On some special-purpose hidden Markov models

Dissertation
presented for the degree of Doctor of Philosophy
at the Faculty of Economics and Business Administration
of the Georg-August-Universität Göttingen

by Roland Langrock (né Solecki)
from Hannover, Germany

Göttingen, 2011

First Examiner:	Prof. Dr. Walter Zucchini
Second Examiner:	Prof. Dr. Iain L. MacDonald
Third Examiner:	Prof. Dr. Tatyana Krivobokova
Thesis defence:	28.04.2011

To Andrea and Johann

Contents

Abbreviations & notation

The following list comprises the most important abbreviations and notation used throughout the thesis.

HMM	hidden Markov model
HSMM	hidden semi-Markov model
EM	expectation-maximization (algorithm)
ML	maximum likelihood (estimation)
SSM	state-space model
MCMC	Markov chain Monte Carlo
MCL	Monte Carlo likelihood
SV	stochastic volatility
VaR	value-at-risk
PoHMM	population hidden Markov model
AR	autoregressive (process)
t.p.m.	transition probability matrix
p.d.f.	probability density function
p.m.f.	probability mass function
c.d.f.	cumulative distribution function
i.i.d.	independently and identically distributed
mllk	minus log likelihood
AIC	Akaike information criterion
BIC	Bayesian information criterion
EEG	electroencephalogram
SDB	sleep disordered breathing
REM	rapid eye movement
SHHS	Sleep Heart Health Study
e.g.	*exempli gratia* ("for example")
cf.	*confer* ("compare")
i.e.	*id est* ("that is")

$\mathbb{P}(A)$	probability of event A
X_t	state-dependent process of HMM or HSMM
S_t	state process of HMM or HSMM
$\mathbb{P}^X$	distribution induced by random variable X
$\mathbb{P}^{X\|Y=y}$	conditional distribution induced by random variable X, given $Y = y$
$\mathbf{\Gamma}$	transition probability matrix of Markov chain
Γ	(with two arguments) gamma distribution
	(with one argument) gamma function
γ_{ij}	transition probability (from state i to state j)
$[0,1]^N$	N-fold cross product of the interval $[0,1]$
$\boldsymbol{\delta}$	initial or stationary distribution of Markov chain
$\mathbf{P}(\mathbf{x})$	diagonal matrix that appears in the HMM likelihood
$\mathbf{1}$	row vector of ones
$\boldsymbol{\alpha}_t$	forward probability
$X_t^\star$	state-dependent process of HMM that approximates HSMM
$S_t^\star$	state process of HMM that approximates HSMM
I_k	state aggregate
i_k^-	smallest element of state aggregate I_k
i_k^+	largest element of state aggregate I_k
y_t	state-dependent process of state-space model
g_t	state process of state-space model
Φ	cumulative distribution function of the standard normal distribution
g_{min}	lower bound of chosen range for g_t
g_{max}	upper bound of chosen range for g_t
Δ_4	unit 4-simplex
$\mathcal{D}(\lambda_1, \ldots, \lambda_N)$	Dirichlet distribution with parameters $\lambda_1, \ldots, \lambda_N$

List of Tables

List of Figures

Acknowledgements

I am sincerely grateful to my principal advisor, Prof. Walter Zucchini, for his constant support and encouragement, for countless fruitful discussions and suggestions, and in general for sharing his widespread expertise and enthusiasm for statistics. Prof. Walter Zucchini is everything one can wish for in a supervisor and I am indebted to him for his commitment.

I wish to extend my thanks to Prof. Rainer Dahlhaus of the University of Heidelberg, who encouraged me to write my PhD thesis on hidden Markov models in Göttingen. In addition, I am especially grateful to Assoc. Prof. Iain MacDonald of the University of Cape Town, for the fruitful collaboration, for many stimulating discussions, and for his alertness to mathematical inaccuracies. Furthermore, I would like to thank Assoc. Prof. Ciprian Crainiceanu and Assoc. Prof. Brian Caffo of Johns Hopkins University for giving me the opportunity to collaborate with them in an interesting field of application. At its very beginning this collaboration was encouraged by Prof. Tatyana Krivobokova, whom I thank not only for her support in that respect, but also for her willingness to act as examiner. I also wish to thank the members of the Institute for Statistics and Econometrics and of the Centre for Statistics for their support in various respects.

On a more personal note, my sincere thanks go to my parents for their unconditional support throughout my academic career. I also thank my son Johann for enriching my life and for being such a peaceful baby — as if he knew that daddy was about to finish his thesis. Lastly, I would like to thank my wonderful wife Andrea for her never-ending understanding. Her faith in me, as well as her continuous support and encouragement, is invaluable to me.

Introduction

"The purpose of models is not to fit the data but to sharpen the questions."

(Samuel Karlin, 1983)

Hidden Markov models (HMMs) provide flexible devices for modelling time series of observations that depend on underlying serially correlated states. They have been successfully applied to a wide range of types of time series: continuous-valued, circular, multivariate, as well as binary data, bounded and unbounded counts and categorical observations (see for example Zucchini and MacDonald 2009).

Originally, HMMs were developed in the field of speech recognition (Rabiner 1989); the underlying state sequence is then given by the spoken sequence of phonemes, while the observations are essentially given by the Fourier transforms of the recorded sounds. Speech recognition tries to decode the spoken sequence from the noisy observations — a typical HMM problem. Apart from speech recognition, HMMs have proved useful in many other application fields such as

- finance (Rydén *et al.* 1998, Banachewicz *et al.* 2008),
- economics (Hamilton 1989),
- biology (Durbin *et al.* 1998, Krogh *et al.* 1994),
- computer vision (Vogler and Metaxas, 1997) and
- environment (Zucchini and MacDonald 2009).

As Rabiner (1989) put it: "the models [HMMs] are very rich in mathematical structure and hence can form the theoretical basis for use in a wide range of applications". This thesis exploits the mathematical structure of HMMs to develop some "special-purpose HMMs", i.e. HMMs that differ from the standard setting and that are designed to address special demands. In particular, we investigate

1) HMMs with nonstandard sojourn times in the hidden states,

2) structured HMMs that are designed to approximate general-type state-space models and

3) HMMs for analysing populations of time series in an application to electroencephalogram data.

The motivation for working on the first topic can be described as follows. The hidden part of an HMM is a sequence of states. The time spent in a given state prior to a switch to a different state is geometrically distributed. This restrictive feature of conventional HMMs can be overcome by considering so-called hidden *semi*-Markov models (HSMMs). However, statistical inference then becomes more involved. In recent years, quite some literature has dealt with HSMMs (see e.g. Sansom and Thomson 2001, Guédon 2003, Yu and Kobayashi 2003 and Bulla 2006). In Chapter 2 we consider an HMM with a special structure that captures the 'semi'-property of HSMMs. The proposed model allows for arbitrary dwell-time distributions in the underlying states. It thus represents a tool for modelling HSMMs using standard HMM methods. For dwell-time distributions with finite support the HMM formulation is exact while for those that have infinite support, e.g. the Poisson, the distribution can be approximated with arbitrary accuracy. A benefit of using the HMM formulation is that it is easy to incorporate covariates, trend and seasonal variation in all components of the model. In addition, the formulae and methods for forecasting, state prediction, decoding and model checking that exist for ordinary HMMs are applicable to the proposed class of models. A seasonal HSMM is used to model daily rainfall occurrence for sites in Bulgaria. Additional applications include models for time series of daily returns and the inter-arrival times of geyser eruptions.

The second main part of the thesis explores ways of applying HMM methods in the context of state-space modelling. The dependence structure of HMMs is the same as that of state-space models (SSMs). The latter models are more general since they allow for infinite state spaces while conventional HMMs only have a finite number of states. SSMs have proved to be useful in numerous applications, the most prominent perhaps being stochastic volatility (SV) modelling. On the other hand, the price to pay for the increased flexibility is that parameter estimation, as well as state decoding, is very challenging for SSMs because their likelihood is given by a high-dimensional multiple integral that cannot be evaluated directly. The available methods are either simple but inefficient, or efficient but computationally demanding and rather difficult to implement. In Chapter 3 we make use of the close relationship between SSMs and

HMMs and investigate an alternative method that is based on the fact that SSMs can be approximated arbitrarily accurately by HMMs. The main benefit of this approach is that the HMM likelihood is easy to compute and to maximize. In addition, and in contrast to competing SSM methods, simple formulae are available for the forecast distributions, for computing appropriately defined residuals, and for decoding. An important benefit of the HMM representation is the ease of implementation, not only for fitting standard SSMs but also for fitting experimental extensions and variants of such models. To illustrate this advantage we first concentrate particularly on SV models. We define a number of nonstandard SV models and examine their performance when these are applied to various series of daily returns on stocks. In particular we assess the out-of-sample performance of the one-step-ahead forecasts of each model during the recent financial crisis. Besides the extensive discussion of the application to SV modelling, we use structured HMMs to model time series of earthquake counts, polio counts, rainfall occurrences and glacial varve thicknesses. The applications were selected in order to cover a wide range of different types of time series.

The third main topic addressed in the thesis evolved from a collaboration which aimed at applying HMMs to sleep electroencephalogram (EEG) data. In specific applications with populations of time series it can be difficult to compare fitted HMMs with each other. This is because the stochastic structure of HMMs is driven by two components: by the transition probabilities of the states and by the conditional observation distribution, given the states. As a consequence, if one faces a population of time series, and if one wishes to fit an individual HMM to each of these series, then fitted HMMs, even with the same design, are essentially incommensurable: the state transition probabilities cannot be compared directly as the states do not imply the same observation distributions, and the observation distributions can not directly be compared as the stochastic structures of the state sequences may be very different. In Chapter 4, we consider a related problem in a specific application of HMMs to sleep EEG data. We consider methods to analyse populations of sleep EEG signals for studying sleep disease using HMMs. An easily implemented method for combining HMM fits in a population is proposed. The method is applied to study sleep disordered breathing (SDB) in the Sleep Heart Health Study (SHHS), a landmark study of SDB and its cardiovascular consequences. We specifically use the entire, longitudinally collected, SHHS cohort to develop the state-dependent parameters of the HMM, which we then apply to obtain subject-specific Markovian predictions. From these predictions we create several indices of interest, such as transition frequencies between latent states.

The thesis is structured as follows. A brief introduction to standard HMM methodology is given in Chapter 1. Basic concepts are illustrated by means of an extensively discussed application to modelling geyser eruption inter-arrival times. Chapter 2 addresses the HMM representation of HSMMs based on Langrock and Zucchini (2011). The approximation of SSMs by structured HMMs is discussed in Chapter 3. Results in this case are aggregated from Langrock, MacDonald and Zucchini (2010) and Langrock (2010). The application of HMMs to sleep EEG data, as discussed in Langrock, Swihart, Caffo, Crainiceanu and Punjabi (2010), is described in Chapter 4. The final chapter summarizes the main results and discusses possible future research. Basically, the individual chapters can be read independently, although parts of Chapter 2 are built on results from Chapter 1, and parts of Chapter 3 analyse data that are introduced in Chapter 2. Most of the data sets appearing in the thesis are available for download[1].

[1] `www.statoek.wiso.uni-goettingen.de/cms/user/index.php?section=institut.team.rlangrock.data`

1 Hidden Markov models

This first chapter introduces the reader to hidden Markov models. First of all, the basic components of HMMs are defined (Section 1.1). This is followed by a review of parameter estimation in HMMs (Section 1.2). Subsequently, some of the most important distributions related to HMMs are given (Section 1.3). Finally, these basics are illustrated by means of two real time series related to the Old Faithful geyser (Sections 1.4.1 and 1.4.2).

1.1 Basics

Mathematically, HMMs comprise two components. The first, an unobserved (hidden) N-state Markov chain, is designed to account for serial correlation. The second is a state-dependent process whose outcomes (i.e. the observations) are assumed to be generated by one of N distributions as determined by the current state of the Markov chain. Each observation thus is modelled as outcome of a mixture distribution with N components, where the sequence of chosen components is a realization of a Markov chain. HMMs constitute a specific class of *dependent mixtures*. For an account of the theory of mixture distributions we refer to Frühwirth-Schnatter (2006).

Unless explicitly stated otherwise, the nonobservable Markov chain in the following is denoted by $\{S_t\}_{t=1,2,\ldots}$, and the observable state-dependent process by $\{X_t\}_{t=1,2,\ldots}$ (from now on, we mostly omit the subscript). Given the current state, the distribution of X_t is assumed to be conditionally independent of previous observations and states, i.e.

$$\mathbb{P}^{X_t|X_{t-1}=x_{t-1},X_{t-2}=x_{t-2},\ldots,S_t=s_t,S_{t-1}=s_{t-1},\ldots} = \mathbb{P}^{X_t|S_t=s_t}\,.$$

Usually, but not necessarily, the Markov chain is assumed to be of first order, i.e.

$$\mathbb{P}^{S_t|S_{t-1}=s_{t-1},S_{t-2}=s_{t-2},\ldots} = \mathbb{P}^{S_t|S_{t-1}=s_{t-1}}\,. \tag{1.1}$$

Figure 1.1 displays the dependence structure of a basic HMM in a directed acyclic graph.

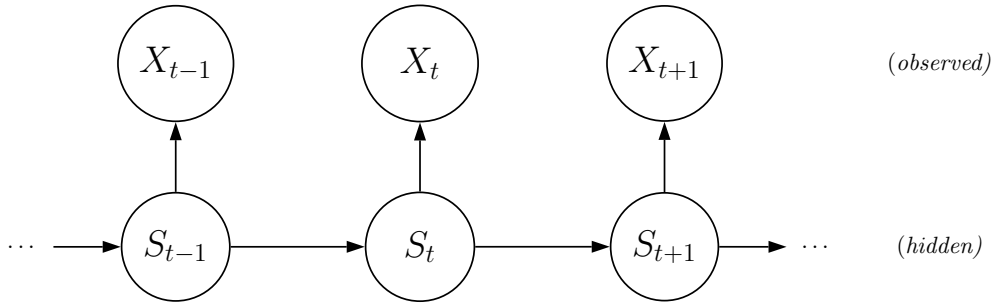

Figure 1.1: *Dependence structure of an HMM.*

At time t, we summarize the probabilities of transitions between the N states of S_t in the $N \times N$ *transition probability matrix* (t.p.m.) $\mathbf{\Gamma}^{(t)} = \{\gamma_{ij}^{(t)}\}$, where

$$\gamma_{ij}^{(t)} = \mathbb{P}\big(S_{t+1} = j \,|\, S_t = i\big), \quad i, j = 1, \ldots, N.$$

The Markov chain is said to be *homogeneous* if the t.p.m. does not depend on the time index t. In that case it is said to have a *stationary distribution* if there exists a row vector $\boldsymbol{\delta} \in [0,1]^N$ that fulfils

$$\boldsymbol{\delta}\mathbf{\Gamma} = \boldsymbol{\delta} \quad \text{subject to} \quad \sum_{i=1}^{N} \delta_i = 1\,.$$

If the Markov chain is in its stationary distribution at time t, i.e. if the (unconditional) distribution of the states at time t is given by $\boldsymbol{\delta}$, then for all subsequent time instants the states will have the same (unconditional) distribution. To see this, consider the (unconditional) probability of state i at time t, $\delta_i^{(t)} := \mathbb{P}(S_t = i)$. Then

$$\delta_j^{(t+1)} = \mathbb{P}(S_{t+1} = j) = \sum_{i=1}^{N} \mathbb{P}(S_{t+1} = j \,|\, S_t = i)\mathbb{P}(S_t = i) = \sum_{i=1}^{N} \gamma_{ij}\delta_i^{(t)}\,,$$

and thus

$$\boldsymbol{\delta}^{(t+1)} = \boldsymbol{\delta}^{(t)}\mathbf{\Gamma}\,,$$

where $\boldsymbol{\delta}^{(t)} := \big(\delta_1^{(t)} \;\ldots\; \delta_N^{(t)}\big)$. The statement now follows by induction. If the Markov chain starts from its stationary distribution, i.e. if $\boldsymbol{\delta}^{(1)} = \boldsymbol{\delta}$, then it is said to be *stationary*. A detailed account of the theory of Markov chains can be found in Brémaud (1999).

In an HMM the Markov chain represents the nonobservable state process that determines the distribution at the observation level, i.e. that of the state-dependent process. In the one-dimensional case we denote the probability mass function (in the discrete

case), or probability density function (in the continuous case), of the state-dependent process, given the underlying Markov chain is in state k, $k \in \{1, 2, \ldots, N\}$, by

$$f_k(x) := \begin{cases} \mathbb{P}\big(X_t = x | S_t = k\big) & \text{(discrete case)} \\ f_{X_t|S_t=k}(x) & \text{(continuous case)} \end{cases} .$$

(Here $f_{X_t|S_t=k}$ denotes the conditional density of X_t, given $S_t = k$.) If the underlying Markov chain is stationary, then the marginal univariate distribution of an HMM — $\mathbb{P}(X_t = x)$ in the discrete case — at each time t is given by

$$\sum_{i=1}^{N} \delta_i f_i(x) \tag{1.2}$$

(this follows by applying the theorem of total probability). Expression (1.2) verifies that the observable part of an HMM, X_t, is a finite mixture. In general the mixture is *dependent* due to the influence of the Markov chain. (Exceptions for instance are HMMs with stationary Markov chain whose components of $\mathbf{\Gamma}$ are column-wise identical.)

For comprehensive accounts of the theory of HMMs see e.g. Ephraim and Merhav (2002), Cappé *et al.* (2005) or Zucchini and MacDonald (2009).

1.2 Parameter estimation via numerical likelihood maximization

The likelihood function of an HMM is available in a form that is easy to compute, the parameters thus can be estimated by direct numerical likelihood maximization. Cappé *et al.* (2005, Chapter 12) show that, under certain regularity conditions, the MLEs of HMM parameters are consistent, asymptotically normal and efficient.

The likelihood can be written as the following product (see e.g. Zucchini and MacDonald 2009):

$$L_T = \boldsymbol{\delta}^{(1)} \mathbf{P}(x_1) \mathbf{\Gamma} \mathbf{P}(x_2) \mathbf{\Gamma} \cdot \ldots \cdot \mathbf{\Gamma} \mathbf{P}(x_{T-1}) \mathbf{\Gamma} \mathbf{P}(x_T) \mathbf{1}^t ,$$

where

$$\mathbf{P}(x) = \operatorname{diag}\big(f_1(x), \ldots, f_N(x)\big) ,$$

$\mathbf{1}$ is a row vector of ones and $x_1, \ldots, x_T$ denote the observations. The above expression applies to the homogeneous case. In some applications either $\mathbf{\Gamma}$ or $\mathbf{P}(x)$ can depend on the time index t. The computational effort required to evaluate the likelihood is linear in the number of observations, T, and quadratic in the number of HMM states, N.

In order to numerically maximize the likelihood with respect to the parameters one needs to take care of some technical problems. Firstly, as the likelihood involves multiple products of probabilities, numerical underflow can occur (especially for long series of observations). Appropriate scaling of the likelihood circumvents underflow in many cases. For more details on scaling see Chapter 3 of Zucchini and MacDonald (2009).

Secondly, if an unconstrained maximization or minimization algorithm, e.g. `nlm()` in **R** (Ihaka and Gentleman 1996), is used, then it is necessary to reparameterize the model in terms of unconstrained parameters. If, say, a Poisson state-dependent distribution is to be fitted, then the parameter λ_k has to be positive. We would then numerically maximize the likelihood with respect to the unconstrained parameter, $\eta_k = \log \lambda_k$, and afterwards obtain the estimate of the constrained parameter by $\lambda_k = \exp \eta_k$. For more details on how to deal with parameter constraints see Chapter 3 of Zucchini and MacDonald (2009).

Thirdly, in the case of continuous observations the likelihood can be unbounded, which renders its maximization impossible. This problem does not arise in the case of discrete observations because probabilities are bounded by one (while densities in general are unbounded). One natural way to circumvent this difficulty is to treat the observations as interval-censored (as opposed to continuous); since each recording of continuous phenomena involves a certain amount of rounding, it is more accurate to replace the precise values (e.g. 14.6 seconds) by suitable intervals (e.g. $(14.55, 14.65]$). Replacing the density values by probabilities of the corresponding intervals leads to a bounded (discrete) likelihood (cf. Zucchini and MacDonald 2009).

Finally, one should be aware that the maximization algorithm might converge to a local rather than the global maximum. The main strategy to deal with this problem is to use a range of suitable starting values.

It is also possible to apply the expectation-maximization (EM) algorithm to estimate the HMM parameters. As EM is not employed in the course of this work, it is not further discussed here. For a comprehensive account of the EM algorithm, including a discussion of advantages and disadvantages of the two possible parameter estimation methods, see Bulla and Berzel (2006).

1.3 Forecasting, decoding and state prediction in HMMs

When applying HMMs, one is often interested in particular conditional or joint distributions, e.g. for

- forecasting future observations,

- decoding the most likely state sequence for a given sequence of observations,
- predicting future states.

For each of these purposes convenient expressions and algorithms are available (cf. Zucchini and MacDonald 2009). To begin with, we recall the expressions for forecasting and state prediction (for details on the derivations we refer to the manuscript just cited). All expressions are given for the case of discrete observations. (The continuous case can be treated analogously; the probabilities then have to be replaced by densities.)

