{j,k,q}^{(i)} B_q^d(x_{j,t}). \tag{2.9}$$

The construction of B-spline bases is illustrated in FIGURE 2.3 for basis functions of different degrees.

To avoid overfitting, which naturally occurs when minimizing the sum of squared residuals between B-splines as given by EQUATION (2.9) and the gradient vectors, we add a roughness penalty based on higher-order differences between adjacent basis function coefficients to the sum of squared residuals. Differences of order m can be evaluated recursively via

$$\begin{aligned} \Delta^1 \phi_{j,k,q}^{(i)} &= \phi_{j,k,q}^{(i)} - \phi_{j,k,q-1}^{(i)}; \\ \Delta^m \phi_{j,k,q}^{(i)} &= \Delta^1 \left(\Delta^{m-1} \phi_{j,k,q}^{(i)} \right), \end{aligned}$$

$q = m+1, \ldots, Q$, which leads to the penalized least squares criterion

$$\hat{h}_{k,j}^{(i)}(x_{j,t}) = \underset{\phi_{j,k,1}^{(i)}, \ldots, \phi_{j,k,Q}^{(i)}}{\operatorname{argmin}} \underbrace{\left(\nabla_{k,t}^{(i)} - h_{k,j}^{(i)}(x_{j,t}) \right)^2}_{\text{goodness of fit}} + \lambda_{j,k}^{(i)} \underbrace{\sum_{q=m+1}^{Q} \left(\Delta^m \phi_{j,k,q}^{(i)} \right)^2}_{\text{smoothness}},$$

where $\lambda_{j,k}^{(i)}$ denotes a smoothing parameter that governs the weight of the penalty term and thus determines the smoothness of the resulting base-learner fit.

P-spline base-learners are typically applied with fixed, low effective degrees of freedom (i.e. strong penalization), which are not tuned for the different boosting iterations (the R package mboost, on which the implementation of the MS-gamboostLSS algorithm is based, uses four effective degrees of freedom as the default option; cf. HOFNER, 2011, for details). However, as the same P-spline base-learner can be selected as the best-performing base-learner and updated in several boosting iterations, the resulting solution can have arbitrarily large complexity (i.e. wiggliness). The complexity increases as the number of boosting iterations increases. More advanced base-learners that can be used are interaction terms (e.g. based on tensor product P-splines) as well as random or spatial effects (e.g.

based on Markov random fields). For an overview of available base-learners, cf. MAYR *et al.* (2012).

2.3.3 Choice of the number of boosting iterations

The stopping iterations, $n_{\text{stop}}^{(i)}$, are the main tuning parameters for boosting algorithms. They control the variable selection properties of the algorithm and the smoothness of the estimated effects. They represent the classical trade-off between bias and variance in statistical modeling: using more boosting iterations leads to larger and more complex models with smaller bias but larger variance, while stopping the algorithm earlier leads to sparser, less complex models with less variance but larger bias. Without early stopping, i.e. running the (gamboostLSS) algorithm (within the M-step) until convergence, the resulting fit converges to the maximum likelihood estimate (MAYR *et al.*, 2012; if this estimate exists for the given model).

Choosing an optimal number of boosting iterations is typically achieved via K-fold cross-validation. For some set $\Lambda = \mathbf{n}_{\text{stop}}^{(1)} \times \cdots \times \mathbf{n}_{\text{stop}}^{(N)} \subset \mathbb{N}^N$ we follow CELEUX AND DURAND (2008) and proceed in the following way: first, we split the data into K distinct partitions (typically, $K \geq 10$), estimate the model based on $K-1$ partitions and compute the out-of-sample log-likelihood for the remaining partition (which is straightforward using the forward algorithm from SECTION 2.3.1, cf. EQUATION (2.5)). This procedure is repeated K times, i.e. until each partition has been out-of-sample once. The score of interest is then the average out-of-sample log-likelihood over all partitions, where the number of boosting iterations corresponding to the highest score is chosen.

2.3.4 Selecting the number of states

The choice of the number of states, N, is a rather difficult task — even though the vast majority of Markov-switching regression models appearing in the literature assume two states without any critical reasoning, there actually exists a variety of different methods for order selection in HMMs, which basically fall in two categories: on the one hand, a cross-validated likelihood approach can be used, as described in SECTION 2.3.3. On the other hand, information criteria such as Akaike's Information Criterion (AIC), the Bayesian Information Criterion (BIC), or the Integrated Completed Likelihood Criterion (BIERNACKI *et al.*, 2000; CELEUX AND DURAND, 2008) can be considered, all of which

result in a compromise between goodness of fit and model complexity.

One problem in practice, however, is that information criteria often tend to favor overly complex models. Real data typically exhibit more structure than can actually be captured by the model, which e.g. is the case if the true state-dependent distributions are too complex to be fully modeled by some (rather simple) parametric distribution or if certain temporal patterns are neglected in the model formulation. In the case of Markov-switching GAMLSS, additional states may be able to capture this further structure. As a consequence, the goodness of fit increases, which may outweigh the higher model complexity. However, as models with too many states are usually difficult to interpret and are therefore often not desired, information criteria should be considered as a rough guidance rather than as a deterministic decision rule, which should be treated with some caution. For an in-depth discussion of pitfalls, practical challenges, and pragmatic solutions regarding order selection in HMMs, we refer to POHLE *et al.* (2017).

2.4 Simulation experiments

To assess the performance of the suggested approach, we present two different simulation experiments, where we consider linear (cf. SECTION 2.4.1) and non-linear (cf. SECTION 2.4.2) relationships between the explanatory variables and the parameters of the response distribution.

2.4.1 Linear setting

For the linear setting, we use simple linear models as base-learners. In each of 100 simulation runs, we simulated 500 realizations from a 2-state Markov chain, $\{S_t\}_{t=1,\ldots,500}$, with t.p.m.

$$\Gamma = \begin{pmatrix} 0.95 & 0.05 \\ 0.05 & 0.95 \end{pmatrix},$$

and initial (stationary) state probabilities $\delta_i = 0.5$, $i = 1, 2$. Based on the simulated state sequence, we then draw 500 observations from a negative binomial distribution with state-

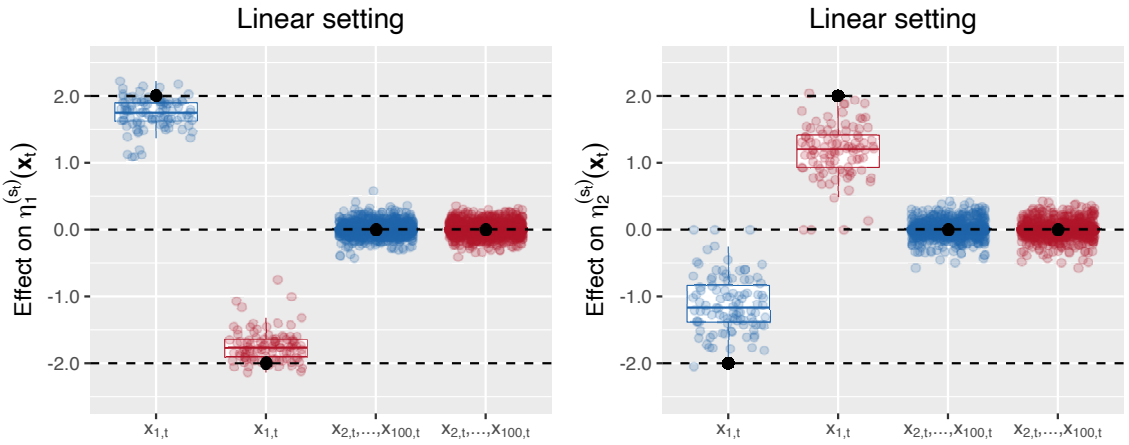

FIGURE 2.4: *Boxplots of the estimated state-dependent coefficients (linear setting). Displayed are the estimated state-dependent coefficients for the mean (left panel) and the dispersion (right panel) for states* 1 *(blue) and* 2 *(red) obtained in* 100 *simulation runs. True parameters (i.e. without shrinkage) are indicated by black dots. The estimated coefficients for all* 99 *non-informative covariates are visualized in a single boxplot for each state.*

dependent p.m.f.

$$f_Y\left(y_t;\theta_{1,t}^{(s_t)},\theta_{2,t}^{(s_t)}\right)=\frac{\Gamma\left(y_t+\theta_{2,t}^{(s_t)}\right)}{\Gamma(y_t+1)\Gamma\left(\theta_{2,t}^{(s_t)}\right)}\frac{\left(\frac{\theta_{1,t}^{(s_t)}}{\theta_{2,t}^{(s_t)}}\right)^{y_t}}{\left(\frac{\theta_{1,t}^{(s_t)}}{\theta_{2,t}^{(s_t)}+1}\right)^{\left(y_t+\theta_{2,t}^{(s_t)}\right)}},$$

where

$$\begin{aligned}
\log\left(\theta_{1,t}^{(1)}\right)&=\eta_1^{(1)}(\mathbf{x}_t)=2+2x_{1,t}+\sum_{j=2}^{100}0x_{j,t};\\
\log\left(\theta_{1,t}^{(2)}\right)&=\eta_1^{(2)}(\mathbf{x}_t)=2-2x_{1,t}+\sum_{j=2}^{100}0x_{j,t};\\
\log\left(\theta_{2,t}^{(1)}\right)&=\eta_2^{(1)}(\mathbf{x}_t)=2x_{1,t}+\sum_{j=2}^{100}0x_{j,t};\\
\log\left(\theta_{2,t}^{(2)}\right)&=\eta_2^{(2)}(\mathbf{x}_t)=-2x_{1,t}+\sum_{j=2}^{100}0x_{j,t},
\end{aligned}$$

and $x_{j,t}\sim\text{uniform}(-1,1)$, $j=1,\ldots,100$, $t=1,\ldots,500$. To assess the variable selection performance, we included 99 non-informative explanatory variables in each predictor. The stopping iterations were chosen via 20-fold cross-validation over the grid $\Lambda=\mathbf{n}_{\text{stop}}^{(1)}\times\mathbf{n}_{\text{stop}}^{(2)}$, $\mathbf{n}_{\text{stop}}^{(1)}=\mathbf{n}_{\text{stop}}^{(2)}=(100,200,400,800)$, where the average chosen number of boosting iterations was 435 (state 1) and 468 (state 2).

The sample means of the estimated off-diagonal t.p.m. entries, $\hat{\gamma}_{1,2}$ and $\hat{\gamma}_{2,1}$, were obtained as 0.047 (standard deviation: 0.020) and 0.047 (0.019), respectively, which apparently is very close to the true values. The estimated state-dependent coefficients obtained in 100 simulation runs are displayed in FIGURE 2.4: for the mean, $\theta_{1,t}^{(s_t)}$, the estimated coefficients are slightly shrunken towards zero, while for the dispersion, $\theta_{2,t}^{(s_t)}$, the shrinkage effect is quite large. The informative covariates were — on average — selected in 98.5 % of the cases (100.0 % for the mean and 97.0 % for the dispersion), while the non-informative ones were — on average — selected in 10.6 % of the cases (13.4 % for the mean and 7.7 % for the dispersion), which indicates on the one hand that the variable selection works quite well but on the other hand that there is a tendency towards too many covariates being included in the model (this apparently is a problem related to boosting in general rather than a specific one related to the MS-gamboostLSS algorithm; c.f. the simulation experiments presented in MAYR *et al.*, 2012).

Using a 3.6 GHz Intel® Core™ i7 CPU, the average computation time was 1.4 minutes for a (single) model (i.e. for a given number of boosting iterations), which is remarkably fast considering the fact that it involves variable selection among 100 potential explanatory variables.

2.4.2 Non-linear setting

Encouraged by the performance in the linear setting, we next present a non-linear setting using P-splines as base-learners, again simulating 500 realizations from a 2-state Markov chain with t.p.m.

$$\Gamma = \begin{pmatrix} 0.95 & 0.05 \\ 0.05 & 0.95 \end{pmatrix},$$

and initial (stationary) state probabilities $\delta_i = 0.5$, $i = 1,2$. We then draw 500 observations from a normal distribution with state-dependent p.d.f.

$$f_Y\left(y_t; \theta_{1,t}^{(s_t)}, \theta_{2,t}^{(s_t)}\right) = \frac{1}{\sqrt{2\pi {\theta_{2,t}^{(s_t)}}^2}} \exp\left(-\frac{\left(y_t - \theta_{1,t}^{(s_t)}\right)^2}{2{\theta_{2,t}^{(s_t)}}^2}\right),$$

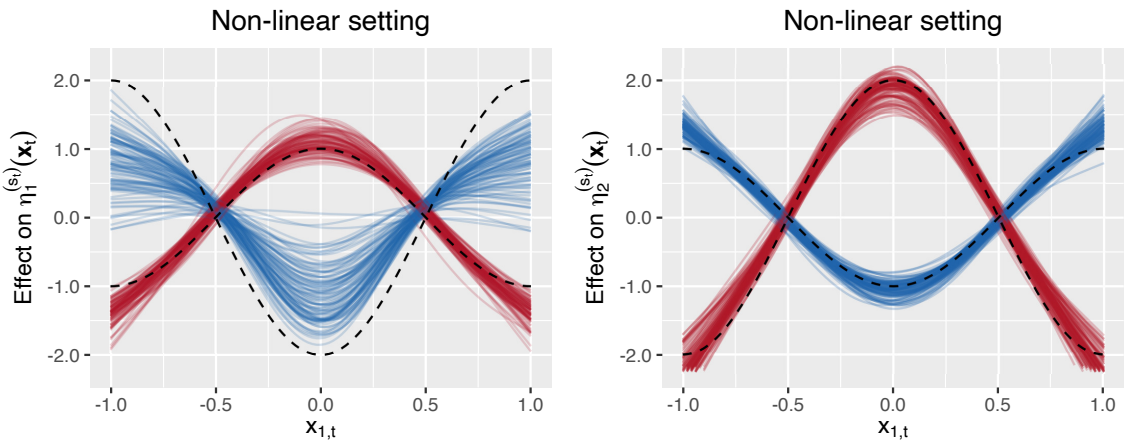

FIGURE 2.5: *Estimated state-dependent effects (non-linear setting). Displayed are the estimated state-dependent effects on the predictor for the mean (left panel) and the dispersion (right panel) for states* 1 *(blue) and* 2 *(red) obtained in* 100 *simulation runs. True effects (i.e. without shrinkage) are indicated by dashed lines. All effects have been centered around zero.*

where

$$\theta_{1,t}^{(1)} = \eta_1^{(1)}(\mathbf{x}_t) = 2 + 2\sin\big(\pi(x_{1,t} - 0.5)\big) + \sum_{j=2}^{100} 0x_{j,t};$$

$$\theta_{1,t}^{(2)} = \eta_1^{(2)}(\mathbf{x}_t) = -2 - \sin\big(\pi(x_{1,t} - 0.5)\big) + \sum_{j=2}^{100} 0x_{j,t};$$

$$\log\big(\theta_{2,t}^{(1)}\big) = \eta_2^{(1)}(\mathbf{x}_t) = \sin\big(\pi(x_{1,t} - 0.5)\big) + \sum_{j=2}^{100} 0x_{j,t};$$

$$\log\big(\theta_{2,t}^{(2)}\big) = \eta_2^{(2)}(\mathbf{x}_t) = -2\sin\big(\pi(x_{1,t} - 0.5)\big) + \sum_{j=2}^{100} 0x_{j,t},$$

and $x_{j,t} \sim \text{uniform}(-1,1)$, $j = 1,\ldots,100$, $t = 1,\ldots,500$. The stopping iterations were again chosen via 20-fold cross-validation over the grid $\Lambda = \mathbf{n}_{\text{stop}}^{(1)} \times \mathbf{n}_{\text{stop}}^{(2)}$, $\mathbf{n}_{\text{stop}}^{(1)} = \mathbf{n}_{\text{stop}}^{(2)} = (25, 50, 100, 200)$, where the average chosen number of boosting iterations was 141.5 (state 1) and 177 (state 2).

The sample means of the estimated off-diagonal t.p.m. entries, $\hat{\gamma}_{1,2}$ and $\hat{\gamma}_{2,1}$, were obtained as 0.050 (0.014) and 0.051 (0.016), respectively. The estimated state-dependent effects obtained in 100 simulation runs are displayed in FIGURE 2.5: as in SECTION 2.4.1, we observe a shrinkage effect (especially for the larger effects, i.e. the effects of $x_{1,t}$ on $\eta_1^{(1)}(\mathbf{x}_t)$ and $\eta_2^{(2)}(\mathbf{x}_t)$). In addition, a smoothing effect can be observed (particularly for very small and large values of $x_{1,t}$). The informative covariates were selected in all cases, while the non-informative ones were — on average — selected in 11.2 % of the cases (7.4

% for the mean and 15.0 % for the dispersion), which again indicates that the variable selection works quite well but apparently is not very conservative (particularly in the case of the dispersion, where the shrinkage effect is considerably smaller than the one for the mean, the average number of non-informative explanatory variables included in the model is fairly large).

For a given number of boosting iterations, model fitting took — on average — 7.8 minutes per (single) model, which again is quite remarkable considering the fact that it does not only involve variable selection among 100 potential covariates (as in the linear setting presented in SECTION 2.4.1) but also results in smooth fits (without relying on a computer-intensive smoothing parameter selection).

2.5 Application to energy prices in Spain

To illustrate the suggested approach in a real-data setting, we model the conditional distribution of the daily average price of energy in Spain (in c per kWh), EnergyPrice_t, over time. Our aim here is to present a simple case-study that provides some intuition and demonstrates the potential of Markov-switching GAMLSS, which is why we focus on a relatively simple model involving only one explanatory variable, the daily oil price (in EUR per barrel), OilPrice_t. The data, which are available in the R package MSwM (SANCHEZ-ESPIGARES AND LOPEZ-MORENO, 2014), cover 1,760 working days between February 4, 2002, and October 31, 2008 (cf. FONTDECABA *et al.*, 2009, for an overview of the data). As in SECTION 2.4.2, we assume a normal distribution for the EnergyPrice_t and fitted two different 2-state Markov-switching GAMLSS with state-dependent predictors for the conditional mean, $\theta_{1,t}^{(s_t)}$, and the conditional variance, $\theta_{2,t}^{(s_t)}$, considering i) simple linear models (linear model), and ii) P-splines (as detailed in SECTION 2.3.2; non-linear model) as base-learners. The stopping iterations were chosen via 20-fold cross-validation over the grid $\Lambda = \mathbf{n}_{\text{stop}}^{(1)} \times \mathbf{n}_{\text{stop}}^{(2)}$, $\mathbf{n}_{\text{stop}}^{(1)} = \mathbf{n}_{\text{stop}}^{(2)} = (25, 50, 100, \ldots, 3{,}200)$, which led to the optimal values $n_{\text{stop}}^{(1)} = 100$, $n_{\text{stop}}^{(2)} = 200$ (linear model) and $n_{\text{stop}}^{(1)} = 1{,}600$, $n_{\text{stop}}^{(2)} = 200$ (non-linear model). For the chosen stopping iterations, the computation times were 0.4 minutes (linear model) and 7.9 minutes (non-linear model).

The estimated state-dependent effects, as well as the locally decoded time series of daily energy prices, are visualized in Figures 2.6 and 2.7: according to both models, the oil price exhibits a (mostly) positive effect on the conditional mean of the energy price distribution, which essentially holds for both states. However, the linear model lacks the flexibility to capture the decreasing effect for $\text{OilPrice}_t \geq 60$ that is revealed by the non-

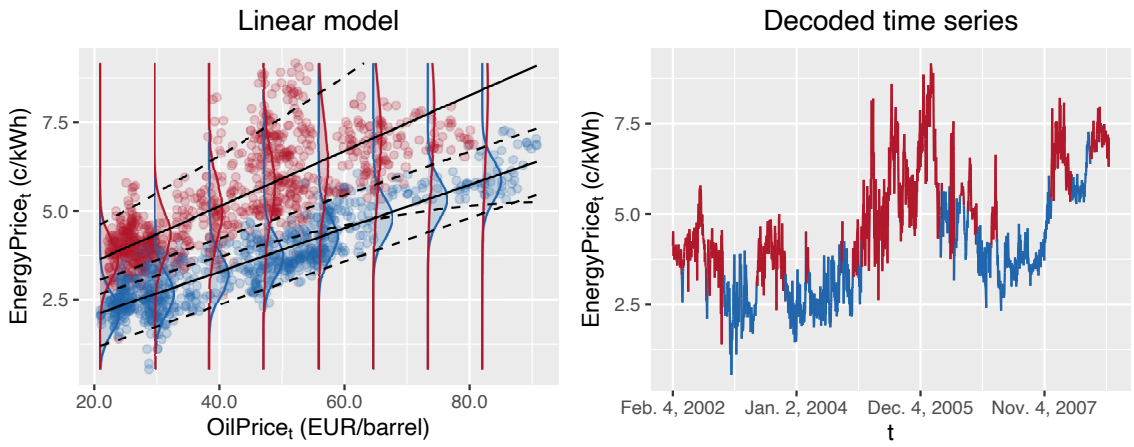

FIGURE 2.6: *Estimated state-dependent effects and decoded time series of daily energy prices (linear model). The plot displayed in the left panel shows the estimated state-dependent effects on the conditional mean (black solid lines) for states* 1 *(blue) and* 2 *(red) and the fitted state-dependent distributions for different oil prices (vertical p.d.f.s), which were computed based on the estimated state-dependent predictors for the conditional variance. Dashed lines indicate the* 0.05 *and* 0.95 *quantiles of the fitted state-dependent distributions. The plot displayed in the right panel shows the locally decoded time series of daily energy prices.*

linear model, which leads to a severe overestimation in that area. The effect on the conditional variance considerably differs across the two states: in state 1, the oil price has only a minor effect, whereas in state 2, the conditional variance is strongly affected by the oil price. As in the case of the conditional mean, the effect on the conditional variance clearly has a non-linear form (the volatility is relatively high for $40 \leq \text{OilPrice}_t < 60$ and relatively low for $40 > \text{OilPrice}_t \geq 60$), which is well-captured by the non-linear model but not captured by the linear model. The consequence is a severe under- (over-) estimation for $40 \leq \text{OilPrice}_t < 60$ ($40 > \text{OilPrice}_t \geq 60$), as indicated by the quantile curves for the linear model depicted in FIGURE 2.6. From an economic point of view, state 1 may be linked to a calm market regime (which implies relatively low prices alongside a moderate volatility). State 2, in contrast, may correspond to a nervous market regime (which implies relatively high prices alongside a high volatility).

The t.p.m. for the linear model was estimated as

$$\hat{\Gamma}_{\text{LM}} = \begin{pmatrix} 0.983 & 0.017 \\ 0.016 & 0.984 \end{pmatrix},$$

which implies the stationary distribution $(0.480, 0.520)$, indicating that about 48.0 % and 52.0 % of the observations were generated in states 1 and 2, respectively. The t.p.m. for

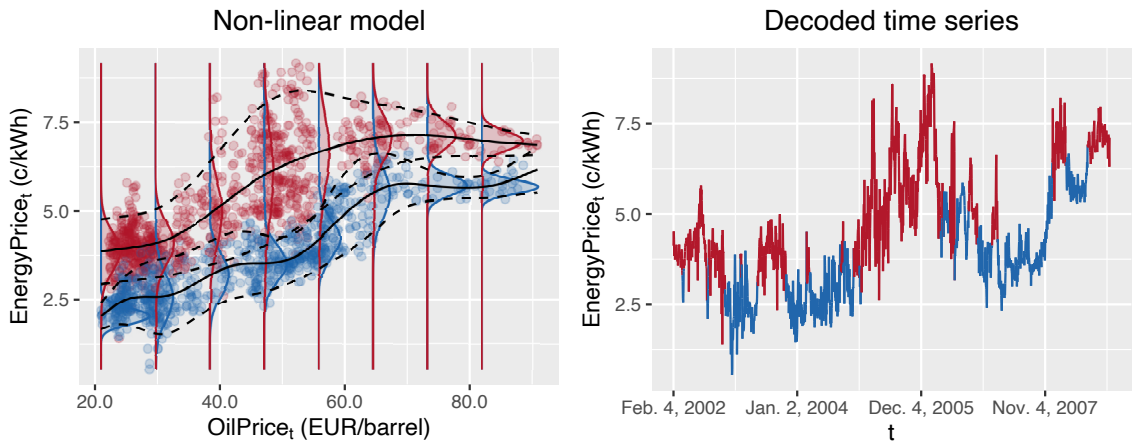

FIGURE 2.7: *Estimated state-dependent effects and decoded time series of daily energy prices (non-linear model). The plot displayed in the left panel shows the estimated state-dependent effects on the conditional mean (black solid lines) for states* 1 *(blue) and* 2 *(red) and the fitted state-dependent distributions for different oil prices (vertical p.d.f.s), which were computed based on the estimated state-dependent predictors for the conditional variance. Dashed lines indicate the* 0.05 *and* 0.95 *quantiles of the fitted state-dependent distributions. The plot displayed in the right panel shows the locally decoded time series of daily energy prices.*

the non-linear model was estimated as

$$\hat{\Gamma}_{\text{NLM}} = \begin{pmatrix} 0.983 & 0.017 \\ 0.016 & 0.984 \end{pmatrix},$$

which implies the stationary distribution $(0.483, 0.517)$, indicating that about 48.3 % and 51.7 % of the observations were generated in states 1 and 2, respectively[4]. In both cases, the estimated state transition probabilities indicate high persistence within the states (according to the fitted models, the average dwell-times within a state were — depending on the model and the state — between 58.8 and 62.5 days).

Quantile-quantile plots (qq-plots) and sample autocorrelation functions (ACFs) of one-step-ahead forecast pseudo-residuals are displayed in FIGURE 2.8. The qq-plots of the pseudo-residuals indicate some minor lack of fit regarding the marginal distribution under the linear model, which clearly improves when using non-linear base-learners. Although, for both models, the sample ACFs indicate some residual correlation, we consider the goodness of fit of the two models as satisfactory. The residual correlation can potentially be reduced using autoregressive terms in the state-dependent process (cf. SECTION 2.6),

[4]The stationary distributions slightly differ across the two models as the estimated state transition probabilities, which are rounded above, are not exactly equal.

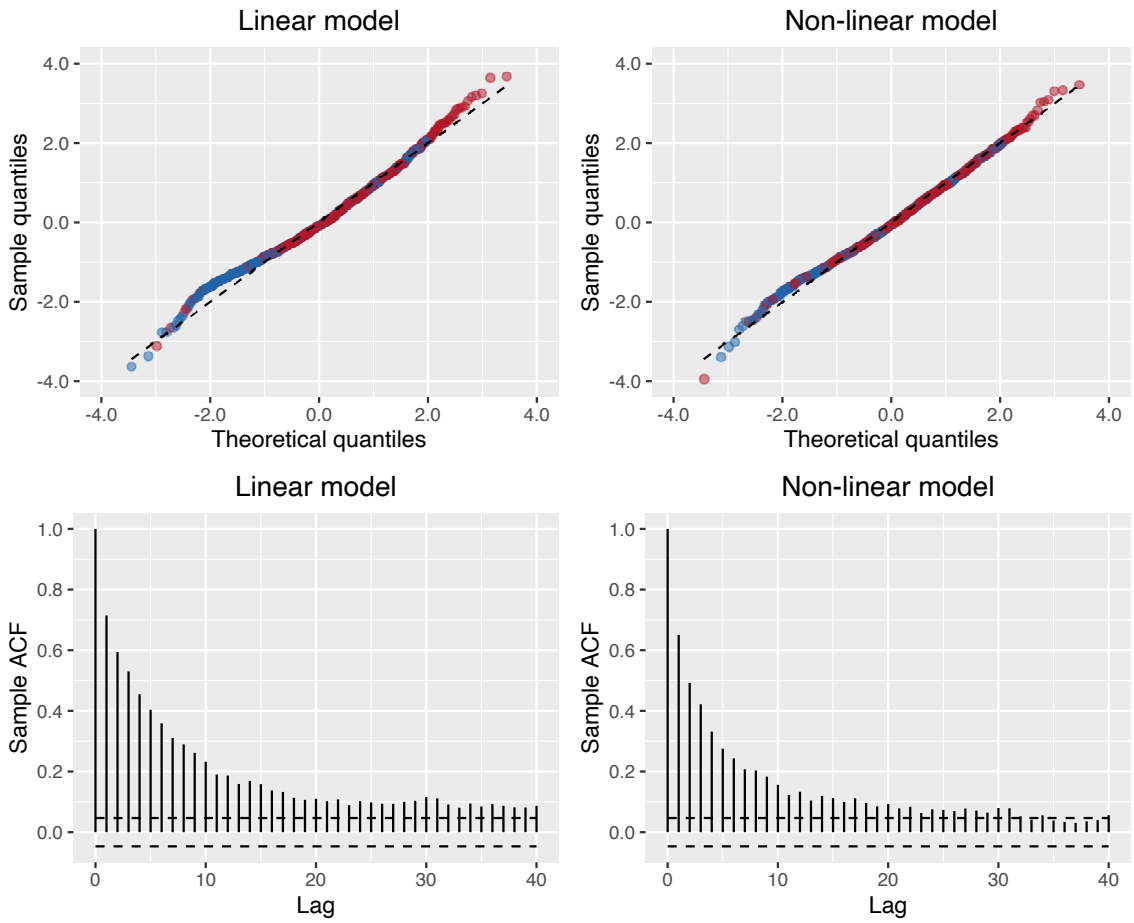

FIGURE 2.8: *Qq-plots and sample ACFs of one-step-ahead forecast pseudo-residuals (linear and non-linear model). Colors indicate the locally decoded states underlying the observed energy prices, where blue refers to state* 1 *and red refers to state* 2*.*

which, however, we refrain from investigating further as our aim here is to present an illustrative case study.