The so-called *forward probabilities* are given by

$$\boldsymbol{\alpha}_t = \boldsymbol{\delta}^{(1)}\mathbf{P}(x_1)\boldsymbol{\Gamma}\mathbf{P}(x_2)\boldsymbol{\Gamma}\cdot\ldots\cdot\boldsymbol{\Gamma}\mathbf{P}(x_{t-1})\boldsymbol{\Gamma}\mathbf{P}(x_t), \qquad t = 1,\ldots,T.$$

(Note that $\boldsymbol{\alpha}_t$ is a row vector.) The forward probabilities can be employed to compute the h-step ahead forecast distribution of an HMM, conditional on all observations up to time T, as follows:

$$\mathbb{P}(X_{T+h} = x | X_T = x_T, X_{T-1} = x_{T-1}, \ldots, X_1 = x_1) = \frac{\boldsymbol{\alpha}_T\boldsymbol{\Gamma}^h\mathbf{P}(x)\mathbf{1}^t}{\boldsymbol{\alpha}_T\mathbf{1}^t}\,. \tag{1.3}$$

(As above $\mathbf{1}$ is a row vector of ones.) Similarly, one obtains the following expression for the conditional probability of a future state, given all observations up to time T:

$$\mathbb{P}(S_{T+h} = k | X_T = x_T, X_{T-1} = x_{T-1}, \ldots, X_1 = x_1) = \frac{\boldsymbol{\alpha}_T\boldsymbol{\Gamma}^h e_k^t}{\boldsymbol{\alpha}_T\mathbf{1}^t}\,, \quad k = 1,\ldots,N, \tag{1.4}$$

with $e_k := (0,\ldots,0,1,0,\ldots,0)$ denoting the unit row vector with kth entry equal to 1. Given homogeneity, aperiodicity and irreducibility (see Brémaud 1999 for definitions) of the Markov chain, both the forecast distribution of the observations, and that of the states, converge to the respective stationary distribution as the forecast horizon h increases, i.e.

$$\begin{aligned} &\lim_{h\to\infty} \mathbb{P}(X_{T+h} = x | X_T = x_T, X_{T-1} = x_{T-1}, \ldots, X_1 = x_1) = \sum_{i=1}^{N} \delta_i f_i(x) \\ \text{and} \quad &\lim_{h\to\infty} \mathbb{P}(S_{T+h} = k | X_T = x_T, X_{T-1} = x_{T-1}, \ldots, X_1 = x_1) = \delta_k\,. \end{aligned} \tag{1.5}$$

One distinguishes between *local* and *global* decoding. Local decoding yields the most likely states at individual time instants, while global decoding yields the most likely state sequence. The latter is usually carried out using the Viterbi algorithm (Viterbi 1967).

1.4 Example: Old Faithful eruptions

Due to its intriguing dynamics, the Old Faithful geyser has attracted much attention in the geophysical and statistical literature; papers or monographs that deal with the Old Faithful geyser include, *inter alia*, Azzalini and Bowman (1990), Weisberg (2005), Silverman (1986), Robert and Titterington (1998), Aston and Martin (2007), Varin and Vidoni (2005) and Zucchini and MacDonald (2009). While focusing on different aspects of the eruption dynamics, these contributions have in common that they all investigate rather short time series related to Old Faithful. The most frequently investigated data set consists of 299 pairs of measurements, namely the eruption durations and the time intervals between consecutive eruptions, dating back to August 1985 (cf. Azzalini and Bowman).

In this section we consider a substantially larger data set that includes almost all eruptions of Old Faithful in 2009. (More precisely, we consider all eruptions after the 2nd of January, a day on which there is a gap in the series due to an ice formation). The data were downloaded from `www.geyserstudy.org`. (Acknowledgements are due to Yellowstone National Park and Ralph Taylor for providing the data.) This data set contains 5768 observations and thus gives a more comprehensive insight in the dynamics of the geyser than do the several well-documented series from the 1980s. The series comprises the time intervals between starts of consecutive eruptions (rounded to the nearest minute). Azzalini and Bowman (1990) argue that consideration of the inter-arrival times while ignoring the eruption duration times does not neglect any important information, as these two measurements are equivalent indicators for the state of Old Faithful. The minimum inter-arrival time in 2009 was 46 minutes, the maximum was 124 minutes.

Figure 1.2 displays the time series of eruption inter-arrival times together with a local polynomial smoother. There is some indication of nonhomogeneity in the data. Indeed, the geyser has been shown to be influenced by covariates, such as precipitation or seismic activity (cf. Hurwitz *et al.* 2008). However, the seasonal variation is rather small, and as this chapter mainly has illustrative purposes we limit ourselves to the consideration of homogeneous models.

Figure 1.3 displays a histogram of all inter-arrival times observed in 2009 (after the 2nd of January). The estimated distribution is bimodal and it is thus plausible that the geyser operates in at least two different modes. In fact, the Yellowstone National Park predicts eruption inter-arrival times one step ahead, and (as of 2010) the prediction is completely determined by whether the current eruption duration is classified as “short” or “long” (source: `www.geyserstudy.org`). Furthermore, Figure 1.4 shows

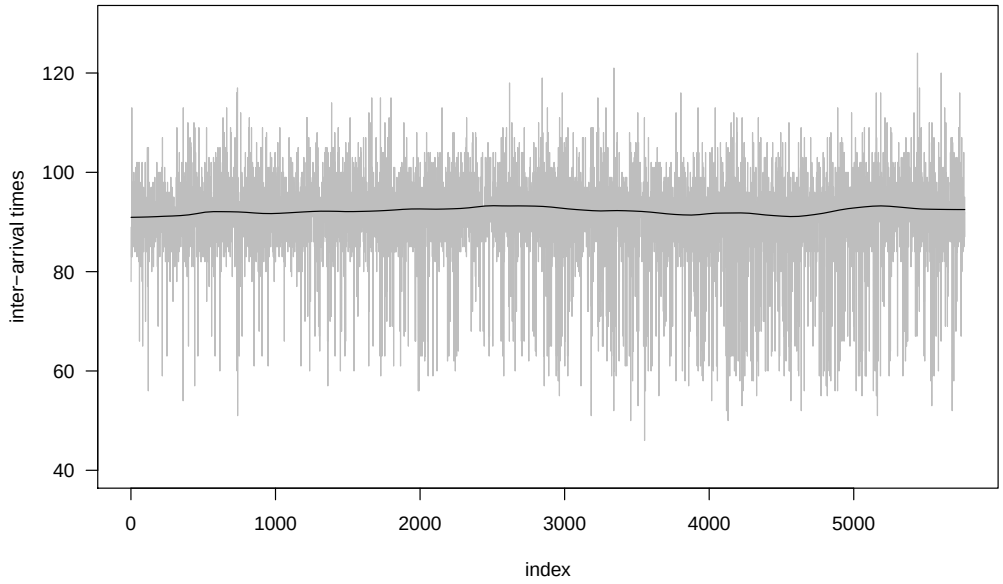

Figure 1.2: *Observed inter-arrival times in 2009 and local polynomial smoother.*

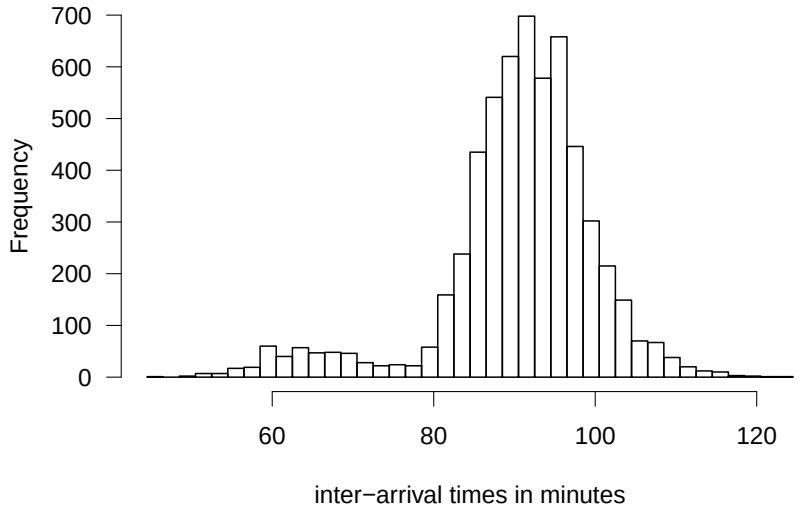

Figure 1.3: *Histogram of eruption inter-arrival times.*

that the consecutive inter-arrival times are serially correlated. Taking into account both the bimodality and the serial dependence, HMMs appear to be reasonable choices for

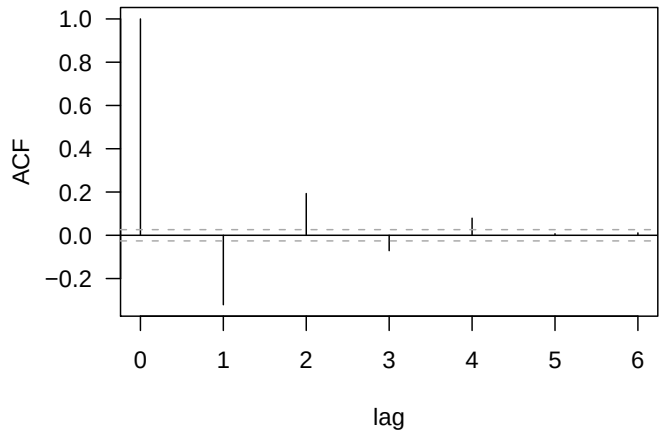

Figure 1.4: *Sample autocorrelation function for the time series of inter-arrival times.*

modelling this time series.

We analyse the time series of eruption inter-arrival times in two ways: in Section 1.4.1 we follow Azzalini and Bowman in dichotomizing the series into short and long inter-arrival times respectively. We model the resulting binary series by means of Markov chains of different orders and by means of HMMs. Subsequently, in Section 1.4.2, we analyse the original time series by means of differently designed HMMs. (In Section 2.5 hidden *semi*-Markov models are fitted to the series.)

1.4.1 Modelling the binary series of long and short inter-arrival times

Azzalini and Bowman (1990) identify two distinct states of Old Faithful: one involves short inter-arrival times (with long subsequent eruption times), and the other long inter-arrival times (followed by short and long eruption times in roughly equal proportions). They argue that a discretization of the observations into either short or long inter-arrival times is reasonable since the most important feature of the geyser is the alternation between these two states. For the moment we follow this suggestion and discretize the series of inter-arrival times accordingly: to inter-arrival times less than 75 minutes we assign the value 0, to those longer than or equal to 75 minutes we assign the value 1. The resulting series contains 401 zeros (short inter-arrival times) and 5367 ones (long inter-arrival times). The series starts and ends with a long inter-arrival time, and short inter-arrival times are always followed by long ones, meaning that there are no consecutive zeros. This feature of Old Faithful was already reported by Azzalini and

Bowman (1990). There are, however, proportionally far fewer short inter-arrival times in the series from 2009 than in that from August 1985.

Markov chains

We begin by considering a first-order Markov chain with two states (labelled 0 and 1 for short and long inter-arrival times, respectively). The conditional maximum likelihood estimate of the t.p.m. (cf. Zucchini and MacDonald 2009), conditioned on the first observation, is given by

$$\widehat{\mathbf{\Gamma}} = \begin{pmatrix} 0 & 1 \\ \frac{401}{5366} & \frac{4965}{5366} \end{pmatrix} \approx \begin{pmatrix} 0 & 1 \\ 0.07 & 0.93 \end{pmatrix}.$$

For long series the unconditional likelihood estimate differs only marginally from the conditional likelihood estimate; we restrict ourselves to the latter.

Azzalini and Bowman (1990) demonstrate that a first-order Markov chain is inadequate because it does not fully capture the observed autocorrelation of the series. They found that a second-order Markov chain provides a better fit. A second-order Markov chain with two states can equivalently be described by a first-order Markov chain with four states. More precisely, if the Markov chain $\{S_t\}$ is of second order, then the Markov chain $\{T_t\} := \{(S_{t-1}, S_t)\}$ is of first order. In this application the states of $\{T_t\}$ are $(0,0)$, $(0,1)$, $(1,0)$ and $(1,1)$. As no two short inter-arrival times occur consecutively, the state $(0,0)$ does not occur. Consequently, a first-order Markov chain with states $(0,1)$, $(1,0)$ and $(1,1)$, in order, can be used to express the second-order Markov chain. By counting the transitions between these three states, the conditional maximum likelihood estimate of the t.p.m., conditioned on the first two observations, is obtained as

$$\widehat{\mathbf{\Gamma}} = \begin{pmatrix} 0 & \frac{83}{401} & \frac{318}{401} \\ 1 & 0 & 0 \\ 0 & \frac{318}{4964} & \frac{4646}{4964} \end{pmatrix} \approx \begin{pmatrix} 0 & 0.21 & 0.79 \\ 1 & 0 & 0 \\ 0 & 0.06 & 0.94 \end{pmatrix}.$$

Note that some entries necessarily equal zero; the transitions $(0,1) \to (0,1)$, $(1,0) \to (1,0)$, $(1,0) \to (1,1)$ and $(1,1) \to (0,1)$ are impossible. According to the fitted model the state pair $(0,1)$ is more likely to be followed by a zero than is the pair $(1,1)$. While short inter-arrival times never occur consecutively, it happens relatively often that only one long inter-arrival time occurs between two short ones. In this respect increasing the memory of the Markov chain improves the fit, since a first-order Markov chain can not reflect this dependence feature.

Bernoulli hidden Markov models

As a next step we fitted stationary HMMs with Bernoulli state-dependent distributions (*Bernoulli HMMs*) to the binary Old Faithful series. In these HMMs the states of the Markov chain S_t are no longer taken to be observations. Instead the value of S_t now represents the unobserved state of the HMM. The observation is assumed to be a realization of a Bernoulli distributed random variable X_t whose parameter depends on the state: if $S_t = i$, with i denoting one out of N possible states of the Markov chain, then the probability of a long inter-arrival time is π_i, i.e. $X_t \sim$ Bern (π_i). The following models are considered:

- a two-state Bernoulli HMM with underlying Markov chain of first order,
- a three-state Bernoulli HMM with underlying Markov chain of first order and
- a two-state Bernoulli HMM with underlying Markov chain of second order.

In the last case the applied model fitting strategy is analogous to the case of the second-order Markov chain described above. Table 1.1 summarizes the results of the model fitting exercise, including both the fitted Markov chains and the fitted HMMs.

Table 1.1: *Minus log likelihood, AIC and BIC for the models fitted to the dichotomized Old Faithful series.*

	mllk	*AIC*	*BIC*
Markov chain of order 1	1425.85	2855.69	2869.01
Markov chain of order 2	1386.06	2780.12	**2806.76**
2-state HMM of order 1	1400.29	2808.58	2835.22
3-state HMM of order 1	1380.72	2779.44	2839.38
2-state HMM of order 2	1382.44	**2776.88**	2816.84

The model suggested by Azzalini and Bowman (1990), namely the second-order Markov chain, outperforms all other models in terms of the BIC. In terms of the AIC, the two-state HMM of second order performs best. The three-state HMM attains the highest likelihood value, while the remaining two models, namely the first-order Markov chain and the two-state HMM of first order, are apparently less suitable.

The estimated parameters of the two-state HMM of second order are

$$\widehat{\boldsymbol{\pi}} = (\widehat{\pi}_1, \widehat{\pi}_2) = (0.45, 1.00)$$

and

$$\widehat{\mathbf{\Gamma}} = \begin{pmatrix} 0.00 & 0.38 & 0.62 \\ 1.00 & 0.00 & 0.00 \\ 0.00 & 0.10 & 0.90 \end{pmatrix},$$

where the three states are, in order, $(1,2)$, $(2,1)$ and $(2,2)$. (According to the fitted model state $(1,1)$ almost surely does not occur.) The main features attributed to the second-order Markov chain reappear in this HMM.

The parameter estimates for the three-state HMM are

$$\widehat{\boldsymbol{\pi}} = (\widehat{\pi}_1, \widehat{\pi}_2, \widehat{\pi}_3) = (0.62, 0.99, 1.00) \tag{1.6}$$

and

$$\widehat{\mathbf{\Gamma}} = \begin{pmatrix} 0.00 & 0.00 & 1.00 \\ 0.00 & 0.89 & 0.11 \\ 0.56 & 0.21 & 0.23 \end{pmatrix}. \tag{1.7}$$

It is worth noting that states 2 and 3 imply very similar properties at the observation level. Intuitively the given state partition suggests that two states suffice. Bearing this in mind, the increase in the likelihood compared to the two-state HMM is remarkable. In Section 2.5.1 we revisit this puzzle.

1.4.2 Modelling the series of inter-arrival times

As described in the previous section, dichotomizing the inter-arrival times series seems plausible in regard of the most striking features of Old Faithful. Nevertheless, it does involve a loss of information. We now analyse the original (nondichotomized) series to see whether the geyser operates in two or in more states.

Independent mixtures

As a first step, independent mixture distributions were fitted to the time series. Even though such models neglect the serial dependence of the data, they help to decide which type of distributions might be considered suitable when fitting HMMs. The p.d.f. of a mixture distribution with N components is given by

$$f(x) = \alpha_1 f_1(x) + \alpha_2 f_2(x) + \ldots + \alpha_N f_N(x)\,,$$

where f_i denotes the p.d.f. of the ith component distribution and $\alpha_1 + \ldots + \alpha_N = 1$. Table 1.2 compares mixtures of normal and of gamma distributions, respectively, in terms of the model selection criteria for the given series of inter-arrival times.

Table 1.2: *Minus log likelihood, AIC and BIC for independent mixtures of normal and of gamma distributions with $N = 2, 3, 4$ components.*

normal mixtures				*gamma mixtures*			
N	*mllk*	*AIC*	*BIC*	*N*	*mllk*	*AIC*	*BIC*
2	20358.3	40726.6	40759.9	*2*	20323.9	40657.7	40691.0
3	20291.4	40598.8	40652.1	*3*	20289.5	40594.9	**40648.2**
4	20283.9	40589.7	40663.0	*4*	20283.2	**40588.3**	40661.6

Except in the case with only $N = 2$ components, the normal and gamma mixture distributions led to similar results. The gamma distribution performed slightly better for each number of components and is a more natural choice anyway: the normal distribution is defined on the whole real line while the observations are, by nature, all positive.

First-order gamma hidden Markov models

The series of eruption inter-arrival times shows significant autocorrelation (cf. Figure 1.4). Thus HMMs can be expected to be more appropriate than independent mixtures. As a first step we fitted stationary *gamma HMMs* (GHMMs) with different numbers of states N. An N-state GHMM is a dependent mixture of N gamma distributions, where the components of the mixture are selected by an N-state Markov chain. For state k, $k = 1, 2, \ldots, N$, the gamma distribution is characterized by the *shape parameter* κ_k and the *scale parameter* θ_k, where $\kappa_k, \theta_k > 0$. The mean of the kth component distribution is $\mu_k = \kappa_k \cdot \theta_k$. The conditional p.d.f. of an observation x, given state k, is

$$f_k(x) = f_{\kappa_k,\theta_k}(x) = x^{\kappa_k - 1} \frac{\exp(\frac{-x}{\theta_k})}{\theta_k^{\kappa_k} \Gamma(\kappa_k)}, \quad x \geq 0. \tag{1.8}$$

The parameters were estimated by numerical maximization of the discrete likelihood (e.g. the density of an observation $x = 78$ was replaced by the probability of the interval $(77.5, 78.5]$; cf. Section 1.2). The parameter estimates for the GHMMs with $N = 2, 3, 4$ and 5 components are given in Appendix A1. Table 1.3 summarizes the model fitting results. In terms of the model selection criteria AIC and BIC, only the models with four and five states, respectively, appear to be competitive. (Noticeably, Robert and Titterington, 1998, as well as Zucchini and MacDonald, 2009, concentrate on three-state HMMs when investigating the shorter Old Faithful series from 1985.)

Figure 1.5 displays the mixture component distributions and the marginal observation distributions of the fitted models with three, four and five states respectively. The

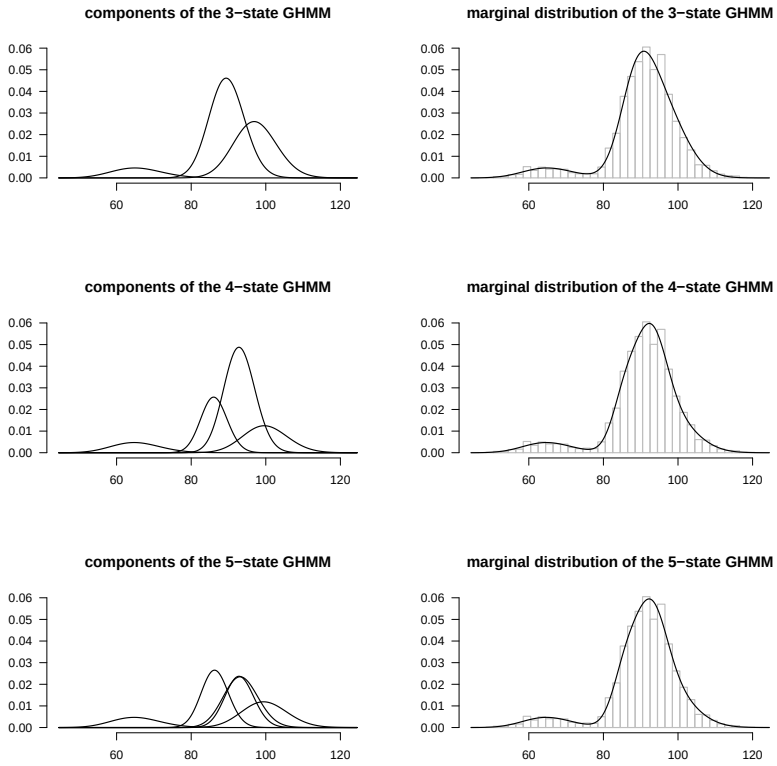

Figure 1.5: *Weighted mixture components (left hand side), and marginal distributions (right hand side, together with histograms of the observations), of the fitted GHMMs with* $N = 3, 4, 5$.