The results presented in this real-data application clearly demonstrate the potential of Markov-switching GAMLSS: by accounting for the state-switching dynamics in the model formulation, they allow to draw a precise picture of the response distribution at any point in time, which may be particularly useful in applications where the focus lies on short-term forecasting. Furthermore, a precise picture of the entire response distribution (which certainly includes not only the mean, but also variance and potentially skewness and kurtosis parameters) is crucial when the focus is shifted from the expected value towards the quantiles, which e.g. is the case in risk measurement and portfolio optimization applications (ACERBI AND TASCHE, 2002): estimating the value-at-risk of a given investment, for instance, requires the prediction of certain quantiles of the corresponding loss distribution (ROCKAFELLAR AND URYASEV, 2002), which could potentially be addressed using Markov-switching GAMLSS.

2.6 Discussion

In this chapter, we introduced Markov-switching GAMLSS as a novel class of flexible latent-state time series regression models, which can be used to model different state-dependent parameters of the response distribution as potentially smooth functions of a given set of explanatory variables. In addition, we demonstrated how gradient boosting can be exploited to avoid overfitting while simultaneously performing variable selection. Limitations of gradient boosting, particularly the fact that the design of the algorithm does not allow to compute standard errors for the effect estimates, also apply to the proposed MS-gamboostLSS algorithm. If the possible number of explanatory variables is small and the interest does not lie on prediction but on unbiased estimates, then the gamboostLSS algorithm in the M-step could be replaced by weighted versions of other algorithms that are commonly used to fit GAMLSS, e.g. those implemented in the R package gamlss (RIGBY AND STASINOPOULOS, 2005; STASINOPOULOS *et al.*, 2017; cf. also LANGROCK *et al.*, 2018, for an application to Markov-switching GAMLSS).

While we have assumed a relatively simple state architecture, the underlying dependence structure could potentially be extended in various ways: i) higher-order Markov chains could be used to allow the states to depend not only on the previous state but on a sequence of multiple previously visited states (ZUCCHINI *et al.*, 2016), ii) semi-Markov state processes could be used to specify arbitrary dwell-time distributions for the states (LANGROCK AND ZUCCHINI, 2011), and iii) hierarchical state processes could be used to infer states at multiple time scales (LEOS-BARAJAS *et al.*, 2017b; ADAM *et al.*, 2019a). Another potential feature of the latter approach is that multiple data streams collected at different temporal resolutions could be incorporated into a joint Markov-switching GAMLSS, which may be particularly useful in economic applications, where data often tend to be collected on a daily, monthly, or quarterly basis (cf. SECTION 4.4.2 for an example of such an application).

A distribution-free alternative to Markov-switching GAMLSS is provided by non-parametric Markov-switching quantile regression models (ADAM *et al.*, 2019e), where maximum likelihood estimation is commonly carried out by assuming an asymmetric Laplace distribution for the response. This approach, in a Bayesian setup, yields posterior consistent estimators even if the observations are not asymmetrically Laplace distributed (SRIRAM *et al.*, 2016). While certainly a useful alternative when the interest lies on only one or two specific quantiles, Markov-switching GAMLSS may be more feasible when the interest lies on the entire response distribution. In comparison with Markov-switching GAMLSS, one of the main disadvantages of non-parametric Markov-switching quantile

regression models is that they require one set of basis function coefficients not only for each state and explanatory variable but also for each quantile of interest, which leads to a large number of parameters to be estimated (and additional challenges that need to be addressed, e.g. quantile crossing). For a detailed comparison of GAMLSS with quantile regression models, we refer to RIGBY *et al.* (2013).

On a final note, we would like to raise awareness of the fact that the flexibility of Markov-switching GAMLSS can be both a blessing and a curse: in some applications, these models could be overparameterized, and models as complex as Markov-switching GAMLSS may not be appropriate even if they fit the data well (particularly in the case of short time series, overfitting may become a severe problem). In that regard, it is therefore worth mentioning that Markov-switching GAMLSS contain other, nested (i.e. less complex) HMM-type models, e.g. simple HMMs (ZUCCHINI *et al.*, 2016) or Markov-switching (generalized) linear and additive models (LANGROCK *et al.*, 2017; LANGROCK *et al.*, 2018). By specifying appropriate base-learners, the proposed MS-gamboostLSS algorithm can be used to fit all these nested special cases: using intercept-only terms (hence neglecting any covariate dependence), for instance, results in simple HMMs, while using simple linear models or P-splines for the conditional mean and intercept-only terms for the other parameters leads to Markov-switching (generalized) linear and additive models, respectively. Since none of the latter classes of models has been incorporated into the gradient boosting framework yet, the proposed MS-gamboostLSS algorithm, which lies at the core of this work, may provide a promising method for model fitting and variable selection not only in Markov-switching GAMLSS but also in a variety of other HMM-type models.

Chapter 3

Non-parametric inference in hidden Markov models for discrete-valued time series

Chapter 3

Non-parametric inference in hidden Markov models for discrete-valued time series[1]

> "If the assumed model is not the correct one, inferences can be worse than useless, leading to grossly misleading interpretations of the data."
>
> — *J.S. Simonoff*

Summary

In this chapter, we propose an effectively non-parametric approach to fitting HMMs to discrete-valued time series. While specifically for time series of counts, the Poisson distribution — or more flexible alternatives such as the negative binomial, zero-inflated, and mixture distributions — is often chosen for the state-dependent distributions, choosing an adequate class of parametric distributions remains difficult in practice, where an inadequate choice can have severe negative consequences. To overcome this problem, we estimate the state-dependent distributions in a completely data-driven way without the need to specify a parametric family of distributions, where a penalty based on higher-order differences between adjacent count probabilities is proposed to prevent overfitting. The suggested approach is assessed in simulation experiments and illustrated in a real-data application, where we model the distribution of the annual number of earthquakes over time. The proposed methodology is implemented in the R package countHMM.

[1] This chapter is based on ADAM *et al.* (2019c) and ADAM *et al.* (2019d).

3.1 Introduction

Over the last decades, HMMs have become increasingly popular for modeling time series where, at each point in time, a hidden state process selects among a finite set of possible distributions for the observations (ZUCCHINI *et al.*, 2016). In economic applications, for instance, the states of the Markov chain underlying the observations, which typically determines the state process, are often good proxies for economic regimes such as recessions or periods of economic growth (cf. GOLDFELD AND QUANDT, 1973; HAMILTON, 1989), while in ecology, they can regularly be linked to an animal's behavioral modes such as resting, foraging, or traveling (LANGROCK *et al.*, 2012b; MCCLINTOCK *et al.*, 2020). Other fields where HMMs are commonly applied include medicine (cf. WANG AND PUTERMAN, 2001; JACKSON AND SHARPLES, 2002), meteorology (cf. ZUCCHINI AND GUTTORP, 1991; PINSON AND MADSEN, 2012), marketing (cf. CHING *et al.*, 2004; NETZER *et al.*, 2008), and sports (cf. GREEN AND ZWIEBEL, 2018; ÖTTING *et al.*, 2020), to name but a few examples.

In various applications, it has been demonstrated that HMMs can be tailored to, *inter alia*, binary data (cf. SCHLIEHE-DIECKS *et al.*, 2012), positive real-valued data (cf. LANGROCK, 2012a), circular data (cf. BULLA *et al.*, 2012), categorical data (cf. MARUOTTI AND ROCCI, 2012), compositional data (cf. LANGROCK *et al.*, 2013b), and count data (cf. LAGONA *et al.*, 2015). In this chapter, we specifically focus on the latter type of time series, i.e. sequences of non-negative integers. For a general introduction to HMMs for discrete-valued time series, including the specific case of time series of counts, we refer to MACDONALD AND ZUCCHINI (1997), while WEIß (2018) provides a comprehensive overview of the various other classes of statistical models for discrete-valued time series. A motivating example for a time series of counts is the number of corporate defaults observed on a monthly, quarterly, or yearly basis: in periods of economic growth, these could be thought of as being generated by some distribution with relatively small mean, whereas during recessions, another distribution with relatively larger mean could be active. Although the economic regime is not directly observed, it still determines the observed corporate default counts (cf. LI AND CHENG, 2015; BERENTSEN *et al.*, 2018). For an application that is similar in spirit, cf. HAMBUCKERS *et al.* (2018), where both the number and the amount of a bank's operational losses are modeled using an HMM-type approach. Beyond economics, HMMs have been applied to time series of counts across a variety of scientific disciplines, including medicine (e.g. multiple sclerosis leisure counts; ALTMAN AND PETKAU, 2005), geology (e.g. volcanic eruption counts; BEBBINGTON, 2007), epidemiology (e.g. poliomyelitis counts; LE STRAT AND CARRAT, 1999), ecol-

ogy (e.g. pilot whale vocalization counts; POPOV *et al.*, 2017), and bioinformatics (e.g. T-lymphocyte counts; MARINO *et al.*, 2018), to name but a few examples.

Specifically for time series of counts, the Poisson distribution is often chosen for the state-dependent distributions, in which case one rate parameter is estimated for each state. While more flexible alternatives such as the negative binomial, zero-inflated, and mixture distributions can also be used — or distributions for bounded counts such as the binomial distribution — choosing an adequate class of parametric distributions remains difficult in practice, with potentially severe negative consequences in case an inadequate choice is made. For the specific case of continuous-valued time series, LANGROCK *et al.* (2015) propose a non-parametric approach to estimating the state-dependent distributions within an HMM based on linear combinations of B-spline basis functions (DE BOOR, 1978), where a penalty based on higher-order differences between adjacent basis function coefficients (cf. EILERS AND MARX, 1996) results in flexible yet smooth p.d.f.s without the need to make any distributional assumptions. For time series of counts that are either naturally bounded or considered to be effectively bounded (defining an upper threshold for the support on which the state-dependent distributions are to be modeled), one can, in principle, avoid such distributional assumptions by directly estimating the values of the state-dependent p.m.f.s on the support considered (this approach is in fact implemented in the R package hmm.discnp; cf. TURNER, 2018). However, without an adequate penalization, such an approach will often lead to overfitting, which severely limits the practical usefulness of corresponding models in particular in scenarios where the interest lies on classification and prediction.

Following SCOTT *et al.* (1980) and SIMONOFF (1983), and similar in spirit to LANGROCK *et al.* (2015), we here suggest to address this problem by considering a penalized likelihood function, where a penalty based on higher-order differences between adjacent count probabilities is proposed. Furthermore, we demonstrate how the suggested penalty can be adjusted in presence of (e.g. zero-) inflated observations, where small differences between corresponding count probabilities and their respective neighbors are not necessarily desirable. This conceptually simple approach is demonstrated to produce reliable estimates of both simple and complex state-dependent distributions, where smoothing parameters are considered to adjust the required flexibility in a completely data-driven way. In slightly different settings, penalized estimation of HMMs has previously been discussed in STÄDLER AND MUKHERJEE (2013), where a penalization approach is proposed to obtain sparse variance-covariance matrices in high-dimensional state-dependent processes, in FARCOMENI (2017), where a penalty based on Jeffrey's prior is considered to ensure that the estimation does not break down in scenarios where the time series to be mod-

eled are relatively short, and in ANDERSON *et al.* (2019), where a penalty is placed on the number of states.

This chapter is structured as follows: in SECTION 3.2, we recall the model formulation and dependence structure of basic HMMs and introduce some notation specifically for the case of discrete-valued time series. Furthermore, we provide an efficient algorithm for evaluating the likelihood and discuss how the model parameters can be estimated in a penalized maximum likelihood framework. In SECTION 3.3, we assess the feasibility of the suggested approach in simulation experiments, where we also compare the performance of the proposed penalized non-parametric approach to its parametric counterpart. In SECTION 3.4, we illustrate the suggested approach in a real-data application, where we model the distribution of the annual number of earthquakes over time. The proposed methodology is implemented in the R package countHMM (ADAM, 2019b).

3.2 Model formulation and model fitting

In this section, we introduce the model formulation of non-parametric HMMs for discrete-valued time series, which incorporates the smoothing approach developed in SIMONOFF (1983)[2] into the HMM framework.

3.2.1 Model formulation and dependence structure

A basic HMM comprises two stochastic processes, only one of which is observed, namely the time series to be modeled, which is denoted by $\{Y_t\}_{t=1,\dots,T}$. The observed state-dependent process is driven by a hidden state process, which is denoted by $\{S_t\}_{t=1,\dots,T}$ and typically modeled by a discrete-time, N-state Markov chain. Specifically, we consider a first-order Markov chain, i.e. we assume the state process to satisfy the Markov property, $\Pr(S_{t+1} = s_{t+1}|S_1 = s_1,\dots,S_t = s_t) = \Pr(S_{t+1} = s_{t+1}|S_t = s_t)$, $t = 1,\dots,T-1$, which is exploited in the likelihood calculations provided in SECTION 3.2.2 and could in fact be relaxed to higher-order Markov chains if deemed necessary (ZUCCHINI *et al.*, 2016). Assuming the Markov chain to be time-homogeneous, the state transition probabilities are

[2] SIMONOFF (1983) proposes a penalty function approach to smoothing large, sparse contingency tables, the idea of which we here adopt to smoothing the non-parametric state-dependent distributions within an HMM.

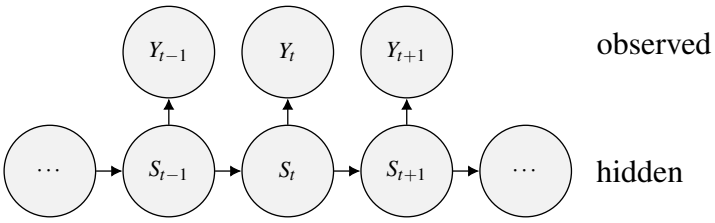

FIGURE 3.1: *Dependence structure of an HMM for discrete-valued time series. Throughout this chapter, the observed process is assumed to be a sequence of non-negative integers, e.g. a time series of counts.*

summarized in the $N \times N$ t.p.m. $\Gamma = (\gamma_{i,j})$, with elements

$$\gamma_{i,j} = \Pr(S_{t+1} = j | S_t = i), \tag{3.1}$$

$i, j = 1, \ldots, N$. The initial state probabilities, i.e. the probabilities of the state process being in the different states at time 1, are summarized in the row vector $\delta = (\delta_i)$, with elements

$$\delta_i = \Pr(S_1 = i), \tag{3.2}$$

$i = 1, \ldots, N$. If the Markov chain is assumed to be stationary, which is reasonable in many applications, then the initial distribution is the stationary distribution, i.e. the solution to the equation system $\delta\Gamma = \delta$ subject to $\sum_{i=1}^{N} \delta_i = 1$ (ZUCCHINI *et al.*, 2016). Otherwise, the initial state probabilities need to be estimated. The state process is completely specified by the initial state probabilities and the state transition probabilities as given by EQUATIONS (3.1) and (3.2), respectively.

The basic dependence structure of an HMM for discrete-valued time series is such that the observations are assumed to be conditionally independent of each other, given the states, where the state process selects which of N possible distributions generates the observation at any time point. This dependence structure is illustrated in FIGURE 3.1. In practice, it is common to assume some parametric class of distributions, such as the Poisson distributions, for the state-dependent distributions (cf. ALTMAN AND PETKAU, 2005; BEBBINGTON, 2007). Here, we do not make any such assumption, and instead assign one state-dependent probability mass to each possible count on the bounded support $\{0, 1, 2, \ldots, K\}$, thus considering the distribution parameters

$$\pi_k^{(i)} = \Pr(Y_t = k | S_t = i), \tag{3.3}$$

$i = 1, \ldots, N$, $k = 0, \ldots, K$. While count data can, in principle, be unbounded, we consider an upper threshold, which is denoted by K, as to obtain a fixed, finite number of

parameters. The support of the state-dependent distributions should be bounded in a reasonable way (specifically, it should at least cover all observed counts; cf. SECTION 3.2.5 for a discussion hereof). The state-dependent process is completely specified by the state-dependent count probabilities as given by EQUATION (3.3). With this model formulation, we thus consider a possibly large number of parameters rather than only e.g. one (as in the case of the Poisson or the binomial distribution) or two (as in the case of the negative binomial distribution).

Although the parameter space in this model formulation is still finite-dimensional, it will usually have a fairly high dimension, with the individual parameters not being of direct interest themselves. As a consequence, we follow TURNER (2018) and call our approach non-parametric. In particular, this label emphasizes that the state-dependent distributions are not determined by a small number of parameters, as would be the case when a distributional family such as the class of Poisson distributions would be considered. In addition, with the given model formulation, we are not restricted to any particular functional shape of the state-dependent distributions, and instead have full flexibility to let the data "speak" for themselves, like with other methods for which the label non-parametric is commonly used in the literature.

3.2.2 Likelihood evaluation

For some given parameter vector, which is denoted by θ and comprises — assuming the initial distribution to be the stationary distribution of the Markov chain — the state transition probabilities as given by EQUATION (3.1) and the state-dependent count probabilities as given by EQUATION (3.3), the likelihood of the non-parametric HMM as formulated in SECTION 3.2.1 can be written as a matrix product,

$$\mathcal{L}(\theta|y_1,\dots,y_T) = \boldsymbol{\delta}\mathbf{P}(y_1)\prod_{t=2}^{T}\boldsymbol{\Gamma}\mathbf{P}(y_t)\mathbf{1}, \tag{3.4}$$

with $N \times N$ diagonal matrix

$$\mathbf{P}(y_t) = \begin{pmatrix} \pi_{y_t}^{(1)} & & 0 \\ & \ddots & \\ 0 & & \pi_{y_t}^{(N)} \end{pmatrix}, \tag{3.5}$$

and $\mathbf{1} \in \mathbb{R}^N$ denoting a column vector of ones. The evaluation of the likelihood as given by EQUATION (3.4) corresponds to the application of the forward algorithm. Defining

the forward probabilities $\alpha_t(i) = \Pr(y_1, \ldots, y_t, S_t = i)$, which are summarized in the row vectors $\boldsymbol{\alpha}_t = (\alpha_t(1), \ldots, \alpha_t(N))$, the recursive scheme

$$\begin{aligned} \boldsymbol{\alpha}_1 &= \boldsymbol{\delta}\mathbf{P}(y_1); \\ \boldsymbol{\alpha}_t &= \boldsymbol{\alpha}_{t-1}\boldsymbol{\Gamma}\mathbf{P}(y_t), \end{aligned} \tag{3.6}$$

$t = 2, \ldots, T$, can be applied to arrive at $\boldsymbol{\alpha}_T$, from which the likelihood can be obtained by the law of total probability as

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}|y_1, \ldots, y_T) &= \sum_{i=1}^{N} \alpha_T(i) \\ &= \boldsymbol{\alpha}_T \mathbf{1} \end{aligned}$$

(ZUCCHINI *et al.*, 2016). Using the recursive scheme as given by EQUATIONS (3.6), evaluating the likelihood requires $\mathcal{O}(TN^2)$ operations, which renders an estimation of the model parameters by numerically maximizing the likelihood (or, in case of numerical underflow, the log-likelihood, which is denoted by $l(\boldsymbol{\theta}|y_1, \ldots, y_T)$), practically feasible even for relatively long time series and a moderately large number of states. Alternatively, the EM algorithm, which also arrives at a (local) maximum of the likelihood, can be used (ZUCCHINI *et al.*, 2016; cf. also SECTION 2.3.1).

For the model formulation considered, an implementation of numerical maximum likelihood estimation using the forward algorithm is provided in the R package countHMM (ADAM, 2019b), while the EM algorithm constitutes the default choice in the R package hmm.discnp (TURNER, 2018).

3.2.3 Roughness penalization

The downside of the above non-parametric and hence very flexible approach is its propensity to overfit any given data. Especially in cases where the length of the time series is short relative to the number of model parameters, which, if the initial distribution is assumed to be the stationary distribution of the Markov chain, is given by $N(N-1)+NK$, the fitted state-dependent distributions will often be anything but smooth, and may even involve isolated spikes with implausible gaps in between (corresponding to a lack of data in regions where observations would in fact be expected to occur in the long run). For short time series, it can in fact easily happen that for specific values well within the plausible range of observations to occur in future, each state-dependent probability is estimated to be zero, namely if no such observations are present in the training data (cf. SECTION 3.4 for an ex-

ample of this problem). The consequence of this would be that the fitted model deems the corresponding values to be impossible to occur in future, which could be problematic in particular in applications where the focus lies on forecasting.

To avoid such kind of overfitting, we add a penalty to the logarithm of the likelihood as given by EQUATION (3.4), which leads to the penalized log-likelihood

$$l^{\text{pen.}}(\theta|y_1,\ldots,y_T) = \underbrace{l(\theta|y_1,\ldots,y_T)}_{\text{goodness of fit}} - \sum_{i=1}^{N} \lambda^{(i)} \underbrace{\sum_{k=m}^{K} \left(\Delta^m \pi_k^{(i)}\right)^2}_{\text{smoothness}}, \tag{3.7}$$

where $\lambda^{(i)}, i = 1,\ldots,N$, denotes a smoothing parameter associated with the i-th state-dependent distribution, and where

$$\begin{aligned} \Delta^1 \pi_k^{(i)} &= \pi_k^{(i)} - \pi_{k-1}^{(i)}; \\ \Delta^m \pi_k^{(i)} &= \Delta^1\left(\Delta^{m-1} \pi_k^{(i)}\right), \end{aligned} \tag{3.8}$$

$k = m,\ldots,K$, denotes the m-th order differences between adjacent count probabilities (cf. SECTION 3.2.5 for a discussion of the choice of the difference order). The inclusion of the penalty term, together with the associated smoothing parameters, allows us to control the variance of the otherwise unrestricted and hence highly variable estimation of the state-dependent distributions. Maximizing the penalized log-likelihood as given by EQUATION (3.7) then amounts to finding a good compromise between the goodness of fit, as measured by the likelihood given by EQUATION (3.4), and the smoothness of the state-dependent distributions, as measured by the m-th order differences between adjacent count probabilities given by EQUATION (3.8).

In presence of zero-inflation, which in practice often occurs when dealing with count data, it can make sense not to penalize differences between probability masses on zero and the adjacent count probabilities (i.e. those on $1,2,3,\ldots,m$), as otherwise the penalization will shrink the estimate of $\pi_0^{(i)}$ and increase its neighboring state-dependent count probabilities as to ensure smoothness of the resulting state-dependent distributions, which in case of a genuine excess of zeros can be undesirable. The penalty in the penalized log-likelihood as given by EQUATION (3.7) can then be replaced by an inflation-adjusted penalty, which leads to the inflation-adjusted penalized log-likelihood

$$l^{\text{infl.-adj. pen.}}(\theta|y_1,\ldots,y_T) = \underbrace{l(\theta|y_1,\ldots,y_T)}_{\text{goodness of fit}} - \sum_{i=1}^{N} \lambda^{(i)} \underbrace{\sum_{k=m+1}^{K} \left(\Delta^m \pi_k^{(i)}\right)^2}_{\text{smoothness}}, \tag{3.9}$$

such that the state-dependent probability masses on zero can be estimated without any constraints related to the smoothness of the resulting state-dependent distributions. Probability masses on other counts, e.g. the upper bound in case of bounded counts, can be excluded from penalization analogously; cf. ADAM *et al.* (2019c) for an example where this is demonstrated.

3.2.4 Model fitting and parameter constraints

Maximum penalized likelihood estimates of the model parameters can be obtained by numerically maximizing the penalized log-likelihood as given by EQUATION (3.7) using some Newton-Raphson-type optimization routine, such as implemented in the R function nlm (R CORE TEAM, 2019). A notorious difficulty with basic HMMs that is likely exacerbated in non-parametric HMMs, where the number of parameters tends to be much larger, is the often complex shape of the surface of the penalized log-likelihood. To increase the chance of having found the global rather than a local maximum, a multiple start point strategy can be applied, where the penalized log-likelihood is maximized from different, possibly randomly selected initial values, where the estimate corresponding to the highest penalized log-likelihood is chosen (ZUCCHINI *et al.*, 2016).

As the model parameters are all probabilities, a number of parameter constraints need to be satisfied, which can be achieved by transforming the constrained parameters into unconstrained ones using multinomial logit link functions, and then maximizing the penalized log-likelihood with respect to the unconstrained parameters. Specifically, to ensure $\gamma_{i,j} \in [0,1]$, $i,j = 1,\ldots,N$, and $\sum_{j=1}^{N} \gamma_{i,j} = 1$, $i = 1,\ldots,N$, the constrained state transition probabilities as given by EQUATION (3.1) can be written as

$$\gamma_{i,j} = \frac{\exp(\gamma'_{i,j})}{\sum_{k=1}^{N} \exp(\gamma'_{i,k})},$$

$i,j = 1,\ldots,N$. Furthermore, to ensure $\delta_i \in [0,1]$, $i = 1,\ldots,N$, and $\sum_{i=1}^{N} \delta_i = 1$, we can write the constrained initial state probabilities given by EQUATION (3.2) as

$$\delta_i = \frac{\exp(\delta'_i)}{\sum_{k=1}^{N} \exp(\delta'_k)},$$

$i = 1,\ldots,N$. To ensure $\pi_k^{(i)} \in [0,1]$, $i = 1,\ldots,N$, $k = 0,\ldots,K$, and $\sum_{k=0}^{K} \pi_k^{(i)} = 1$, $i = 1,\ldots,N$, the constrained state-dependent count probabilities as given by EQUATION (3.3)

can be written as

$$\pi_k^{(i)} = \frac{\exp\left(\pi_k'^{(i)}\right)}{\sum_{l=0}^{K}\exp\left(\pi_l'^{(i)}\right)},$$

$i = 1,\ldots,N$, $k = 0,\ldots,K$. After having maximized the penalized log-likelihood with respect to the unconstrained parameters, which are denoted by $\gamma'_{i,j}$, δ'_i, and $\pi_k'^{(i)}$ (fixing one state transition probability for each row of the t.p.m., one initial state probability, and one state-dependent count probability for each of the state-dependent distributions at zero to ensure the model to be identifiable), the constrained parameters can be obtained by applying the above transformations.

Regarding identifiability in general, and in particular with regard to the very flexible model formulation considered here, ALEXANDROVICH *et al.* (2016) show that for an HMM to be identifiable it is sufficient if the t.p.m. has full rank and the state-dependent distributions are distinct, conditions that can be expected to be satisfied in most practical scenarios where HMMs seem to be natural candidate models.

3.2.5 Choice of the tuning parameters

The difference order, m, is a tuning parameter that we recommend to be chosen pragmatically depending on the data at hand, and validated based on a close inspection of the goodness of fit resulting from different choices, e.g. based on pseudo-residual analyses. With $m =$ 1-st order differences, a uniform distribution is obtained as $\lambda^{(i)} \to \infty$ (the penalty term vanishes if all count probabilities are equal), whereas for $m =$ 2-nd order differences, a triangular distribution is obtained in the limit (the penalty term vanishes if all count probabilities lie on a straight line with arbitrary slope). Based on our experience, $m =$ 3-rd or even higher order differences produce the most reliable estimates in a number of settings with varying complexity, especially in scenarios where the state-dependent distributions have complex functional shapes (cf. the simulation experiments presented in SECTION 3.3).