Table 1.3: *Minus log likelihood, AIC and BIC for* N*-state gamma HMMs fitted to the nondichotomized Old Faithful series.*

N	*mllk*	*AIC*	*BIC*
2	20292.19	40596.38	40636.34
3	19948.57	39921.13	40001.05
4	19837.98	39715.96	**39849.16**
5	19822.86	**39705.73**	39905.53

fitted mixture components on the left have been weighted with their proportion in the mixture according to the stationary distribution of the Markov chain (i.e. the fitted component $f_k(x)$ has been multiplied by δ_k). On the right hand side of Figure 1.5, the marginal distribution of the observations is obtained as the sum of these weighted components according to (1.2). Evidently, in the five-state GHMM, two of the states

are hardly distinguishable at the observation level. In view of the similarity of these states, the likelihood gap between the four- and the five-state model can be expected to be due either to *(i)* distinctive features of the state dwell-time distributions or to *(ii)* correlation features that can not be captured by the four-state GHMM. The subsequent paragraph investigates *(ii)*, while *(i)* is discussed in Section 2.5.2.

Four-state second-order gamma hidden Markov model

So far, the most promising stochastic model for Old Faithful's eruption inter-arrival times series is the four-state GHMM. We now investigate whether increasing the memory of the underlying Markov chain can further improve the fit. Thus we now consider a four-state GHMM with underlying second-order Markov chain. The value of the minimized minus log likelihood for this model is 19800.69, the AIC and BIC values are 39713.38 and 40086.34, respectively. Based on model selection criteria other models are superior (cf. Table 1.3). However, the likelihood value is, by a substantial gap, the largest of all models that have been considered (in particular it is larger than that of a five-state GHMM). This gives motivation for further investigating the model.

The estimated parameters for the four-state second-order GHMM are given in Appendix A1. Some features of the estimated t.p.m. and the corresponding stationary distribution are noteworthy; e.g.

- states 1 and 2, respectively, do not occur twice in a row (almost surely),
- state 1 is never followed by state 2 (a.s.),
- transitions from state 4 to state 1, as well as transitions from state 3 to state 1, occur more than three times as often as do transitions from state 2 to state 1,
- state 3 is the only state that can occur three times or more in a row (a.s.),
- state 4, which involves the longest inter-arrival times, can occur twice in a row, but only following either state 1 or 2 (a.s.),
- if there is a switch from state 1 to state 3, then neither state 1 nor state 2 will occur next (a.s.),
- the likelihood of a switch from state 4 to state 1 or vice versa is relatively high if the opposite switch has just taken place.

Note that the last four properties listed above can not be captured by an HMM with first-order Markov chain. In particular the last-mentioned property is interesting: the

stationary probability of state 1 is 0.076, short inter-arrival times thus occur relatively seldom. However, given that a short inter-arrival time (state 1) occurs at time instant t, the probability that there will be another short one at time $t+2$ is 0.230, a probability that is substantially larger than the stationary probability. The geyser apparently operates in such a way that extreme observations, i.e. very short or very long inter-arrival times, are followed by extreme observations of the opposite type with a relatively high probability. This confirms the findings from 1.4.1 and shows that, to some extent, short inter-arrival times tend to appear in (small) clusters.

Both the substantial increase in the likelihood and the new insights offered by the second-order property constitute reasons to regard the four-state second-order GHMM as a suitable model, even though the model selection criteria considered select other models. Thus we use this model to illustrate forecasting, decoding and state prediction for HMMs.

Using the four-state second-order GHMM, the first six inter-arrival times at the beginning of the year 2010 are forecast *one-step-ahead* using (1.3). Figure 1.6 displays the forecast probabilities for the inter-arrival times to fall into intervals of one minute length (e.g. $(93.5, 94.5]$).

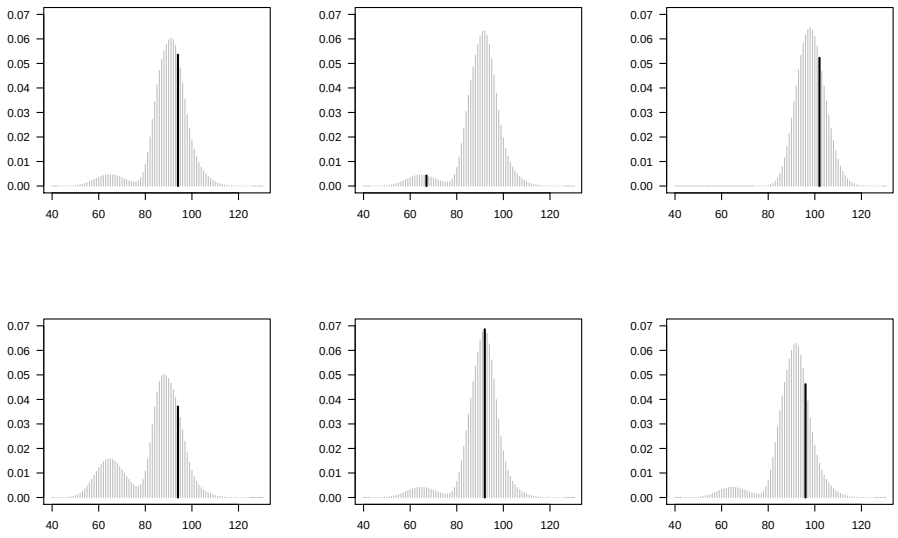

Figure 1.6: *One-step-ahead forecast distributions for the first six eruption inter-arrival times in 2010 (to be read row-wise), and actually observed inter-arrival time (bold bar).*

A part of a Viterbi path for the four-state second-order GHMM is depicted in Figure

1.7. It gives the final 50 eruption inter-arrival times from 2009 and the associated (globally) decoded sequence of states.

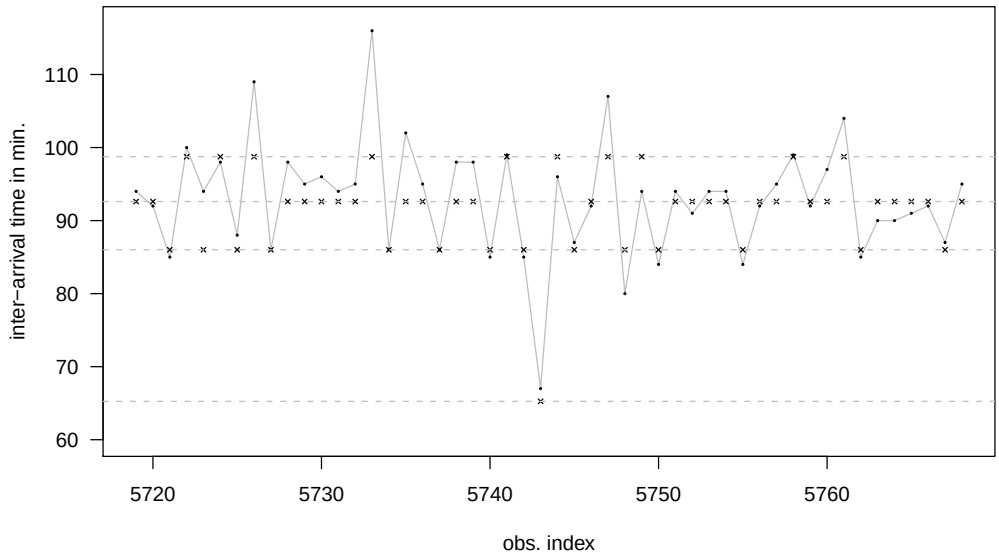

Figure 1.7: *State-dependent means (dashed horizontal lines), means according to the globally decoded state sequence (crosses) and actual observations (filled circles, connected by grey line).*

As a next step, Expression (1.4) was employed to perform state prediction. Table 1.4 gives the probabilities of the possible state pairs associated with the first six eruption inter-arrival times in 2010. All observations from 2009, $x_1, \ldots, x_{5768}$, were used to obtain the predictions. In contrast to the forecasts in Figure 1.6, the state predictions are not one-step-ahead (except of the first one). The state pairs $(1,1)$, $(1,2)$ and $(2,2)$ almost surely do not occur. Evidently, as the forecast horizon h increases, the probabilities converge quite fast towards the stationary probabilities (cf. Equation (1.5)). Note that it is straightforward to derive the probabilities of the single states from that of the state pairs: e.g. the probability of state 3 is the sum of the probabilities of the pairs $(1,3)$, $(2,3)$, $(3,3)$ and $(4,3)$.

Lastly, Figure 1.8 compares the original time series with a series of same length that was simulated from the fitted four-state second-order GHMM. The dashed lines in the top graphs give the empirical 0.05-, 0.25-, 0.75- and 0.95-quantiles respectively. In addition, histograms of the observations and the sample autocorrelation functions are given. The fitted model seems to be able to reproduce the most striking features of Old Faithful.

Table 1.4: *Predicted state probabilities, multiplied by* 100*, of* $\{T_t\} := \{(S_{t-1}, S_t)\}$ *for the first six observations in 2010; the bottom line gives the stationary distribution.*

$(1,3)$	$(1,4)$	$(2,1)$	$(2,3)$	$(2,4)$	$(3,1)$	$(3,2)$	$(3,3)$	$(3,4)$	$(4,1)$	$(4,2)$	$(4,3)$	$(4,4)$
0	0	0	1	0	5	11	41	14	3	19	6	1
0	7	1	16	12	3	11	25	8	2	6	7	0
1	5	1	9	8	3	9	28	9	4	15	8	1
1	7	1	13	10	3	9	26	8	3	12	7	1
1	6	1	11	9	3	9	26	8	4	13	8	1
1	7	1	12	9	3	9	26	8	3	13	7	1
1	7	1	12	9	3	9	26	8	3	13	7	1

1.5 Concluding remarks

In this first chapter we covered some of the basic concepts needed in the chapters to follow. We introduced HMMs and their ingredients, in particular Markov chains. Well-established methodology for HMMs, including parameter estimation and forecasting, was reviewed briefly. Although we have illustrated how model selection criteria can be applied to select the 'best' model, we have not yet discussed how to check the adequacy of a selected HMM, and how to identify outliers relative to a fitted model. This can be done by the use of residuals; Section 3.2.3.2 contains more details.

The practical implementation of many of the basic HMM concepts was demonstrated by means of an extensively discussed application to two time series related to the Old Faithful geyser. The investigation of the series of eruption inter-arrival times also illustrated the considerable flexibility of HMMs. In this application the first decision to be made is whether or not to dichotomize the series. If so, then Bernoulli HMMs are plausible models. If, instead, the nondichotomized series is to be modelled, then gamma HMMs provide reasonable results. One also has to choose the number of states of the hidden process. In case of the nondichotomized Old Faithful series four distinct states were identified. Lastly, we have shown that standard HMMs with memory of one time lag can not capture all features of the series; the consideration of an HMM with underlying second-order Markov chain offered further insights.

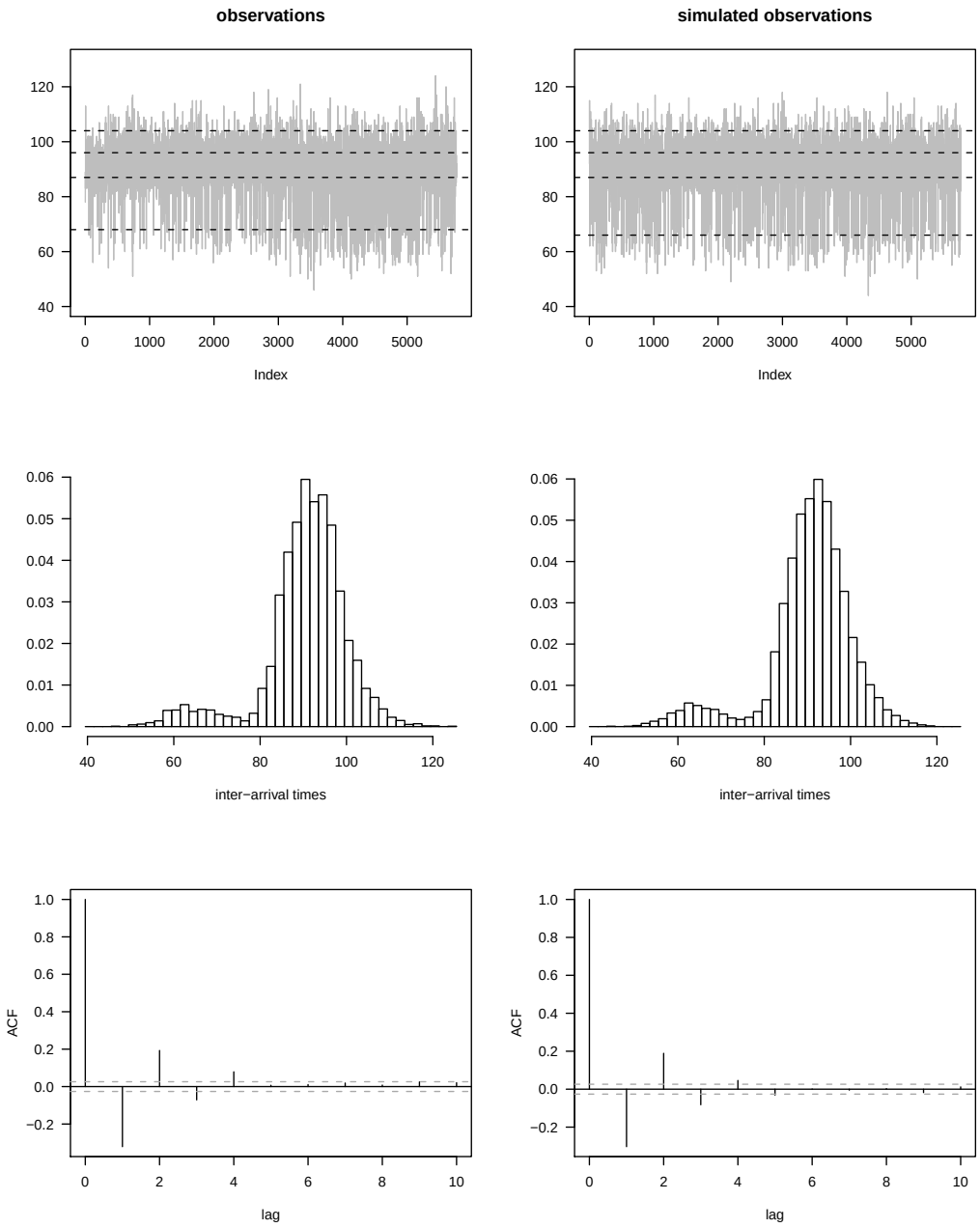

Figure 1.8: *Observed (plots on the left) and simulated (plots on the right) series, as well as corresponding histograms and autocorrelation functions.*

2 Hidden Markov models with arbitrary state dwell-time distributions[1]

The number of consecutive time points that the Markov chain spends in a given state (the dwell time) follows a geometric distribution. Therefore, for example, the modal dwell time for every state of an HMM is one. Hidden semi-Markov models (HSMMs) are designed to relax this restrictive condition; the dwell time in each state can follow any discrete distribution on the natural numbers. HSMMs and their applications are discussed, inter alia, in Ferguson (1980), Sansom and Thomson (2001), Guédon (2003) and Yu and Kobayashi (2003).

The additional generality offered by HSMMs carries a computational cost: they are more demanding to apply than are HMMs. Furthermore, in HSMMs state changes and state dwell-time distributions are modelled separately, meaning that the embedded Markov chain operates on a non-uniform time scale, and consequently the Markov property is lost. In some applications this can be regarded as natural, e.g. in the modelling of breakpoint rainfall data in Sansom and Thomson (2001). However, in general it can be unnatural and it leads to difficulties if one wants to do prediction or if one wishes to include covariates, trend or seasonality in the model. Covariate modelling in HMMs, in the state-dependent process as well as in the Markov chain, has been broadly explored and is fairly standard (see e.g. Part Two in Zucchini and MacDonald 2009). For HMMs the inclusion of random effects in both components of the model is also feasible (see Altman 2007). The generalization of Altman's MHMMs (mixed hidden Markov models) to HSMMs is straightforward if the covariates and random effects merely influence the state-dependent process (see Chaubert-Pereira *et al.* 2008). However, the incorporation of covariates or random effects in the latent part of the HSMM, i.e. in the semi-Markov chain, is much more difficult due to the timing problems arising from the separate modelling of state changes and state sojourns. This problem can be circumvented by using HMMs with arbitrary state dwell-time distributions — the existing HMM methodology then becomes available. In particular, and in contrast to the HSMM case, simple expressions for forecasts and residuals are available for

[1] This chapter is based on Langrock and Zucchini (2011).

HMMs. Furthermore, in the HSMM literature it is generally assumed that the start of the series coincides with a state switch (see e.g. Sansom and Thomson 2001, Guédon 2003 and Bulla *et al.* 2009). This assumption is expected to have a relatively minor impact on the estimates when the series are long; nevertheless it is both arbitrary and often unrealistic. In general it makes the HSMM non-stationary. Indeed it turns out that fitting stationary (as opposed to only homogeneous) HSMMs is not as easy as in the case of HMMs. Finally, the elaborateness of the EM algorithm, which usually is applied to fit HSMMs, increases unless one makes the simplifying assumption that there is a state switch at the end of the series (see Guédon 2003).

In this chapter we consider specially structured HMMs that capture the 'semi'-property of HSMMs, i.e. that can fit, at least approximately, any desired state dwell-time distribution, both parametric and nonparametric ones. The idea is to use so-called "state aggregates" which in similar, but not identical, ways have also been discussed e.g. in Russell and Cook (1987) and Guédon (2005). A good overview of the various existing approaches can be found in Johnson (2005) who also argues that the usage of this kind of models is "almost certainly a much better practical choice for duration modelling than development and implementation of more complex and computationally expensive models with explicit modifications to handle duration probabilities". Our HMM formulation is designed to fit, at least approximately, *any* given dwell-time distribution. It has exactly the same number of parameters as the corresponding HSMM and, in theory, the approximation of any dwell-time distribution can be made arbitrarily accurate. Furthermore, there are important subclasses of distributions for which the representation of the dwell-time distributions in the HMM formulation is exact. The topology of the state aggregates in our model is convenient since it is the same whatever dwell-time distribution we want to represent. Using our HMM formulation and direct likelihood maximization it is straightforward to fit stationary HMMs (with arbitrary state dwell-time distributions), or to incorporate trend, seasonality and covariates in the models, either in the hidden process or in the observed process. Indeed the whole standard HMM methodology including state prediction, local and global decoding, forecasting and model checking is applicable.

In Section 2.1 we describe how to structure an HMM so that it has the desired state dwell-time distribution. We illustrate the HMM formulation using a number of examples. A brief simulation study in Section 2.2 compares the estimation of the approximating HMM (via maximum likelihood) with the estimation of the approximated HSMM (via EM). In Sections 2.3, 2.4 and 2.5 three different applications are discussed, namely the modelling of daily rainfall occurrence, of daily returns on shares and of times between Old Faithful's eruptions.

2.1 Model description

We start by giving the definition of an HSMM. Subsequently, we show how an HSMM can be approximated by a suitably structured HMM. An HSMM comprises an observable output process $\{X_t\}_{t=1,2,\ldots}$, where the distribution of X_t is determined by the state, S_t, of an unobserved (hidden) N-state semi-Markov process $\{S_t\}_{t=1,2,\ldots}$. In general, the process S_t does *not* satisfy the Markov property (1.1). (For a general reference about semi-Markov processes see Kulkarni 1995.) Conditional on the states the observations of the output process are assumed to be independent.

Compared to conventional HMMs, HSMMs increase the flexibility of the state process S_t. More precisely, they allow for arbitrary state dwell-time distributions, whereas in conventional HMMs the state dwell-times are necessarily geometrically distributed. Let p_k denote the probability mass function (p.m.f.) of the dwell time in state $k \in \{1,\ldots,N\}$ and let F_k denote its distribution function. The support of p_k is $\mathbb{N}$, the set of natural numbers, or some subset of $\mathbb{N}$.

Consider the subsequence of $\{S_t\}_{t=1,2,\ldots}$ comprising the first occurrences of states in each run. (For example the subsequence corresponding to 1, 1, 2, 2, 2, 1, 3, 3, 3, 3 is 1, 2, 1, 3.) We assume that this subsequence is generated by an irreducible Markov chain (the 'embedded Markov chain') having transition probability matrix (t.p.m.) $\mathbf{\Omega} = \{\omega_{ij}\}$, where

$$\omega_{ij} = \mathbb{P}\big(S_{t+1} = j | S_t = i, S_{t+1} \neq i\big), \quad i,j = 1,\ldots,N,\ i \neq j,$$

$\sum_j \omega_{ij} = 1$, $\omega_{ii} = 0$, and that the initial probabilities for this Markov chain are given by $\delta_i^{(1)}$, $i = 1,\ldots,N$. For theoretical and computational convenience some authors (e.g. Sansom and Thomson 2001) also assume that the time instants $t = 1$ and $t = T$ are state boundaries, in the sense that $S_0 \neq S_1$ and $S_T \neq S_{T+1}$ (see discussion below). An HMM is the special case of an HSMM for which p_k is the p.m.f. of the geometric distribution.