The size of the support on which the state-dependent distributions are to be estimated, K, is a tuning parameter that must be chosen greater than or at least equal to the highest count observed. However, with unbounded counts, it does in fact make sense to choose a somewhat larger size, as the absence of values greater than the highest count observed in the training data does not guarantee that such values will not occur in future. While without penalization, a non-parametric approach would estimate the probability of such

future events to be zero, the penalized approach will place positive probability on counts slightly larger than the maximal value observed due to the enforced smoothing. Again, we recommend to validate the choice of the size of the support based on a close inspection of the goodness of fit resulting from different choices, e.g. based on pseudo-residual analyses.

An adequate choice of the smoothing parameters, $\lambda^{(i)}$, is crucial for finding a good balance between goodness of fit and estimator variance. We here adopt the K-fold cross-validation approach proposed in LANGROCK *et al.* (2015), where the optimal vector of smoothing parameters, $\boldsymbol{\lambda} = (\lambda^{(1)}, \ldots, \lambda^{(N)})$, from some pre-specified grid, $\boldsymbol{\Lambda} = \lambda^{(1)} \times \cdots \times \lambda^{(N)} \subset \mathbb{R}^N$, can be found using a greedy search algorithm: first, we choose an initial vector $\boldsymbol{\lambda}^{[0]} = (\lambda^{(1)[0]}, \ldots, \lambda^{(N)[0]}) \subset \boldsymbol{\Lambda}$ from the grid and set z to zero. Then, we compute the average out-of-sample log-likelihood for the current smoothing parameter vector $\boldsymbol{\lambda}^{[z]} = (\lambda^{(1)[z]}, \ldots, \lambda^{(N)[z]})$ and each direct neighbor on the grid, from wich we then choose the updated smoothing parameter vector $\boldsymbol{\lambda}^{[z+1]} = (\lambda^{(1)[z+1]}, \ldots, \lambda^{(N)[z+1]})$ as the one that yields the highest out-of-sample log-likelihood averaged across folds. We then increase z by one and repeat the previous step until the obtained smoothing parameters do not change anymore, i.e. until $\boldsymbol{\lambda}^{[z+1]} = \boldsymbol{\lambda}^{[z]}$.

Following ZUCCHINI *et al.* (2016), the out-of-sample log-likelihood can be evaluated by treating the out-of-sample observations as missing data for model training using maximum penalized likelihood estimation, hence replacing the corresponding diagonal matrices in the likelihood as given by EQUATION (3.5) by identity matrices. The out-of-sample log-likelihood can then be calculated analogously, now treating the in-sample observations as missing data and using the estimated model parameters for evaluating the out-of-sample unpenalized log-likelihood.

3.3 Simulation experiments

To assess the performance of the suggested approach, we present the following two simulation experiments: in each of 200 simulation runs, we simulated i) 200 (short time series setting) and ii) 500 (long time series setting) realizations from a 2-state Markov chain, $\{S_t\}_{t=1,\ldots,200}$ and $\{S_t\}_{t=1,\ldots,500}$, with t.p.m.

$$\boldsymbol{\Gamma} = \begin{pmatrix} 0.95 & 0.05 \\ 0.05 & 0.95 \end{pmatrix},$$

and initial (stationary) state probabilities $\delta_i = 0.5$, $i = 1, 2$. Conditional on the simulated state sequences, the observations were then drawn from either of the following two dis-

tributions: a Conway-Maxwell-Poisson distribution (when the state process was in state 1), or a two-component mixture of a Poisson and a Conway-Maxwell-Poisson distribution (when in state 2; cf. FIGURES 3.2 and 3.4 for an illustration of the state-dependent distributions).

In comparison with the Poisson distribution, the Conway-Maxwell-Poisson distribution comprises an additional parameter that allows to model under- and overdispersion relative to the Poisson distribution. As the marginal distribution of the data simulated from this distribution could be fairly well captured by a two-component mixture of Poisson distributions, a 2-state Poisson HMM would seem to provide a natural choice. However, the underlying state-dependent distributions do in fact substantially deviate from a Poisson distribution, exhibiting some underdispersion in state 1 and strong overdispersion as well as bimodality in state 2. This complex model formulation was chosen to demonstrate the full potential of the suggested approach, but also to highlight potential pitfalls that can occur when choosing too simplistic parametric models based in particular on a visual inspection of the marginal distribution of the data.

Initially, the performance of the suggested approach was assessed by visually comparing the empirical distributions of the estimated distribution parameters. In addition, to formally compare the performance of the proposed methodology with alternative approaches, we considered the following measures: first, we computed the Kullback-Leibler divergences (KLDs) between the true and the estimated state-dependent distributions, averaged across 200 simulation runs,

$$\mathrm{KLD}\big(\hat{\pi}^{(i)}\big) = \frac{1}{200}\sum_{r=1}^{200}\sum_{k=0}^{40}\pi_k^{(i)}\log\left(\frac{\pi_k^{(i)}}{\hat{\pi}_k^{(i)[r]}}\right),$$

$i = 1,2$, with $\hat{\pi}_k^{(i)[r]}$ denoting the estimate of $\pi_k^{(i)}$ obtained in the r-th simulation run. Furthermore, the mean absolute errors (MAEs) of the estimated off-diagonal t.p.m. entries obtained in 200 simulation runs were computed as

$$\mathrm{MAE}(\hat{\gamma}_{i,j}) = \frac{1}{200}\sum_{r=1}^{200}\sqrt{\big(\hat{\gamma}_{i,j}^{[r]} - \gamma_{i,j}\big)^2},$$

$i, j = 1,2$, $i \neq j$, and $\hat{\gamma}_{i,j}^{[r]}$ denoting the estimate of $\gamma_{i,j}$ obtained in the r-th simulation run. Lastly, we computed the state misclassification rates (SMRs), averaged across 200 simulation runs,

$$\mathrm{SMR}(\hat{\mathbf{s}}) = \frac{1}{200}\sum_{r=1}^{200}\frac{1}{T}\sum_{t=1}^{T}\mathbb{1}_{\{\hat{s}_t^{[r]} \neq s_t^{[r]}\}},$$

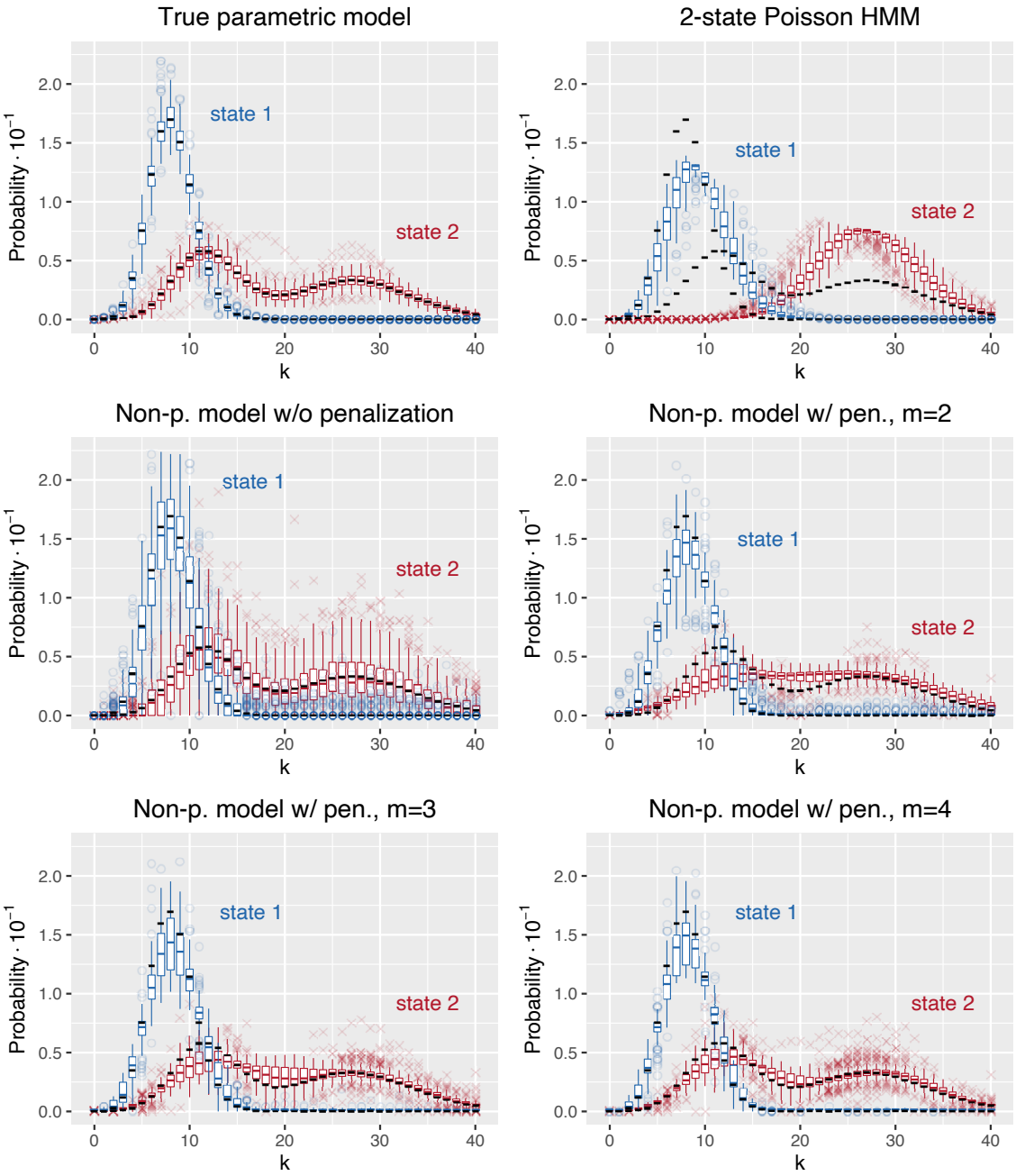

FIGURE 3.2: *Boxplots of the estimated state-dependent distribution parameters (short time series setting) for states* 1 *(blue) and* 2 *(red) obtained in* 200 *simulation runs under the true parametric model, the* 2*-state Poisson HMM, the non-parametric model fitted without penalization, and the non-parametric model fitted with penalization. Estimates that lie outside* 1.5 *times the interquartile range are visualized by dots, while true state-dependent distributions are indicated by dashed lines.*

where $T = 200$ (short time series setting) and 500 (long time series setting), with $\hat{s}_t^{[r]}$ denoting the globally decoded state at time t, where the Viterbi algorithm (VITERBI, 1967) was used for state decoding, and $s_t^{[r]}$ being the true realization of the simulated state sequence at time t obtained in the r-th simulation run.

TABLE 3.3: *Results of the simulation experiments (short time series setting). Displayed are the KLDs and the MAEs for states* 1 *and* 2 *as well as the SMRs obtained in* 200 *simulation runs under the true parametric model, the* 2*-state Poisson HMM, the non-parametric model fitted without penalization, and the non-parametric models fitted with penalization, respectively.*

Model specification	KLD($\hat{\pi}^{(1)}$)	KLD($\hat{\pi}^{(2)}$)	MAE($\hat{\gamma}_{1,2}$)	MAE($\hat{\gamma}_{2,1}$)	SMR($\hat{\mathbf{s}}$)
True parametric model	0.016	0.026	0.023	0.022	0.033
2-state Poisson HMM	0.138	2.121	0.130	0.401	0.244
Non-p. mod. w/o pen.	1.107	3.276	0.020	0.019	0.062
Non-p. mod. w/ pen., $m = 2$	0.114	0.111	0.021	0.028	0.066
Non-p. mod. w/ pen., $m = 3$	0.090	0.059	0.016	0.019	0.049
Non-p. mod. w/ pen., $m = 4$	0.076	0.048	0.014	0.017	0.045

In each simulation run, we fitted the following models: i) the true parametric model, as a benchmark only, noting that in practice, a model as complex as the given one effectively can usually not be guessed based on a visual inspection of the marginal distribution of the data, ii) a 2-state Poisson HMM, which, as discussed above, would seem to provide a reasonable choice based on a visual inspection the marginal distribution of the data, iii) the unpenalized non-parametric model, to demonstrate the need for roughness penalization, as well as the suggested non-parametric model fitted with iv) $m = 2$-nd, v) $m = 3$-rd, and vi) $m = 4$-th order difference penalties, respectively. The size of the support on which the state-dependent distributions were estimated was chosen as $K = 40$ (or, alternatively, the highest count observed in case this was greater than 40). The smoothing parameters were selected via 20-fold cross-validation over the grid $\Lambda = \lambda^{(1)} \times \lambda^{(2)}, \lambda^{(1)} = \lambda^{(2)} = (10, 100, 1{,}000, \ldots, 10^8)$.

The empirical distributions of the estimated state-dependent distribution parameters, as obtained under the models considered in 200 simulation runs, are visualized in FIGURES 3.2 (short time series setting) and 3.4 (long time series setting), respectively. It can be seen that the non-parametric models fitted with penalization produced estimates very close to those obtained when using the true parametric model (especially for high difference orders and long time series; cf. the bottom-right and the top-left panel in FIGURE 3.4). For small difference orders and short time series, however, there is some underestimation of the peaks and some overestimation of the troughs (cf. the middle-right and the bottom-left panel in FIGURE 3.2). Given that the true parametric model is unknown in practice, these first impressions regarding the performance of the non-parametric models fitted with penalization are encouraging. Regarding the other two competitors, the 2-state Poisson HMM clearly lacks the flexibility to capture the functional shapes of the true state-

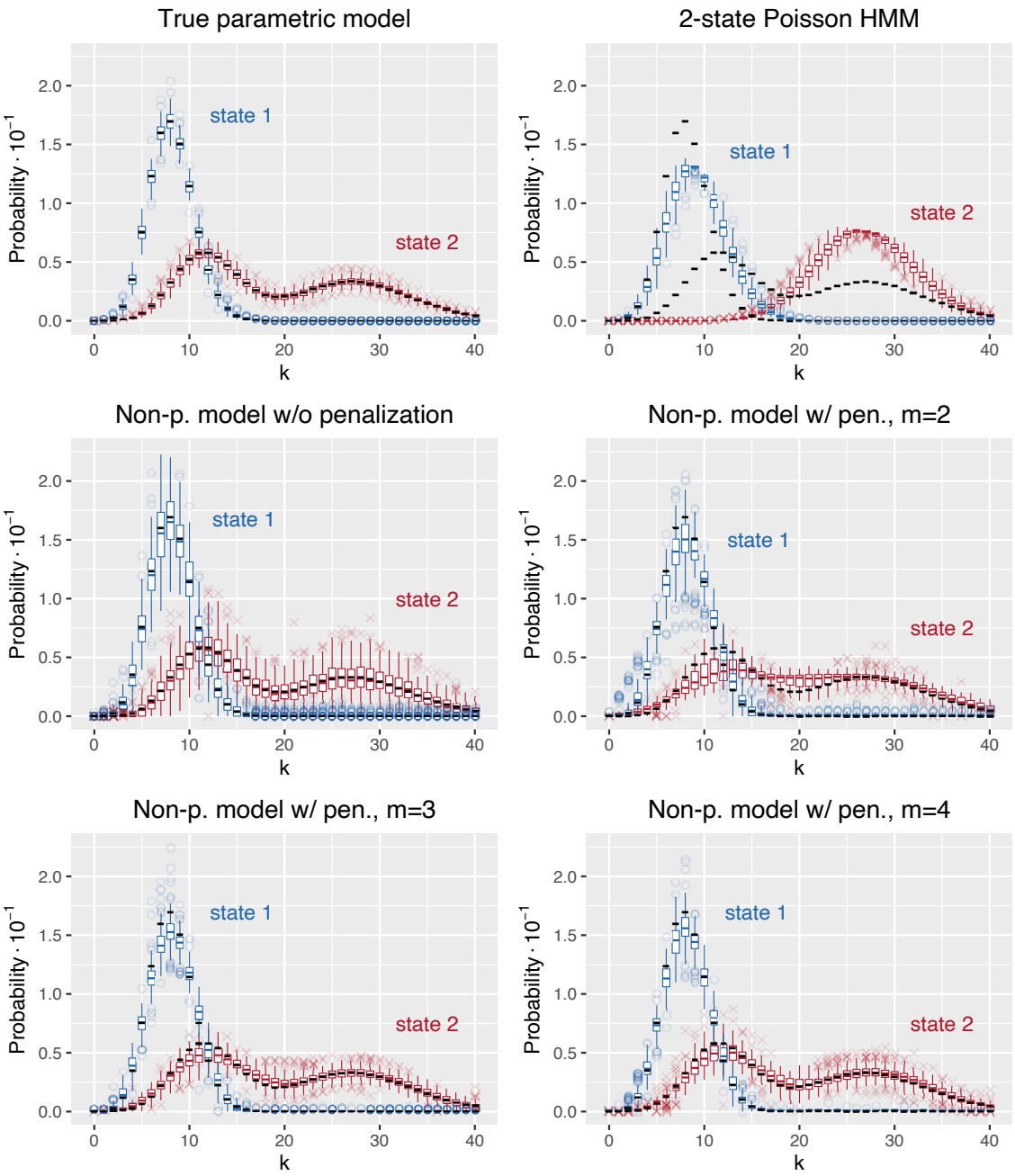

FIGURE 3.4: *Boxplots of the estimated state-dependent distribution parameters (long time series setting) for states* 1 *(blue) and* 2 *(red) obtained in* 200 *simulation runs under the true parametric model, the* 2*-state Poisson HMM, the non-parametric model fitted without penalization, and the non-parametric model fitted with penalization. Estimates that lie outside* 1.5 *times the interquartile range are visualized by dots, while true state-dependent distributions are indicated by dashed lines.*

dependent distributions and thus exhibits a strong bias (cf. the top-right panel in FIGURES 3.2 and 3.4, respectively), while the non-parametric model fitted without penalization leads to a much higher estimator variance, which can be attributed to overfitting.

TABLE 3.5: *Results of the simulation experiments (long time series setting). Displayed are the KLDs and the MAEs for states* 1 *and* 2 *as well as the SMRs obtained in* 200 *simulation runs under the true parametric model, the* 2*-state Poisson HMM, the non-parametric model fitted without penalization, and the non-parametric models fitted with penalization, respectively.*

Model specification	$\text{KLD}(\hat{\boldsymbol{\pi}}^{(1)})$	$\text{KLD}(\hat{\boldsymbol{\pi}}^{(2)})$	$\text{MAE}(\hat{\gamma}_{1,2})$	$\text{MAE}(\hat{\gamma}_{2,1})$	$\text{SMR}(\hat{\mathbf{s}})$
True parametric model	0.005	0.009	0.012	0.013	0.033
2-state Poisson HMM	0.112	2.085	0.122	0.395	0.240
Non-p. mod. w/o pen.	0.403	0.652	0.012	0.012	0.040
Non-p. mod. w/ pen., $m = 2$	0.032	0.039	0.013	0.015	0.040
Non-p. mod. w/ pen., $m = 3$	0.020	0.023	0.013	0.014	0.036
Non-p. mod. w/ pen., $m = 4$	0.014	0.019	0.012	0.014	0.035

As expected based on the considerations made above, the 2-state Poisson HMM shows — regardless of the length of the time series — the (overall) worst performance, with large KLDs ($\text{KLD}(\hat{\boldsymbol{\pi}}^{(1)}) = 0.138$ and $\text{KLD}(\hat{\boldsymbol{\pi}}^{(2)}) = 2.121$ in the short time series setting; cf. TABLE 3.3, and $\text{KLD}(\hat{\boldsymbol{\pi}}^{(1)}) = 0.112$ and $\text{KLD}(\hat{\boldsymbol{\pi}}^{(2)}) = 2.085$ in the long time series setting; cf. TABLE 3.5), which is due to the lack of flexibility to capture the functional shapes of the true state-dependent distributions. This obviously also results in high SMRs, as most of the observations in the interval $[5, 15]$ were assigned to state 1, although a considerable number of them were actually generated in state 2. The 2-state Poisson HMM also yields large MAEs of the estimated off-diagonal t.p.m. entries, with nearly every fourth globally decoded state differing from the true state, which again is an obvious consequence of the large proportion of observations in the interval $[5, 15]$ being incorrectly allocated to state 1.

The non-parametric model fitted without penalization, on the one hand, shows a much better performance than the 2-state Poisson HMM, which is due to its flexibility to capture the shapes of the true state-dependent distributions, in particular the bimodality in state 2, but, on the other hand, suffers from a high variance of the estimators, which manifests itself in large KLDs ($\text{KLD}(\hat{\boldsymbol{\pi}}^{(1)}) = 1.107$ and $\text{KLD}(\hat{\boldsymbol{\pi}}^{(2)}) = 3.276$ in the short time series setting; cf. TABLE 3.3, and $\text{KLD}(\hat{\boldsymbol{\pi}}^{(1)}) = 0.403$ and $\text{KLD}(\hat{\boldsymbol{\pi}}^{(2)}) = 0.652$ in the long time series setting; cf. TABLE 3.5). Due to the substantial reduction of the estimators' variances, the roughness penalization further considerably improves the performance, in particular the average deviation from the true state-dependent distributions, as measured by the KLDs, where the importance of the penalization is expected to increase as the length of the time series considered decreases. Although, in the simulation experiments presented here, the performance of the non-parametric model fitted with penalization continually improves as

the difference order increases, we would like to note that it will not generally be the case that a higher difference order will result in a better fit. In fact, when chosen too large, only deviations from very complex distributional shapes of the state-dependent distributions are penalized, which, as a consequence, can lead to overfitting, regardless of the weight of the penalty term.

Using a 3.6 GHz Intel® Core™ i7 CPU and the R function nlm (R CORE TEAM, 2019) to numerically maximize the penalized log-likelihood, the average computation time was — depending on the difference order — between 5.6 and 6.1 seconds (short time series setting) and 10.6 and 10.9 seconds (long time series setting) for a (single) model (i.e. for a given set of smoothing parameters).

3.4 Application to earthquake counts

To illustrate the suggested approach in a real-data setting, we model the distribution of the annual number of earthquakes with magnitude ≥ 7, EarthquakeCount_t, over time. The data, which are available in the R package countHMM (ADAM, 2019b), cover the period from 1900 to 2006, thus comprising 107 years in total. Our aim here is to present a simple case-study that provides some intuition and demonstrates the potential of non-parametric HMMs for modeling discrete-valued time series, which is why we refrain from comparing the goodness of fit resulting from different choices of the tuning parameters and instead choose $m = 3$-rd-order difference penalties and the highest count observed, $K = 41$, as an upper bound for the support on which the state-dependent distributions are to be estimated[3]. The smoothing parameters were selected via 20-fold cross-validation over the grid $\boldsymbol{\Lambda} = \boldsymbol{\lambda}^{(1)} \times \boldsymbol{\lambda}^{(2)}, \boldsymbol{\lambda}^{(1)} = \boldsymbol{\lambda}^{(2)} = (1{,}000, 10{,}000, 100{,}000, \ldots, 10^{10})$, which led to the optimal values $\lambda^{(1)} = 10^8$ and $\lambda^{(2)} = 10^9$. For the chosen smoothing parameters, the computation time was 1.6 seconds.

As a benchmark for the non-parametric model fitted with penalization, we consider the 2-state Poisson HMM presented in ZUCCHINI *et al.* (2016), which, as discussed above, provides a natural choice for the state-dependent distributions when modeling time series of counts (cf. MACDONALD AND ZUCCHINI, 1997). In addition, to demonstrate the need

[3]For a comparison of the goodness of fit resulting from different choices of the difference orders and the upper bounds of the support on which the state-dependent distributions are to be estimated, we refer to the online supplementary material provided in ADAM *et al.* (2019c).

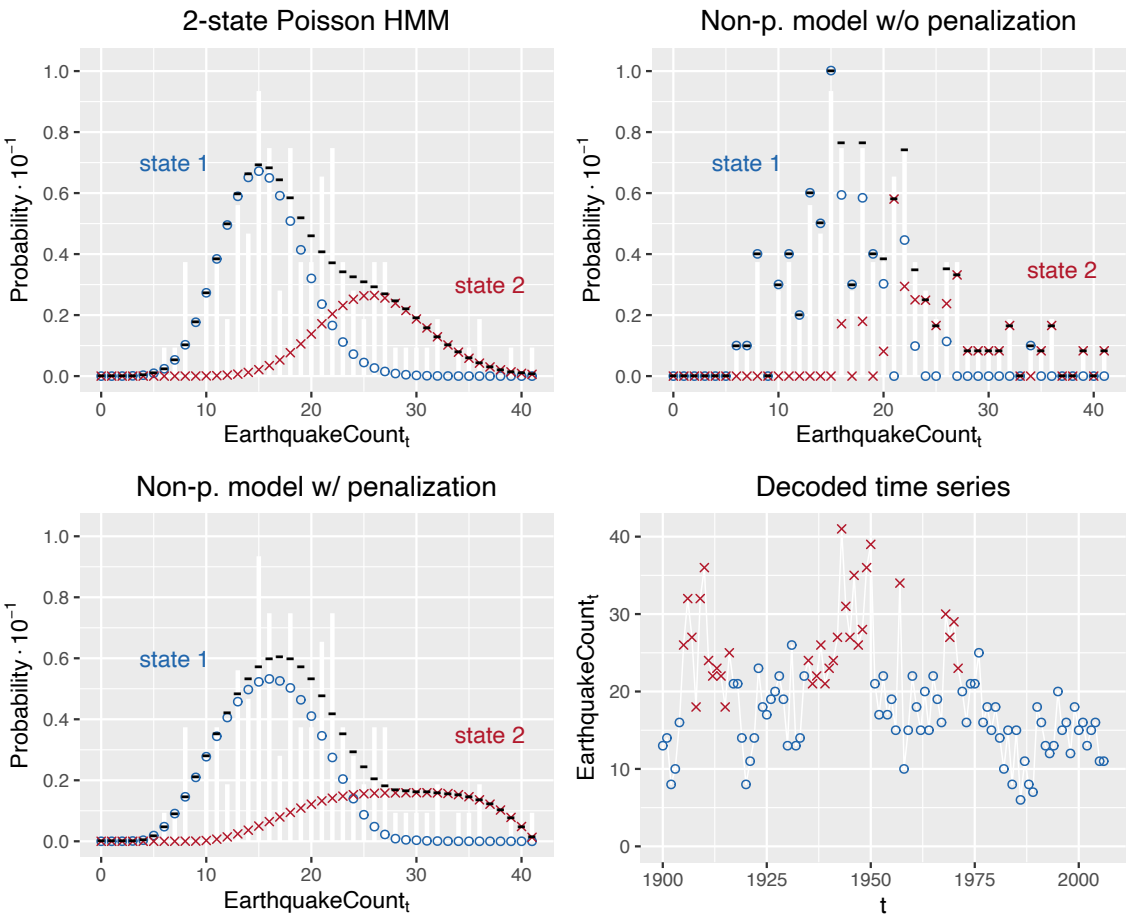

FIGURE 3.6: *Estimated state-dependent distributions of annual earthquake counts for states* 1 *(blue) and* 2 *(red) obtained under the* 2*-state Poisson HMM, the non-parametric model fitted without penalization, and the non-parametric model fitted with penalization. White bars indicate the empirical distribution of the observations, while dashed lines indicate the marginal distributions under the fitted models. The plot displayed in the bottom-right panel shows the globally decoded time series of annual earthquake counts under the non-parametric model fitted with penalization.*

for roughness penalization, we also compare the approach developed in this work with the non-parametric model fitted without penalization, as implemented in the R package hmm.discnp (TURNER, 2018).

The t.p.m. for the non-parametric model fitted with penalization was estimated as

$$\hat{\Gamma}_{\text{NPM}} = \begin{pmatrix} 0.934 & 0.066 \\ 0.128 & 0.872 \end{pmatrix},$$

which implies the stationary distribution $(0.660, 0.340)$, indicating that about 66.0 % (71 years) and 34.0 % (36 years) of the observations were generated in states 1 and 2, respectively. The estimated state transition probabilities under the non-parametric model fitted

with penalization are very close to those obtained under the 2-state Poisson HMM (cf. ZUCCHINI *et al.*, 2016) and the non-parametric model fitted without penalization, indicating that, unlike in the simulation experiments presented in SECTION 3.3, here the 2-state Poisson HMM is able to capture the state-switching dynamics.

However, as indicated by FIGURE 3.6, the estimated state-dependent distributions do clearly differ. In particular, the 2-state Poisson HMM lacks the flexibility to account for the overdispersion that is present in the data, particularly in state 2, as captured by the two non-parametric models considered. However, due to the short length of the time series, the non-parametric model fitted without penalization heavily overfits the data (cf. the top-right panel in FIGURE 3.6), which clearly demonstrates the need for roughness penalization in such a setting. In particular, the values 9 and 33, for instance, are both assigned a conditional probability of exactly zero in either of the two states — simply because these two values did not occur between 1900 and 2006 — such that, according to the fitted model, these values also cannot occur in future years, which obviously seems to be implausible and is problematic especially for prediction. The non-parametric model fitted with penalization (cf. the bottom-left panel in FIGURE 3.6), in contrast, avoids this severe overfitting and produces smooth functional shapes of the estimated state-dependent distributions. Despite these differences in the estimated state-dependent distributions, the three models considered identify essentially the same patterns, where state 1 can be linked to a calm geo-physical regime that is characterized by relatively low seismic activity, whereas state 2 corresponds to periods exhibiting relatively high seismic activity. This is further illustrated by means of the decoded state sequence underlying the observed earthquake counts under the non-parametric model fitted with penalization, which is displayed in the bottom-right panel of FIGURE 3.6.

For the non-parametric model fitted with penalization, a qq-plot and the sample ACF of normal ordinary pseudo-residuals are displayed in FIGURE 3.7. The qq-plot of the normal ordinary pseudo-residuals does not reveal any problematic lack of fit, and the sample ACF indicates only little residual correlation in the earthquake counts' series. Overall, the non-parametric model fitted with penalization shows a satisfactory goodness of fit. In particular, there is no indication that a third state needs to be included in the model, as it is the case when choosing Poisson state-dependent distributions (cf. ZUCCHINI *et al.*, 2016): while the empirical variance of the observations is 50.573, the variance of the marginal distribution under the 2-state Poisson HMM is only 44.523. This can likely be attributed to the inflexibility of the Poisson distribution, the rate parameter of which determines both the mean and the variance. The marginal distributions under the 3- and 4-state HMMs, in contrast, have variances 50.709 and 49.837, respectively, indicating that, when choosing

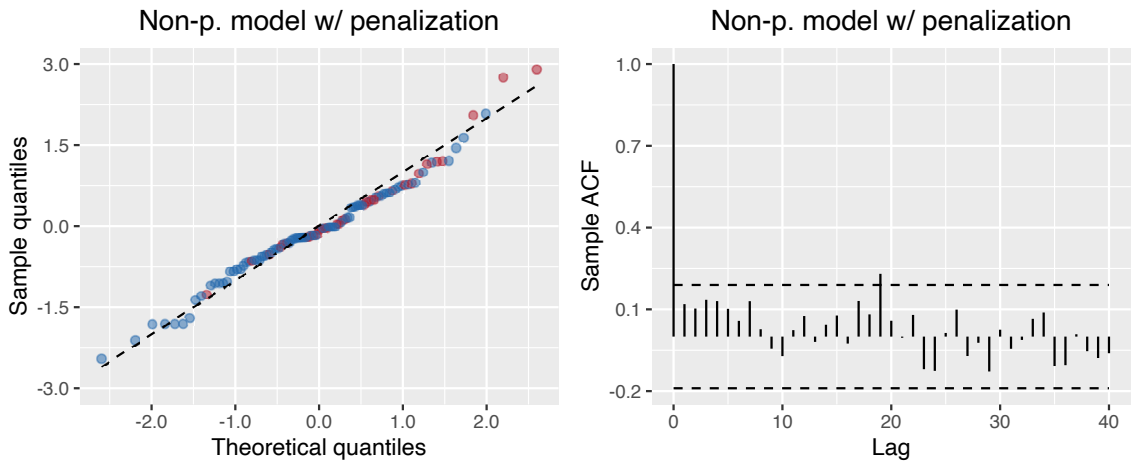

FIGURE 3.7: *Qq-plot and sample ACF of normal ordinary pseudo-residuals (non-parametric model fitted with penalization). Colors indicate the globally decoded states underlying the observed earthquake counts, where blue refers to state* 1 *and red refers to state* 2*.*

Poisson state-dependent distributions, then at least three states are needed to adequately capture the marginal distribution of the observations (cf. ZUCCHINI *et al.*, 2016), a problem that can be avoided using the suggested non-parametric approach.

3.5 Discussion

In this chapter, we introduced an effectively non-parametric approach to fitting HMMs to discrete-valued time series, where we focused on the specific case of time series of counts. The proposed estimation framework was demonstrated to provide a promising alternative to parametric HMMs, and may be superior in cases where simple, parametric state-dependent distributions are not able to capture some of the features that are present in the data. Specifically, the increased flexibility to capture complex distributional shapes can improve the accuracy of time series forecasts, reduce state misclassification rates, and help to avoid making biased inference related e.g. to the dynamics of the state process. In any case, the suggested approach can also be regarded as an exploratory tool, which can be applied in cases where it is unclear which parametric family is to be chosen for the state-dependent distributions. However, a simple parametric family is to be preferred whenever appropriate, as it will usually be much easier to implement and to interpret, and also the computational effort will be much lower than when using the suggested penalized non-parametric approach.

A notorious difficulty with HMMs, which can partly be addressed using the suggested approach, is the selection of the number of states. When using model selection criteria to

choose an adequate number of states, then these often tend to point to models with more states than can plausibly be interpreted (cf. POHLE *et al.*, 2017, for an in-depth discussion of pitfalls, practical challenges, and pragmatic solutions regarding order selection in HMMs). Inflexibility of simple, parametric state-dependent distributions to capture certain features is a common cause of this problem, as such inflexibility can always be compensated for by including additional states that will then usually have no meaningful interpretation (LANGROCK *et al.*, 2015). In that respect, the effectively unlimited flexibility of the non-parametric approach can help to reduce the required number of states, which will often substantially improve interpretability. In fact, with the proposed methodology, a single state is required to capture the marginal distribution of the data, and potential additional states only need to be included if they help to capture the dependence structure. On the other hand, model selection using information criteria is in fact more difficult within a non-parametric estimation framework, as it is necessary to derive the effective number of parameters that were used to fit the model (LANGROCK *et al.*, 2018), a challenge we did not address in this work.

In the simulation experiments and real-data applications presented in SECTIONS 3.3 and 3.4, respectively, the size of the support on which the state-dependent distributions were to be modeled was moderate, where the largest size considered was 41 (cf. SECTION 3.4). If state-dependent distributions on much larger supports, e.g. with size 5,000, are to be modeled, then the high dimensionality of the parameter space can become problematic, especially with regard to the computing time. In those instances, the parameter space can potentially be decreased by treating the data as stemming from a continuous distribution, constructing the state-dependent distributions based on linear combinations of B-spline basis functions, where higher-order differences between adjacent basis function coefficients are penalized (EILERS AND MARX, 1996; LANGROCK *et al.*, 2015; LANGROCK *et al.*, 2018; cf. also SECTION 2.3.2). Alternatively, the data could be binned, e.g. in intervals $\mathbf{I}_1 = \{1,2,3,\ldots,10\}$, $\mathbf{I}_2 = \{11,12,13,\ldots,20\}$, $\mathbf{I}_3 = \{21,22,23,\ldots,30\}$, …, $\mathbf{I}_{500} = \{4{,}991,4{,}992,4{,}993,\ldots,5{,}000\}$, then estimating the state-dependent distributions defined on the intervals instead of directly on the counts (in the example given above thus reducing the size of the support by a factor ten).

On a final note, we would like to highlight that the considerations made above imply that the approach developed in this thesis is by no means restricted to modeling time series of counts. Instead, essentially any type of time series data where the observations are at least of ordinal scale can, in principle, be modeled using the proposed methodology (in the same spirit as presented for large sparse contingency tables in SIMONOFF, 1983). This we believe could be relevant in particular for modeling longitudinal time series on Likert-

type scales, which are particularly common in social sciences (cf. SCOTT *et al.*, 2005). Lastly, our penalization approach could potentially also be adapted to different types of conventional models for discrete-valued time series: in that regard, one may e.g. think of combining it with the non-parametric estimation approach for integer-valued autoregressive models proposed in DROST *et al.* (2009). The approach developed in this work is thus not only readily applicable to basic time series of counts, but also provides a promising starting point for future research into modeling various types of discrete-valued time series using non-parametric modeling techniques.

Chapter 4

Joint modeling of multi-scale time series using hierarchical hidden Markov models

Chapter 4

Joint modeling of multi-scale time series using hierarchical hidden Markov models[1]

"We are drowning in information and starving for knowledge."

— R.D. Rogers

Summary

In this chapter, we propose hierarchical HMMs as a versatile class of statistical models for multi-scale time series. While conventional HMMs are restricted to modeling single-scale data, in practice variables are often observed at different temporal resolutions. An economy's gross domestic product, for instance, is typically observed on a yearly, quarterly, or monthly basis, whereas stock prices are available daily or at even finer resolutions. Step lengths performed by an animal, to give another example, are often observed on a daily or hourly basis, whereas accelerations obtained from accelerometers are available at much higher frequencies, with observations typically made several times per second. To incorporate such multi-scale data into a joint HMM, we regard the observations as stemming from multiple, connected state processes, each of which operates at the time scale at which the corresponding variables were observed. The suggested approach is illustrated in two real-data applications, where we jointly model the distribution of i) daily horizontal movements and ten-minute vertical displacements of an Atlantic cod and ii) monthly trade volumes and daily log-returns of the Goldman Sachs stock, respectively.

[1]This chapter is based on ADAM *et al.* (2019a) and ADAM AND OELSCHLÄGER (2020).

4.1 Introduction

Over the last decades, HMMs have emerged as a versatile class of statistical models for time series (ZUCCHINI *et al.*, 2016). In ecological applications, for instance, HMMs are commonly used to infer behavioral modes and their drivers from various types of telemetry data (MICHELOT *et al.*, 2016; WHORISKEY *et al.*, 2017; GRECIAN *et al.*, 2018), where a typical aim is to identify and understand the key patterns in an animal's movement through space, the factors, both intrinsic and extrinsic, that affect movements, and ultimately how individual behavior scales to population-level processes. In such applications, it is often of particular interest to make inference related to the influence of environmental covariates, e.g. regarding the behavioral response of blue whales to sonar exposure (DERUITER *et al.*, 2017), the effect of wind speed on Verreaux's eagles' flying dynamics (LEOS-BARAJAS *et al.*, 2017a), or diel variation in Florida panther movements (LI AND BOLKER, 2017). Beyond ecology, HMMs have proven useful e.g. in economics, where they are routinely used to model economic time series such as share returns, oil prices, or bond yields, which are driven by hidden economic regimes such as periods of high and low economic growth, inflation, or unemployment, respectively (HAMILTON, 1989).

In the recent past, the ability to remotely track individual animals has revolutionized the field of movement ecology, especially via Global Positioning System (GPS) technology (RUTZ AND HAYS, 2009; HUSSEY *et al.*, 2015). To make sense of the corresponding new types of data, various statistical models have been developed and are now routinely applied by ecologists (MORALES *et al.*, 2004; JONSEN *et al.*, 2005; JOHNSON *et al.*, 2008; PATTERSON *et al.*, 2009). Over the last few years, however, we have witnessed a second wave of advancements in bio-logging technology, most notably accelerometry, which, on the one hand, provide great opportunities for statistical inference but, on the other hand, also pose new methodological challenges (LEOS-BARAJAS *et al.*, 2017a). In general, we are now able to remotely track and monitor individual animals at increasingly long time scales but at the same time also at increasingly fine temporal resolutions. A similar revolution with regard to the temporal resolution at which time series are now available could be observed in economics, where e.g. stock prices are now available once per second or at even finer temporal resolutions (cf. O'HARA, 2015; KIRILENKO *et al.*, 2017). In any such application, the temporal resolution of the data strongly affects what kind of inference can be made.

While the observations can be multivariate, conventional HMMs have the limitation that all variables need to be equally spaced in time (or, alternatively, to follow some other regular sampling protocol). This, however, is not always given in practice. In ecology, for

instance, recent advances in bio-logging technology have led to a variety of novel telemetry sensors that often collect data from the same individual simultaneously at different time scales. Typical examples are hourly step lengths obtained from GPS tags, dive depths collected by time-depth recorders once per dive, and accelerations recorded by accelerometers several times per second. Since different types of behaviors can manifest themselves at different time scales (LEOS-BARAJAS *et al.*, 2017b; MICHELOT *et al.*, 2017), being able to collect such multiple data streams, with differing temporal resolutions, offers various opportunities for ecological inference. Similarly, economic variables are also often observed at different time scales, ranging from yearly data such as economic indices to high-frequency stock market data. Incorporating multiple such variables into a joint modeling framework can help us to draw a more comprehensive picture of the stock market's dynamics, in particular with regard to short-term vs. long-term patterns. Furthermore, by considering multiple time series observed at different time scales, joint models of such multi-scale data can contribute to reducing the effect of the often arbitrarily chosen time intervals between observations.

However, as the state process of a conventional HMM operates on the same time scale as the state-dependent process, HMMs do not readily accommodate such multi-scale data. What usually would be done to model such data within an HMM framework is either to down-sample the observations from the different data streams to the coarsest of the different time scales (e.g. by processing hourly observations into daily means of these observations, which, however, can lead to a substantial loss of information that is actually contained in the raw data; cf. GRIFFITHS *et al.*, 2018), or by fitting separate models for the different variables, which conceptually is clearly inferior to formulating and fitting a joint model for the different variables, in particular with regard to identifying states that affect multiple observed variables simultaneously. In this work, we demonstrate that these problems can to some extent be overcome using hierarchical HMMs, where the observations are regarded as stemming from multiple, connected state processes, each of which operates at the time scale at which the corresponding variables were observed.

Hierarchical HMMs originate from supervised machine learning, where they were introduced as a versatile tool for pattern recognition applications. In handwriting or voice recognition, for instance, different scales may be single letters or syllables, words, and sentences (FINE *et al.*, 1998). The hierarchy in those instances results from the fact that multiple letters or syllables taken together constitute a word, multiple words taken together constitute a sentence, and so forth. Within hierarchical HMMs, these different levels are modeled using distinct state processes that are correlated with each other. Hierarchical HMMs have been previously proposed for modeling animal movement data in LEOS-

BARAJAS *et al.* (2017b), considering, however, only a single observed process. Here we extend the proposed model formulation such that it allows for multiple state-dependent processes observed at different time scales. By incorporating multiple such data streams, collected at different temporal resolutions, corresponding models allow us to draw a more comprehensive picture e.g. of animal behavior, with clear implications for ecological inference and conservation actions. Similarly, by providing a more comprehensive picture e.g. of the stock market's dynamics, hierarchical HMMs can help to more accurately assess short- vs. long-term risks and, ultimately, to make better informed investment decisions.

This chapter is structured as follows: in SECTION 4.2, we introduce the different components of hierarchical HMMs, discuss the underlying dependence assumptions, and provide some details on the evaluation of the likelihood. In SECTION 4.3, we discuss how the model's parameters can be estimated in a maximum likelihood framework and give a brief overview of related topics, including model selection, model checking, and state decoding. In SECTION 4.4, we illustrate the feasibility of the proposed methodology in two real-data applications, where we jointly model the distribution of i) daily horizontal movements and ten-minute vertical displacements of an Atlantic cod as well as ii) monthly trade volumes and daily log-returns of the Goldman Sachs stock, respectively.

4.2 Model formulation and dependence structure

In this section, we introduce the different components of hierarchical HMMs, discuss the underlying dependence assumptions, and provide some details on the evaluation of the likelihood. The proposed model formulation constitutes an extension of the closely related hierarchical HMM proposed in LEOS-BARAJAS *et al.* (2017b)[2].

4.2.1 Multivariate hidden Markov models

Multivariate HMMs comprise two stochastic processes: an observed state-dependent process, which is denoted by $\{\mathbf{Y}_t\}_{t=1,\dots,T}$, $\mathbf{Y}_t = (Y_{1,t}, \dots, Y_{P,t})$, with P denoting the number of variables included in the model (these could e.g. be daily step lengths and turning angles or

[2]Hierarchical HMMs as proposed in LEOS-BARAJAS *et al.* (2017b) extend conventional HMMs to multiple state processes operating at different time scales. However, they do not accommodate multi-scale time series.

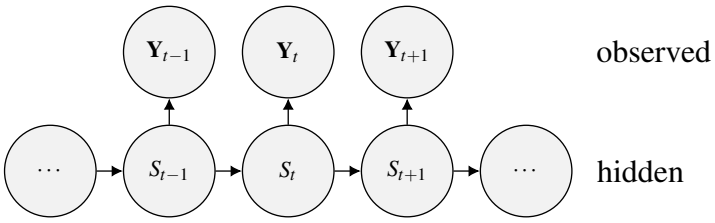

FIGURE 4.1: *Dependence structure of a multivariate HMM in its basic form. The state-dependent process is driven by a single state process. While the observations can be multivariate, conventional HMMs have the limitation that all variables need to be observed at the same temporal resolution.*

monthly trade volumes and log-returns, in which case $P = 2$), and a hidden state process, which is denoted by $\{S_t\}_{t=1,\ldots,T}$. The state process is typically modeled by a discrete-time, N-state Markov chain with $N \times N$ t.p.m. $\Gamma = (\gamma_{i,j})$, with elements

$$\gamma_{i,j} = \Pr(S_{t+1} = j | S_t = i),$$

$i, j = 1, \ldots, N$, denoting the probability of switching from state i at time t to state j at time $t+1$, and initial distribution vector $\delta = (\delta_i)$, with elements

$$\delta_i = \Pr(S_1 = i),$$

$i = 1, \ldots, N$, denoting the probability of state i being active at time $t = 1$ (the initial state probabilities can either be estimated or assumed to be the stationary state probabilities of the Markov chain; cf. ZUCCHINI *et al.*, 2016, for details).

Conditional on $S_t = i$, i.e. on state i being active at time t, the observation vector, $\mathbf{Y}_t$, is drawn from a state-dependent distribution associated with state i, defined by the P-dimensional p.d.f. (or, in the discrete case, p.m.f.) $f_{\mathbf{Y}}(\mathbf{y}_t; \theta^{(i)})$. Conditional on the entire state sequence, the observations are assumed to be independent of each other. In addition, it is convenient to also assume the P variables at time t to be conditionally independent of each other, given the state at time t, S_t, such that the joint p.d.f. (or, in the discrete case, p.m.f.) can be written as a product of univariate densities or probabilities, i.e.

$$f_{\mathbf{Y}}\left(\mathbf{y}_t; \theta^{(i)}\right) = \prod_{k=1}^{P} f_Y\left(y_{k,t}; \theta^{(i)}\right).$$

The Markov property and the assumption of conditional independence across time and variables substantially facilitate statistical inference, but can in certain scenarios be unrealistic and may then need to be relaxed (ZUCCHINI *et al.*, 2016; cf. SECTION 4.5 for an overview of possible model extensions). The dependence structure of a multivariate HMM

in its basic form is illustrated in FIGURE 4.1.

Under the dependence assumptions stated above, the likelihood of a multivariate HMM can be written as a matrix product,

$$\mathcal{L}(\theta|\mathbf{y}_1,\ldots,\mathbf{y}_T) = \boldsymbol{\delta}\mathbf{P}(\mathbf{y}_1)\prod_{t=2}^{T}\boldsymbol{\Gamma}\mathbf{P}(\mathbf{y}_t)\mathbf{1}, \tag{4.1}$$

with $N \times N$ diagonal matrix

$$\mathbf{P}(\mathbf{y}_t) = \begin{pmatrix} f_{\mathbf{Y}}\left(\mathbf{y}_t;\theta^{(1)}\right) & & 0 \\ & \ddots & \\ 0 & & f_{\mathbf{Y}}\left(\mathbf{y}_t;\theta^{(N)}\right) \end{pmatrix},$$

and $\mathbf{1} \in \mathbb{R}^N$ denoting a column vector of ones. The evaluation of the likelihood as given by EQUATION (4.1) corresponds to applying the forward algorithm, which constitutes a powerful tool that renders likelihood-based inference in HMMs fast and convenient and allows to estimate the model's parameter using numerical optimization techniques (ZUCCHINI *et al.*, 2016; cf. SECTION 4.3.1 for details on numerical likelihood maximization).

4.2.2 Hierarchical hidden Markov models

To extend the multivariate HMM introduced in SECTION 4.2.1 such that it allows for joint inference at multiple time scales, we first distinguish between state- and state-dependent processes operating on a coarse and a fine scale, respectively. The observed coarse-scale P-dimensional state-dependent process, which is denoted by $\{\mathbf{Y}_t\}_{t=1,\ldots,T}$ (these could e.g. be daily step lengths and turning angles, in which case $P = 2$, or monthly trade volumes, in which case $P = 1$), is driven by a hidden coarse-scale state process, which is denoted by $\{S_t\}_{t=1,\ldots,T}$. The observed fine-scale P'-dimensional state-dependent process, which is denoted by $\{\mathbf{Y}'_{t,t'}\}_{t'=1,\ldots,T'}$ (these could e.g. be ten-minute vertical displacements or daily log-returns, in which cases $P' = 1$), is driven by a hidden fine-scale state process, which is denoted by $\{S'_{t,t'}\}_{t'=1,\ldots,T'}$.

We then segment the fine-scale observations into T distinct chunks, each of length T', such that each chunk contains all fine-scale observations that were observed during the t-th sampling of the coarse-scale state-dependent process (e.g. all $T' = 144$ ten-minute vertical movements that were observed during the t-th sampling of daily step lengths and turning angles or all $T' = 21$ daily log-returns that were observed during the t-th sampling of monthly trade volumes). Each chunk of fine-scale observations is then connected to one

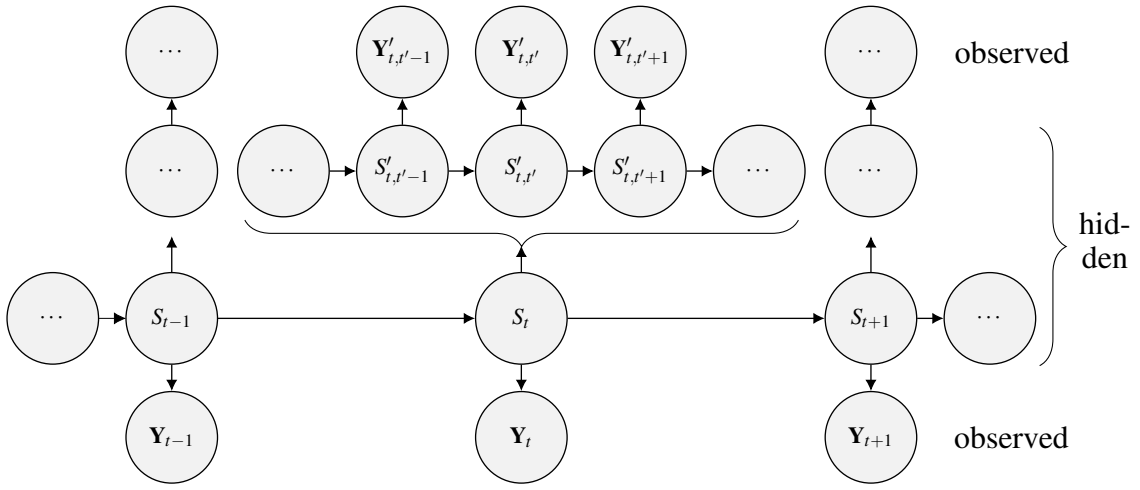

FIGURE 4.2: *Dependence structure of an hierarchical HMM. In contrast to the multivariate HMM introduced in* SECTION *4.2.1, here the observations are driven by multiple, connected state processes, each of which operates at the time scale at which the corresponding variables were observed.*

of N possible HMMs, each of which is determined by its own parameter vector, which is denoted by $\theta'^{(i)}$, $i = 1,\ldots,N$. Specifically, each fine-scale HMM has its own $N' \times N'$ t.p.m. $\Gamma'^{(i)} = ({\gamma'_{k,l}}^{(i)})$, with elements

$$ {\gamma'_{k,l}}^{(i)} = \Pr(S'_{t,t'+1} = l | S'_{t,t'} = k, S_t = i), $$

$k,l = 1,\ldots,N'$, and initial distribution vector $\delta'^{(i)} = (\delta'^{(i)}_k)$, with elements

$$ \delta'^{(i)}_k = \Pr(S'_{t,1} = k | S_t = i), $$

$k = 1,\ldots,N'$. The state of the coarse-scale state process that is active at time t, $S_t = i$, thus selects one of N possible state-dependent distributions for the observations at the coarse scale as well as one of N possible HMMs that generates the fine-scale observations during the t-th sampling of the coarse-scale state process.

Assuming conditional independence across variables, the state-dependent p.d.f. (or, in the discrete case, p.m.f.) of the fine-scale observations can be written as

$$ f_{\mathbf{Y}'}\left(\mathbf{y}'_{t,t'}; \theta'^{(i,l)}\right) = \prod_{k=1}^{P'} f_{\mathbf{Y}'}\left(y'_{k,t,t'}; \theta'^{(i,l)}\right), \tag{4.2} $$

$i = 1,\ldots,N$, $l = 1,\ldots,N'$, with $f_{\mathbf{Y}'}(y'_{k,t,t'}; \theta'^{(i,l)})$ denoting the density (or, in the discrete case, probability) of the k-th fine-scale variable being observed at time t' during the t-th sampling of the coarse-scale state-dependent process. The dependence structure of an

hierarchical HMM is illustrated in FIGURE 4.2.

We assume both state processes to be of first order (Markov property), and both state-dependent processes to satisfy the two conditional dependence assumptions (across time and variables) as detailed in SECTION 4.2.1. In ecological applications, the two state processes can often be thought of as proxies for behavioral modes, or movement strategies, relevant at shorter term (fine-scale state process) and longer term (coarse-scale state process), respectively. Similarly, in economic applications, the two state processes can typically be related to different economic regimes, relevant at shorter term (fine-scale state process) and longer term (coarse-scale state process), respectively. By incorporating several such state- and state-dependent processes into a joint modeling framework, hierarchical HMMs thus allow for joint inference at multiple time scales.

Analogously to the likelihood of a multivariate HMM as given by EQUATION (4.1), the likelihood of an hierarchical HMM can be written as a matrix product,

$$\mathcal{L}(\theta|\mathbf{y}_1,\ldots,\mathbf{y}_T,\mathbf{y}'_1,\ldots,\mathbf{y}'_T) = \boldsymbol{\delta}\mathbf{P}(\mathbf{y}_1,\mathbf{y}'_1)\prod_{t=2}^{T}\boldsymbol{\Gamma}\mathbf{P}(\mathbf{y}_t,\mathbf{y}'_t)\mathbf{1}, \tag{4.3}$$

with $N \times N$ diagonal matrix

$$\mathbf{P}(\mathbf{y}_t,\mathbf{y}'_t) = \begin{pmatrix} \mathcal{L}(\theta'^{(1)}|\mathbf{y}'_t) f_{\mathbf{Y}}\left(\mathbf{y}_t;\theta^{(1)}\right) & & 0 \\ & \ddots & \\ 0 & & \mathcal{L}(\theta'^{(N)}|\mathbf{y}'_t) f_{\mathbf{Y}}\left(\mathbf{y}_t;\theta^{(N)}\right) \end{pmatrix},$$

and $\mathcal{L}(\theta'^{(i)}|\mathbf{y}'_t)$ denoting the likelihood of the t-th chunk of fine-scale observations being generated by the i-th fine-scale HMM. A recursive algorithm to efficiently evaluate the logarithm of the likelihood as given by EQUATION (4.3), which renders a numerical maximization of the likelihood fast and convenient while simultaneously preventing numerical underflow, is provided in APPENDIX A.