Our aim now is to show how one can structure an HMM such that it is a reformulation of any given HSMM, i.e. such that it comprises any desired dwell-time distribution. Let $m_1, m_2, \ldots, m_N \in \mathbb{N}$, $m_0 := 0$, and consider an HMM with state-dependent process $\{X_t^\star\}_{t=1,2,\ldots}$ (observable) and Markov chain $\{S_t^\star\}_{t=1,2,\ldots}$ (unobservable) with states $\{1, 2, \ldots, \sum_{i=1}^N m_i\}$. We refer to the sets

$$I_k := \Big\{ n \mid \sum_{i=0}^{k-1} m_i < n \leq \sum_{i=0}^{k} m_i \Big\}, \quad k = 1,\ldots,N,$$

as *state aggregates*, and define $i_k^- := \min(I_k)$ and $i_k^+ := \max(I_k)$. We assume that each state of I_k is associated with the same distribution of the state-dependent process,

namely the distribution associated with the kth state in the HSMM defined above. In other words, the distribution of X_t^*, given state $S_t^* = l$, is the same for all $l \in I_k$, and is also the same as that of X_t given $S_t = k$, i.e.

$$\mathbb{P}^{X_t^\star | S_t^\star \in I_k} = \mathbb{P}^{X_t | S_t = k}, \quad t = 1, \ldots, T,\ k = 1, \ldots, N\,. \tag{2.1}$$

We denote the t.p.m. of $\{S_t^*\}_{t=1,2,\ldots}$ by $\mathbf{\Gamma} = \{\gamma_{ij}\}$, where $\gamma_{ij} = \mathbb{P}\big(S_{t+1}^\star = j | S_t^\star = i\big)$, $i, j = 1, \ldots, \sum_{i=1}^N m_i$, and structure it as follows:

$$\mathbf{\Gamma} = \begin{pmatrix} \mathbf{\Gamma}_{11} & \cdots & \mathbf{\Gamma}_{1N} \\ \vdots & \ddots & \vdots \\ \mathbf{\Gamma}_{N1} & \cdots & \mathbf{\Gamma}_{NN} \end{pmatrix}, \tag{2.2}$$

where the $m_i \times m_i$ diagonal blocks $\mathbf{\Gamma}_{ii}$ $(i = 1, \ldots, N)$ are defined, for $m_i \geq 2$, as

$$\mathbf{\Gamma}_{ii} := \left(\begin{array}{c|ccccc} 0 & 1 - c_i(1) & 0 & \cdots & & 0 \\ \vdots & 0 & \ddots & & & \vdots \\ & \vdots & & & & 0 \\ 0 & 0 & \cdots & 0 & & 1 - c_i(m_i - 1) \\ \hline 0 & 0 & \cdots & 0 & & 1 - c_i(m_i) \end{array}\right), \tag{2.3}$$

$\mathbf{\Gamma}_{ii} := 1 - c_i(1)$ for $m_i = 1$, and the $m_i \times m_j$ off-diagonal matrices $\mathbf{\Gamma}_{ij}$ $(i, j = 1, \ldots, N,$ $i \neq j)$ as

$$\mathbf{\Gamma}_{ij} := \begin{pmatrix} \omega_{ij} c_i(1) & 0 & \cdots & 0 \\ \omega_{ij} c_i(2) & 0 & \cdots & 0 \\ \vdots & & & \\ \omega_{ij} c_i(m_i) & 0 & \cdots & 0 \end{pmatrix}. \tag{2.4}$$

(In case of $m_j = 1$ the zeros disappear.) Here, for $r = 1, 2, 3, \ldots$,

$$c_k(r) := \begin{cases} \frac{p_k(r)}{1 - F_k(r-1)} & \text{for } F_k(r-1) < 1, \\ 1 & \text{for } F_k(r-1) = 1. \end{cases}$$

Note first that the functions c_k — the *hazard rates* of the dwell-time distributions — play the key role in our HMM formulation since they are responsible for rendering the desired dwell-time distributions p_k. Different dwell-time distributions lead to different c_k's, while the ω_{ij}'s and the structure of the t.p.m. remain unaffected. The c_k's are generated solely from the parameters of the dwell-time distributions; no additional parameters or constraints occur in comparison to the HSMM.

Note also that the matrix $\mathbf{\Gamma}$ indeed constitutes a t.p.m. since the entries all lie in the interval [0,1] and the row sums are equal to one. Although $\mathbf{\Gamma}$ appears to be somewhat complicated, its structure is not difficult to interpret: all transitions within state aggregate I_k are governed by the diagonal block $\mathbf{\Gamma}_{kk}$ which thus determines the dwell-time distribution of that state aggregate. The off-diagonal matrices determine the probabilities of transitions between different state aggregates. For example, for $k \neq l$, the matrix $\mathbf{\Gamma}_{kl}$ contains the probabilities of all possible transitions between the state aggregates I_k and I_l. Note that in this construction a transition from I_k to I_l must enter I_l in state i_l^-. We now show that this choice of $\mathbf{\Gamma}$ yields an HMM that is a reformulation of the given HSMM.

We denote the probability of a transition from state aggregate I_i to I_j by $\omega_{ij}^\star := \mathbb{P}\big(S_{t+1}^\star \in I_j | S_t^\star \in I_i, S_{t+1}^\star \notin I_i\big)$. This is analogous to the transition probability ω_{ij} in the HSMM. It is shown in the appendix that:

Proposition 1. *For $i \neq j$, $1 \leq i, j \leq N$,*

$$\omega_{ij}^\star = \omega_{ij}$$

We focus now on the accuracy of the representation of the dwell-time distributions p_k in our HMM formulation of the HSMM. We denote the p.m.f. of the dwell-time distribution in state aggregate I_k by $p_k^\star$ $(k = 1, \ldots, N)$. With the possible exception of the state aggregate that is active at $t = 1$, the stay in a given aggregate I_k begins in state i_k^-. By $p_k^\star$ we refer to the distributions of those dwell times that do start in state i_k^-.

Proposition 2. *For any $k \in \{1, \ldots, N\}$*

$$p_k^\star(r) = \begin{cases} p_k(r) & \text{for } r \leq m_k, \\ p_k(m_k)(1 - c_k(m_k))^{r-m_k} & \text{for } r > m_k. \end{cases}$$

The two p.m.f.'s thus differ only for $r > m_k$, i.e. in the right tail. Clearly, the difference between p_k and $p_k^\star$ can be made arbitrarily small by choosing m_k sufficiently large. It also follows that, for *any dwell-time distribution with finite support*, we can ensure that $p_k^\star(r) = p_k(r)\ \forall r$ by choosing m_k to be the maximum dwell time in state k having non-zero probability.

The use of state aggregates to allow for non-geometric dwell-time distributions while preserving the Markovian unit time scale has been suggested before, but in ways that are different to that proposed here. Well-known is the so-called *method of stages*, which can be divided into the *stages in series* and the *stages in parallel* (see Cox and Miller 1965). The former method leads to a distribution which is a convolution of

geometric dwell-time distributions and can be used to fit HSMMs with negative binomial dwell-time distributions (see Guédon 2005). On the other hand the method of stages in parallel corresponds to mixtures of geometric distributions which involve greater dispersion than that of a single geometric. Using combinations of both methods, one can approximate any distribution on the natural numbers (see Cox and Miller 1965). However, as pointed out by Kleinrock (1975), a good approximation can require a "horribly complicated" state-aggregate structure. Consequently this theoretical result is likely to be of limited usefulness in applications except in some special cases, e.g. where the dwell-time distribution is a negative binomial, a mixture of geometric distributions, or some simple combination of these two possibilities. In contrast the HMM formulation presented here has the same simple state-transition diagram whatever distribution we want to approximate. Another approach using state aggregates utilizes the Ferguson topology and is designed to allow for arbitrary dwell-time distributions having finite support (see Russell and Cook 1987). This is equivalent to the special case of our model in which no self-transition in the last visited state of the aggregates is allowed. An overview of the existing approaches can be found in Johnson (2005).

The following examples illustrate the approximation using our HMM formulation.

Example 1. Let p_k be the p.m.f. of a shifted Poisson distribution:

$$p_k(r) = \exp(-\lambda_k)\frac{\lambda_k^{r-1}}{(r-1)!}, \quad r = 1, 2, \dots \ . \tag{2.5}$$

Then, according to Proposition 2,

$$p_k^\star(r) = \begin{cases} p_k(r) & \text{for } r \le m_k, \\ p_k(m_k) z^{r-m_k} & \text{for } r > m_k, \end{cases}$$

where $z = 1 - p_k(m_k)/\big(1 - F_k(m_k - 1)\big)$ is independent of r. Although the functions $p_k^\star(r)$ and $p_k(r)$ differ for $r > m_k$, the discrepancy between them becomes small as m_k increases. This is illustrated in Figure 2.1, which displays a shifted Poisson distribution with parameter $\lambda = 5$ and the corresponding $p_k^\star(r)$ for $m_k = 4, 6, 8$.

Example 2. Let p_k be the p.m.f. of the geometric distribution:

$$p_k(r) = \pi_k(1-\pi_k)^{r-1}, \quad r = 1, 2, \dots \ .$$

Then, for $r \ge 2$,

$$\begin{aligned} c_k(r) &= \frac{p_k(r)}{1 - \sum_{s=1}^{r-1} p_k(s)} = \frac{\pi_k(1-\pi_k)^{r-1}}{\sum_{s=r}^{\infty} \pi_k(1-\pi_k)^{s-1}} \\ &= \frac{\pi_k}{\sum_{s=0}^{\infty} \pi_k(1-\pi_k)^{s}} = \pi_k \,. \end{aligned}$$

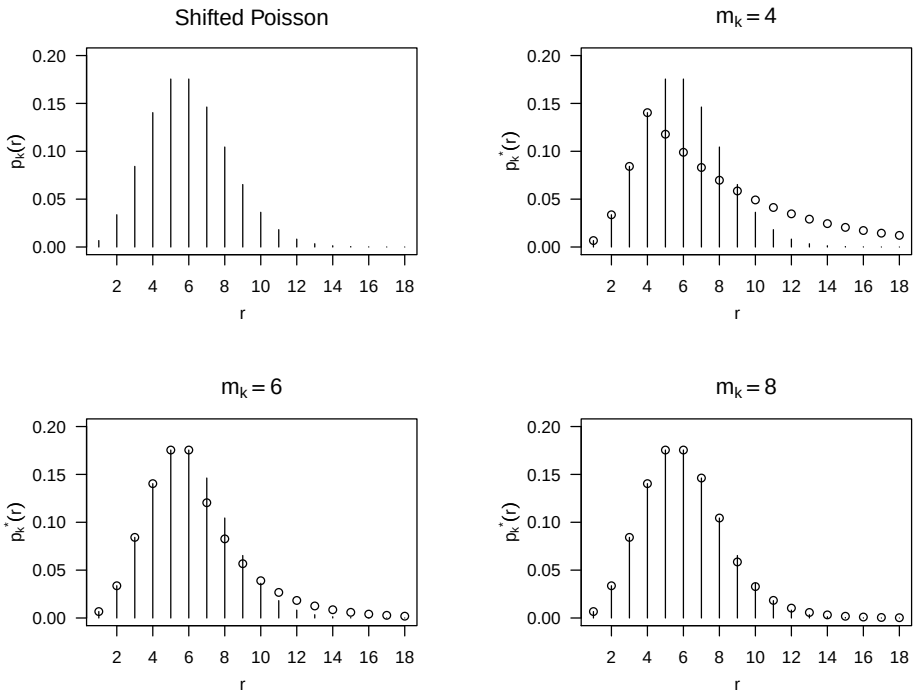

Figure 2.1: *The functions $p_k(r)$ (vertical lines), and $p_k^\star(r)$ (circles) for $m_k = 4, 6, 8$, from Example 1.*

By Proposition 2 we have $p_k^\star(r) = p_k(r)$ for $r \leq m_k$ and, for $r > m_k$, that

$$\begin{aligned} p_k^\star(r) &= p_k(m_k)\big(1 - c_k(m_k)\big)^{r-m_k} = \pi_k(1-\pi_k)^{m_k-1}(1-\pi_k)^{r-m_k} \\ &= \pi_k(1-\pi_k)^{r-1} = p_k(r)\,. \end{aligned}$$

In this example the choice of m_k thus does not play any role; the HSMM reduces to an HMM.

Example 3. Consider a shifted binomial distribution with p.m.f.

$$p_k(r) = \binom{n_k}{r-1} \pi_k^{r-1}(1-\pi_k)^{n_k-(r-1)}, \quad r = 1, \ldots, n_k + 1.$$

Since $\sum_{i=1}^{n_k+1} p_k(i) = 1$, we have, for $m_k = n_k + 1$, that $c_k(m_k) = p_k(n_k+1)/(1 - \sum_{i=1}^{n_k} p_k(i)) = 1$. For $r > m_k$, application of Proposition 2 yields

$$p_k^\star(r) = p_k(m_k)\big(1 - c_k(m_k)\big)^{r-m_k} = 0 = p_k(r).$$

Proposition 2 also guarantees that $p_k^\star(r) = p_k(r)$ for $r \leq m_k$. Thus choosing $m_k = n_k+1$ ensures that $p_k^\star(r) = p_k(r)\ \forall r$.

In summary, Propositions 1 and 2, together with (2.1), imply that our HMM formulation is capable of representing, at least approximately, any given dwell-time distribution. In other words, with respect to HSMMs, the distribution of $\{X_t\}_{t=1,2,\ldots}$ can be approximated by that of $\{X_t^\star\}_{t=1,2,\ldots}$, where the approximation can be made arbitrarily accurate by choosing the m_k sufficiently large.

We now consider situations in which it can be beneficial to use the HMM formulation instead of HSMMs. Firstly, the separate modelling of state changes and state dwell-time distributions in the hidden part of HSMMs can lead to difficulties when introducing covariates, including trend and seasonality, in the latent process. Up to now hardly any work on seasonal HSMMs has been published, except the contributions of Sansom and Thomson (2007, 2010). On the other hand covariate modelling in HMMs was considered numerous times in the literature (see e.g. Altman 2007, Bartolucci 2009 and Part Two of Zucchini and MacDonald 2009). In principle it is straightforward to fit HMMs involving covariates, both in the state-dependent process and in the latent process. A further benefit of using HMMs with arbitrary state dwell-time distributions is that it enables the transfer of the very general MHMM approach of Altman (2007) to the HSMM setting. In addition, standard HMM techniques, e.g. for forecasting and model checking, are available.

Another question of interest concerns the choice of the initial distribution of the process. Suppose that the HSMM is in state k at time $t = 1$. Now unless the state process *entered* state k at $t = 1$ (i.e. unless $S_0 \neq k$) the distribution of the first dwell time will differ from p_k in general. It is to circumvent the difficulties of taking this difference into account that Sansom and Thomson (2001), as well as Guédon (2003) and Bulla *et al.* (2009), assume that the time instant $t = 1$ is a state boundary. The initial distribution $\boldsymbol{\delta}^{(1)}$ is then explicitly modelled and can feasibly be estimated (see Guédon 2003). However, although this assumption is unlikely to have much effect on parameter estimation except for short series, it is clearly unrealistic in general. More serious, the enforced state switch at the start of the series in general impedes stationarity of the HSMM. In fact, there is no simple procedure yet available to fit a stationary HSMM. In contrast, the assumption of a state switch at the start of the series is easily avoidable using our model by allowing the initial state probabilities to be nonzero also for states i with $i \notin \{i_1^-, i_2^-, \ldots, i_N^-\}$. Moreover, it is straightforward to fit a stationary HMM: the initial distribution $\boldsymbol{\delta}^{(1)}$ of a stationary HMM with t.p.m. $\boldsymbol{\Gamma}$ is the solution to the linear system $\boldsymbol{\delta}^{(1)}\boldsymbol{\Gamma} = \boldsymbol{\delta}^{(1)}$ subject to $\sum_i \delta_i^{(1)} = 1$ (cf. Section 1.1), and thus can be expressed in terms of the parameters that determine $\boldsymbol{\Gamma}$. Thus, in effect, the model formulation considered here allows one to fit stationary HSMMs. On the other hand this formulation can also deal with the case where the user wishes to assume that there

is a state boundary at time $t = 1$.

The equally unrealistic assumption that there is a state change at the end of the series impacts on the forecast distribution as well as on the likelihood. Furthermore, it excludes the possibility of absorbing states in the semi-Markov chain (Guédon 2003). Both theory (Guédon 2003) and software (Bulla *et al.* 2009) have been developed to fit HSMMs that avoid such an assumption. On the other hand it is a convenient feature of the HMM formulation that no such modification of the standard algorithm is required.

The following two examples illustrate other possible variants of the model.

Example 4. To model rainfall data, Sansom and Thomson (2001) proposed an HSMM having dwell-time p.m.f. of the following form:

$$p(r) = \begin{cases} \theta_r & \text{for } r \leq d, \\ \theta(1-\theta)^{r-d} & \text{for } r > d, \end{cases}$$

where $\theta := \theta_1 + \ldots + \theta_d$. This distribution on the positive integers has an unstructured start and a geometric tail. We say that the start is *of order* $d - 1$, motivated by the fact that in comparison to the geometric distribution $d - 1$ additional parameters are considered. (Note that the case $d = 1$ (order 0) reduces to a geometric distribution.) In the manner of (2.2), this HSMM can be formulated as an HMM which has the identical probability structure if we choose the corresponding m_i equal to d.

Example 5. Guédon (2005) studied *hybrid models* in which some of the states are Markovian and others are semi-Markovian. These models can easily be described by using our model by choosing $m_i = 1$ and thus

$$\mathbf{\Gamma}_{ii} = 1 - c_i(1) = 1 - \pi_i\,, \quad \mathbf{\Gamma}_{ij} = \omega_{ij} c_i(1) = \omega_{ij}\pi_i\,,$$

for each Markovian state i, where π_i is the parameter of the corresponding geometric distribution, and by defining $\mathbf{\Gamma}_{kk}$ and $\mathbf{\Gamma}_{kl}$ according to (2.3) and (2.4), respectively, for each semi-Markovian state k.

The HMM formulation can also render dwell-time distributions from *different families* in one model; we just define the block matrices according to (2.3) and (2.4). If all the dwell-time distributions have either finite support or geometric tails then the HMM formulation is equivalent, otherwise it is approximate. In the latter case the approximation can be made arbitrarily accurate.

2.2 Simulation study

The HMM method for fitting HSMMs is now analysed using simulated data. We assess the performance of the method and, in particular, compare it to existing software for HSMMs. The R-package `hsmm` by Bulla *et al.* (2009) is designed to fit HSMMs using the EM algorithm as described by Guédon (2003). It needs to be kept in mind that the package `hsmm` makes extensive use of C code whereas the code for the HMM method described above was written purely in R.

The computational effort required to evaluate the likelihood of an HMM with arbitrary state dwell-time distribution is linear in the number of observations, T, and quadratic in the number of HMM states, $M = \sum_{i=1}^{N} m_i$. Likelihood evaluation for HSMMs on the other hand is quadratic in T in the worst case, if one uses standard HSMM methodology (see Guédon 2003 for reference). The additional computational effort due to the explicit modelling of the dwell-time distributions hence in the two approaches is expressed in different terms.

To begin with, we generated simulations from a two-state HSMM with shifted Poisson state dwell-time distributions (cf. (2.5)) and normal state-dependent distributions. Table 2.1 summarizes the results of applying the two competing methods for different numbers of observations. The same starting values were used for both methods. In this particular example numerical maximum likelihood estimation of the approximating HMM outperformed the estimation of the HSMM via EM: the parameter estimates are almost identical, but the computing time is significantly smaller for the HMM method and the gap increases as T, the number of observations, increases.

However, it turns out the HMM method is not always faster than EM. Indeed, the performance of the two competing algorithms depends on, to a large extent, *(i)* the given combination of true model parameters and, in the case of the HMM method, also on *(ii)* the size of the state aggregates that are needed to provide a good approximation. Concerning *(i)* it turns out the EM algorithm converges particularly fast if the observations can easily be assigned to individual states, i.e. if the states are clearly distinguishable at the observation level. This is illustrated in Table 2.2. For the EM algorithm the computing time decreases as the gap between the state-dependent means increases. The same holds for the HMM method, but the decrease is slower. For some parameter combinations the HMM method and for others the EM algorithm is faster. Both methods yielded approximately the same estimates for all given parameter combinations.

Referring to *(ii)*, Table 2.3 illustrates the role of the sizes of the state aggregates that are involved when using the HMM method: increasing the sizes of the state aggregates, and thus the number of HMM states, slows down the computation. On the other hand

Table 2.1: *Estimation of the approximating HMM (via ML) vs. estimation of the HSMM (via EM): parameter estimates and computing time for different numbers T of simulated observations (size of the state aggregates: $m_1 = m_2 = 20$)*

	time	State level		Observation level			
	(sec.)	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$
$T = 1000$							
approx. HMM	12	2.25	5.50	1.76	3.91	1.86	2.73
HSMM	17	2.26	5.49	1.76	3.91	1.86	2.73
$T = 5000$							
approx. HMM	54	2.78	5.21	2.00	3.97	1.91	3.15
HSMM	102	2.79	5.19	2.00	3.97	1.91	3.15
$T = 20000$							
approx. HMM	227	2.94	4.92	2.04	4.00	1.98	3.11
HSMM	647	2.95	4.90	2.04	4.00	1.98	3.11
true para.		3	5	2	4	2	3

Table 2.2: *Estimation of the approximating HMM (via ML) vs. estimation of the HSMM (via EM): computing times for different values of μ_1; the other true parameters are fixed at the values given in Table 2.1; $T = 2000$, $m_1 = m_2 = 20$.*

	$\mu_1 = -1$	$\mu_1 = 0$	$\mu_1 = 1$	$\mu_1 = 2$	$\mu_1 = 2.5$	$\mu_1 = 3$
approx. HMM	18.6	22.1	20.2	22.1	24.4	27.4
HSMM	6.1	11.5	23.3	30.0	58.9	162.9

the state aggregates must not be chosen too small, compared to the range of the state dwell-time distributions, as otherwise the estimates deteriorate. Clearly, the HMM method for fitting HSMMs will be advantageous when the number of states in the state aggregates is not large, i.e. when the modes of the dwell-time distributions are rather small. The method becomes less efficient as the size of the state aggregates increases,

Table 2.3: *Estimation of the approximating HMM: estimated parameter of the dwell-time distribution in state 2 for different true parameters λ_2 and different state aggregate sizes m_2; computing times (in sec.) in parentheses; all true parameters except λ_2 are fixed at the values given in Table 2.1; $T = 4000$.*

	$\lambda_2 = 3$		$\lambda_2 = 6$		$\lambda_2 = 12$		$\lambda_2 = 20$	
$m_2 = 5$	3.10	(29.1)	4.49	(31.0)	6.89	(31.8)	7.74	(41.8)
$m_2 = 10$	3.28	(32.8)	5.55	(35.6)	10.64	(37.7)	13.04	(43.9)
$m_2 = 20$	3.28	(44.9)	5.56	(45.3)	10.99	(53.4)	19.76	(52.6)
$m_2 = 50$	3.28	(90.7)	5.56	(93.4)	10.99	(95.7)	19.90	(106.9)

since the matrices to be multiplied become larger and also require more memory space.