4.2.3 Incorporating covariates into the model

Covariates can be incorporated into hierarchical HMMs by expressing (some of) the model's parameters as functions of covariates. In principle, covariates can be incorporated both into the different state-dependent processes, where they determine the parameters of the state-dependent distributions, and into the different state processes, where they determine the state transition probabilities. While the former was done in the case of Markov-

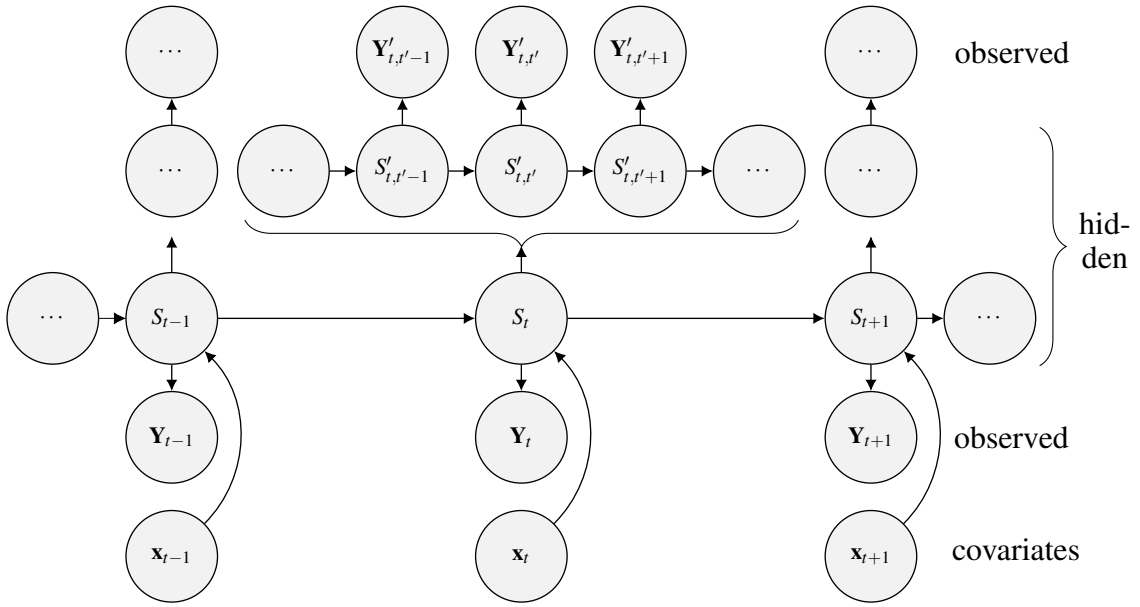

FIGURE 4.3: *Dependence structure of an hierarchical HMM with covariate-dependent coarse-scale state process. In contrast to the hierarchical HMM introduced in* SECTION *4.2.2, here the coarse-scale state process depends on covariates.*

switching GAMLSS (cf. SECTION 2.2.2), here we focus on the latter, i.e. incorporating covariates into the different state processes.

Therefore, we express the state transition probabilities as a function of a predictor, which is denoted by $\eta^{(i,j)}(\mathbf{x}_t)$, with $\mathbf{x}_t = (x_{1,t}, \ldots, x_{P,t})$ denoting a P-dimensional covariate vector. Using multinomial logit links to ensure the parameter constraints $\gamma_{i,j}(\mathbf{x}_t) \in [0,1]$, $i, j = 1, \ldots, N$, and $\sum_{j=1}^{N} \gamma_{i,j}(\mathbf{x}_t) = 1$, $i = 1, \ldots, N$, to be satisfied, we obtain the t.p.m. $\Gamma(\mathbf{x}_t) = (\gamma_{i,j}(\mathbf{x}_t))$, with elements

$$\gamma_{i,j}(\mathbf{x}_t) = \frac{\exp(\eta^{(i,j)}(\mathbf{x}_t))}{\sum_{k=1}^{N} \exp(\eta^{(i,k)}(\mathbf{x}_t))}, \tag{4.4}$$

where the predictor can be written as

$$\eta^{(i,j)}(\mathbf{x}_t) = \begin{cases} \beta_0^{(i,j)} + \sum_{k=1}^{P} \beta_k^{(i,j)} x_{k,t} & \text{if } i \neq j; \\ 0 & \text{otherwise,} \end{cases} \tag{4.5}$$

$i, j = 1, \ldots, N$. Instead of estimating the state transition probabilities directly, we then maximize the likelihood of the hierarchical HMM as given by EQUATION (4.3) with respect to the coefficients contained in EQUATION (4.5), which are denoted by $\beta_k^{(i,j)}$, $i, j = 1 \ldots, N$, $i \neq j$, $k = 0, \ldots, P$.

In some applications, it is of particular interest to model seasonal or within-day varia-

tion, i.e. the time is considered as a deterministic rather than as a stochastic covariate. To account for the corresponding periodic effects, trigonometric functions can be used, where the predictor in EQUATION (4.4) can be written as

$$\eta^{(i,j)}(\mathbf{x}_t) = \begin{cases} \beta_0^{(i,j)} + \beta_1^{(i,j)} \sin\left(\frac{2\pi t}{r}\right) + \beta_2^{(i,j)} \cos\left(\frac{2\pi t}{r}\right) & \text{if } i \neq j; \\ 0 & \text{otherwise,} \end{cases} \tag{4.6}$$

with r denoting the length of the period of interest (e.g. $t = 365$ in case of seasonal variation and daily observations or $r = 24$ when modeling within-day variation and hourly observations). For more flexibility, additional sine and cosine terms with shorter cycles can be added to the predictor.

Incorporation of covariates into the fine-scale state process is analogous, but note that in this case we have one t.p.m., which is denoted by $\Gamma'^{(i)}(\mathbf{x}'_t) = (\gamma'^{(i)}_{k,l}(\mathbf{x}'_t))$, for each state of the coarse-scale state process, i.e. N such matrices to be expressed as functions of covariates. The dependence structure of an hierarchical HMM where the coarse-scale state process depends on covariates is illustrated in FIGURE 4.3, while an example of a covariate-dependent fine-scale state process is given in SECTION 4.4.1.

4.3 Some remarks on model fitting and related topics

In this section, we provide some details on maximum likelihood estimation of the model's parameters and briefly outline further topics related to hierarchical HMMs, including model selection, model checking, and state decoding.

4.3.1 A note on likelihood maximization

Using the forward algorithm proposed in SECTION 4.2.2, the evaluation of the likelihood as given by EQUATION (4.3) requires $\mathcal{O}(NT'N'^2 + TN^2)$ operations, which renders a numerical maximization of the likelihood using some Newton Raphson-type optimization routine, such as implemented in the R function nlm (R CORE TEAM, 2019), practically feasible even for relatively long time series and a moderately large number of states. To increase the speed of the likelihood evaluation and, consequently, the maximization, the dependence structure of hierarchical HMMs can be exploited to apply parallel computing techniques, where the evaluation of the likelihoods of the fine-scale HMMs associated with the different coarse-scale states can be distributed across multiple cores, which can

reduce the computation time up to a factor N.

Typical challenges that are inherent to numerical likelihood maximization in conventional HMMs, particularly parameter constraints, numerical underflow, and local maxima of the likelihood, also apply to hierarchical HMMs. Specifically, to account for parameter constraints, we can transform the constrained parameters into unconstrained ones using some one-to-one transformation and then maximize the likelihood with respect to the unconstrained parameters (cf. SECTION 3.2.4). To avoid numerical underflow, which can occur when multiplying a large number of small probabilities in the likelihood calculations, we can maximize the log-likelihood and evaluate all quantities on the log-scale; cf. the implementation of the forward algorithm provided in APPENDIX A. As the numerical maximization can yield a local rather than the global maximum of the likelihood, using appropriate initial values for the search is crucial. To increase the chance of finding the global maximum, we advise to run the search from a range of different, possibly randomly selected initial values and then select the model corresponding to the highest likelihood.

4.3.2 Model selection and model checking

Model selection in HMM-type models primarily involves the specification of the state-dependent distributions, selecting the number of states, and variable selection (in case of covariates being included in the model), but could also extend to investigations of possible assumption violations, particularly with regard to the dependence structure. The state-dependent distributions are typically determined by the data type of the variables considered: for positive continuous-valued variables (e.g. step lengths performed by an animal or trade volumes of a stock), for instance, gamma distributions provide a natural choice, whereas for circular variables (e.g. turning angles performed by an animal or wind directions), von Mises or wrapped Cauchy distributions are commonly used (LANGROCK *et al.*, 2012b). More flexible, non-parametric state-dependent distributions could be constructed based on linear-combinations of B-spline basis functions (EILERS AND MARX, 1996; LANGROCK *et al.*, 2015; LANGROCK *et al.*, 2018; cf. also SECTION 2.3.3), though parametric state-dependent distributions are generally to be preferred if they fit the data sufficiently well.

Information criteria, such as AIC or the BIC, provide a natural approach to order selection in hierarchical HMMs when fitted via maximum likelihood estimation. However, it has been demonstrated that these criteria often tend to favor overly complex HMMs, with more states than seem plausible, when fitted to noisy data with complex features (cf.

LI AND BOLKER, 2017; POHLE *et al.*, 2017). As these practical problems are inevitably exacerbated by the more complex structure of hierarchical HMMs, we advise against over-reliance on such criteria. Instead, we recommend a more pragmatic approach to finding a suitable model, where expert knowledge, a thorough exploratory data analysis, and a close inspection of how well different candidate models, with differing numbers of states, capture the key patterns of interest and relevance, together guide (and justify) the model selection process.

Model checking in HMM-type models is typically done based on pseudo-residuals, which use the probability integral transformation to assess whether any given observation is well explained by the fitted model. For the coarse-scale observations, the evaluation of the pseudo-residuals proceeds as in basic HMMs (cf. ZUCCHINI *et al.*, 2016). For the fine-scale observations, it is convenient to first decode the coarse-scale Markov chain using the Viterbi algorithm (VITERBI, 1967; cf. also SECTION 4.3.3 for details) and then to compute the pseudo-residuals separately for each chunk of fine-scale observations conditional on the fine-scale HMM that is active according to the decoded coarse-scale states. However, we would like to raise awareness of the fact that it will not usually be feasible to make a simple, binary decision on whether or not a model is suitable: for data as complex as those that will typically be modeled using hierarchical HMMs, any simple model will likely be deemed inadequate, and unlike in basic HMMs, model checking in hierarchical HMMs applies to different layers, which further complicates a decision on the model's suitability. Alternative strategies for model checking include comparisons of the empirical distribution of the observed variables and the corresponding marginal distribution as implied under the fitted model, or simulating observations from the fitted model to check whether it can reproduce the key patterns found in the data (cf. LANGROCK *et al.*, 2013a).

4.3.3 A note on state decoding

In many applications, it is of particular interest to decode the hidden states, i.e. to compute the most likely sequence of states that may have given rise to the observations under the fitted model. The simplest and most convenient approach to state decoding in hierarchical HMMs is to first decode the coarse-scale states, $s_1, \ldots, s_T$ (taking into account both the coarse-scale and the fine-scale observations), and then, for any time t of the coarse-scale state process, to decode the fine-scale states, $s'_{t,1}, \ldots, s'_{t,T'}$, conditional on the most likely coarse-scale state to be active at time t (taking into account only the fine-scale observations). The decoding can be done either locally, considering each time point in isolation,

or globally, considering the time series as a whole. In practice, global decoding, which can conveniently be carried out using the Viterbi algorithm (VITERBI, 1967), is usually the default choice and therefore used throughout this chapter. Details on both local and global decoding transfer directly from conventional HMMs to hierarchical HMMs (cf. ZUCCHINI *et al.*, 2016, for details).

4.4 Real-data applications

In this section, we illustrate the suggested approach in two real-data applications, where we jointly model the distribution of i) daily horizontal movements and ten-minute vertical displacements of an Atlantic cod (cf. SECTION 4.4.1) as well as ii) monthly trade volumes and daily log-returns of the Goldman Sachs stock (cf. SECTION 4.4.2), respectively.

4.4.1 Application to Atlantic cod movement

Atlantic cod is a commercially valuable demersal fish species found throughout the shelf seas surrounding the British Isles (RIGHTON *et al.*, 2001; NEAT *et al.*, 2014). To facilitate informed conservation actions, information about when, where, and how individuals move and undertake key life-history events are essential (HUSSEY *et al.*, 2015; HAYS *et al.*, 2019). In this real-data application, we are particularly interested in understanding diel and circatidal patterns in the cod's fine-scale vertical movements and how these are driven by its coarse-scale horizontal movements. As demersal fish rarely swim in surface waters (which is a pre-requisite for satellite tags; RUTZ AND HAYS, 2009), tagging was achieved using an archival data storage tag. Data storage tags are typically pre-programmed to record the depth at regular time intervals for the duration of deployment (here every ten minutes). From these depth records, we calculated log-vertical displacements, $\text{LogVerticalDisplacement}_{t,t'}$, $t = 1,\ldots,291$, $t' = 1,\ldots,144$, and estimated daily geo-positions using a single-state version of the tidal geo-location model proposed in PEDERSEN *et al.* (2008)[3], which were then processed to give daily step lengths, StepLength_t, and turning angles, TurningAngle_t, $t = 1,\ldots,291$. The data, which are available on the CEFAS Data Hub (RIGHTON *et al.*, 2019), cover 291 days (about ten months) between

[3]The method was adapted to ensure that the underlying diffusion model operates under a fixed diffusivity parameter (30 km per day^2) and does not switch between two based on the presence or absence of a tidal signal.

March 25, 2005, and January 9, 2006. As some (more precisely, one or two per day) of each day's 144 depth observations were used to produce the cod's daily geo-positions, some minor conditional dependence between the two movement rates is expected, which for simplicity is neglected in the model formulation. Thus, we ended up with two separate time series, which were sampled at different temporal resolutions: vertical displacements at ten-minute intervals and horizontal movements at daily intervals, i.e. for each of the $T = 291$ daily horizontal step lengths and turning angles, we have $T' = 144$ ten-minute vertical displacements. Previous work has overcome this difference in sampling by either gaining meaningful inference from a single dimension (cf. HOBSON *et al.*, 2007) or, in the case of GRIFFITHS *et al.* (2018), who analyze movement in both dimensions, by simplifying the vertical dimension at the daily scale. In this work, we demonstrate how hierarchical HMMs can be used to jointly analyze movement in both dimensions while retaining the vertical dimension at the ten-minute scale.

Based on an exploratory data analysis and a comparison of fitted models with different numbers of states, we chose $N = 3$ states for the coarse-scale state process, as a visual inspection of the data revealed two different types of horizontal movements, one of which corresponds to again two different vertical movement patterns, which can only be captured if a third state is considered at the coarse scale. Each of the coarse-scale states was then associated with an HMM with $N' = 3$ fine-scale states (thus resulting in nine fine-scale states in total), which allows us to draw a relatively nuanced yet not overly complex picture of the cod's vertical movements. To model diel variation in the vertical displacements, the transition probabilities of the Markov chains determining the fine-scale state processes were estimated as functions of the time of day, TimeOfDay_t, with the predictors specified as given by EQUATION (4.6). The coarse-scale state transition probabilities were assumed to be constant over time. For the step lengths and vertical displacements, we assumed gamma state-dependent distributions (with an additional point mass on zero in case of the vertical displacements to account for the zeros observed), while for the turning angles, von Mises state-dependent distributions were considered. The computation time required to fit the model was 6.1 hours, where the likelihood was evaluated in C++ and numerically maximized using the R function nlm (R CORE TEAM, 2019) on a 3.6 GHz Intel® Core™ i7 CPU.

The estimated state-dependent distributions of coarse-scale step lengths and turning angles are displayed in FIGURE 4.4. Coarse-scale states 1 and 2 capture short, slightly less directed horizontal movements, where we interpret coarse-scale state 1 as a resident or foraging behavior and coarse-scale state 2 as a more mobile foraging behavior. Although these two states are very similar in terms of horizontal movements, they differ

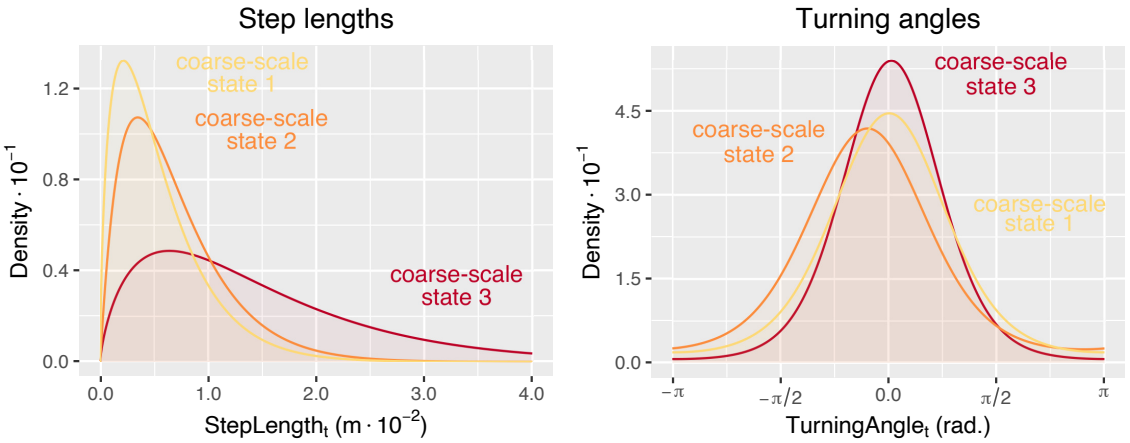

FIGURE 4.4: *Estimated state-dependent distributions of daily step lengths (left panel) and turning angles (right panel) of an Atlantic cod. Coarse-scale states* 1 *and* 2 *can be interpreted as resting or foraging and more mobile foraging behavior, respectively, which are very similar in terms of horizontal movements but differ substantially in the corresponding vertical movement patterns, while coarse-scale state* 3 *can be linked to a more traveling- or migratory-like behavior.*

substantially in the corresponding vertical movement patterns (cf. the considerations below). Coarse-scale state 3 relates to relatively longer, slightly more directed horizontal movements, which can be linked to a traveling or migrating behavior.

The t.p.m. of the coarse-scale state process was estimated as

$$\hat{\Gamma}_{\text{AC}} = \begin{pmatrix} 0.945 & 0.000 & 0.055 \\ 0.064 & 0.777 & 0.160 \\ 0.098 & 0.075 & 0.827 \end{pmatrix},$$

which implies the stationary distribution $(0.618, 0.096, 0.286)$, indicating that about 61.8 % (180 days), 9.6 % (28 days), and 28.6 % (83 days) of the observations were generated in coarse-scale state 1, 2, and 3, respectively.

The estimated state-dependent distributions of fine-scale vertical displacements and the associated stationary distributions of the corresponding fine-scale state processes as functions of the time of day, along with 95 % confidence intervals (CIs)[4], are displayed in FIGURE 4.5. When the cod was in coarse-scale state 1 (resting or foraging), then the vertical displacements were generated by the three state-dependent distributions displayed in the top-left panel. The level of vertical activity was fairly low (according to the stationary

[4]The uncertainty was quantified based on the inverse of the Hessian matrix of the likelihood at its maximum; cf. ZUCCHINI *et al.* (2016) for details.

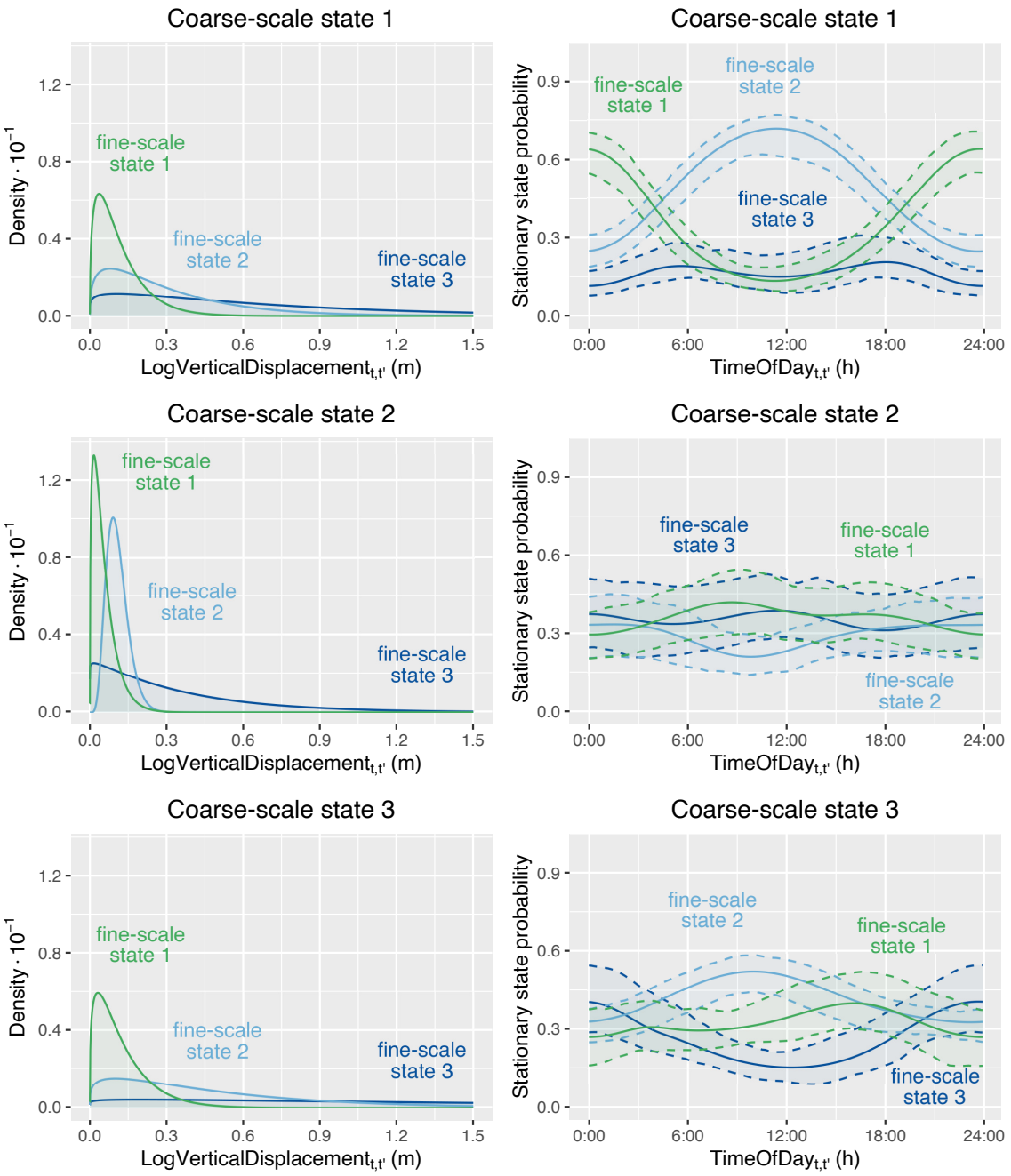

FIGURE 4.5: *Estimated state-dependent distributions of ten-minute vertical displacements of an Atlantic cod (left panel) and stationary distributions of the corresponding fine-scale state processes as functions of the time of day (right panel). Dashed lines indicate* 95 % *CIs associated with the stationary distributions. Fine-scale states* 1*,* 2*, and* 3 *represent relatively low, moderate, and high levels of vertical movement, respectively, where the corresponding levels differ substantially across the different coarse-scale states (the means of the state-dependent distributions for fine-scale state* 3*, for instance, vary from* 0.355 *in coarse-scale state* 2 *over* 0.689 *in coarse-scale state* 1 *to* 1.983 *in coarse-scale state* 3*).*

distribution, the cod was in fine-scale state 3, which corresponds to a relatively high level of vertical activity, less than 20 % of the time) and slightly increased during the day (where

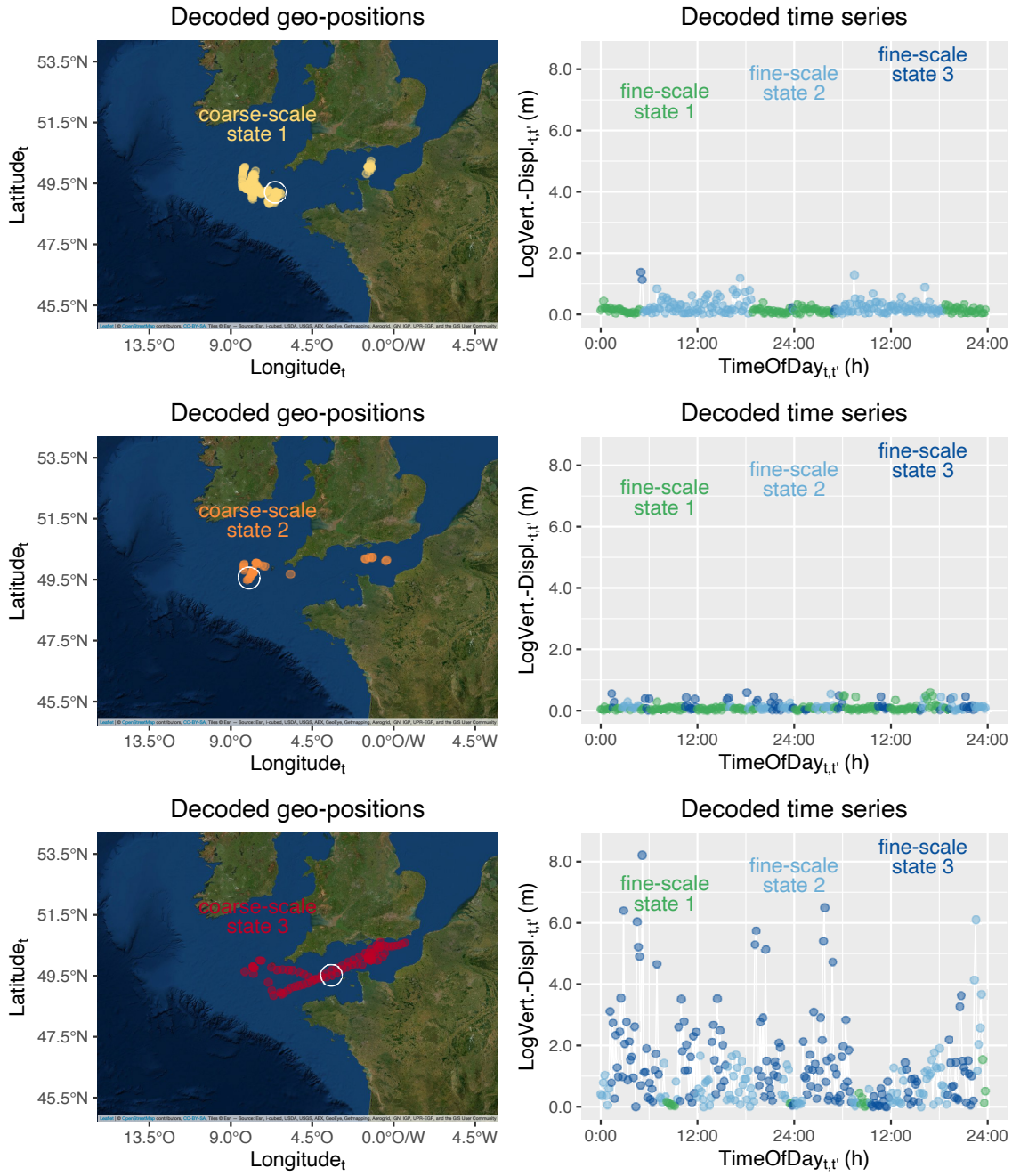

FIGURE 4.6: *Decoded time series of daily geo-positions (left panel) and ten-minute vertical displacements (right panel). The decoded time series of vertical displacements displayed in the right panel correspond to three example sequences of length* 48 *hours, one for coarse-scale state* 1 *(resting or foraging),* 2 *(more mobile foraging), and* 3 *(traveling or migrating), respectively. Circles indicate the days that correspond to the example sequences displayed in the right panel.*

it spent up to 75 % of the time in fine-scale state 2, which corresponds to a moderate level of vertical activity, and less than 25 % of the time in fine-scale state 1, which corresponds to a relatively low level of vertical activity). When the cod was in coarse-scale state 2 (more mobile foraging), then the vertical displacements were generated by the three state-

dependent distributions displayed in the middle-left panel, which correspond to a very low level of vertical activity (all three state-dependent distributions have considerably smaller means than those corresponding to coarse-scale states 1 and 3), where state occupancy (as indicated by the associated stationary distributions displayed in the middle-right panel) is not much affected by the time of day. When the cod was in coarse-scale state 3 (traveling or migrating), then the vertical displacements were generated by the three state-dependent distributions displayed in the bottom-left panel. Here, the opposite could be observed: the level of vertical activity was much higher relative to coarse-scale states 1 and 2 (fine-scale state 3, whose state-dependent distribution has mean 1.983 and therefore captures much higher vertical activity than those corresponding to fine-scale state 3 within the HMMs corresponding to coarse-scale states 1 (0.689) and 2 (0.355), was — depending on the time of day — active between 15 % and 45 % of the time) and slightly decreased during the day (but note that due to the fairly high uncertainty associated with the stationary distributions these results should be treated with some caution). For details on the estimated coefficients that determine the corresponding predictors, which were used to compute the stationary distributions as functions of the time of day displayed in the right panel of FIGURE 4.6, we refer to APPENDIX B.