Up to now, we have simulated and estimated merely two-state HSMMs. The case with two states is relatively simple, as no transition probabilities between state aggregates have to be estimated. However, as Proposition 1 essentially states, the HMM method for fitting HSMMs can also deal with more than two semi-Markovian states. To illustrate this we simulated 5000 observations from a three-state HSMM with negative binomial state dwell-time distributions and Poisson state-dependent distributions. The p.d.f. of the negative binomial distribution is given by

$$p(r) = \binom{r+k-2}{r-1} \pi^k (1-\pi)^{r-1}, \qquad r = 1, 2, 3, \ldots .$$

This distribution has the two parameters $k \in \{1, 2, 3, \ldots\}$ and $\pi \in [0, 1]$. It represents a generalization of the geometric distribution: $p(r)$ gives the probability that $r - 1$ "failures" occur before k "successes" have occurred, and thus we are back in the case of a geometric distribution if $k = 1$. Using the gamma function, $p(\cdot)$ can be extended to allow for any positive real-valued k:

$$p(r) = \frac{\Gamma(r+k-1)}{(r-1)!\Gamma(k)} \pi^k (1-\pi)^{r-1}, \qquad r = 1, 2, 3, \ldots . \tag{2.6}$$

The following parameters were used to generate the observations:

$$\begin{aligned} k_1 = 0.5, \quad \pi_1 = 0.1, \quad & \text{(state 1)} \\ k_2 = 2, \quad \pi_2 = 0.3, \quad & \text{(state 2)} \\ k_3 = 10, \quad \pi_3 = 0.6 \quad & \text{(state 3)} \end{aligned}$$

$$\text{and} \quad \mathbf{\Omega} = \begin{pmatrix} 0 & 0.5 & 0.5 \\ 0.7 & 0 & 0.3 \\ 0.8 & 0.2 & 0 \end{pmatrix}$$

at the state level, and

$$\lambda_1 = 6, \quad \lambda_2 = 12 \quad \text{and} \quad \lambda_3 = 18$$

at the observation level. (Note that $\mathbf{\Omega} = \{\omega_{ij}\}$, where $\omega_{ij} = \mathbb{P}\big(S_{t+1} = j | S_t = i, S_{t+1} \neq i\big)$.) Fitting an HSMM to the simulated data via the HMM method, with state aggregates of size $m_1 = m_2 = m_3 = 20$, yields the following parameter estimates:

$$\begin{aligned} \hat{k}_1 = 0.53, \quad \hat{\pi}_1 = 0.11, \quad &\text{(state 1)} \\ \hat{k}_2 = 2.68, \quad \hat{\pi}_2 = 0.36, \quad &\text{(state 2)} \\ \hat{k}_3 = 11.50, \quad \hat{\pi}_3 = 0.63, \quad &\text{(state 3)} \end{aligned}$$

$$\hat{\mathbf{\Omega}} = \begin{pmatrix} 0 & 0.51 & 0.49 \\ 0.73 & 0 & 0.27 \\ 0.81 & 0.19 & 0 \end{pmatrix},$$

$$\hat{\lambda}_1 = 6.03, \quad \hat{\lambda}_2 = 11.96, \quad \hat{\lambda}_3 = 17.94\,.$$

The `hsmm`-package produces approximately the same estimates. At the state level the parameters are less accurately estimated than are those at the observation level. There was no indication of sensitivity to starting values; ten different combinations of initial values for the parameters all yielded the same results. This example verifies that the HMM approximaton method for HSMMs also works in the case of more than two states of the semi-Markov chain.

2.3 Application to daily rainfall occurrences

There exists a substantial literature on the stochastic modelling of daily precipitation (for reviews see Woolhiser 1992 or Srikanthan and McMahon 2001). Many of the proposed models are constructed from two submodels: the first describes rainfall occurrence (whether a particular day is wet or dry) and the second the rainfall amount on wet days. We restrict our attention to rainfall occurrence and fit a variety of models, in particular HMMs with non-geometric dwell-time distributions, to binary sequences of dry and wet days. A similar application can be found in MacDonald and Zucchini (1997); these authors consider Markov Chains and conventional HMMs.

Our main objective is to illustrate the application of the class of HMMs described in the preceding sections. As do Sansom and Thomson (2007) we include seasonality in

the models. We show how this can be done in two different ways and thereby illustrate that it is easy to incorporate covariates in the latent process as well as in the observed process.

The data considered here comprise binary series of dry and wet days over a period of about 47 years at five sites in Bulgaria, namely Zlatograd, Plovdiv, Kurdjali, Ihtiman and Ivailo. As is usually done with hydrological series, in order to avoid the complication arising from having 366 days on leap years, we discard observations for February 29. We mainly concentrate on the daily rainfall series from Zlatograd (a town in the Rhodope mountains). As can be expected, the series shows significant seasonality (cf. Table 2.4).

Table 2.4: *Zlatograd series: sample probability of a rainy day per month.*

Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
0.20	0.18	0.21	0.29	0.35	0.34	0.31	0.31	0.33	0.39	0.41	0.27

Markov chains

We begin by considering a simple model for the daily rainfall series from Zlatograd, namely a two-state Markov chain in which the transition probabilities are allowed to vary seasonally (*Model 1*). This can be regarded as nonhomogeneous HMM with deterministic state-dependent process. Let S_t $(t = 1, 2, \ldots, 16951)$ be a Markov chain representing rainfall occurrence, where $S_t = 1$ if day t is dry, and $S_t = 2$ if day t is wet. The t.p.m. for *Model 1* depends on t:

$$\mathbf{\Gamma}_t = \begin{pmatrix} \mathbb{P}(S_{t+1} = 1 \,|\, S_t = 1) & \mathbb{P}(S_{t+1} = 2 \,|\, S_t = 1) \\ \mathbb{P}(S_{t+1} = 1 \,|\, S_t = 2) & \mathbb{P}(S_{t+1} = 2 \,|\, S_t = 2) \end{pmatrix}.$$

To take care of seasonality the logit transforms of the diagonal elements of the t.p.m. are modelled as linear combinations of trigonometric functions:

$$\alpha_0 + \sum_{k=1}^{q} \left(\alpha_k \sin\left(\frac{2\pi k t}{365}\right) + \beta_k \cos\left(\frac{2\pi k t}{365}\right) \right), \tag{2.7}$$

where q is usually chosen using a model selection criterion, and is seldom greater than 2. We return to the choice of q later.

Hidden Markov models

As a next step we fit an HMM incorporating seasonality in the transition probabilities to the series. Now the Markov chain S_t is no longer taken to be an observation; it represents the unobserved state on day t. The observation on day t is regarded as a realization of a Bernoulli random variable whose parameter is determined by the state: the probability that day t is wet is π_1 if day t is in state 1, and is π_2 if day t is in state 2 (*Model 2*). *Model 1* is the special case of *Model 2* with $\pi_1 = 0$ and $\pi_2 = 1$.

An alternative way of introducing seasonality in the HMM is to assume that the entries of the t.p.m. are constant (i.e. do not depend on t), but that the Bernoulli parameters depend on t instead. Suppose now that the logit transforms of $\pi_1(t)$ and of $\pi_2(t)$ each have the form given in (2.7) (*Model 3*).

Table 2.5: *Minus log lik. and AIC for Models 1-3 fitted to the Zlatograd series ($q = 1$).*

Model	*mllk*	*AIC*
1	9468.44	18948.87
2	9445.84	18907.67
3	9423.27	**18862.55**

Table 2.5 gives the minus log likelihood and the AIC for *Models 1-3* with $q = 1$. In terms of the AIC, the HMM with seasonality in the Bernoulli parameters provides the best fit by a substantial margin.

Hidden semi-Markov models

As the next step, we extend *Model 3* by allowing the dwell-time distribution in each state to have an unstructured start and a geometric tail, as in Example 4. For the orders of the unstructured starts we try the values 0 (Markovian), 1, 2 and 3 for both the state belonging to the dry periods and that belonging to the wet periods. We also try different orders of seasonality, i.e. $q = 1, 2, 3$. Considering all possible combinations of orders of the unstructured starts and the seasonality another 47 different models emerge (one of the combinations is *model 3* above).

Out of these models the AIC would select the HMM which has dwell-time distribution with unstructured start of order 2 for the dry periods, unstructured start of order 3 for the wet periods and order of seasonality $q = 2$ in the Bernoulli parameters. The fitted transition probability matrix is given by

$$\left(\begin{array}{ccc|cccc} 0 & 0.94 & 0 & 0.06 & 0 & 0 & 0 \\ 0 & 0 & 0.73 & 0.27 & 0 & 0 & 0 \\ 0 & 0 & 0.85 & 0.15 & 0 & 0 & 0 \\ \hline 0.15 & 0 & 0 & 0 & 0.85 & 0 & 0 \\ 0.37 & 0 & 0 & 0 & 0 & 0.63 & 0 \\ 0.34 & 0 & 0 & 0 & 0 & 0 & 0.66 \\ 0.27 & 0 & 0 & 0 & 0 & 0 & 0.73 \end{array}\right).$$

(Note that the zeros are structural zeros.) Here the state aggregate $I_1 = \{1, 2, 3\}$ is associated with low probability of precipitation (dry periods), and $I_2 = \{4, 5, 6, 7\}$ with high probability of precipitation (wet periods). The upper left block matrix determines the dwell-time distribution in the dry periods, the lower right block that in the wet periods. The estimated p.m.f.'s of the dwell-time distributions in the two state aggregates are displayed in Figure 2.2. The deviation from the p.m.f. of a geometric distribution is evident, in particular the modal dwell time is not one in either case.

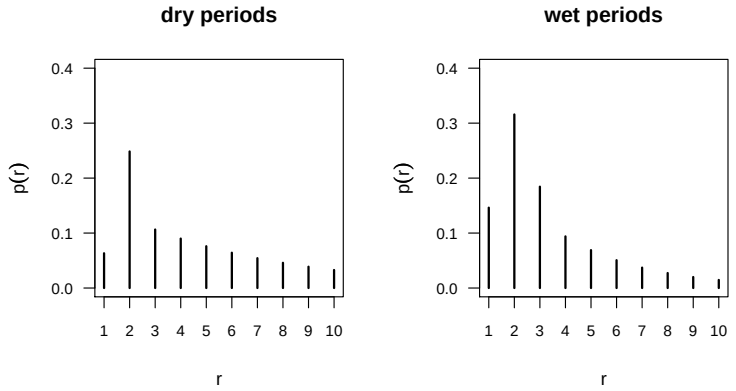

Figure 2.2: *Estimated p.m.f.'s for dry and wet periods.*

The estimated Bernoulli parameter functions $\pi_1(t)$ or $\pi_2(t)$, which have period 365, are displayed in Figure 2.3. State aggregate I_1 can be regarded as the "dry" HSMM state; the probability of rain is generally low but peaks slightly in early November. In state aggregate I_2, the "wet" HSMM state, the probability of rain is generally high but ranges between 0.5 and 0.8. The stationary probabilities for state aggregates I_1 and I_2 are 0.63 and 0.37 respectively. In other words, the system is in the "dry" state 63% of the time, and in the "wet" state 37% of the time.

The AIC value for the chosen model is 18806.04, while it is 18832.25 for the HMM with geometric dwell-time distributions and the same order of seasonality ($q = 2$). The

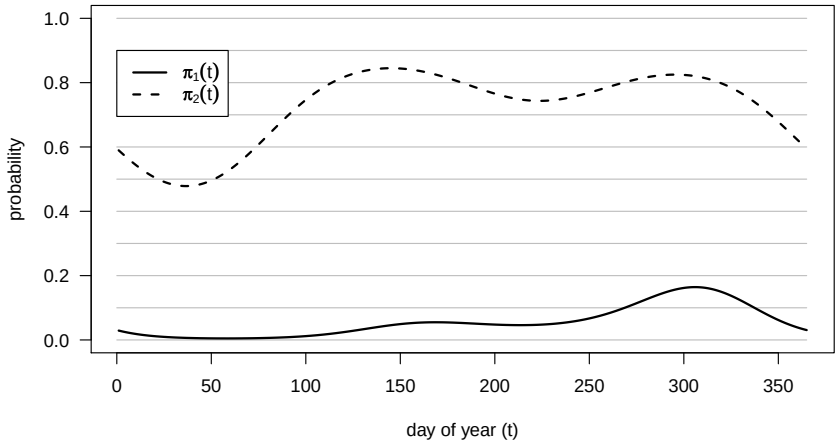

Figure 2.3: *Estimated Bernoulli parameter functions for the Zlatograd series.*

additional flexibility provided by allowing for non-geometric dwell-time distributions leads to a substantial improvement in the fit.

Table 2.6: *Selected models for five sites.*

Site	*Order dry periods*	*Order wet periods*	*Order of seasonality*
Zlatograd	2	3	2
Plovdiv	2	1	2
Kurdjali	2	1	3
Ihtiman	0	1	2
Ivailo	2	1	2

The model selection exercise was repeated using the data from sites in Plovdiv, Kurdjali, Ihtiman and Ivailo with all the above models. Table 2.6 lists the models that led to the lowest AIC in each case. Recall that order 0 means a Markovian state; the selected model for Ihtiman is thus a hybrid HMM/HSMM (see Example 5 in Section 2.1). The selected models are not the same for all sites, but there is no reason to suppose that there exists a single model structure that is appropriate for all sites.

2.4 Application to Dow Jones returns

We consider another application of HSMMs, namely the modelling of daily return series. More precisely, we apply the HMM approximation method to fit HSMMs to a series of

returns of the Dow Jones Industrial Average index. The adjusted closing prices, p_t, for the period 02.01.1980 – 31.12.2009, were downloaded from 'finance.yahoo.com', and the daily returns were computed as $y_t = \log(p_t/p_{t-1})$, $t = 2, ..., T$. The series was corrected for one outlier, namely the return on the "Black Monday" (19th October 1987), which was -0.256. After exclusion of this outlier, 7571 observations remain.

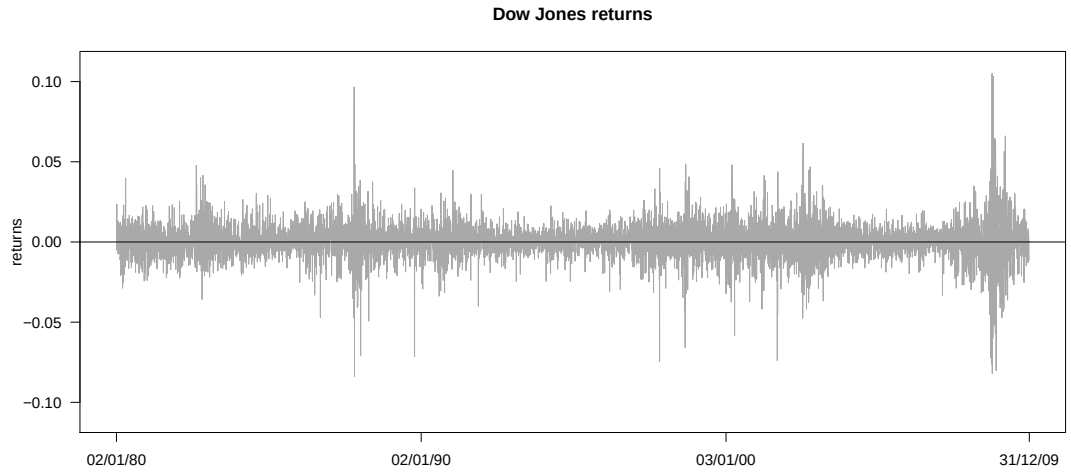

Figure 2.4: *Dow Jones returns from 2nd of January 1980 until 31st of December 2009 (corrected for the "Black Monday").*

Figure 2.4 displays the considered series of returns. The most extreme returns, i.e. the smallest and the largest observations, occurred in 1987 and in 2008. The former are associated with the 1987 crash (which also involves the Black Monday), the latter were caused by the collapse of Lehman Brothers and the subsequent breakout of the recent financial crisis. Some other extreme returns occurred due to events such as the Asian financial crisis in 1997 and the September 11 attacks in 2001.

One of the stylized facts commonly attributed to return series is corroborated; namely *volatility clustering*, i.e. the tendency for high (or low) volatilities to occur in clusters. This is illustrated by Figure 2.5, which displays the sample autocorrelation functions for the two series y_t and y_t^2 ($t = 1, \ldots, 7571$), respectively. The variance of the returns is persistent, while the returns themselves show no indication of correlation over time. Another stylized fact of daily return series, *excess kurtosis*, is illustrated by Figure 2.6. The plot displays a histogram of the observed Dow Jones returns together with a fitted normal distribution (obtained by maximum-likelihood estimation). The fit of the normal distribution is unsatisfactory since the distribution of the observations is highly leptokurtic (the sample kurtosis is approximately 11.4).

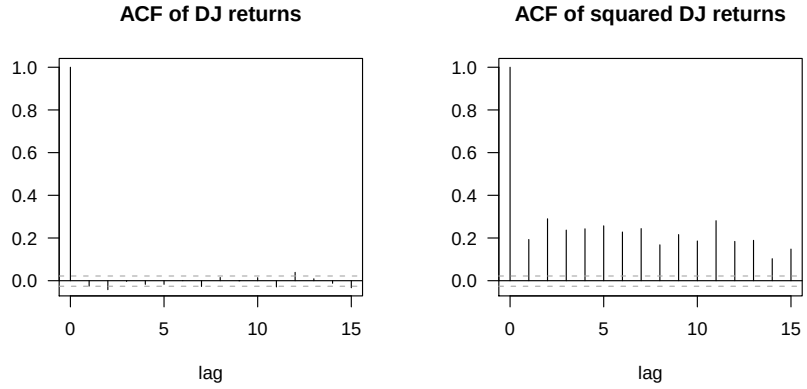

Figure 2.5: *Sample autocorrelation function for the series of returns (y_t, left plot) and for the series of squared returns (y_t^2, right plot).*

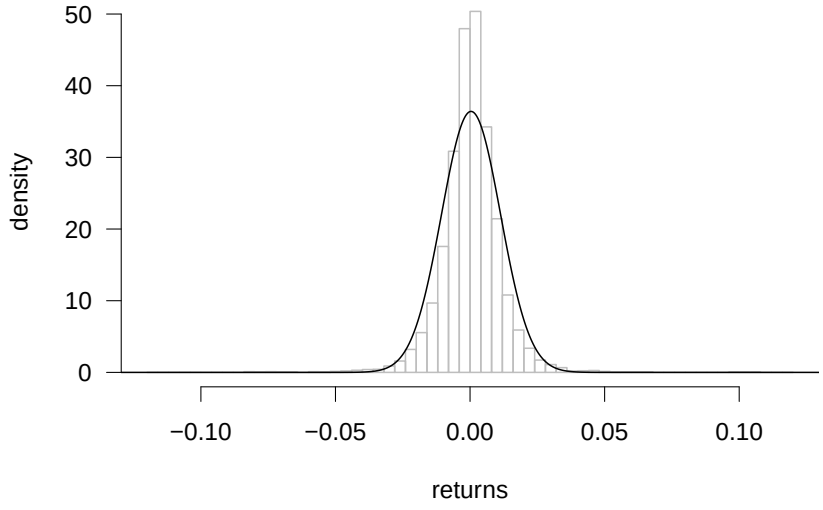

Figure 2.6: *Histogram of the DJ returns from 02.01.1980 – 31.12.2009 and fitted normal distribution (solid line).*

HMMs can accommodate both the volatility clustering and the excess kurtosis in a natural way. Following Rydén *et al.* (1998) and Bulla and Bulla (2006), who also fitted HMMs to return series, we consider only two-state HMMs and HSMMs; parameter estimates for models with more states are strongly influenced by extreme observations (cf. Rydén *et al.*, 1998). The following models were fitted to the Dow Jones return series:

- a two-state HMM where, conditional on $S_t = k$, $X_t \sim \mathcal{N}(0, \sigma_k^2)$ (normal HMM),
- a two-state HMM where, conditional on $S_t = k$, $\sigma_k^{-1} X_t \sim t_{\eta_k}$ (t-HMM),
- a two-state HSMM where, conditional on $S_t = k$, $X_t \sim \mathcal{N}(0, \sigma_k^2)$, and the state dwell-times are negative binomially distributed (normal HSMM), and
- a two-state HSMM where, conditional on $S_t = k$, $\sigma_k^{-1} X_t \sim t_{\eta_k}$, and the state dwell-times are negative binomially distributed (t-HSMM).

(The p.m.f. of the dwell-time distributions for the HSMMs is given by (2.6).)

Table 2.7: *Minus log likelihood, AIC and BIC for HMMs and HSMMs fitted to the Dow Jones return series.*

	mllk	*AIC*	*BIC*
normal HMM	-24433.15	-48858.29	-48830.56
t-HMM	-24571.54	-49131.08	-49089.49
normal HSMM	-24501.87	-48991.75	-48950.16
t-HSMM	-24591.18	**-49166.36**	**-49110.90**

Table 2.7 summarizes the model fitting results. The models with t-distributions at the observation level attain substantially higher likelihood values than those with normal distributions. Furthermore, the additional flexibility gained by allowing for non-geometric dwell-time distributions (here: negative binomial dwell-time distributions) improves the fit substantially. Both model selection criteria select the t-HSMM, which confirms the findings of Bulla and Bulla (2006).

The parameter estimates for the t-HSMM are given by

$$\begin{aligned} 100 \cdot \hat{k}_1 &= 0.687, \quad 100 \cdot \hat{\pi}_1 = 0.010, \quad \text{(state 1)} \\ 100 \cdot \hat{k}_2 &= 4.586, \quad 100 \cdot \hat{\pi}_2 = 1.402, \quad \text{(state 2)} \end{aligned}$$

for the dwell-time distributions and

$$\begin{aligned} 100 \cdot \hat{\sigma}_1 &= 0.626, \quad \hat{\eta}_1 = 15.105, \quad \text{(state 1)} \\ 100 \cdot \hat{\sigma}_2 &= 1.363, \quad \hat{\eta}_2 = \ 5.583, \quad \text{(state 2)} \end{aligned}$$

for the state-dependent distributions. On average approximately 94% of the dwell times in state 1 of the fitted model are of length one; in those cases the process directly switches back to state 2, i.e. the state associated with relatively high volatility.

Nonetheless, due to the long tail of the fitted distribution the mean dwell-time in state 1 is quite large, namely 69.7. The deviation from the geometric case is evident. In state 2 approximately 82% of the dwell times are of length one, on average, and the mean dwell-time is 4.2. The state process mostly switches back and forth between the two states, and occasionally remains in one state for a relatively long period. A conventional HMM, with geometric state dwell-time distributions, can not accommodate both of these attributes.

2.5 Application to Old Faithful eruptions

In this section we investigate whether HSMMs can improve the fit to the Old Faithful series that was already analysed in Section 1.4. We start again by looking at the dichotomized, binary series of long and short inter-arrival times (Section 2.5.1). Subsequently, in Section 2.5.2, the nondichotomized time series is analysed.

2.5.1 Modelling the binary series via HSMMs

Consider again the series of Old Faithful's eruption inter-arrival times that have been discretized into either "short" or "long" (cf. Section 1.4.1). Up to now the most suitable models for this series have been identified to be a second-order Markov chain, a two-state second-order Bernoulli HMM and a three-state Bernoulli HMM. The latter model yielded the highest likelihood of all considered models, but it was surprising to find that two of its states are almost equivalent at the observation level. Consider again the parameter estimates for the three-state Bernoulli HMM, given by equations (1.6) and (1.7). As the Bernoulli parameters of states 2 and 3 are approximately equal, the set of states $\{2, 3\}$ can be regarded as a kind of state aggregate. The time the Markov chain spends in this state aggregate is not geometrically distributed. The main reason for the improved likelihood, compared to the two-state Bernoulli HMM, may thus well be the departure from the assumption of geometric dwell-time distributions. This motivates the use of HSMMs in this particular application.

In Section 1.4.1 we have seen that the dwell-time in the state associated with short inter-arrival times almost surely is of length one (across all models). The dwell-time of this state is thus not worth being modelled by a non-geometric distribution. We consider the following two-state Bernoulli hybrid HMM/HSMM:

- given either of the two states, the observations are Bernoulli distributed,
- state 1 involves a geometric dwell-time distribution and

- the dwell-time in state 2 is negative binomially distributed.

The maximum of the log likelihood for this model is -1380.00, the AIC and BIC are 2770.00 and 2803.30, respectively. In terms of these criteria the model performs better than any of the models that were considered in Section 1.4.1 (cf. Table 1.1). The estimated Bernoulli parameter vector is

$$\widehat{\boldsymbol{\pi}} = (0.59, 1.00)\,.$$

Dwell-times in state 1 of the fitted hybrid HMM/HSMM are of length one (almost surely). The parameter estimates of the negative binomial dwell-time distribution in state 2 (cf. (2.6)) are

$$\hat{k} = 0.261 \text{ and } \hat{\pi} = 0.063\,.$$

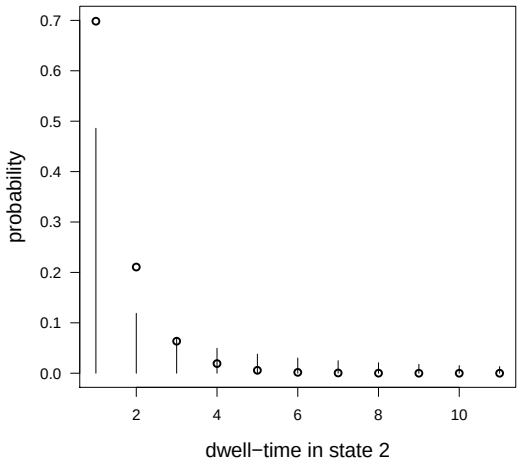

Figure 2.7: *Fitted dwell-time distributions for the state associated with long eruption inter-arrival times, for the two-state HMM (circles) and for the two-state hybrid HMM/HSMM (bars).*

Figure 2.7 compares the dwell-time distribution in state 2 of the fitted hybrid HMM/HSMM with that in state 2 of the two-state Bernoulli HMM from Section 1.4.1. One should be aware that the states are not synchronized across the models; those labelled "state 1" do not lead to the same distribution at the observation level. For the displayed distribution associated with the hybrid HMM/HSMM the deviation from the geometric case is evident. Allowing for more flexibility in the state dwell-time distributions leads to an improved fit in this example.

2.5.2 Modelling the series of inter-arrival times via HSMMs

Now consider again the original, nondichotomized series of eruption inter-arrival times (cf. Section 1.4.2). One of the most promising models that was fitted to this series is the four-state HMM with gamma state-dependent distributions (cf. Table 1.3). We found that a further increase in the number of states improved the likelihood significantly, even though two of the states in the fitted five-state model were hardly distinguishable at the observation level. Intuitively the two components with mean approximately 93 would not be regarded as two separate states (cf. Figure 1.5). These two states resemble once more a state aggregate, at least in an approximate sense. In view of the likelihood gap between the four-state and the five-state model it seems worthwhile to investigate whether the fit of the four-state model can be improved by employing non-geometric, more flexible dwell-time distributions.

Consider again the four-state GHMM that was fitted in Section 1.4.2 (the parameter estimates are given in Appendix A1). The estimated state-dependent means of this model for states 1, 2, 3, and 4 are 65.35, 86.16, 92.99 and 99.94, respectively. The estimated parameters of the (geometric) state dwell-time distributions, $\hat{\pi}_i$ for state i, are $\hat{\pi}_1 = 1$, $\hat{\pi}_2 = 1$, $\hat{\pi}_3 = 0.43$ and $\hat{\pi}_4 = 0.98$. As sojourn times in states 1, 2 and 4 are infrequently longer than 1, it does not seem worthwhile to replace the dwell-time distributions of these states by non-geometric ones. When moving from the four-state to the five-state model, it is essentially the remaining state 3 that is split up into two states, which are almost equivalent at the observation level. We thus consider the following four-state gamma hybrid HMM/HSMM:

- given any of the four states, the observations are gamma distributed,
- states 1, 2 and 4 have geometric dwell-time distributions and
- the dwell-time in state 3 is assumed to have an unstructured start of order 4 (cf. Example 4).

(The states are ordered in terms of increasing means of the state-dependent distributions.) The reason for using a distribution with unstructured start, rather than the negative binomial distribution, is the small essential range of the dwell-time distribution in the given application; not many additional parameters are needed to obtain sufficient flexibility. The maximum of the log likelihood of the suggested model is -19834.14, which yields AIC = 39716.28 and BIC = 39876.12. The additional flexibility in the dwell-time of state 3 only led to a minor improvement of the fit. This is confirmed by Figure 2.8, which displays the dwell-time distributions in state 3, for the conventional

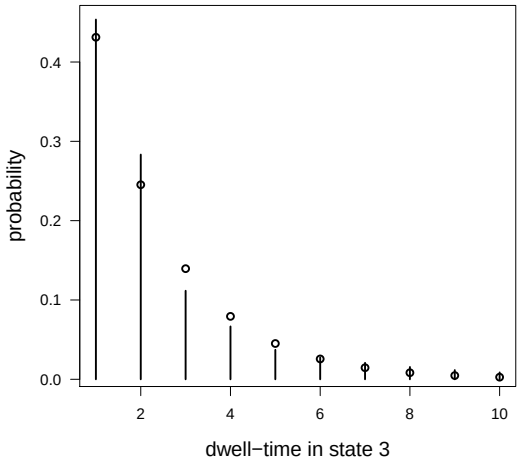

Figure 2.8: *Fitted dwell-time distributions for the state associated with the second-largest mean of the state-dependent distribution: for the four-state gamma HMM (circles) and for the four-state gamma hybrid HMM/HSMM (bars).*

HMM and for the hybrid HSMM/HMM, respectively. The discrepancy between the two distributions is small.

Bearing in mind the findings from Section 1.4.2, it can be concluded that the length of the memory of the state process, rather than a possible semi-Markovian behaviour, is responsible for the likelihood gap between the four-state and the five-state GHMM.

2.6 Concluding remarks

In this chapter we considered a class of HMMs that capture the 'semi'-property of HSMMs, i.e. that can represent any given state dwell-time distribution, either exact or approximately, where the approximation can be made arbitrarily accurate. The motivation for doing so was to take advantage of the well-established methodology that is available for HMMs. Key advantages of using HMMs with arbitrary dwell-time distributions (rather than HSMMs) include the ease with which it is possible to incorporate covariate information in different ways, and that fitting stationary models is straightforward. The applications given to rainfall occurrences, daily returns and geyser eruption inter-arrival times illustrate the feasibility of the proposed method.

3 State-space modelling by means of structured hidden Markov models[1]

Despite their considerable flexibility, there are some drawbacks of HMMs related to the state architecture. While in some examples the choice of the number of states and the interpretation of the states can be obvious, in general this need not be the case. Although model selection criteria and residual analyses can be applied, the determination of the number of states often remains critical. Furthermore, even if an HMM captures most of the stylized facts in a given application, the assumption that the number of states is finite may seem counterintuitive and can lead to difficulties in the interpretation. In such cases an HMM is no more and no less than a tool for modelling dependence in a given series of observations (cf. Cappé *et al.* 2005). Another possible drawback of HMMs is that the number of parameters can get very large as the number of states increases. For the sake of better interpretability, and for parsimony in terms of parameters, general-type state-space models (SSMs) sometimes provide a more appropriate alternative.

In this chapter we consider SSMs, that need be neither linear nor Gaussian, of the form

$$\begin{aligned} y_t &= a(g_t, \epsilon_t), \\ g_t &= b(g_{t-1}, \eta_t), \end{aligned}$$

where $\{g_t\}_{t=1,2,\ldots}$ denotes the nonobservable state process, and $\{y_t\}_{t=1,2,\ldots}$ denotes the observable state-dependent process[2]. The innovations ϵ_t and η_t are i.i.d. sequences. The functions a and b may be nonlinear but are known. By their very nature, SSMs are closely related to HMMs. However, in contrast to the HMMs considered up to now, the state space here can be continuous and thus infinite. SSMs allow for various models for the succession of states, e.g. autoregressive processes. Although HMMs are nested in the broader family of SSMs, it is nevertheless reasonable to distinguish these model families as the available methodology for statistical inference is rather different. Extensive accounts of state-space models and their application can be found in Harvey

[1] This chapter is based on Langrock, MacDonald and Zucchini (2010) and Langrock (2010).
[2] The notation y_t and g_t is used to conform with that in the literature on SSMs.

(1989) and Durbin and Koopman (2001).

The likelihood of an SSM is given by a high-order multiple integral that in general cannot be evaluated directly. Linear and Gaussian SSMs can be tackled by applying the Kalman filter (see e.g. Harvey 1989). On the other hand, the nonlinear and non-Gaussian case is more involved. Possible methods for estimating the parameters of nonlinear and non-Gaussian SSMs include

- the extended Kalman filter (see e.g. Welch and Bishop 1995),
- unscented Kalman filtering (see e.g. Julier and Uhlmann 2004),
- the generalized method of moments (see e.g. Melino and Turnbull 1990),
- numerical integration (see e.g. Kitagawa 1987) and
- Monte Carlo methods (see e.g. Carlin *et al.* 1992, Durbin and Koopman 1997).

In this chapter, we proceed along the lines of Kitagawa (1987) and, in doing so, convert the estimation problem for a nonlinear and non-Gaussian SSM to that of an HMM. The general idea can be summarized as follows: by discretizing the state space of an SSM, one obtains an approximation of the likelihood that can be made arbitrarily accurate. The approximated likelihood matches the likelihood of a suitably structured HMM, and thus the whole HMM methodology becomes applicable.

Although the idea of approximating the likelihood is not original (see e.g. Kitagawa 1987, Fridman and Harris 1998), the relation to HMMs has not been thoroughly discussed in the literature. Particular advantages of the proposed estimation method by means of HMMs are that simple explicit formulae exist for the residuals and the forecast distributions of an HMM, and that estimates of the latent process can be obtained by using the Viterbi algorithm. In addition, the HMM formulation makes it particularly simple to consider a variety of nonstandard SSMs that are easy to implement. The discussed method is feasible for one-dimensional state spaces. However, it suffers the so-called "curse of dimensionality" and thus is rather difficult to apply in case of high-dimensional state spaces (Kitagawa 1996). In such cases one might need to resort to MCMC methods.

Section 3.1 discusses how to estimate SSMs using suitably structured HMMs. Probably the most prominent nonlinear SSMs are the so-called *stochastic volatility models*, which are standard tools for modelling the variance of return series. In Section 3.2, standard and nonstandard stochastic volatility models are discussed and then fitted to a number of daily return series. This section also contains a modest simulation study in which the HMM approximation method is compared to a Monte Carlo approach (in

the context of stochastic volatility modelling). Besides stochastic volatility modelling, there are numerous other possible applications of the proposed estimation method. A further four applications are discussed in Sections 3.3 – 3.6, namely the modelling of earthquake counts, polio counts, rainfall occurrence data and glacial varve thicknesses. The applications were selected in order to cover a wide range of possible outcome variables; discrete (with and without seasonality), continuous (with support $\mathbb{R}_{\geq 0}$ and with support $\mathbb{R}$) and binary outcomes are considered. Each of the considered SSMs is non-linear and non-Gaussian, and the HMM approximation method is employed to estimate the SSM parameters. The results for the fitted SSMs are compared to those of their (finite-state) HMM counterparts.

3.1 Model fitting strategy

State-space models are characterized by two processes: a continuous-valued Markov state process, g_t, and an observation process, y_t, whose realizations are assumed to be conditionally independent, given the states. Figure 3.1 displays the dependence structure of an SSM in a directed acyclic graph. HMMs have precisely the same structure, see Figure 1.1, except that the Markov process is discrete-valued instead of continuous-valued. By appropriately discretizing the state space of an SSM into a finite number of states, N, the model can be approximated by an HMM. The point of using such an approximation is that, whereas the likelihood of the SSM involves a multiple integral and is difficult to compute, that of an HMM is easy to compute and to maximize.

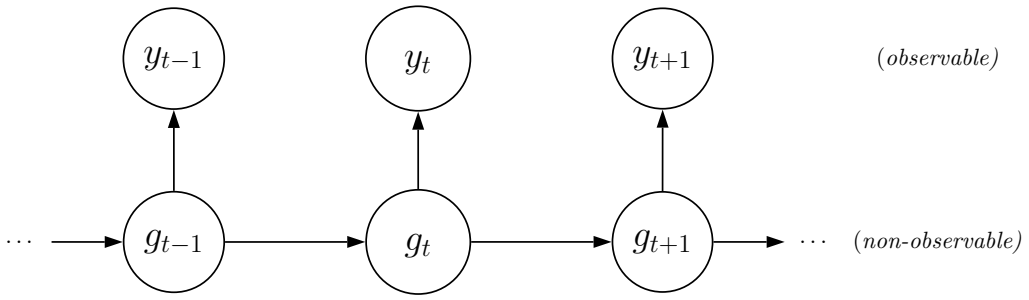

Figure 3.1: *Dependence structure of an SSM.*

In detail, the procedure is as follows. Let the essential range of possible g_t-values be split into N equally-sized intervals $B_i := (b_{i-1}, b_i)$, $i = 1, \ldots, N$. We denote by b_i^* a representative point in B_i, e.g. the midpoint. Making use of the dependence

structure of an SSM, and applying numerical integration, the likelihood of the SSM can be approximated as follows:

$$\begin{aligned}
\mathcal{L} &= \int \dots \int f(\mathbf{y}, \mathbf{g}) \, \mathrm{d}\mathbf{g} \\
&= \int \dots \int f(g_1) f(y_1|g_1) \prod_{t=2}^{T} f(g_t|g_{t-1}) f(y_t|g_t) \, \mathrm{d}g_T \dots \mathrm{d}g_1 \\
&\overset{(*)}{\approx} \sum_{i_1=1}^{N} \dots \sum_{i_T=1}^{N} \mathbb{P}(g_1 \in B_{i_1}) f(y_1|g_1 = b^*_{i_1}) \\
&\qquad \cdot \prod_{t=2}^{T} \mathbb{P}(g_t \in B_{i_t}|g_{t-1} = b^*_{i_{t-1}}) f(y_t|g_t = b^*_{i_t}) \,, \qquad (3.1)
\end{aligned}$$

where f is used as a general symbol for a density and T denotes the number of observations. Note that this approximation is not the same as those of Fridman and Harris (1998) and Bartolucci and De Luca (2003). These authors approximate the integrals in $(*)$ by replacing both g_{t-1} *and* g_t, in the expression $f(g_t|g_{t-1})$, by the corresponding interval midpoints. Thus, we conduct a slightly modified numerical integration[3].

The numerical integration essentially implies a discretization of the state space. We now demonstrate that the approximated likelihood (3.1) indeed matches that of a suitably structured HMM. To see this, regard the midpoints b^*_i, $i = 1, \dots, N$, as possible values of an N-state Markov chain h_t with transition probability matrix $\mathbf{\Gamma} = (\gamma_{ij})$, where

$$\gamma_{ij} := \mathbb{P}(g_t \in B_j | g_{t-1} = b^*_i) \,,$$

and initial distribution $\boldsymbol{\delta}^{(1)}$, where $\delta^{(1)}_i := \mathbb{P}(g_1 \in B_i)$. The transition probabilities γ_{ij}, $i, j = 1, \dots, N$, are determined by the state equation of the SSM. For instance, if the state process is a Gaussian AR(1) with parameters ϕ and σ, then

$$\gamma_{ij} = \Phi\left(\frac{b_j - \phi b^*_i}{\sigma}\right) - \Phi\left(\frac{b_{j-1} - \phi b^*_i}{\sigma}\right) , \qquad (3.2)$$

where Φ denotes the cumulative distribution function of the standard normal distribution. We further define

$$\mathbf{P}(y_t) := \mathrm{diag}\left(f(y_t|g_t = b^*_1), \dots, f(y_t|g_t = b^*_N)\right) ,$$

[3]More precisely, we use an approximation of the type

$$\int_a^b f(x)g(x)dx \approx \int_a^b f(x)dx \cdot g\left(\frac{a+b}{2}\right)$$

rather than

$$\int_a^b f(x)g(x)dx \approx (b-a) f\left(\frac{a+b}{2}\right) g\left(\frac{a+b}{2}\right).$$

where $f(y_t|g_t = b_i^*)$ is determined by the observation equation of the SSM. Now we can rewrite the approximate likelihood (3.1) in HMM notation as follows:

$$\mathcal{L} \approx \boldsymbol{\delta}^{(1)}\mathbf{P}(y_1)\boldsymbol{\Gamma}\mathbf{P}(y_2)\boldsymbol{\Gamma}\cdots\boldsymbol{\Gamma}\mathbf{P}(y_{T-1})\boldsymbol{\Gamma}\mathbf{P}(y_T)\mathbf{1}^t\,. \tag{3.3}$$

In summary, the likelihood of the SSM can be approximated using numerical integration, and the approximate likelihood is precisely that of a suitably structured HMM, namely that determined by the Markov chain h_t and the state-dependent probability functions (densities) $f(y_t|g_t = b_i^*)$ (=: $f(y_t|h_t = b_i^*)$). Clearly, as N increases the intervals become narrower and the approximation consequently improves. The approximating HMM is stationary if the initial distribution $\boldsymbol{\delta}^{(1)}$ is the stationary distribution implied by the transition probability matrix $\boldsymbol{\Gamma}$, i.e. if $\boldsymbol{\delta}^{(1)}\boldsymbol{\Gamma} = \boldsymbol{\delta}^{(1)}$ subject to $\sum_{i=1}^{N}\delta_i^{(1)} = 1$.