The decoded horizontal movement track as well as three example sequences of fine-scale vertical displacements for the different coarse-scale states are displayed in FIGURE 4.6, where the decoding was performed using the Viterbi algorithm (VITERBI, 1967) as described in SECTION 4.2.3. The cod spent most of its time (178 days) in coarse-scale state 1, where reduced rates of horizontal movement indicate prolonged periods of resting or localized foraging. This was then interspersed by two traveling or migrating periods associated with coarse-scale state 3 (81 days) as the cod traversed the English Channel, and some periods of time spent in coarse-scale state 2 (32 days).

Throughout the time spent in coarse-scale state 1 (resting or foraging), the associated fine-scale state process exhibited clear diurnal patterns (similar trends can be found in LØKKEBORG, 1998). During the day, the level of vertical movement increased, as the cod was more likely to switch from fine-scale state 1 to fine-scale state 2. This may be interpreted as more localized foraging, as cod frequently move off the seafloor to pursue benthic-dwelling prey via visual predation (ADLERSTEIN AND WELLEMAN, 2000; HOBSON *et al.*, 2009). The increased probability of switching back to fine-scale state 1 during the night points towards a much more resting-like behavior as the cod returns to the seafloor (as e.g. observed in HOBSON *et al.*, 2007). Coarse-scale state 2 (more mobile foraging), in comparison, involves a much lower level of vertical movement, which is not much affected by the time of day. This could indicate an intermediate behavioral

mode, where the cod was foraging and remained close to the seabed while being slightly more mobile in the horizontal dimension relative to coarse-scale state 1 (i.e. it was not in a resident mode). Throughout the time spent in coarse-scale state 3, the cod was clearly migrating, exhibiting increased rates of horizontal movement, slightly more uniform directionality, and elevated rates of vertical movement as it transits the English Channel. Greater vertical displacements during migration periods could indicate the use of circatidal selective tidal stream transport, as the cod moves up off the seabed into the water column during favorable tides and uses the tide's velocity to efficiently cruise in the desired direction. Selective tidal stream transport is more commonly seen in flatfish such as European plaice (HUNTER *et al.*, 2004), however, cod have also been shown to use this highly efficient means of transport during migration periods in the North Sea (cf. RIGHTON *et al.*, 2007).

Two findings are noteworthy: first, the diel variation in the fine-scale state process associated with coarse-scale state 3 (traveling or migrating), which illustrates that vertical activity is relatively higher during the night (as illustrated in the bottom-right panel of FIGURE 4.6), and second, that coarse-scale state 2 (more mobile foraging) mostly occurs during post-spawning migration, as the cod transits from spawning grounds in the southern North Sea to feeding grounds in the eastern English Channel. The fine-scale patterns of vertical movement identified during coarse-scale state 1 (resting or foraging) are indicative of cods' ability to vary their feeding and foraging patterns in relation to prey availability, whether prey are available by day, by night, or only during crepuscular periods. The variable patterns of vertical movement across coarse-scale states 2 and 3 suggest that cod are capable of migrating quickly to reach spawning grounds after a summer spent foraging on rich feeding grounds or moving more slowly and taking advantage of food resources to recover energy after the spawning period. Such adaptive migratory behavior could likely be overlooked by studies that limit their inquiries to movement in only one dimension (cf. HOBSON *et al.*, 2007) or when considering movement only at a daily scale (cf. GRIFFITHS *et al.*, 2018), which highlights the potential of the suggested approach in particular for ecological applications.

Qq-plots and sample ACFs of normal ordinary pseudo-residuals for coarse-scale step lengths and turning angles, as well as three example sequences of fine-scale vertical displacements, each computed as described in SECTION 4.3.2, are displayed in FIGURE 4.7. The plots indicate some minor lack of fit regarding the marginal distributions of the different variables, and some residual correlation in the step lengths' series. Overall, the magnitude of the lack of fit found here is anything but unusual for movement modeling exercises, which is due to the fairly complex patterns typically found in such data. Thus,

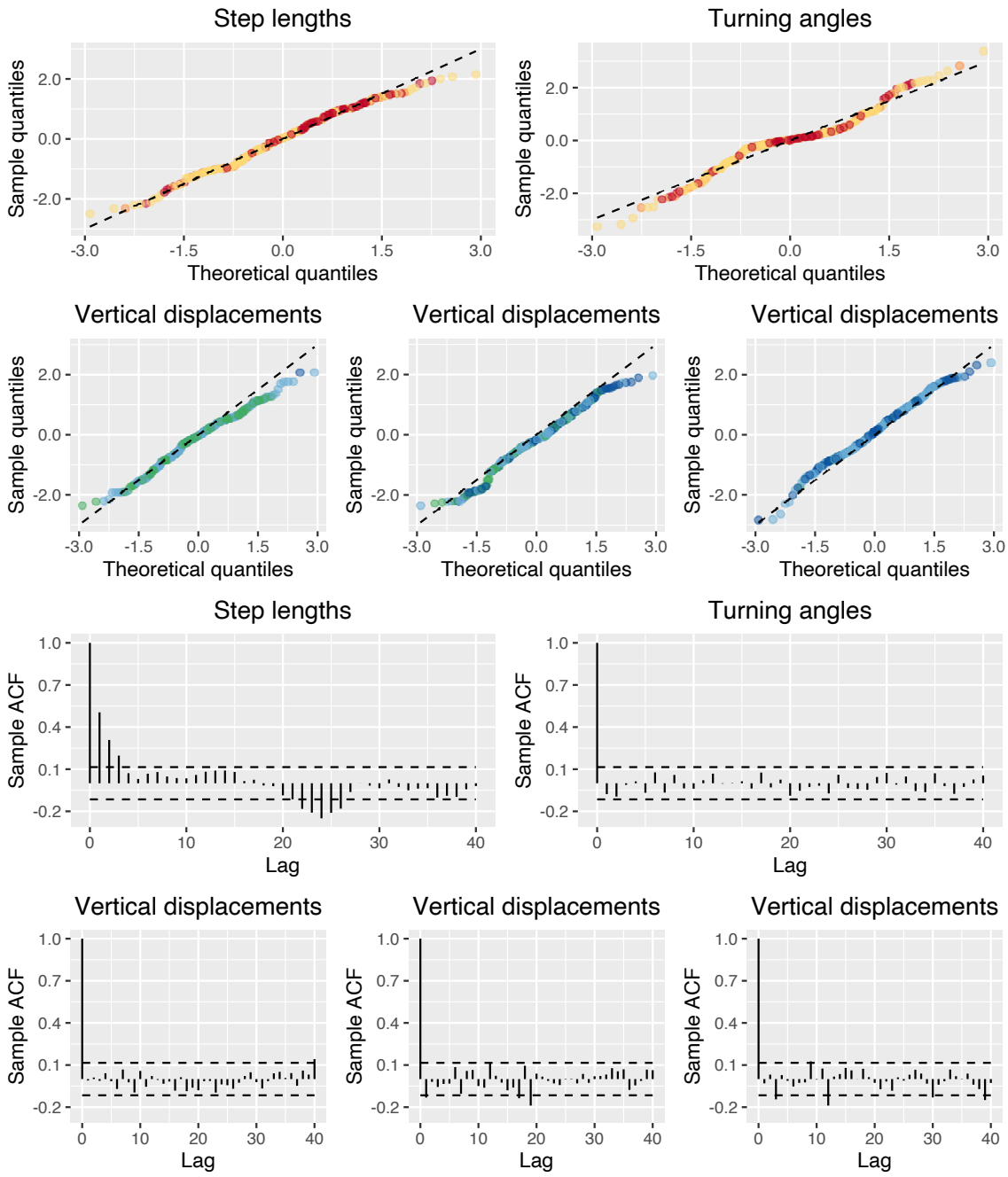

FIGURE 4.7: *Qq-plots (top panels) and sample ACFs (bottom panels) of normal ordinary pseudo-residuals for daily step lengths and turning angles as well as three example sequences of length* 48 *hours for ten-minute vertical displacements, one for coarse-scale state* 1 *(resting or foraging),* 2 *(more mobile foraging), and* 3 *(traveling or migrating), respectively.*

we consider the goodness of fit of our model to be satisfactory. In principle, more flexible state-dependent distributions such as mixture distributions or non-parametric distributions based on linear combinations of B-spline basis functions can be used to improve the fit (EILERS AND MARX, 1996; LANGROCK *et al.*, 2015; LANGROCK *et al.*, 2018; cf. also SECTION 2.3.2), which, however, we refrain from investigating further as our aim here is

to present an illustrative case study, thus trading some relatively minor lack of fit against a more complex model formulation, which would complicate the interpretation of the fitted model.

4.4.2 Application to stock market data

In a second case study, we demonstrate how hierarchical HMMs can be applied to stock market data, where we aim at investigating stock market dynamics at different time scales. Specifically, we jointly model 16 years of monthly trade volumes (in USD), Volume_t, $t = 1, \ldots, 192$, and daily log-returns, $\text{LogReturn}_{t,t'}$, $t = 1, \ldots, 192, t' = 1, \ldots, T'$, where T' varies between 19 and 23 (depending on the number of working days for the given month), of the Goldman Sachs stock. The data, which were downloaded from Yahoo Finance[5], cover 4,026 working days (i.e. 192 months) between January 1, 2004, and December 31, 2019. Thus, for each of the $T = 192$ monthly trade volumes, we have — on average — $T' = 21$ daily log-returns. While such data could potentially be modeled within an HMM framework either by down-sampling the log-returns to the monthly scale (thus focusing on the long-term dynamics), which can lead to a substantial loss of information that is actually contained in the raw data, or by fitting separate models for the two variables, which does not account for state processes operating at the coarse-scale (such as the economic regime), which also affect the fine-scale observations, we here demonstrate how hierarchical HMMs can be used to jointly model the two variables while retaining their respective time scales.

Based on an exploratory analysis of the data and a comparison of fitted models with different numbers of states, we chose $N = 3$ states for the coarse-scale state process, each of which was then connected to an HMM with $N' = 2$ states for the fine-scale observations (thus resulting in a total of six fine-scale states). For the trade volumes, which can take on positive continuous values, we assumed gamma state-dependent distributions, while for the log-returns, state-dependent scaled t-distributions (as preferred over normal distributions by AIC) with means fixed at zero were considered. These choices were validated using pseudo-residual analyses (cf. FIGURE 4.9 and the discussion below). The computation time required to fit the model was 2.8 minutes.

The estimated state-dependent distributions of monthly trade volumes, as displayed in the top-left panel of FIGURE 4.8, reveal three different market regimes. Coarse-scale

[5]https://finance.yahoo.com/quote/GS/history?p=GS. The data were downloaded on January 31, 2020.

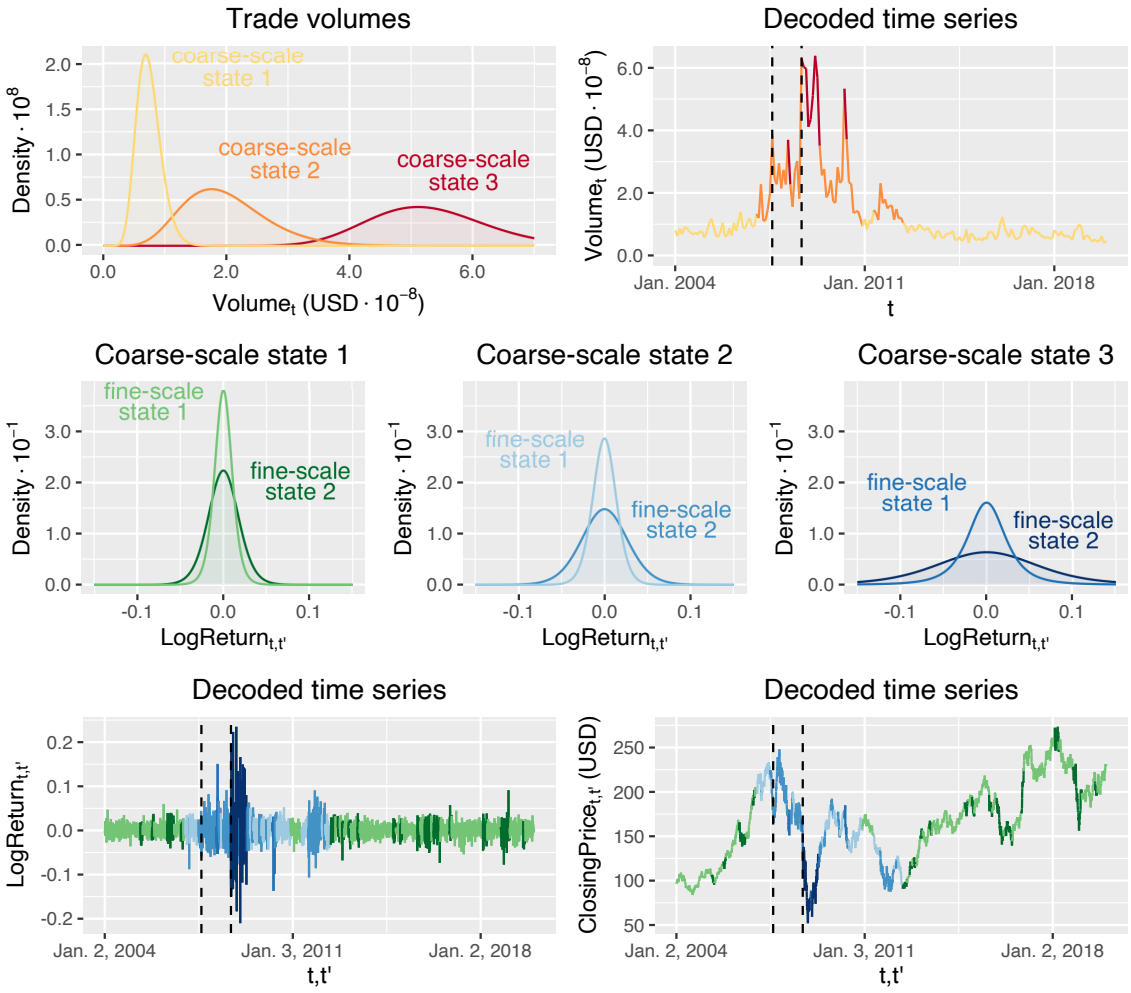

FIGURE 4.8: *Estimated state-dependent distributions and decoded time series of monthly trade volumes (top panel) as well as daily log-returns and closing prices (middle and bottom panel) of the Goldman Sachs stock. Dashed lines on August* 9, 2007*, and September* 15, 2008*, indicate important events associated with the global financial crisis, namely a sudden increase in interest rates for inter-bank credits and the collapse of Lehman Brothers, respectively.*

states 1 and 2 (which are colored in yellow and orange, respectively) capture low and moderately high trade volumes, respectively, thus indicating inactive and moderately active market phases. Coarse-scale state 3 (which is colored in red), in contrast, relates to high trade volumes (which can thus be interpreted as an active market regime). Notably, some switches between the different coarse-scale states can be linked to important events associated with the global financial crisis. In 2007, for instance, when a sudden increase in interest rates for inter-bank credits marked the beginning of the global financial crisis (cf. GUILLÉN, 2009), the decoded time series of monthly trade volumes displayed in the top-right panel of FIGURE 4.8 reveals a switch from coarse-scale state 1 (inactive market) to 2 (moderately active market; cf. the first dashed line). Furthermore, in September 2008,

when the Lehman Brothers collapse marked the peak of the global financial crisis (cf. SWEDBERG, 2010), we observe another switch from coarse-scale state 2 (moderately active market) to 3 (active market; cf. the second dashed line), indicating that trading activity on stock markets substantially increases during financial crises.

The t.p.m. associated with the coarse-scale state process was estimated as

$$\hat{\Gamma}_{\text{GS}} = \begin{pmatrix} 0.984 & 0.016 & 0.000 \\ 0.043 & 0.900 & 0.057 \\ 0.000 & 0.282 & 0.718 \end{pmatrix},$$

which implies the stationary distribution $(0.687, 0.261, 0.053)$, indicating that about 68.7 % (132 months), 26.1 % (50 months), and 5.3 % (10 months) of the observations were generated in coarse-scale states 1, 2, and 3, respectively.

The estimated state-dependent distributions of daily log-returns are displayed in the middle panel of FIGURE 4.8. Depending on the coarse-scale state that is active in month t, the log-returns' volatility is determined by the fine-scale HMM associated with the two state-dependent distributions displayed either in the left, the middle, or the right panel, respectively. According to the fitted model, when coarse-scale state 1 (inactive market) is active (which is the case in about 68.7 % of the time), then the marginal distribution of the log-returns under the fitted model has standard deviation 0.013. When coarse-scale state 3 (active market) is active (which is the case in about 5.3 % of the time), then the log-returns' volatility is about five times higher: the corresponding marginal distribution has standard deviation 0.065.

The t.p.m.s associated with the fine-scale state processes that determine the switches between the state-dependent distributions of the fine-scale HMMs were estimated as

$$\hat{\Gamma}_{\text{GS}}^{\prime(1)} = \begin{pmatrix} 0.993 & 0.007 \\ 0.034 & 0.966 \end{pmatrix}, \hat{\Gamma}_{\text{GS}}^{\prime(2)} = \begin{pmatrix} 0.993 & 0.007 \\ 0.024 & 0.976 \end{pmatrix}, \hat{\Gamma}_{\text{GS}}^{\prime(3)} = \begin{pmatrix} 0.915 & 0.085 \\ 0.029 & 0.971 \end{pmatrix},$$

which imply the stationary distributions $(0.823, 0.177)$, $(0.779, 0.221)$, and $(0.255, 0.745)$, respectively.

These results indicate that coarse-scale market dynamics, as characterized by different levels of trade volumes, strongly affect the stochastic properties of other processes operating at finer scales. By explicitly modeling such multi-scale processes, hierarchical HMMs can help us to draw a more comprehensive picture of the stock market's dynamics, to more accurately quantify risks conditional on the coarse-scale market regime, and ultimately to improve our understanding of the market agents' behavior. As the volatility of a stock

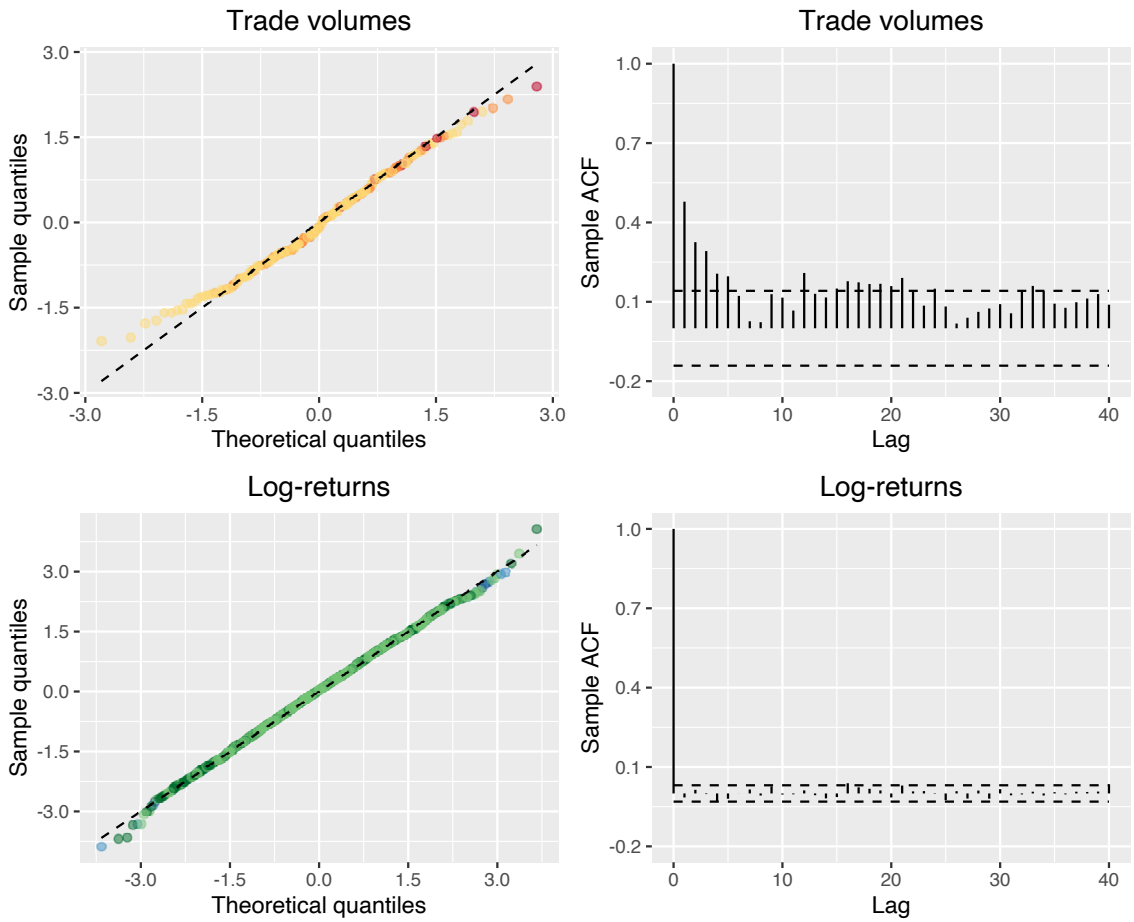

FIGURE 4.9: *Qq-plots (left panel) and sample ACFs (right panel) of normal ordinary pseudo-residuals for monthly trade volumes (top panel) and daily log-returns (bottom panel). Overall, the plots indicate some lack of fit with regard to the marginal distribution of the trade volumes and the serial correlation in the trade volumes' series.*

is often subject to state-switching over time (RYDÉN *et al.*, 1998; BULLA AND BULLA, 2006), potentially driven by complex short- and long-term patterns, hierarchical HMMs provide a useful tool especially for short-term forecasting, and is clearly superior e.g. to fitting a single scaled t-distribution to a time series of log-returns, i.e. neglecting any state-switching dynamics, or fitting a conventional HMM, i.e. without taking the coarse-scale economic regime into account.

Qq-plots and sample ACFs of ordinary normal pseudo-residuals for monthly trade volumes and daily log-returns, each computed as described in SECTION 4.3.2, are displayed in FIGURE 4.9. While indicating some lack of fit regarding the marginal distribution of the trade volumes and some residual correlation in the trade volumes' series, the lack of fit found here is anything but unusual when modeling economic time series. While, in principle, more flexible state-dependent distributions, especially for the trade volumes, could be used to improve the fit, or autoregressive terms in the coarse-scale state-dependent process

to reduce the trade volumes' autocorrelation that is not captured by the model, we consider the goodness of fit of the fitted model to be satisfactory and again trade some relatively minor lack of fit against a more complex model formulation to facilitate the interpretation of the fitted model.

4.5 Discussion

In this chapter, we introduced hierarchical HMMs as a versatile class of statistical models for multi-scale time series. The suggested approach was illustrated in two real-data applications, where we jointly modeled the distribution of i) daily horizontal movements and ten-minute vertical displacements of an Atlantic cod as well as ii) monthly trade volumes and daily log-returns of the Goldman Sachs stock, respectively. A key aspect in any such analysis is the temporal resolution at which the observations are made. A coarse resolution, as often obtained by GPS tags, can be suitable when the focus lies on traveling or migration patterns, whereas fine-scale data, as often collected by time-depth recorders or accelerometers, can reveal detailed information up to individual foraging attempts (PATTERSON *et al.*, 2017). While high-resolution data seem, in principle, to be more informative, we here argue that some of the corresponding short-term decisions made by an animal have to be seen relative to the current context. For example, as demonstrated in SECTION 4.4.1, complex fine-scale movement patterns, such as the effect of the time of day on vertical movements, could not have been revealed without taking the coarse-scale behavioral context into consideration. Vice versa, to obtain a more detailed understanding of movement patterns that appear to manifest themselves at coarser scales, it will often be helpful to be able to additionally "zoom in" at a much finer scale.

Fortunately, new types of remote sensing data, in particular such that result from outfitting animals with multiple telemetry sensors, give us the great opportunity to draw a more comprehensive picture of an animal's behavior. However, these new types of data are very challenging from a statistical perspective (LEOS-BARAJAS *et al.*, 2017b). Due to their intuitive appeal, their versatility in accommodating various dependence structures and essentially any type of time series, and the relative ease with which they can be implemented, hierarchical HMMs seem well-suited to handle such data and allow for comprehensive ecological inference from multi-stream and multi-scale data. Unlike previous approaches based on SSMs (JONSEN *et al.*, 2005; AUGER-MÉTHÉ *et al.*, 2016; AUGER-MÉTHÉ *et al.*, 2020), in the model formulations considered in this work we do not explicitly account for measurement error, which can in fact be large in particular for the geo-positional data used

in SECTION 4.4.1, which is itself the output of a geo-location model. Depending on the magnitude of the error, failing to propagate this uncertainty through to the model can affect state predictions, and hence ultimately also biological inferences. Despite this caveat, we see strong potential for hierarchical HMMs to become increasingly important in the future, especially due to the ongoing progress in bio-logging technology.

Furthermore, we demonstrated the potential of hierarchical HMMs for economic applications, where coarse-scale market dynamics can strongly affect the stochastic properties of other processes operating at finer scales. While hierarchical HMMs as proposed in this work are limited to modeling state processes with discrete state-spaces, they could potentially be extended in that a coarse-scale state process, modeled by a discrete-time, N-state Markov chain, selects among N possible SSMs with continuous state space for the fine-scale observations. This possible extension could be particularly useful in economic applications, where the coarse-scale states can often be linked to discrete economic regimes (such as recessions or periods of economic growth), whereas the fine-scale states (such as the level of the market agents' nervousness) sometimes gradually change over time, which can be naturally accounted for using SSMs (cf. FRIDMAN AND HARRIS, 1998; LANGROCK *et al.*, 2012c). In such scenarios, the synergy of an HMM operating at the coarse scale and multiple SSMs operating at the fine scale thus offers great opportunities for statistical inference.

On a final note, we would like to point out that, in analogy to speech recognition, the model formulation could be extended to more than two temporal resolutions: there could, for instance, be three connected state processes, which could be thought of as corresponding to the presence or absence of migratory behavior or the economic regime (coarsest scale), resting, foraging, and traveling behavior or monthly trade volumes (medium scale), and movements of individual body parts or daily or even intra-daily stock returns (finest scale). Being able to fit such complete models of animal movement or stock market dynamics seems to be intriguing. However, they would certainly not be as straightforward to implement and to handle, and the interpretation of such models would be more involved: while in basic HMMs, it is often straightforward to link the model's states to biologically or economically meaningful states, this is more difficult within hierarchical HMMs, where interpretations ought to be made at different time scales. In such extensions, but also for the models presented in this work, an important question is that of the optimal statistical design. Specifically, it would be of great interest to provide general recommendations as to which temporal resolution is needed at either time scale in order to answer the research questions at hand, which, however, is beyond the scope of this work and provides a promising avenue for future research.