It is then a routine matter to evaluate (3.3) and to maximize it numerically with respect to the parameters (cf. Section 1.2). In practice one has to decide what value of N, the number of states, will be adequate, and what range of g_t-values to allow for. The minimum and maximum values (g_{min} and g_{max}) for g_t have to be chosen sufficiently large to cover the essential domain of the state process. Here it is advisable to examine the stationary distribution of g_t; in the Gaussian AR(1)-case this is given by $\mathcal{N}\big(0, \sigma^2/(1-\phi^2)\big)$. Fridman and Harris (1998) suggest using $-g_{min} = g_{max} = 3\sigma_g$, where σ_g denotes the stationary standard deviation of g_t. The choice of N has a strong influence on the accuracy of the approximation. The accuracy improves as N increases, but the size of the matrices in (3.3) also increases, which slows down the evaluation of the likelihood. Note that, although N needs to be large enough to provide a good approximation, the number of model parameters does not depend on N; the entries of the $N \times N$ matrix Γ depend only on those determining the state equation of the SSM.

The approximating HMM can also be used, *inter alia*, for forecasting, decoding or model checking of the SSM. Indeed, all standard HMM methods are applicable. In particular, the Viterbi algorithm can be used for state decoding (see e.g. Section 3.3), and pseudo-residuals can be used for model checking (cf. Section 3.2.3.2). Furthermore, simple closed-form expressions for forecasts are available (cf. Section 3.2.3.2).

3.2 Application to stochastic volatility modelling

The standard discrete-time stochastic volatility model, without leverage, for returns y_t on an asset can be written in several different forms, e.g.

$$y_t = \varepsilon_t\beta\exp(g_t/2), \qquad g_{t+1} = \phi g_t + \sigma\eta_t, \qquad (t = 1,\ldots,T) \tag{3.4}$$

and

$$y_t = \varepsilon_t \exp(g_t/2), \qquad g_{t+1} = \mu + \phi(g_t - \mu) + \sigma\eta_t, \qquad (t = 1, \ldots, T) \tag{3.5}$$

where, in both (3.4) and (3.5), $|\phi| < 1$ and $\{\varepsilon_t\}$ and $\{\eta_t\}$ are independent sequences of independent standard normal random variables; see e.g. Shephard (1996). We use the model definition (3.4), and, following Chib *et al.* (2002), label it SV_0. A common extension of the basic model assumes for ε_t a Student-t distribution with ν degrees of freedom, and $\nu > 0$ is then treated as an additional parameter. Again following Chib *et al.* (2002), we label this extension *SVt*.

Over the past two decades, stochastic volatility models such as SV_0 and *SVt* have attracted much attention in the finance literature as a competitor to, *inter alia*, GARCH models (cf. Broto and Ruiz 2004). SV models mimic several of the stylized facts attributed to asset returns: kurtosis of returns in excess of 3, zero autocorrelation of returns, and dependence of returns as revealed by the nonzero autocorrelations of squared returns (cf. Section 2.4). For a discussion of these stylized facts, see Taylor (2005, Chapter 4). Danielsson (1994) reported better model-fitting results for even the basic SV model than for any EGARCH model.

On the other hand SV models belong to the class of nonlinear SSMs, and thus are not as easy to fit as GARCH models. In the past two decades much ingenuity has been applied in the derivation of estimation methods for SV models; for a comprehensive overview of the existing methodology we recommend Broto and Ruiz (2004). Some of the most important methods are the generalized method of moments (GMM, see Melino and Turnbull 1990), quasi-maximum-likelihood (QML, see Harvey *et al.* 1994), Markov chain Monte Carlo (MCMC, see Jacquier *et al.* 1994) and Monte Carlo likelihood (MCL, see Sandmann and Koopman 1998). According to Shephard (2005), the methods can be categorized into those that are relatively simple but inefficient (like GMM and QML), and those that attempt to evaluate the likelihood, which are efficient but computer-intensive and rather difficult to implement (like MCMC and MCL).

In this section we apply the approximation method via structured HMMs (see Section 3.1). In the context of stochastic volatility modelling it was applied by Fridman and Harris (1998), and by Bartolucci and De Luca (2003). (See also Section 13.3 of Zucchini and MacDonald (2009), where this approach is applied to an SV model with leverage.) We propose a number of new nonstandard SV models, in particular models with log-volatility processes $\{g_t\}$ which differ from that in (3.4), and which appear to have some advantages.

We start with a brief simulation study that compares the proposed HMM estimation method to the MCL method described by Sandmann and Koopman (1998). Section

3.2.2 then introduces four nonstandard SV models. Some of these are generalizations of the basic models with additional parameters in the log-volatility process (or volatility process), which is assumed to belong to the class of conditional linear AR(1) models described by Grunwald *et al.* (2000). In Section 3.2.3 each of the six SV models considered is fitted to ten series of daily returns, and the relative merits of the models are assessed in terms of the AIC and their out-of-sample performance, especially the accuracy of their forecast distributions.

3.2.1 Simulation study

Table 3.1: *SV_0 model: parameter estimates and computing times for MCL and HMM method ($-g_{min} = g_{max} = 4$, true parameters: $\phi = 0.98$, $\sigma = 0.2$, $\beta = 0.05$); 95% confidence intervals in parentheses.*

	N	time (sec.)	$\widehat{\phi}$	$\widehat{\sigma}$	$\widehat{\beta}$
MCL		90	0.976	0.208	0.047
			(0.969;0.981)	(0.190;0.229)	(0.044;0.052)
HMM	30	21	0.975	0.199	0.047
			(0.968;0.980)	(0.178;0.223)	(0.044;0.051)
	50	32	0.975	0.208	0.047
			(0.968;0.980)	(0.188;0.231)	(0.044;0.051)
	100	78	0.975	0.212	0.047
			(0.968;0.980)	(0.192;0.234)	(0.044;0.052)
	200	245	0.975	0.212	0.047
			(0.968;0.980)	(0.192;0.234)	(0.044;0.052)

Table 3.1 gives an indication of the influence of N on accuracy and computing time. The SV_0 model was fitted to a simulated series of $T = 10\,000$ observations by means of *(i)* the MCL method, which is implemented in Ox in `ssfpack` (Koopman *et al.* 1999), and then *(ii)* the HMM method, implemented in R, for different values of N. The parameters were set at $\phi = 0.98$, $\sigma = 0.2$ and $\beta = 0.05$; the starting values were set at $\phi_0 = 0.9$, $\sigma_0 = 0.3$ and $\beta_0 = 0.2$ for both methods. In the estimation by means of the HMM method the model was reparameterized in terms of unconstrained "working parameters" (cf. Section 1.2); approximate confidence intervals for the constrained parameters, ϕ, σ and β, were obtained by first estimating confidence intervals for the working parameters from the inverse of the estimated information matrix, and then

applying the corresponding inverse transformations to the interval boundaries for the working parameters. Alternatively the parametric bootstrap could have been applied. In case of the MCL method `ssfpack` provides confidence intervals.

The results in Table 3.1, as well as those obtained for many observed series of returns, and for generated series, lead us to conclude that the parameter estimates obtained by the HMM method stabilize for N-values somewhere between 50 and 100. Secondly, for values of $N \leq 100$ the HMM method is comparable with the MCL method in terms of computing time. However, an important motivation for applying the HMM formulation is that all kinds of extensions of the standard SV model, and of state-space models in general, are easy to implement by simply modifying a few lines of code for the computation of $\mathbf{\Gamma}$ and $\mathbf{P}(y_t)$ in expression (3.3). This convenient feature of the HMM formulation is exploited in Section 3.2.3 in order to fit the nonstandard SV models that we introduce in Section 3.2.2.

3.2.2 Some nonstandard SV models

Shifting the volatility process

The models SV_0 and SVt can be generalized by introducing a lower bound to the volatility of the observed process. For instance, the observation equation in the model (3.4) can be replaced by

$$y_t = \varepsilon_t(\beta \exp(g_t/2) + \xi) \; . \tag{3.6}$$

The additional parameter ξ (≥ 0) does appear to be worthwhile (cf. Section 3.2.3), and is plausible on the grounds that some baseline volatility is always present. In all the models that are presented in the subsequent paragraphs we incorporate this additional parameter. Of course, the model with $\xi = 0$ is in all cases nested in the model with $\xi \geq 0$.

In all models covered in this section, ε_t is assumed to follow a Student-t distribution with ν degrees of freedom.

SVMt — mixture of AR(1) processes in the log-volatility

The SVt model can be generalized by using a mixture of two normal distributions in the conditional distribution of g_{t+1} given g_t. Let y_t be given by (3.6), but now assume that, given g_t, g_{t+1} is distributed either $\mathrm{N}(\phi_1 g_t, \sigma_1^2)$ (with probability α) or $\mathrm{N}(\phi_2 g_t, \sigma_2^2)$

(with probability $1-\alpha$). Equivalently,

$$g_{t+1} = \begin{cases} \phi_1 g_t + \sigma_1 \eta_t & \text{with probability } \alpha \\ \phi_2 g_t + \sigma_2 \eta_t & \text{with probability } 1-\alpha, \end{cases} \tag{3.7}$$

with the innovations η_t being independent standard normal. This model, hereafter labelled *SVMt*, allows for abrupt changes in the state process and thus offers additional flexibility. The *SVt* model is nested in *SVMt*: consider the case $\alpha = 1$ and $\xi = 0$. One could also consider using a mixture with more than two AR(1) components, but that generalization will not be pursued here.

Wong and Li (2000) give the following necessary and sufficient condition for second-order stationarity of $\{g_t\}$:

$$\alpha\phi_1^2 + (1-\alpha)\phi_2^2 < 1\ . \tag{3.8}$$

Note that it is possible for one of the AR(1) processes to be 'explosive' (e.g. $\phi_2 = 1.4$) without necessarily destroying the second-order stationarity of the mixed process. The stationary mean of $\{g_t\}$ is 0 and the stationary variance of $\{g_t\}$ is

$$\sigma_g^2 = \frac{\alpha\sigma_1^2 + (1-\alpha)\sigma_2^2}{1 - \left(\alpha\phi_1^2 + (1-\alpha)\phi_2^2\right)}.$$

(These, and other moments given below, are derived in Appendix A3.) As there is no closed-form expression for the marginal distribution of g_t, the exact stationary variance and kurtosis of the observed process, $\{y_t\}$, are not available. Second-order Taylor approximations for these moments are given by

$$\mathrm{var}(y_t) \approx \frac{\nu}{\nu-2}\left((\beta+\xi)^2 + c_1\sigma_g^2\right) \tag{3.9}$$

and

$$\mathrm{kurtosis}(y_t) \approx 3\,\frac{\nu-2}{\nu-4}\,\frac{(\beta+\xi)^4 + c_2\sigma_g^2}{\left((\beta+\xi)^2 + c_1\sigma_g^2\right)^2}\ , \tag{3.10}$$

where $c_1 = 0.5\beta^2 + 0.25\beta\xi$ and $c_2 = 3\beta^2\xi^2 + 2\beta^4 + 4.5\beta^3\xi + 0.5\beta\xi^3$.

MSSVt — Markov-switching innovations in the log-volatility

The model *SVMt* can be extended by replacing the independent mixture of AR components in the log-volatility by a dependent mixture, for example a Markov switching model. Such a generalization allows the sojourn times in each state of the process to be (stochastically) longer than those implied by the *SVMt* model. It is designed to further accommodate volatility clustering, i.e. the tendency for high (or low) volatilities to occur in clusters.

Let y_t be given by (3.6) and assume that

$$g_{t+1} = \phi g_t + \sigma_{\alpha_t} \eta_t \,,$$

with $\{\alpha_t\}$ being a two-state stationary Markov chain described by the 2×2 transition probability matrix $\mathbf{\Gamma}^{(\alpha)} = (\gamma_{ij}^{(\alpha)})$. We thereby allow for two different variances in the innovations — σ_1^2 and σ_2^2 — that are selected by the Markov chain $\{\alpha_t\}$. The model, henceforth labelled *MSSVt*, is similar to that of So *et al.* (1998). However, those authors assume the innovation variance σ^2 to be constant and instead model a reparameterized β nonhomogeneously via $\{\alpha_t\}$. The *SVt* model is again nested in our model: consider the case $\gamma_{11}^{(\alpha)} = \gamma_{21}^{(\alpha)} = 1$ and $\xi = 0$.

Stationarity of $\{g_t\}$ and $\{y_t\}$ holds if and only if $|\phi| < 1$. The stationary variance of $\{g_t\}$ is

$$\sigma_g^2 = \frac{\delta_1^{(\alpha)} \sigma_1^2 + \delta_2^{(\alpha)} \sigma_2^2}{1 - \phi^2} \,,$$

where $\boldsymbol{\delta}^{(\alpha)}$ is the stationary distribution of the Markov chain $\{\alpha_t\}$; $\delta_i^{(\alpha)}$ is therefore the expected proportion of time that $\{\alpha_t\}$ spends in state i. The expressions (3.9) and (3.10) give the approximate stationary variance and kurtosis of $\{y_t\}$.

To fit the model we consider the process

$$z_t := \begin{pmatrix} g_t \\ \alpha_t \end{pmatrix},$$

which is a Markov process on $\mathbb{R} \times \{0, 1\}$. The component $\{g_t\}$ is discretized into N states, as described in Section 3.1, and $\{\alpha_t\}$ takes on one of two values, so the number of states of $\{z_t\}$, after discretization, is $2N$. Writing the t.p.m. of $\{z_t\}$ in terms of the model parameters, it is then straightforward to maximize the likelihood, which is given by (3.3).

In principle the model can be further generalized in a number of ways that we mention but, in the interests of brevity, we will not discuss in detail. One can allow the parameter ϕ also to depend on the state of $\{\alpha_t\}$. However, stationarity conditions then become more involved, and parameter estimation for this model proved to be unstable in practice. Some of the other obvious generalizations are more straightforward. For example, one can allow the parameters β, ν and ξ to depend on the current state of $\{\alpha_t\}$. The extension to more than two states for $\{\alpha_t\}$ is also easy to implement. Of course an increase in the number of states leads to an increase in the size of the t.p.m., and hence in the computational burden.

SVVt — nonhomogeneous innovations in the log-volatility

The models *SVMt* and *MSSVt* provide for more flexibility in the log-volatility process. However, both incorporate a total of five parameters merely in the determination of the log-volatility process — a considerable increase compared to the two parameters in the models SV_0 and *SVt*. The model presented in this section has three parameters in the log-volatility process and in this sense represents a compromise.

Let y_t be given by (3.6), but now assume that

$$g_{t+1} = \phi g_t + \sigma_t \eta_t \,, \tag{3.11}$$

where $\sigma_t = \sqrt{\omega + \gamma \exp(g_t)}$ with $\omega, \gamma > 0$ and $\eta_t \overset{iid}{\sim} \mathrm{N}(0,1)$. The motivation for this model is as follows. In the standard SV model (3.4) the innovations η_t can be interpreted as shocks to the intensity of the news flow (see Franses and van Dijk 2000). Model (3.11) allows for possible influence of g_t, the (log-)volatility at time t, on the magnitude of such shocks at time $t+1$. High volatility at time t indicates that the markets are turbulent which, in turn, could impact on the flow of news at time $t+1$. The parameter σ_t measures the uncertainty about future volatility, and this uncertainty can be expected to increase if the markets are nervous. The model (3.11) is henceforth labelled *SVVt*; clearly the simpler model *SVt* is nested in it (take $\gamma = 0$).

The nonlinear influence of g_t on the variance of the innovations makes it very challenging to derive necessary and sufficient conditions for second-order stationarity of the *SVVt* model. A Taylor expansion provides two approximate necessary conditions for second-order stationarity of $\{g_t\}$: $|\phi| < 1$ and $\gamma < 2(1-\phi^2)$. Simulation experiments suggest that these conditions provide useful approximations but that the stated range for γ is slightly conservative. This is theoretically unsatisfactory, but fortunately it is straightforward to check for stationarity of the discretized *SVVt* model, i.e. the one obtained after discretization of $\{g_t\}$. The log-volatility process is then a Markov chain with finite state space, and so stationarity holds if the initial distribution of the Markov chain, $\boldsymbol{\delta}^{(1)}$, is such that $\boldsymbol{\delta}^{(1)}\boldsymbol{\Gamma} = \boldsymbol{\delta}^{(1)}$ subject to $\sum_{i=1}^{N} \delta_i^{(1)} = 1$. In this way the discretized *SVVt* model can be checked for stationarity.

Using a Taylor expansion we can obtain the approximate stationary variance of $\{g_t\}$:

$$\sigma_g^2 \approx \frac{\omega + \gamma}{1 - 0.5\gamma - \phi^2}.$$

For the stationary variance and kurtosis of $\{y_t\}$ the approximate expressions (3.9) and (3.10) are again applicable.

GSVt — gamma distributed volatility

All the models presented up to now, both standard and nonstandard, involve Gaussian innovations in the log-volatility process. However, variation of the innovation distribution does not lead to additional difficulties in the model-fitting exercise; the following non-Gaussian alternative can be implemented equally easily by using the HMM method.

Let y_t be defined by

$$y_t = \varepsilon_t \beta \sqrt{g_t + \xi}, \tag{3.12}$$

with ε_t again denoting a Student-t distribution, and, conditional on g_t, let g_{t+1} have a gamma distribution with shape parameter $\kappa = \phi g_t + \lambda$ and scale parameter $\theta = 1$:

$$g_{t+1} \sim \Gamma(\kappa = \phi g_t + \lambda, \theta = 1).$$

The parameters β, ϕ, λ and ξ are all taken to be positive. We refer to this model as *GSVt*. If $\{g_t\}$ is stationary, its stationary mean is

$$\mu_g = \frac{\lambda}{1-\phi},$$

and the corresponding stationary variance is

$$\sigma_g^2 = \frac{\lambda}{(1-\phi)(1-\phi^2)}.$$

Provided $\{g_t\}$ is stationary, one obtains

$$\operatorname{var}(y_t) = \beta^2 \big(\mu_g + \xi\big) \frac{\nu}{\nu - 2}$$

and

$$\operatorname{kurtosis}(y_t) = 3\,\frac{\nu-2}{\nu-4}\left(1 + \frac{\mu_g}{(\mu_g+\xi)^2(1-\phi^2)}\right).$$

A sufficient condition for stationarity is that $\phi \in [0, 1)$; see Proposition 3 of Grunwald *et al.* (2000).

3.2.3 Model fitting results for a number of return series

3.2.3.1 Model comparisons based on ten series of returns

The HMM approximation method for SSMs, in this case in particular for SV models, was applied to model the daily returns for ten stocks on the New York Stock Exchange, namely Sony Corporation, Time Warner, Toyota Motor Corporation, The Travelers Companies, British Petroleum plc, Royal Dutch Shell plc, Bank of America

Corporation, Citigroup Inc., Deutsche Bank AG and Morgan Stanley. The adjusted closing prices, p_t, for the period 02.01.1997 – 01.03.2010, were downloaded from 'finance.yahoo.com', and the daily returns were computed as $y_t = \log(p_t/p_{t-1})$, $t = 2, ..., T$. Summary statistics of the resulting ten series are given in Table 3.2. Not surprisingly, in view of the recent financial crisis, the sample standard deviations and kurtoses are high for stocks in the financial sector. (See also Figure 3.2.)

Table 3.2: *Summary statistics for the daily returns of ten stocks on the New York Stock Exchange for the period 02.01.1997 – 01.03.2010.*

	T	min.	max.	std. dev.	kurtosis
Sony	3304	−0.155	0.169	0.024	7.8
Time Warner	3310	−0.188	0.165	0.031	7.6
Toyota	3304	−0.181	0.133	0.020	8.7
Trav. Comp.	3304	−0.200	0.228	0.022	14.2
BP	3310	−0.122	0.147	0.018	9.7
Roy. D. Sh.	3303	−0.121	0.161	0.019	9.4
Bank of Am.	3310	−0.342	0.302	0.033	26.8
Citigroup	3310	−0.495	0.457	0.036	36.2
Deu. Bank	3293	−0.210	0.222	0.028	13.1
Morgan St.	3310	−0.299	0.626	0.036	41.2

The standard models SV_0 and SVt, as well as the four nonstandard SV models covered in Section 3.2.2, were fitted to each of the ten series. The maximum likelihood estimates are given in Tables A.1 – A.6 in Appendix A5. Several things are noteworthy regarding the parameter estimates (including some that are not given in the tables).

- It is striking that, for all series, one of the AR(1) components of the $SVMt$ model is nonstationary, i.e. has $\phi > 1$, although the mixture (3.7), and hence also the observed process, is stationary.
- With a single exception, the estimates of the parameter ξ, which constitutes a lower bound on the volatility, are all well above zero. (Fitting $SVMt$ to the Citigroup series yielded $\hat{\xi} \approx 0$.) This is an indication that the inclusion of a lower bound for the volatility seems worthwhile.
- All models were fitted with both Gaussian and Student-t distributions for ε_t. The latter consistently led to a substantially higher likelihood. The estimates of

the parameter ν, the number of degrees of freedom, range from 7 to 23 across all series and models. This generalization of Gaussian SV models appears to be particularly fruitful.

- The diagonal entries of the estimated t.p.m. $\widehat{\mathbf{\Gamma}}^{(\alpha)}$ of the Markov chain $\{\alpha_t\}$ in the *MSSVt* model are usually close to one. (An exception is the estimate 0.643 obtained for the Sony series.) This indicates that the two states, which reflect high and low uncertainty about future volatility, are usually strongly persistent.