Chapter 5

Conclusions

Chapter 5

Conclusions

> "The numbers have no way to speak for themselves. We speak for them. We imbue them with meaning."
>
> — *N. Silver*

5.1 Summary and outlook

In this thesis, we discussed three particular problems related to HMMs and proposed corresponding extensions of the basic model, namely i) Markov-switching GAMLSS (cf. CHAPTER 2), ii) non-parametric HMMs for discrete-valued time series (cf. CHAPTER 3), and iii) hierarchical HMMs for multi-scale time series (cf. CHAPTER 4). In simulation experiments and real-data examples, primarily focusing on applications from economics and ecology, we demonstrated how the methods developed can be used in particular i) to model different state-dependent parameters of the response distribution as potentially smooth functions of a given set of covariates, ii) to estimate the state-dependent distributions of an HMM for discrete-valued time series in a completely data-driven way without the need to specify a parametric family of distributions, and iii) to jointly model multiple variables that were observed at different temporal resolutions. In this last chapter, we conclude with a brief outlook on potential avenues for future research related to the different methods and provide some final remarks.

From a methodological perspective, it would be conceptually straightforward to combine the proposed extensions with each other: hierarchical state architectures as discussed in CHAPTER 4, for instance, could be incorporated into Markov-switching GAMLSS (cf. CHAPTER 2). In such a model, an N-state Markov chain operating on the coarse scale

could be thought of as selecting one of N possible Markov-switching GAMLSS that generates the observations at the fine-scale. While the energy prices modeled in SECTION 2.5 were collected on a daily scale, such an extension could, for instance, be used to incorporate intra-day prices, or, similar to the stock market application presented in SECTION 4.4.2, to incorporate monthly economic indicators, which could help us to draw a more comprehensive picture of the energy market's dynamics. Furthermore, the penalization approach that was proposed for non-parametric HMMs for discrete-valued time series in CHAPTER 3 could, for instance, be used to estimate the state-dependent distributions of hierarchical HMMs in a completely data-driven way without needing to specify a parametric family of distributions[1].

Another possible direction for future research could be to incorporate some of the tools developed in this work into other HMM-type models: the design of the EM algorithm presented in SECTION 2.3.1, for instance, could be adapted to exploit the gradient boosting framework for parameter estimation and variable selection not only in the state-dependent process (as it was done in this work for the case of Markov-switching GAMLSS), but also in the state process, where, for each row of the t.p.m., one multinomial logistic regression model could be used to model the state transitions obtained in the E-step as potentially smooth functions of a given set of covariates. This could be particularly useful in computational biology, where the set of potential covariates is typically large relative to the number of informative ones (e.g. when modeling gene expressions; cf. GUPTA *et al.*, 2007). Furthermore, the penalization approach proposed for non-parametric HMMs for discrete-valued time series in CHAPTER 3 could also be incorporated into HMMs with arbitrary state dwell-time distributions (which are also referred to as hidden semi-Markov models; cf. LANGROCK AND ZUCCHINI, 2011), where the state dwell-time distributions could be modeled in a completely data-driven way without the need to specify a parametric family of distributions. As an implicit assumption of basic HMMs is that the state dwell-times follow a geometric distribution with mode one, such an extension could also prove useful as an exploratory tool that can be used to investigate possible assumption violations. Finally, the likelihood-based inferential framework of hierarchical HMMs, as presented in CHAPTER 4, could also be extended towards combinations of an N-state HMM operating on the coarse scale whose Markov chain selects one of N possible models for the observations at the fine scale. These could, for instance, be SSMs (as discussed in SECTION

[1] While being conceptually straightforward, it requires further research to assess the extent to which such extensions would be feasible in practice, particularly as fitting these complex models may become difficult from a numerical perspective, cf. the discussion in SECTION 5.2.

4.5), but also other classes of statistical models for time series where the parameters can be estimated in a likelihood-based framework.

Taking the above ideas one step further, it would be conceptually appealing to unify the flexible extensions proposed in this work with the various other tools that are available in a modular "Lego toolbox" that can be used to build custom HMMs[2]. The building blocks of such a modeling framework could be thought of as "Lego bricks", encompassing i) various state architectures (e.g. simple Markov chains, semi-Markov chains, and hierarchical state processes), ii) various types of state-dependent distributions (e.g. discrete, continuous, parametric and non-parametric distributions as well as distributions whose parameters can be modeled as linear or smooth functions of a given set of covariates), and iii) different estimation techniques (e.g. numerical likelihood maximization, the EM algorithm, and gradient boosting). Depending on the data at hand, these "Lego bricks" could be recombined in various ways and thereby help to adequately address specific modeling challenges. While the MS-gamboostLSS algorithm proposed in SECTION 2.3.1, which can be used for variable selection and parameter estimation not only in Markov-switching GAMLSS but also in a variety of other HMM-type models (cf. the discussion in SECTION 2.6), provides a first step towards such a "Lego toolbox" for HMMs, an implementation that encompasses a larger set of "Lego bricks" is not yet available and therefore provides a promising avenue for future research.

5.2 Discussion and final remarks

In conclusion, we would like to summarize the considerations made above by noting that the statistical tools developed in this thesis are not to be regarded as closed, self-contained modeling frameworks that are limited to the different applications presented in this work. Instead, they — or, more precisely, the ideas contained therein — should be regarded as an extension to the previously available HMM toolbox that can also be combined with other HMM-type models, which will hopefully inspire the statistical community to develop new flexible extensions of the basic HMM that will help us to address research questions from a new statistical perspective, to separate the signal from the noise, and, ultimately, to extract information from data.

[2]Similar such "Lego toolboxes" were e.g. proposed for flexible Bayesian regression modeling (cf. UMLAUF *et al.*, 2019) and structured additive distributional regression models (cf. KNEIB *et al.*, 2019).

On a final note, we would like to raise awareness of the fact that the flexibility that comes along with the methods proposed in this work can — beside the opportunities mentioned above — also be a curse: especially in HMM-type models, there is often a trade-off between model complexity and numerical stability, with challenges such as local maxima of the likelihood likely being exacerbated as the number of parameters and the complexity of the model formulation increases. Further investigating the statistical properties of the methods proposed in this work, including providing general guidelines that can be used to increase the numerical stability of the estimation, is therefore an important direction for future research that we believe should always complement the development of new statistical techniques.

Looking to the future, the ever-increasing complexity of the data that is likely being collected over the next decades yields major challenges but at the same time offers great opportunities for statistical modeling in the 21st century. Challenges, on the one hand, primarily arise from the fact that conventional statistical methods sometimes have their difficulties in keeping pace with the available new types of data and, as a consequence, can fail to fully exploit the information contained therein. Great opportunities, on the other hand, lie in that novel statistical techniques can help to make sense of these complex types of data and, thereby, — referring to the words of N. Silver — "to imbue them with meaning", which can guide us towards new conclusions that could not have been drawn using previously available statistical methods. In that regard, this work provides a small contribution to the toolbox of statistical modeling techniques.

Appendix A

A forward algorithm for likelihood evaluation in hierarchical hidden Markov models

Appendix A

A forward algorithm for likelihood evaluation in hierarchical hidden Markov models

In this appendix to SECTION 4.2.2, we provide some details on likelihood evaluation in hierarchical HMMs. Specifically, we present a forward algorithm that can be used to efficiently evaluate the likelihood while simultaneously preventing numerical underflow.

To evaluate the logarithm of the likelihood as given by EQUATION (4.3), we proceed as follows: first, we evaluate the log-likelihoods of the fine-scale observations, i.e. the log-likelihood of each of the T chunks of fine-scale observations being generated by each of the N fine-scale HMMs (as selected by the coarse-scale state process), which is denoted by $\mathcal{L}(\theta'^{(i)}|\mathbf{y}'_t)$, $i = 1,\dots,N, t = 1,\dots,T$. Therefore, we define the fine-scale log-forward probabilities under the i-th fine-scale HMM as

$$\phi'^{(i,l)}_{t,t'} = \log\big(f(\mathbf{y}'_{t,1},\dots,\mathbf{y}'_{t,t'}, s'_{t,t'} = l | s_t = i)\big),$$

$l = 1,\dots,N'$. The fine-scale log-forward probabilities can be evaluated recursively via the forward algorithm, which amounts to applying the recursion

$$\begin{aligned}
\phi'^{(i,l)}_{t,1} &= \log\big(\delta'^{(i)}_l f_{\mathbf{Y}'}(\mathbf{y}'_{t,1}; \theta'^{(i,l)})\big) \\
&= \log\big(\delta'^{(i)}_l\big) + \log\big(f_{\mathbf{Y}'}(\mathbf{y}'_{t,1}; \theta'^{(i,l)})\big); \\
\phi'^{(i,l)}_{t,t'} &= \log\left(\sum_{k=1}^{N'} \exp\big(\phi'^{(i,k)}_{t,t'-1}\big)\gamma'^{(i)}_{k,l} f_{\mathbf{Y}'}\big(\mathbf{y}'_{t,t'}; \theta'^{(i,l)}\big)\right) \\
&= \log\left(\sum_{k=1}^{N'} \exp\left(\phi'^{(i,k)}_{t,t'-1} + \log\big(\gamma'^{(i)}_{k,l}\big) - c'_{t,t'-1}\right)\right) + c'_{t,t'-1} + \log\big(f_{\mathbf{Y}'}(\mathbf{y}'_{t,t'}; \theta'^{(i,l)})\big),
\end{aligned} \tag{A.1}$$

$t' = 2,\dots,T'$, where $c'_{t,t'} = \max(\phi'^{(i,1)}_{t,t'},\dots,\phi'^{(i,N')}_{t,t'})$ is a constant that is used within the

log-sum-of-exponentials function to prevent numerical underflow, which can occur when exponentiating large negative numbers.

Since, by the law of total probability, $\mathcal{L}(\theta'^{(i)}|\mathbf{y}'_t) = f_{\mathbf{Y}'}(\mathbf{y}'_{t,1},\ldots,\mathbf{y}'_{t,T'};\theta'^{(i)}) = \sum_{l=1}^{N'} f(\mathbf{y}'_{t,1},\ldots,\mathbf{y}'_{t,T'},s'_{t,T'}=l|s_t=i)$, the log-likelihood of the t-th chunk of fine-scale observations being generated by the i-th fine-scale HMM follows as

$$l(\theta'^{(i)}|\mathbf{y}'_t) = \log\left(\sum_{l=1}^{N'} \exp\left(\phi_{t,T'}^{\prime(i,l)} - c'_{t,T'}\right)\right) + c'_{t,T'}. \tag{A.2}$$

After having evaluated the log-likelihood for each of the T chunks of fine-scale observations and N fine-scale HMMs as given by EQUATION (A.2), we proceed with evaluating the coarse-scale log-forward probabilities,

$$\phi_t^{(j)} = \log\big(f(\mathbf{y}_1,\ldots,\mathbf{y}_t,\mathbf{y}'_1,\ldots,\mathbf{y}'_t,s_t=j)\big),$$

$j=1,\ldots,N$, which can be obtained in a similar way as given by EQUATIONS (A.1) by applying the recursion

$$\begin{aligned}
\phi_1^{(j)} &= \log\big(\delta_j \mathcal{L}(\theta^{(j)}|\mathbf{y}'_1) f_{\mathbf{Y}}(\mathbf{y}_1;\theta^{(j)})\big)\\
&= \log(\delta_j) + \log\big(\mathcal{L}(\theta^{(j)}|\mathbf{y}'_1)\big) + \log\big(f_{\mathbf{Y}}(\mathbf{y}_1;\theta^{(j)})\big);\\
\phi_t^{(j)} &= \log\left(\sum_{i=1}^{N} \exp(\phi_{t-1}^{(i)})\gamma_{i,j}\mathcal{L}(\theta^{(j)}|\mathbf{y}'_t) f_{\mathbf{Y}}\big(\mathbf{y}_t;\theta^{(j)}\big)\right) + \log\big(\mathcal{L}(\theta^{(j)}|\mathbf{y}'_t)\big)\\
&= \log\left(\sum_{i=1}^{N} \exp\big(\phi_{t-1}^{(i)} + \log(\gamma_{i,j}) - c_{t-1}\big)\right) + c_{t-1} + \log\big(\mathcal{L}(\theta^{(j)}|\mathbf{y}'_t)\big)\\
&\quad + \log\big(f_{\mathbf{Y}}(\mathbf{y}_t;\theta^{(j)})\big),
\end{aligned}$$

$t=2,\ldots,T$, where $c_t = \max(\phi_t^{(1)},\ldots,\phi_t^{(N)})$.

Since, by the law of total probability, $\mathcal{L}(\theta|\mathbf{y}_1,\ldots,\mathbf{y}_T,\mathbf{y}'_1,\ldots,\mathbf{y}'_T) = f_{\mathbf{Y},\mathbf{Y}'}(\mathbf{y}_1,\ldots,\mathbf{y}_T,\mathbf{y}'_1,\ldots,\mathbf{y}'_T) = \sum_{j=1}^{N} f(\mathbf{y}_1,\ldots,\mathbf{y}_T,\mathbf{y}'_1,\ldots,\mathbf{y}'_T,s_T=j)$, the log-likelihood of the hierarchical HMM follows as

$$l(\theta|\mathbf{y}_1,\ldots,\mathbf{y}_T,\mathbf{y}'_1,\ldots,\mathbf{y}'_T) = \log\left(\sum_{j=1}^{N} \exp\big(\phi_T^{(j)} - c_T\big)\right) + c_T.$$

Appendix B

Estimated coefficients for the fine-scale state transition probabilities

Appendix B

Estimated coefficients for the fine-scale state transition probabilities

In this appendix to SECTION 4.4.1, we provide some details on the estimated coefficients that determine the corresponding predictors for the fine-scale state transition probabilities. These were used to compute the stationary distributions as functions of the time of day, which are displayed in the right panel of FIGURE 4.5.

Using the multinomial logit link as detailed for the coarse-scale state transition probabilities in SECTION 4.2.3, the fine-scale state transition probabilities for the model presented in SECTION 4.4.1 can be written as

$$\gamma_{k,l}^{\prime(i)}(\text{TimeOfDay}_{t,t'}) = \frac{\exp(\eta^{\prime(i,k,l)}(\text{TimeOfDay}_{t,t'}))}{\sum_{m=1}^{N'}\exp(\eta^{\prime(i,k,m)}(\text{TimeOfDay}_{t,t'}))},$$

where the predictor can be written as

$$\eta^{\prime(i,k,l)}(\text{TimeOfDay}_{t,t'}) = \begin{cases} \beta_0^{\prime(i,k,l)} + \beta_1^{\prime(i,k,l)} \sin\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right) & \\ +\beta_2^{\prime(i,k,l)} \cos\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right) & \text{if } k \neq l; \\ 0 & \text{otherwise,} \end{cases} \tag{B.1}$$

$i = 1,\ldots,N$, $k,l = 1,\ldots,N'$. Note that predictors were estimated only for the off-diagonal t.p.m. entries; predictors for the diagonal t.p.m. entries were set to zero to ensure identifiability (cf. SECTION 4.2.3 for details).

The coefficients contained in EQUATION (B.1) associated with coarse-scale state 1 (resting or foraging), which determine the fine-scale state transition probabilities that were used to compute the stationary distributions displayed in the top-right panel of FIGURE

4.5, were estimated as

$$\hat{\eta}'^{(1,1,2)}(\text{TimeOfDay}_{t,t'}) = -2.569 - 0.030\sin\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right) - 0.663\cos\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right);$$

$$\hat{\eta}'^{(1,1,3)}(\text{TimeOfDay}_{t,t'}) = -3.397 + 0.266\sin\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right) - 0.266\cos\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right);$$

$$\hat{\eta}'^{(1,2,1)}(\text{TimeOfDay}_{t,t'}) = -2.767 - 0.192\sin\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right) + 0.679\cos\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right);$$

$$\hat{\eta}'^{(1,2,3)}(\text{TimeOfDay}_{t,t'}) = -4.369 + 0.477\sin\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right) + 0.534\cos\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right);$$

$$\hat{\eta}'^{(1,3,1)}(\text{TimeOfDay}_{t,t'}) = -4.850 + 0.271\sin\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right) + 2.812\cos\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right);$$

$$\hat{\eta}'^{(1,3,2)}(\text{TimeOfDay}_{t,t'}) = -2.567 + 0.385\sin\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right) + 0.071\cos\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right).$$

The coefficients that determine the fine-scale state transition probabilities associated with coarse-scale state 2 (more mobile foraging), which were used to compute the stationary distributions displayed in the middle-right panel of FIGURE 4.5, were estimated as

$$\hat{\eta}'^{(2,1,2)}(\text{TimeOfDay}_{t,t'}) = -2.945 + 0.274\sin\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right) + 0.776\cos\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right);$$

$$\hat{\eta}'^{(2,1,3)}(\text{TimeOfDay}_{t,t'}) = -2.409 + 0.069\sin\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right) + 0.038\cos\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right);$$

$$\hat{\eta}'^{(2,2,1)}(\text{TimeOfDay}_{t,t'}) = -2.152 - 0.096\sin\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right) + 0.138\cos\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right);$$

$$\hat{\eta}'^{(2,2,3)}(\text{TimeOfDay}_{t,t'}) = -3.552 - 0.182\sin\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right) - 1.106\cos\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right);$$

$$\hat{\eta}'^{(2,3,1)}(\text{TimeOfDay}_{t,t'}) = -3.140 + 0.816\sin\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right) - 0.226\cos\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right);$$

$$\hat{\eta}'^{(2,3,2)}(\text{TimeOfDay}_{t,t'}) = -2.858 - 0.702\sin\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right) - 0.383\cos\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right).$$

The coefficients that determine the fine-scale state transition probabilities associated with coarse-scale state 3 (traveling or migrating), which were used to compute the stationary distributions displayed in the bottom-right panel of FIGURE 4.5, were estimated as

$$\hat{\eta}'^{(3,1,2)}(\text{TimeOfDay}_{t,t'}) = -2.547 + 0.621\sin\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right) - 0.042\cos\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right);$$

$$\hat{\eta}'^{(3,1,3)}(\text{TimeOfDay}_{t,t'}) = -4.219 - 0.157\sin\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right) + 1.187\cos\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right);$$

$$\hat{\eta}'^{(3,2,1)}(\text{TimeOfDay}_{t,t'}) = -2.520 + 0.222\sin\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right) + 0.309\cos\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right);$$

$$\hat{\eta}'^{(3,2,3)}(\text{TimeOfDay}_{t,t'}) = -2.793 - 0.071\sin\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right) + 0.596\cos\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right);$$

$$\hat{\eta}'^{(3,3,1)}(\text{TimeOfDay}_{t,t'}) = -12.988 + 6.847\sin\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right) + 5.024\cos\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right);$$

$$\hat{\eta}'^{(3,3,2)}(\text{TimeOfDay}_{t,t'}) = -2.125 - 0.034\sin\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right) - 0.045\cos\left(\frac{2\pi\text{TimeOfDay}_{t,t'}}{24}\right).$$

Bibliography

ACERBI, C. AND TASCHE, D. (2002): Expected shortfall: a natural coherent alternative to value at risk. *Economic Notes*, **31**(2), 379–388.

ADAM, T., MAYR, A., AND KNEIB, T. (2017a): Gradient boosting in Markov-switching generalized additive models for location, scale, and shape. *arXiv*, 1710.02385 (submitted to *Econometrics and Statistics, Part B: Statistics*).

ADAM, T., LEOS-BARAJAS, V., LANGROCK, R., AND VAN BEEST, F.M. (2017b): Using hierarchical hidden Markov models for joint inference at multiple temporal scales. *Proceedings of the 32nd International Workshop on Statistical Modelling*, **2**, 181–184.

ADAM, T., MAYR, A., KNEIB, T., AND LANGROCK, R. (2018): Statistical boosting for Markov-switching distributional regression models. *Proceedings of the 33rd International Workshop on Statistical Modelling*, **1**, 30–35.

ADAM, T., GRIFFITHS, C.A., LEOS-BARAJAS, V., MEESE, E.N., LOWE, C.G., BLACKWELL, P.G., RIGHTON, D., AND LANGROCK, R. (2019a): Joint modelling of multi-scale animal movement data using hierarchical hidden Markov models. *Methods in Ecology and Evolution*, **10**(9), 1536–1550.

ADAM, T. (2019b): countHMM: penalized estimation of flexible hidden Markov models for time series of counts. *R package*, version 0.1.0. https://CRAN.R-project.org/ package=countHMM.

ADAM, T., LANGROCK, R., AND WEIß, C.H. (2019c): Penalized estimation of flexible hidden Markov models for time series of counts. *METRON*, **77**(2), 87–104.

ADAM, T., LANGROCK, R., AND WEIß, C.H. (2019d): Nonparametric inference in hidden Markov models for time series of counts. *Proceedings of the 34th International Workshop on Statistical Modelling*, **1**, 135–140.

ADAM, T., LANGROCK, R., AND KNEIB, T. (2019e): Model-based clustering of time series data: a flexible approach using nonparametric state-switching quantile regression

models. *Book of Short Papers of the 12th Scientific Meeting on Classification and Data Analysis*, 8–11.

ADAM, T. AND OELSCHLÄGER, L. (2020): Hidden Markov models for multi-scale time series: an application to stock market data. Available on request (submitted to the *Proceedings of the 35th International Workshop on Statistical Modelling*).

ADLERSTEIN, S.A. AND WELLEMAN, H.C. (2000): Diel variation of stomach contents of North Sea cod (*Gadus morhua*) during a 24-h fishing survey: an analysis using generalized additive models. *Canadian Journal of Fisheries and Aquatic Sciences*, **57**(12), 2363–2367.

ALBERT, P.S. (1991): A two-state Markov mixture model for a time series of epileptic seizure counts. *Biometrics*, **47**(4), 1371–1381.

ALEXANDROVICH, G., HOLZMANN, H., AND LEISTER, A. (2016): Nonparametric identification and maximum likelihood estimation for hidden Markov models. *Biometrika*, **103**(2), 423–434.

ALTMAN, R.M. AND PETKAU, A.J. (2005): Application of hidden Markov models to multiple sclerosis lesion count data. *Statistics in Medicine*, **24**(15), 2335–2344.

ALTMAN, R.M. (2007): Mixed hidden Markov models: an extension of the hidden Markov model to the longitudinal data setting. *Journal of the American Statistical Association*, **102**(477), 201–210.

ANDERSON, G., FARCOMENI, A., PITTAU, M.G., AND ZELLI, R. (2019): Rectangular latent Markov models for time-specific clustering, with an analysis of the well being of nations. *Journal of the Royal Statistical Society, Series C*, **68**(3), 603–621.

ASAI, K., HAYAMIZU, S., AND HANDA, K.I. (1993): Prediction of protein secondary structure by the hidden Markov model. *Bioinformatics*, **9**(2), 141–146.

AUGER-MÉTHÉ, M., FIELD, C., ALBERTSEN, C.M., DEROCHER, A.E., LEWIS, M.A., JONSEN, I.D., AND MILLS FLEMMING, J. (2016): State-space models' dirty little secrets: even simple linear Gaussian models can have estimation problems. *Scientific Reports*, **6**(1), 1–10.

AUGER-MÉTHÉ, M., NEWMAN, K., COLE, D., EMPACHER, F., GRYBA, R., KING, A.A., LEOS-BARAJAS, V., MILLS FLEMMING, J., NIELSEN, A., PETRIS, J., AND THOMAS, L. (2020): An introduction to state-space modeling of ecological time series. *arXiv*, 2002.02001.

BAHL, L. AND JELINEK, F. (1975): Decoding for channels with insertions, deletions, and substitutions with applications to speech recognition. *IEEE Transactions on Information Theory*, **21**(4), 404–411.

BAKER, J. (1975): The DRAGON system — an overview. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **23**(1), 24–29.

BAUM, L.E. AND PETRIE, T. (1966): Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics*, **37**(6), 1554–1563.

BAUM, L.E., PETRIE, T., SOULES, G., AND WEISS, N. (1970): A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, **41**(1), 164–171.

BEBBINGTON, M.S. (2007): Identifying volcanic regimes using hidden Markov models. *Geophysical Journal International*, **171**(2), 921–942.

BERENTSEN, G.D., BULLA, J., MARUOTTI, A., AND STØVE, B. (2018): Modelling corporate defaults: a Markov-switching Poisson log-linear autoregressive model. *arXiv*, 1804.09252.

BEYERLEIN, A., FAHRMEIR, L., MANSMANN, U., AND TOSCHKE, A.M. (2008): Alternative regression models to assess increase in childhood BMI. *BMC Medical Research Methodology*, **8**(1), 59.

BIERNACKI, C., CELEUX, G., AND GOVEART, G. (2000): Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(7), 719–725.

BULLA, J. AND BULLA, I. (2006): Stylized facts of financial time series and hidden semi-Markov models. *Computational Statistics and Data Analysis*, **51**(4), 2192–2209.

BULLA, J., LAGONA, F., MARUOTTI, A., AND PICONE, M. (2012): A multivariate hidden Markov model for the identification of sea regimes from incomplete skewed and circular time series. *Journal of Agricultural, Biological and Environmental Statistics*, **17**(4), 544–567.

CELEUX, G. AND DURAND, J.B. (2008): Selecting hidden Markov model state number with cross-validated likelihood. *Computational Statistics*, **23**(4), 541–564.

CHEN, M.Y., KUNDU, A., AND ZHOU, J. (1994): Off-line handwritten word recognition using a hidden Markov model type stochastic network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **16**(5), 481–496.

CHEN, F.S., FU, C.M., AND HUANG, C.L. (2003): Hand gesture recognition using a real-time tracking method and hidden Markov models. *Image and Vision Computing*, **21**(8), 745–758.

CHING, W.K., NG, M.K., AND WONG, K.K. (2004): Hidden Markov models and their applications to customer relationship management. *IMA Journal of Management Mathematics*, **15**(1), 13–24.

DE BOOR, C. (1978): *A practical guide to splines*. Springer, New York.

DE CASTRO, M., CANCHO, V.G., AND RODRIGUES, J. (2010): A hands-on approach for fitting long-term survival models under the GAMLSS framework. *Computer Methods and Programs in Biomedicine*, **97**(2), 168–177.

DE SOUZA, C.P.E. AND HECKMAN, N.E. (2014): Switching nonparametric regression models. *Journal of Nonparametric Statistics*, **26**(4), 617–637.

DE SOUZA, C.P.E., HECKMAN, N.E., AND XU, F. (2017): Switching nonparametric regression models for multi-curve data. *The Canadian Journal of Statistics*, **45**(4), 442–460.

DEMPSTER, A.P., LAIRD, N.M., AND RUBIN, D.B. (1977): Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, **39**(1), 1–22.

DERUITER, S.L., LANGROCK, R., SKIRBUTAS, T., GOLDBOGEN, J.A., CALAMBOKIDIS, J., FRIEDLAENDER, A.S., AND SOUTHALL, B.L. (2017): A multivariate mixed hidden Markov model for blue whale behaviour and responses to sound exposure. *The Annals of Applied Statistics*, **11**(1), 362–392.

DROST, F.C., VAN DEN AKKER, R., AND WERKER, B.J.M. (2009): Efficient estimation of auto-regression parameters and innovation distributions for semiparametric integer-valued AR (p) models. *Journal of the Royal Statistical Society, Series B*, **71**(2), 467–485.

DURBIN, R., EDDY, S.R., KROGH, A., AND MITCHISON, G. (1998): *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge.

DURBIN, J. AND KOOPMAN, S.J. (2012): *Time series analysis by state space methods*. Oxford University Press, Oxford.

EDDY, S.R. (1996): Hidden Markov models. *Current Opinion in Structural Biology*, **6**(3), 361–365.

EILERS, P.H.C. AND MARX, B.D. (1996): Flexible smoothing with B-splines and penalties. *Statistical Science*, **11**(2), 89–102.

FAHRMEIR, L., KNEIB, T., LANG, S., AND MARX, B. (2013). *Regression*. Springer, Berlin, Heidelberg.

FARCOMENI, A. (2017): Penalized estimation in latent Markov models, with application to monitoring serum calcium levels in end-stage kidney insufficiency. *Biometrical Journal*, **59**(5), 1035–1046.

FINE, S., SINGER, Y., AND TISHBY N. (1998): The hierarchical hidden Markov model: analysis and applications. *Machine Learning*, **32**(1), 41–62.

FONTDECABA, S., MUÑYOZ, M.P., AND SÀNCHEZ, J.A. (2009): Estimating Markovian switching regression models in R. An application to model energy price in Spain. *The Use R Conference 2009*.

FRIDMAN, M. AND HARRIS, L. (1998): A maximum likelihood approach for non-Gaussian stochastic volatility models. *Journal of Business and Economic Statistics*, **16**(3), 284–291.

GOLDFELD, S.M. AND QUANDT, R.E. (1973): A Markov model for switching regressions. *Journal of Econometrics*, **1**(1), 3–16.

GRECIAN, W.J., LANE, J.V., MICHELOT, T., WADE, H.M., AND HAMER, K.C. (2018): Understanding the ontogeny of foraging behaviour: insights from combining marine predator bio-logging with satellite-derived oceanography in hidden Markov models. *Journal of the Royal Society Interface*, **15**(143), 20180084.

GREEN, B. AND ZWIEBEL, J. (2018): The hot-hand fallacy: cognitive mistakes or equilibrium adjustments? Evidence from major league baseball. *Management Science*, **64**(11), 5315–5348.

GREWAL, M.S. AND ANDREWS, A.P. (2010): Applications of Kalman filtering in aerospace 1960 to the present. *IEEE Control Systems Magazine*, **30**(3), 69–78.

GRIFFITHS, C.A., PATTERSON, T.A., BLANCHARD, J.L., RIGHTON, D., WRIGHT, S.R., PITCHFORD, J.W., AND BLACKWELL, P.G. (2018): Scaling marine fish movement behavior from individuals to populations. *Ecology and Evolution*, **8**(14), 7031–7043.

GUILLÉN, M.F. (2009): The global economic and financial crisis: a timeline. *The Lauder Institute*, University of Pennsylvania, 1–91.

GUPTA, M., QU, P., AND IBRAHIM, J.G. (2007): A temporal hidden Markov regression model for the analysis of gene regulatory networks. *Biostatistics*, **8**(4), 805–820.

HAMBUCKERS, J., KNEIB, T., LANGROCK, R., AND SILBERSDORFF, A. (2018): A Markov-switching generalized additive model for compound Poisson processes, with applications to operational loss models. *Quantitative Finance*, **18**(10), 1679–1698.

HAMILTON, J.D. (1989): A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, **57**(2), 357–384.

HASSAN, M.R. AND NATH, B. (2005): Stock market forecasting using hidden Markov model: a new approach. *Proceedings of 5th International Conference on Intelligent Systems Design and Applications*, 192–196.

HAYS, G.C., BAILEY, H., BOGRAD, S.J., DON BOWEN, W., CAMPAGNA, C., . . . , AND SEQUEIRA, A.M.M. (2019): Translating marine animal tracking data into conservation policy and management. *Trends in Ecology and Evolution*, **34**(5), 459–473.

HELLER, G.Z., STASINOPOULOS D.M., RIGBY R.A., AND DE JONG, P. (2007): Mean and dispersion modeling for policy claims costs. *Scandinavian Actuarial Journal*, **4**, 281–292.

HOBSON, V.J., RIGHTON, D., METCALFE, J.D., AND HAYS, G.C. (2007): Vertical movements of North Sea cod. *Marine Ecology Progress Series*, **347**, 101–110.

HOBSON, V.J., RIGHTON, D., METCALFE, J.D., AND HAYS, G.C. (2009): Link between vertical and horizontal movement patterns of cod in the North Sea. *Aquatic Biology*, **5**(2), 133–142.

HOFNER, B. (2011): *Boosting in structured additive models*. Doctoral dissertation, Ludwig-Maximilians-Universität, Munich.

HOFNER, B., MAYR, A., AND SCHMID, M. (2016): gamboostLSS: an R package for model building and variable selection in the GAMLSS framework. *Journal of Statistical Software*, **74**(1), 1–31.

HU, J., BROWN, M.K., AND TURIN, W. (1996): HMM based online handwriting recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **18**(10), 1039–1045.

HUDSON, I.L. (2010): Interdisciplinary approaches: towards new statistical methods for phenological studies. *Climatic Change*, **100**(1), 143–171.

HUGHEY, R. AND KROGH, A. (1996): Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Bioinformatics*, **12**(2), 95–107.

HUNTER, E., METCALFE, J.D., O'BRIEN, C.M., ARNOLD, G.P., AND REYNOLDS, J.D. (2004): Vertical activity patterns of free-swimming adult plaice in the southern North Sea. *Marine Ecology Progress Series*, **279**, 261–273.

HUSSEY, N.E., KESSEL, S.T., AARESTRUP, K., COOKE, S.J., COWLEY, P.D., FISK, A.T., HARCOURT, R.G., HOLLAND, K.N., IVERSON, S.J., KOCIK, J.F., MILLS FLEMMING, J., AND WHORISKEY, F.G. (2015): Aquatic animal telemetry: a panoramic window into the underwater world. *Science*, **348**(6240), 1255642.

JACKSON, C.H. AND SHARPLES, L.D. (2002): Hidden Markov models for the onset and progression of bronchiolitis obliterans syndrome in lung transplant recipients. *Statistics in Medicine*, **21**(1), 113–128.

JELINEK, F. (1969): A fast sequential decoding algorithm using a stack. *IBM Journal of Research and Development*, **13**(6), 675–685.

JELINEK, F., BAHL, R., AND MERCER, R. (1975): Design of a linguistic statistical decoder for the recognition of continuous speech. *IEEE Transactions on Information Theory*, **21**(3), 250–256.

JELINEK, F. (1976): Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, **64**(4), 532–556.

JOHNSON, D.S., LONDON, J.M., LEA, M.A., AND DURBAN, J.W. (2008): Continuous-time correlated random walk model for animal telemetry data. *Ecology*, **89**(5), 1208–1215.

JONSEN, I.D., MILLS FLEMMING, J., AND MYERS, R.A. (2005): Robust state-space modeling of animal movement data. *Ecology*, **86**(11), 2874–2880.

KIM, C.J. AND NELSON, C.R. (1999): *State-space models with regime switching: classical and Gibbs-sampling approaches with applications*. MIT Press, Cambridge.

KIM, C.J., PIGER, J., AND STARTZ, R. (2008): Estimation of Markov regime-switching regression models with endogenous switching. *Journal of Econometrics*, **143**(2), 263–273.

KIRILENKO, A., KYLE, A.S., SAMADI, M., AND TUZUN, T. (2017): The flash crash: high-frequency trading in an electronic market. *The Journal of Finance*, **72**(3), 967–998.

KNEIB, T., KLEIN, N., LANG, S., AND UMLAUF, N. (2019): Modular regression — a Lego system for building structured additive distributional regression models with tensor product interactions. *Test*, **28**(1), 1–39.

KROGH, A., MIAN, I.S., AND HAUSSLER, D. (1994a): A hidden Markov model that finds genes in E. coli DNA. *Nucleic Acids Research*, **22**(22), 4768–4778.

KROGH, A., BROWN, M., MIAN, I.S., SJOLANDER, K., AND HAUSSLER, D. (1994b): Hidden Markov models in computational biology. Applications to protein modeling. *Journal of Molecular Biology*, **235**(5), 1501–1531.

LAGONA, F., MARUOTTI, A., AND PADOVANO, F. (2015): Multilevel multivariate modelling of legislative count data, with a hidden Markov chain. *Journal of the Royal Statistical Society, Series A*, **178**(3), 705–723.

LANGROCK, R. AND ZUCCHINI, W. (2011): Hidden Markov models with arbitrary state dwell-time distributions. *Computational Statistics and Data Analysis*, **55**(1), 715–724.

LANGROCK, R. (2012a): Flexible latent-state modelling of Old Faithful's eruption interarrival times in 2009. *Australian and New Zealand Journal of Statistics*, **54**(3), 261–279.

LANGROCK, R., KING, R., MATTHIOPOULOS, J., THOMAS, L., FORTIN, D., AND MORALES, J.M. (2012b): Flexible and practical modeling of animal telemetry data: hidden Markov models and extensions. *Ecology*, **93**(11), 2336–2342.

LANGROCK, R., MACDONALD, I.L., AND ZUCCHINI, W. (2012c): Some nonstandard stochastic volatility models and their estimation using structured hidden Markov models. *Journal of Empirical Finance*, **19**(1), 147–161.

LANGROCK, R., MARQUES, T.A., BAIRD, R.W., AND THOMAS, L. (2013a): Modeling the diving behavior of whales: a latent-variable approach with feedback and semi-Markovian components. *Journal of Agricultural, Biological and Environmental Statistics*, **19**(1), 82–100.

LANGROCK, R., SWIHART, B.J., CAFFO, B.S., PUNJABI, N.M., AND CRAINICEANU, C.M. (2013b): Combining hidden Markov models for comparing the dynamics of multiple sleep electroencephalograms. *Statistics in Medicine*, **32**(19), 3342–3356.

LANGROCK, R., KNEIB, T., SOHN, A., AND DERUITER, S.L. (2015): Nonparametric inference in hidden Markov models using P-splines. *Biometrics*, **71**(2), 520–528.

LANGROCK, R., KNEIB, T., GLENNIE, R., AND MICHELOT, T. (2017): Markov-switching generalized additive models. *Statistics and Computing*, **27**(1), 259–270.

LANGROCK, R., ADAM, T., LEOS-BARAJAS, V., MEWS, S., MILLER, D.L., AND PAPASTAMATIOU, Y.P. (2018): Spline-based nonparametric inference in general state-switching models. *Statistica Neerlandica*, **72**(3), 179–200.

LE STRAT, Y. AND CARRAT, F. (1999): Monitoring epidemiologic surveillance data using hidden Markov models. *Statistics in Medicine*, **18**(24), 3463–3478.

LEOS-BARAJAS, V., PHOTOPOULOU, T., LANGROCK, R., PATTERSON, T.A., WATANABE, Y.Y., MURGATROYD, M., AND PAPASTAMATIOU, Y.P. (2017a): Analysis of animal accelerometer data using hidden Markov models. *Methods in Ecology and Evolution*, **8**(2), 161–173.

LEOS-BARAJAS, V., GANGLOFF, E.J., ADAM, T., LANGROCK, R., VAN BEEST, F.M., NABE-NIELSEN, J., AND MORALES, J.M. (2017b): Multi-scale modeling of animal movement and general behavior data using hidden Markov models with hierarchical structures. *Journal of Agricultural, Biological and Environmental Statistics*, **22**(3), 232–248.

LI, L. AND CHENG, J. (2015): Modeling and forecasting corporate default counts using hidden Markov model. *Journal of Economics, Business and Management*, **3**(5), 493–497.

LI, M. AND BOLKER, B.M. (2017): Incorporating periodic variability in hidden Markov models for animal movement. *Movement Ecology*, **5**(1).

LIU, X. AND CHENG, T. (2003): Video-based face recognition using adaptive hidden Markov models. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, **1**, 1.

LØKKEBORG, S. (1998): Feeding behaviour of cod, *Gadus morhua*: activity rhythm and chemically mediated food search. *Animal Behaviour*, **56**(2), 371–378.

MACDONALD, I.L. AND RAUBENHEIMER, D. (1995): Hidden Markov models and animal behaviour. *Biometrical Journal*, **37**(6), 701–712.

MACDONALD, I.L. AND ZUCCHINI, W. (1997): *Hidden Markov and other models for discrete-valued time series*. Chapman and Hall/CRC, Boca Raton.

MARINO, M.F., TZAVIDIS, N., AND ALFÒ, M. (2018): Mixed hidden Markov quantile regression models for longitudinal data with possibly incomplete sequences. *Statistical Methods in Medical Research*, **27**(7), 2231–2246.

MARUOTTI, A. (2011): Mixed hidden Markov models for longitudinal data: an overview. *International Statistical Review*, **79**(3), 427–454.

MARUOTTI, A. AND ROCCI, R. (2012): A mixed non-homogeneous hidden Markov model for categorical data, with application to alcohol consumption. *Statistics in Medicine*, **31**(9), 871–886.

MAYR, A., FENSKE, N., HOFNER, B., KNEIB, T., AND SCHMID, M. (2012): Generalized additive models for location, scale, and shape for high dimensional data — a flexible approach based on boosting. *Journal of the Royal Statistical Society, Series C*, **61**(3), 403–427.

MAYR, A., BINDER, H., GEFELLER, O., AND SCHMID, M. (2014): The evolution of boosting algorithms — from machine learning to statistical modelling. *Methods of Information in Medicine*, **53**(6), 419–427.

MAYR, A., SCHMID, M., PFAHLBERG, A., UTER, W., AND GEFELLER, O. (2017): A permutation test to analyse systematic bias and random measurement errors of medical devices via boosting location and scale models. *Statistical Methods in Medical Research*, **26**(3), 1443–1460.

MCCLINTOCK, B.T., LANGROCK, R., GIMENEZ, O., CAM, E., BORCHERS, D.L., GLENNIE, R., AND PATTERSON, T.A. (2020): Uncovering ecological state dynamics with hidden Markov models. *arXiv*, 2002.10497v1.

MICHELOT, T., LANGROCK, R., AND PATTERSON, T.A. (2016): moveHMM: an R package for the statistical modelling of animal movement data using hidden Markov models. *Methods in Ecology and Evolution*, **7**(11), 1308–1315.

MICHELOT, T., LANGROCK, R., BESTLEY, S., JONSEN, I.D., PHOTOPOULOU, T., AND PATTERSON, T.A. (2017): Estimation and simulation of foraging trips in land-based marine predators. *Ecology*, **98**(7), 1932–1944.

MICHELOT, T. AND BLACKWELL, P.G. (2019): State-switching continuous-time correlated random walks. *Methods in Ecology and Evolution*, **10**(5), 637–649.

MORALES, J.M., HAYDON, D.T., FRAIR, J., HOLSINGER, K.E., AND FRYXELL, J.M. (2004): Extracting more out of relocation data: building movement models as mixtures of random walks. *Ecology*, **85**(9), 2436–2445.

MUNCH, K. AND KROGH, A. (2006): Automatic generation of gene finders for eukaryotic species. *BMC Bioinformatics*, **7**(1), 263.

NEAT, F.C., BENDALL, V., BERX, B., WRIGHT, P.J., Ó CUAIG, M., TOWNHILL, B., SCHÖN, P.-J., LEE, J., AND RIGHTON, D. (2014): Movement of Atlantic cod around the British Isles: implications for finer scale stock management. *Journal of Applied Ecology*, **51**(6), 1564–1574.

NEFIAN, A.V. AND HAYES, M.H. (1998): Hidden Markov models for face recognition. *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*, **5**, 2721–2724.

NETZER, O., LATTIN, J. M., AND SRINIVASAN, V. (2008): A hidden Markov model of customer relationship dynamics. *Marketing Science*, **27**(2), 185–204.

O'HARA, M. (2015): High frequency market microstructure. *Journal of Financial Economics*, **116**(2), 257–270.

ÖTTING, M., LANGROCK, R., DEUTSCHER, C., AND LEOS-BARAJAS, V. (2020): The hot hand in professional darts. *Journal of the Royal Statistical Society, Series A*, **183**(2), 565–580.

PACHTER, L., ALEXANDERSSON, M., AND CAWLEY, S. (2002): Applications of generalized pair hidden Markov models to alignment and gene finding problems. *Journal of Computational Biology*, **9**(2), 389–399.

PATTERSON, T.A., BASSON, M., BRAVINGTON, M.V., AND GUNN, J.S. (2009): Classifying movement behaviour in relation to environmental conditions using hidden Markov models. *Journal of Animal Ecology*, **78**(6), 1113–1123.

PATTERSON, T.A., PARTON, A., LANGROCK, R., BLACKWELL, P.G., THOMAS, L., AND KING, R. (2017): Statistical modelling of individual animal movement: an overview of key methods and a discussion of practical challenges. *AStA Advances in Statistical Analysis*, **101**(4), 399–438.

PEDERSEN, M.W., RIGHTON, D., THYGESEN, U.H., ANDERSEN, K.H., AND MADSEN, H. (2008): Geolocation of North Sea cod (*Gadus morhua*) using hidden Markov models and behavioural switching. *Canadian Journal of Fisheries and Aquatic Sciences*, **65**(11), 2367–2377.

PINSON, P. AND MADSEN, H. (2012): Adaptive modelling and forecasting of offshore wind power fluctuations with Markov-switching autoregressive models. *Journal of Forecasting*, **31**(4), 281–313.

PLÖTZ, T. AND FINK, G.A. (2009): Markov models for offline handwriting recognition: a survey. *International Journal on Document Analysis and Recognition*, **12**(4), 269.

POHLE, J., LANGROCK, R., VAN BEEST, F.M., AND SCHMIDT, N.M. (2017): Selecting the number of states in hidden Markov models — pragmatic solutions illustrated using animal movement. *Journal of Agricultural, Biological and Environmental Statistics*, **22**(3), 270–293.

POPOV, V., LANGROCK, R., DERUITER, S.L., AND VISSER, F. (2017): An analysis of pilot whale vocalization activity using hidden Markov models. *Journal of the Acoustical Society of America*, **141**(1), 159–171.

R CORE TEAM (2019): *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. https://www.R-project.org.

RABINER, L.R. AND JUANG, B.H. (1986): An introduction to hidden Markov models. *IEEE ASSP Magazine*, **3**(1), 4–16.

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

RIDGEWAY, G. (1999): The state of boosting. *Computing Science and Statistics*, **31**, 172–181.

RIGBY, R.A. AND STASINOPOULOS, D.M. (2005): Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society, Series C*, **54**(3), 507–554.

RIGBY, R.A. AND STASINOPOULOS, D.M. (2006): Using the Box-Cox t distribution in GAMLSS to model skewness and kurtosis. *Statistical Modelling*, **6**(3), 209–229.

RIGBY, R.A., STASINOPOULOS, D.M., AND VOUDOURIS, V. (2013): Discussion: a comparison of GAMLSS with quantile regression. *Statistical Modelling*, **13**(4), 335–348.

RIGBY, R.A., STASINOPOULOS, D.M., HELLER, G.Z., AND DE BASTIANI, F. (2019): *Distributions for modeling location, scale, and shape: using GAMLSS in R*. Chapman and Hall/CRC, Boca Raton.

RIGHTON, D., METCALFE, J.D., AND CONNOLLY, P. (2001): Different behaviour of North and Irish Sea cod. *Nature*, **411**(6834), 156.

RIGHTON, D., QUAYLE, V.A., HETHERINGTON, S., AND BURT, G. (2007): Movements and distribution of cod (*Gadus morhua*) in the southern North Sea and English Channel: results from conventional and electronic tagging experiments. *Journal of the Marine Biological Association of the United Kingdom*, **87**(2), 599–613.

RIGHTON, D., WRIGHT, S., GRIFFITHS, C.A., AND ADAM, T. (2019): Horizontal and vertical movement data derived from a data storage tag deployed on a single Atlantic cod in the English Channel from 2005 to 2006. *Cefas*, https://doi.org/10.14466/Cefas DataHub.71.

ROCKAFELLAR, R.T. AND URYASEV, S. (2002): Conditional value-at-risk for general loss distributions. *Journal of Banking and Finance*, **26**(7), 1443–1471.

ROGERS, R.D. (1985): Quote from an interview by C. Campbell, in: Torrent of print strains the fabric of libraries. *The New York Times*, February 25, 1985, **A**, 10.

RUTZ, C. AND HAYS, G.C. (2009): New frontiers in biologging science. *Biology Letters*, **5**(3), 289–292.

RYDÉN, T., TERÄSVIRTA, T., AND ÅSBRINK, S. (1998): Stylized facts of daily return series and the hidden Markov model. *Journal of Applied Econometrics*, **13**(3), 217–244.

SANCHEZ-ESPIGARES, J.A. AND LOPEZ-MORENO, A. (2014): MSwM: fitting Markov-switching models. *R package*, version 1.2. http://CRAN.R-project.org/package =MSwM.

SCHLIEHE-DIECKS, S., KAPPELER, P.M., AND LANGROCK, R. (2012): On the application of mixed hidden Markov models to multiple behavioural time series. *Interface Focus*, **2**(2), 180–189.

SCHUSTER-BÖCKLER, B. AND BATEMAN, A. (2007): An introduction to hidden Markov models. *Current Protocols in Bioinformatics*, **18**(1), A-3A.

SCOTT, D.W., TAPIA, R.A., AND THOMPSON, J.R. (1980): Nonparametric probability density estimation by discrete maximum penalized-likelihood criteria. *The Annals of Statistics*, **8**(4), 820–832.

SCOTT, S.L., JAMES, G.M., AND SUGAR, C.A. (2005): Hidden Markov models for longitudinal comparisons. *Journal of the American Statistical Association*, **100**(470), 359–369.

SILVER, N. (2012): *The Signal and the Noise*. Penguin, London.

SIMONOFF, J.S. (1983): A penalty function approach to smoothing large sparse contingency tables. *The Annals of Statistics*, **11**(1), 208–218.

SIMONOFF, J.S. (1996): *Smoothing Methods in Statistics*. Springer, New York.

SPIEGELHALTER, D. (2019): *The Art of Statistics*. Penguin, London.

SRIRAM, K., RAMAMOORTHI, R.V., AND GHOSH, P. (2016): On Bayesian quantile regression using a pseudo-joint asymmetric Laplace likelihood. *Sankhya A*, **78**(1), 87–104.

STÄDLER, N. AND MUKHERJEE, S. (2013): Penalized estimation in high-dimensional hidden Markov models with state-specific graphical models. *The Annals of Applied Statistics*, **7**(4), 2157–2179.

STANKE, M., SCHÖFFMANN, O., MORGENSTERN, B., AND WAACK, S. (2006): Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics*, **7**(1), 62.

STASINOPOULOS, D.M., RIGBY, R.A., HELLER, G.Z., VOUDOURIS, V., AND DE BASTIANI, F. (2017): *Flexible regression and smoothing: using GAMLSS in R*. Chapman and Hall/CRC, Boca Raton.

STATHOPOULOS, A. AND KARLAFTIS, M.G. (2003): A multivariate state space approach for urban traffic flow modeling and prediction. *Transportation Research Part C: Emerging Technologies*, **11**(2), 121–135.

SWEDBERG, R. (2010): The structure of confidence and the collapse of Lehman Brothers. *Research in the Sociology of Organizations*, **30**(A), 71–114.

THOMAS, J., MAYR, A., BISCHL, B., SCHMID, S., SMITH, A., AND HOFNER, B. (2018): Gradient boosting for distributional regression: faster tuning and improved variable selection via noncyclical updates. *Statistics and Computing*, **28**(3), 673–687.

TURNER, R. (2018): hmm.discnp: hidden Markov models with discrete non-parametric observation distributions. *R package*, version 2.1–12. https://cran.r-project.org/package=hmm.discnp.

UMLAUF, N., KLEIN, N., SIMON, T., AND ZEILEIS, A. (2019): bamlss: a Lego toolbox for flexible Bayesian regression (and beyond). *arXiv*, 1909.11784.

VILLARINI, G., SMITH, J.A., AND NAPOLITANO, F. (2010): Nonstationary modeling of a long record of rainfall and temperature over Rome. *Advances in Water Resources*, **33**(10), 1256–1267.

VISSER, I., RAIJMAKERS, M.E.J., AND MOLENAAR, P. (2002): Fitting hidden Markov models to psychological data. *Scientific Programming*, **10**(3), 185–199.

VITERBI, A.J. (1967): Error bounds for convolutional codes and an asymptotically optimal decoding algorithm. *IEEE Transactions on Information Theory*, **13**(2), 260–269.

VOUDOURIS, V., STASINOPOULOS, D.M., RIGBY, R.A., AND DI MAIO, C. (2011): The ACEGES laboratory for energy policy: exploring the production of crude oil. *Energy Policy*, **39**(9), 5480–5489.

WANG, P. AND PUTERMAN, M.L. (2001): Analysis of longitudinal data of epileptic seizure counts — a two-state hidden Markov model. *Biometrical Journal*, **43**(8), 941–962.

WEIß, C.H. (2018): *An introduction to discrete-valued time series*. Wiley and Sons, Chichester.

WELCH, L.R. (2003): Hidden Markov models and the Baum-Welch algorithm. *IEEE Information Theory Society Newsletter*, **53**(4), 10–13.

WHORISKEY, K., AUGER-MÉTHÉ, M., ALBERTSEN, C.M., WHORISKEY, F.G., BINDER, T.R., KRUEGER, C.C., AND MILLS FLEMMING, J. (2017): A hidden Markov movement model for rapidly identifying behavioral states from animal tracks. *Ecology and Evolution*, **7**(7), 2112–2121.

WILSON, A.D. AND BOBICK, A.F. (1999): Parametric hidden Markov models for gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **21**(9), 884–900.

YOON, B.J. (2009): Hidden Markov models and their applications in biological sequence analysis. *Current Genomics*, **10**(6), 402–415.

ZENG, Y. AND WU, S. (Editors) (2013): *State-space models: applications in economics and finance*. Springer, New York.

ZUCCHINI, W. AND GUTTORP, P. (1991): A hidden Markov model for space-time precipitation. *Water Resources Research*, **27**(8), 1917–1923.

ZUCCHINI, W., MACDONALD, I.L., AND LANGROCK, R. (2016): *Hidden Markov models for time series: an introduction using R*, 2nd Edition. Chapman and Hall/CRC, Boca Raton.

Short curriculum vitae

Academic education

- October 2013–May 2016: Studies of Business Administration and Economics (M.Sc.), Bielefeld University, Germany.
- August 2014–July 2015: Year abroad at the University of Copenhagen, Denmark (funded by an ERASMUS scholarship awarded by the European Union).
- October 2010–September 2013: Studies of Business Administration and Economics (B.Sc.), Bielefeld University, Germany.

Publications and preprints

- ADAM, T. AND OELSCHLÄGER, L. (2020): Hidden Markov models for multi-scale time series: an application to stock market data. Available on request (submitted to the *Proceedings of the 35th International Workshop on Statistical Modelling*).
- ADAM, T., GRIFFITHS, C.A., LEOS-BARAJAS, V., MEESE, E.N., LOWE, C.G., BLACKWELL, P.G., RIGHTON, D., AND LANGROCK, R. (2019): Joint modelling of multi-scale animal movement data using hierarchical hidden Markov models. *Methods in Ecology and Evolution*, **10**(9), 1536–1550.
- ADAM, T., LANGROCK, R., AND WEIß, C.H. (2019): Penalized estimation of flexible hidden Markov models for time series of counts. *METRON*, **77**(2), 87–104.
- ADAM, T. (2019): countHMM: penalized estimation of flexible hidden Markov models for time series of counts. *R package*, version 0.1.0. https://CRAN.R-project.org/package=countHMM.

- ADAM, T., LANGROCK, R., AND WEIß, C.H. (2019): Nonparametric inference in hidden Markov models for time series of counts. *Proceedings of the 34th International Workshop on Statistical Modelling*, **1**, 135–140.

- ADAM, T., LANGROCK, R., AND KNEIB, T. (2019): Model-based clustering of time series data: a flexible approach using nonparametric state-switching quantile regression models. *Book of Short Papers of the 12th Scientific Meeting on Classification and Data Analysis*, 8–11.

- LANGROCK, R., ADAM, T., LEOS-BARAJAS, V., MEWS, S., MILLER, D.L., AND PAPASTAMATIOU, Y.P. (2018): Spline-based nonparametric inference in general state-switching models. *Statistica Neerlandica*, **72**(3), 179–200.

- ADAM, T., MAYR, A., KNEIB, T., AND LANGROCK, R. (2018): Statistical boosting for Markov-switching distributional regression models. *Proceedings of the 33rd International Workshop on Statistical Modelling*, **1**, 30–35.

- LEOS-BARAJAS, V., GANGLOFF, E.J., ADAM, T., LANGROCK, R., VAN BEEST, F.M., NABE-NIELSEN, J., AND MORALES, J.M. (2017): Multi-scale modeling of animal movement and general behavior data using hidden Markov models with hierarchical structures. *Journal of Agricultural, Biological and Environmental Statistics*, **22**(3), 232–248.

- ADAM, T., LEOS-BARAJAS, V., LANGROCK, R., AND VAN BEEST, F.M. (2017): Using hierarchical hidden Markov models for joint inference at multiple temporal scales. *Proceedings of the 32nd International Workshop on Statistical Modelling*, **2**, 181–184.

- ADAM, T., MAYR, A., AND KNEIB, T. (2017): Gradient boosting in Markov-switching generalized additive models for location, scale, and shape. *arXiv*, 1710.02385 (submitted to *Econometrics and Statistics, Part B: Statistics*).