Comparing the models in terms of their AIC values, given in Table 3.3, the main results from the model-fitting exercise are as follows:

Table 3.3: *For the model SV_0 the AIC is given. The remaining entries in the table are AIC deviations from the AIC of the SV_0 model for the corresponding series. For example, in the case of Sony, the AIC for SVt is given by $-16179-33=-16212$. Entries displayed in bold font indicate the model with the lowest AIC.*

	SV_0	*SVt*	*SVMt*	*MSSVt*	*SVVt*	*GSVt*
# parameters	3	4	8	8	6	6
Sony	−16179	−33	−31	**−35**	−33	−17
Time Warner	−15472	−29	**−41**	−30	−33	−3
Toyota	−17321	−13	−19	**−22**	−15	6
Trav. Comp.	−17324	−35	−50	**−52**	−46	−27
BP	−18043	−9	−30	**−33**	−28	8
Roy. D. Sh.	−17721	−7	**−32**	−27	−27	18
Bank of Am.	−17080	−29	−47	**−50**	−46	90
Citigroup	−16249	−29	**−53**	−31	−43	111
Deuts. Bank	−15955	−46	−49	**−55**	−51	1
Morgan St.	−14955	−19	**−37**	−33	−35	58

- For each of the ten series, SV_0 is inferior to the models *SVt*, *SVMt*, *MSSVt* and *SVVt*.

- In every case either *SVMt* or *MSSVt* performed best.

- For stocks that were relatively mildly affected by the financial crisis (Sony, Time Warner, Toyota), there is relatively little difference in the performance of the models *SVt*, *SVMt*, *MSSVt* and *SVVt*.
- For stocks that were more strongly affected by the crisis, the nonstandard models *SVMt*, *MSSVt* and *SVVt* outperformed their simpler competitors.
- The *GSVt* model mainly yielded poor fits, and worse than SV_0 in seven cases.

3.2.3.2 Forecast pseudo-residuals for three series

We concentrate now on the analysis of three selected series from Table 3.3, namely the series for Sony Corporation, Morgan Stanley and BP plc. The observation period is 02.01.97 – 01.03.10, as before, but the data are now divided into a calibration and a validation sample:

- *Calibration sample (in-sample period)*: 02.01.97 – 08.08.07,
- *Validation sample (out-of-sample period)*: 09.08.07 – 01.03.10.

The dividing date (09.08.07) has been referred to as the beginning of the current financial crisis (see e.g. Swiss National Bank 2008). That date was chosen in order to assess how well the different SV models would have performed during the crisis, a period of unusually high volatility. The three series, shown in Figure 3.2, were selected to illustrate the behaviour of the models for different types of stocks: one (Morgan Stanley) from a sector that was strongly affected by the crisis, and two (Sony Corporation and BP plc) that were less dramatically affected.

As a first step, each of the six models was fitted to the calibration sample of each series. This was done using the HMM method with $N = 200$, a value that is large enough to ensure that any anomalies that may occur could not be attributed to inaccuracies in the approximation of the likelihood. Then, for each of the 644 observations in the validation sample, the (one-step-ahead forecast) pseudo-residual was computed as follows:

$$r_t = \Phi^{-1}\big(F(y_t \mid y_{t-1}, y_{t-2}, \ldots)\big)\,,$$

where F is the c.d.f. of the one-step-ahead forecast distribution on day $t-1$, i.e. the conditional distribution of the return on day t, given all previous observations. The distribution function F is easy to compute for an HMM; it is given by:

$$F(y_t \mid y_{t-1}, y_{t-2}, \ldots) \approx \sum_{i=1}^{N} \zeta_i F(y_t \mid b_i^*)\,, \tag{3.13}$$

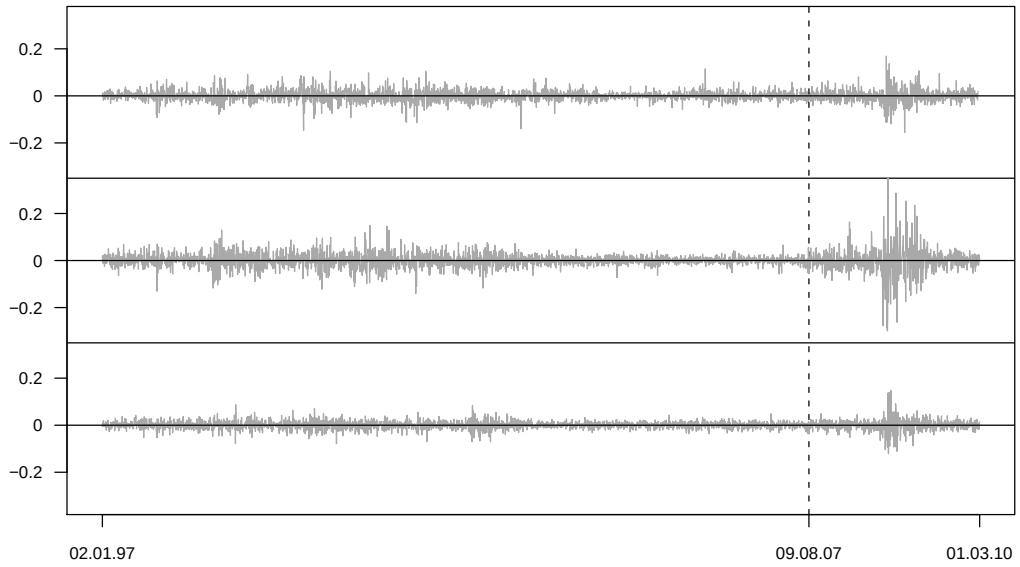

Figure 3.2: *Return series for Sony, Morgan Stanley and BP (from top to bottom). The dashed line shows the boundary between the calibration and validation samples. One observation for Morgan Stanley (0.63 on 13.08.08) fell outside the range of the graph.*

where ζ_i is the ith entry of the vector $\boldsymbol{\alpha}_t\boldsymbol{\Gamma}/(\boldsymbol{\alpha}_t\mathbf{1}^t)$, obtained from the forward probabilities:

$$\boldsymbol{\alpha}_t = \boldsymbol{\delta}\mathbf{P}(y_1)\boldsymbol{\Gamma}\mathbf{P}(y_2)\boldsymbol{\Gamma}\cdots\boldsymbol{\Gamma}\mathbf{P}(y_t)\,.$$

In the context of SV models such residuals were first used by Kim *et al.* (1998). It follows immediately from a result of Rosenblatt (1952) that, if the fitted model is correct, the pseudo-residuals are distributed standard normal. (See also Chapter 6 in Zucchini and MacDonald 2009.) Thus forecast pseudo-residuals can be used to monitor time series; extreme values can be identified, and the continued suitability of the model can be checked by using, for example, qq-plots or formal tests for normality. The qq-plots for the three series investigated here are given in Figures 3.3 – 3.5; the p-values for the Jarque-Bera test are listed in Table 3.4.

The index plots and qq-plots of pseudo-residuals for the Sony Corporation series, displayed in Figure 3.3, show no substantial deviations from normality for any of the six models, except perhaps for the upper tail in the case of SV_0. None of the p-values from the Jarque-Bera tests leads to a rejection of the hypothesis of normality. These findings

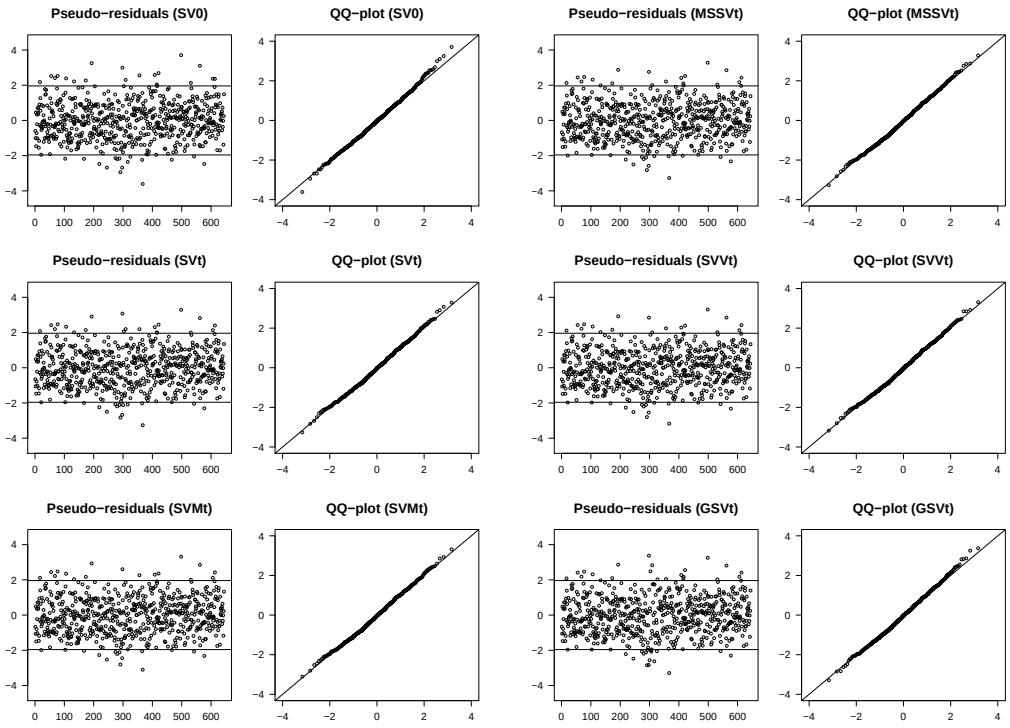

Figure 3.3: *Forecast pseudo-residuals for the Sony series.*

are consistent with the results of the model-fitting exercise, where the likelihood values of the different models turned out not to differ much.

Table 3.4: *p-values of Jarque-Bera tests applied to one-step-ahead forecast pseudo-residuals.*

	SV_0	SVt	$SVMt$	$MSSVt$	$SVVt$	$GSVt$
Sony	0.235	0.323	0.232	0.253	0.262	0.395
Morgan St.	7×10^{-5}	0.052	0.380	0.346	0.807	1×10^{-13}
BP	0.008	0.093	0.218	0.276	0.244	0.001

The pattern changes for the Morgan Stanley series (Figure 3.4). Except for $MSSVt$ and $SVVt$ the qq-plots indicate a lack of fit in the tails; the models were unable to capture the extreme returns that were observed during the financial crisis. The fit of the SV_0 and $GSVt$ models appear to be especially poor, an impression that is confirmed in both cases by the p-values of the Jarque-Bera test. The performance of the SVt model is

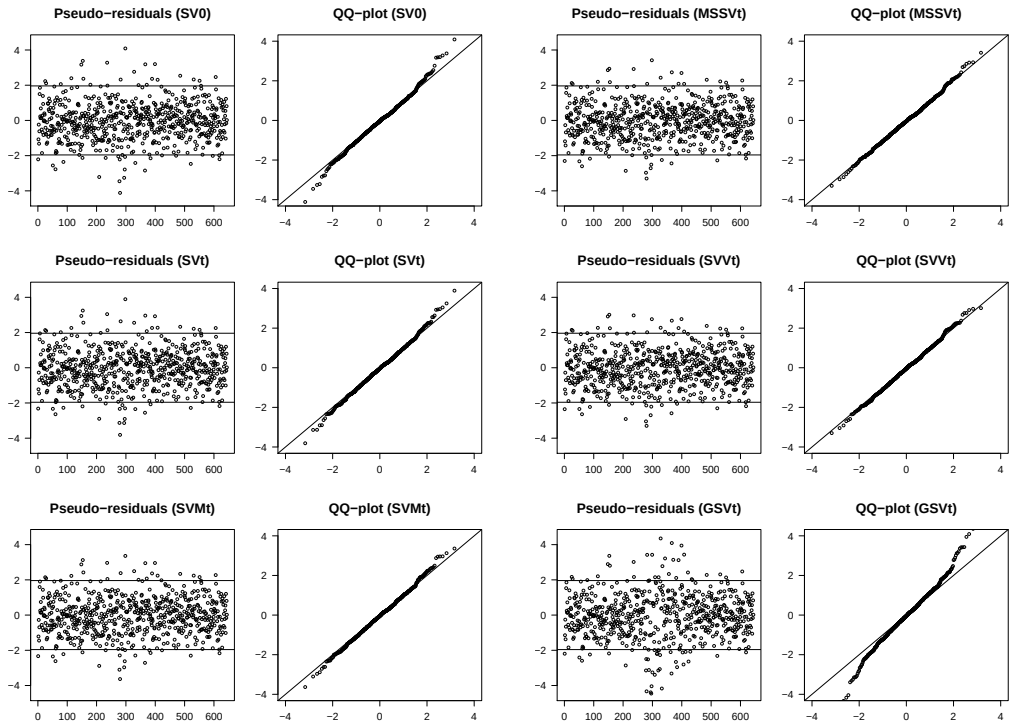

Figure 3.4: *Forecast pseudo-residuals for the Morgan Stanley series.*

somewhat better but still unsatisfactory; at the 10% level of significance, normality of the residuals for this model is also rejected by the Jarque-Bera test. In contrast, the p-values for the nonstandard SV models *SVMt*, *MSSVt* and *SVVt* are well above the conventional significance levels. In particular the results for the six-parameter model *SVVt* are surprisingly good, considering the turbulent behaviour of the Morgan Stanley return during the crisis.

The pseudo-residuals for the BP plc series (Figure 3.5) show tendencies similar to those of the Morgan Stanley series. The models SV_0, *SVt* and *GSVt* show deficiencies in the fits of the tails, whereas the performance of the *SVMt*, *MSSVt* and *SVVt* models is again better in that respect. The p-values of the Jarque-Bera test of normality of the pseudo-residuals for those three models are all above 0.2.

3.2.3.3 Backtesting

The plots and tests discussed in the previous section are useful for assessing the overall fit of a model but, for the purposes of assessing market risk, it is the extreme left tail of the forecast distribution that is of particular interest. It determines the value-

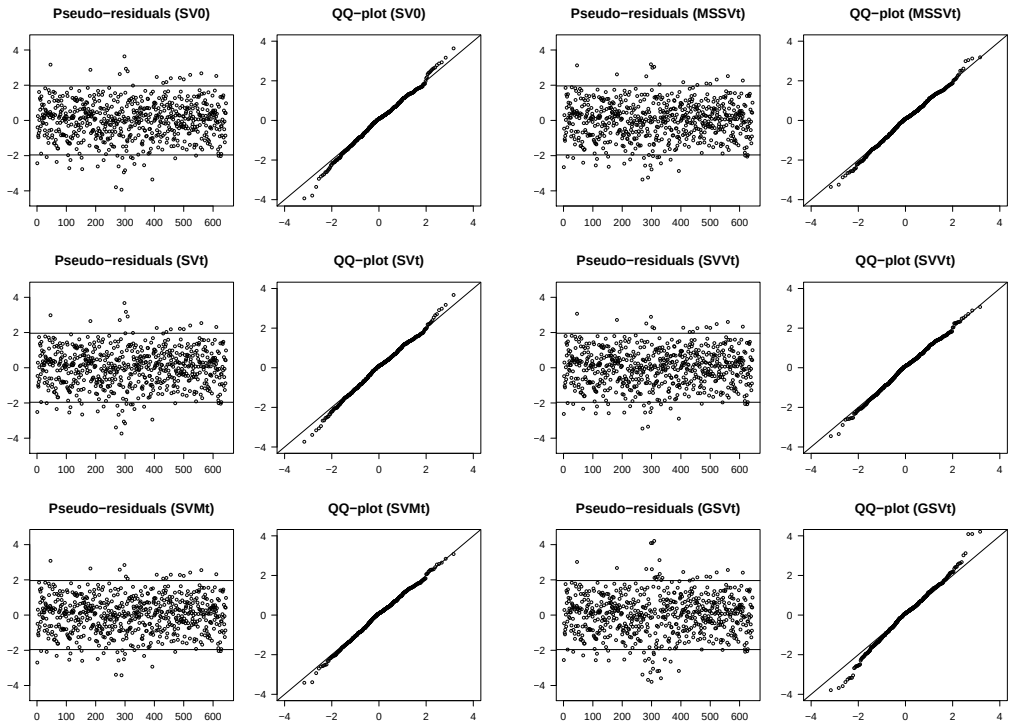

Figure 3.5: *Forecast pseudo-residuals for the BP series.*

at-risk (VaR), defined as the maximum possible loss of a portfolio (over a specified period) at a given confidence level. For example, the one-day $1\% - \text{VaR}$ is computed from the 0.01-quantile of the one-day-ahead forecast distribution. Whenever the return falls below that quantile an *exception* is said to have occurred. If the model used for forecasting is correct, then, using a $100\,\alpha\%-\text{VaR}$, the number of exceptions, X, in n days is binomially distributed with parameters n and α. This distributional result makes it possible to implement *backtesting*, a procedure applied by major central banks and regulatory authorities in terms of the Basel Accords (Basel Committee on Banking Supervision 2006). Specifically (Annex 10a, pp. 310–319), the adequacy of the time series model is assessed by the number of exceptions, X. Three zones are defined:

- *green zone:* X falls below the $95th$ percentile of its distribution, in which case the model is regarded as accurate;
- *red zone:* X falls above its $99.99th$ percentile, in which case the model is regarded as inaccurate;
- *yellow zone:* X falls between the above percentiles, in which case "*the supervisor should encourage a bank to present additional information about its model before*

taking action [...]" (p. 315).

(For a recent account of backtesting see Wong 2010.) Table 3.5 lists the number of exceptions in $n = 644$ days with $\alpha = 0.01$ for each of the ten return series, and for each model fitted to the calibration samples as outlined above. Backtesting was applied "out-of-sample", using the conditional forecast distribution computed according to (3.13). In each case it is indicated whether the observed counts fell into the green, red or yellow zone.

Table 3.5: *Backtesting: the number of exceptions in* $n = 644$ *out-of-sample daily returns with* $\alpha = 0.01$. *Counts less than 11 fall in the green zone; those above 17 (marked **) fall in the red zone; the remainder (marked *) fall in the yellow zone.*

	SV_0	SVt	$SVMt$	$MSSVt$	$SVVt$	$GSVt$
Sony	7	4	4	5	4	6
Time Warner	5	4	4	3	3	3
Toyota	10	11*	10	9	9	11*
Trav. Comp.	12*	8	8	7	8	14*
BP	13*	13*	9	9	9	18**
Roy. D. Sh.	14*	10	10	9	9	21**
Bank of Am.	19**	13*	11*	14*	10	46**
Citigroup	14*	13*	15*	10	11*	39**
Deuts. Bank	16*	12*	10	13*	11*	31**
Morgan St.	11*	10	9	8	8	21**

Certain aspects of the backtests in Table 3.5 are noteworthy. All models are in the green zone for Sony and Time Warner. For all other series and models the number of exceptions exceeded 6.4, the expected value under the hypothesis that the model is correct. In retrospect this underestimation of the market risk is not surprising in view of the enormous increase in volatility in the out-of-sample period, which had been preceded by a prolonged period of low volatility. (See Figure 3.2.) The highest number of exceptions occurred in the financial sector. Of the models considered, the $SVVt$ model led to the "least poor" results, with two outcomes (Citibank and Deutsche Bank) on the lower boundary of the yellow zone, and all others in the green zone. The models $MSSVt$ and $SVMt$ also led to two outcomes in the yellow zone and eight in the green

zone but, on the whole, the numbers of exceptions were a little higher than for $SVVt$. The next best, in terms of backtesting, was the model SVt, which led to five outcomes in the yellow zone and five in the green zone. The model SV_0 led to even more exceptions, although all but one of the outcomes were in the green or yellow zone. Model $GSVt$, with six outcomes in the red zone, is clearly unsuitable. The results indicate that, for the purposes of assessing market risk, it is possible to improve on SV_0 and SVt.

3.3 Application to earthquake counts

We now consider the time series of earthquake counts that was analysed by Zucchini and MacDonald (2009). It contains annual counts of major earthquakes (worldwide), namely earthquakes of magnitude 7 or higher.

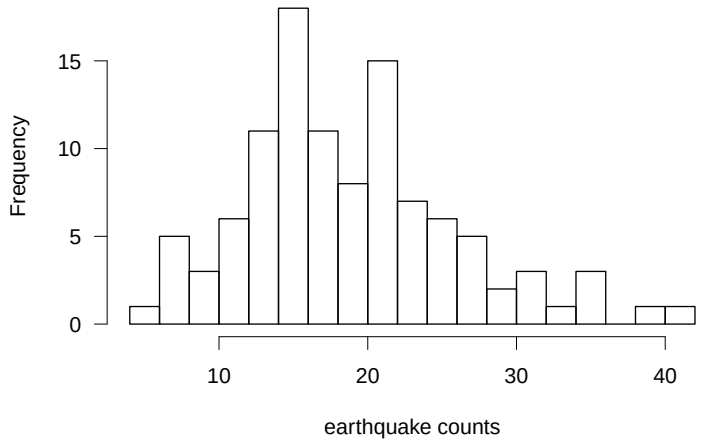

Figure 3.6: *Histogram of earthquake counts.*

Figure 3.6 displays a histogram of the observations; the observations are given in Figure 3.7. The observations are clearly overdispersed: sample mean and sample variance are $\hat{\mu} = 19.36$ and $\hat{\sigma}^2 = 51.57$. Furthermore the series exhibits significant positive autocorrelation (cf. Figure 2.1 in Zucchini and MacDonald 2009). The combination of overdispersion and serial dependence renders SSMs (as well as HMMs) plausible candidates for modelling this series. The standard candidate for modelling unbounded counts is the Poisson distribution.

We consider a *Poisson SSM*, whose time-dependent means, λ_t, for the Poisson random

variable y_t, are assumed to be generated by an AR(1)-process as follows:

$$\log(\lambda_t) - \mu = \phi(\log(\lambda_{t-1}) - \mu) + \sigma\eta_t \,,$$

with $|\phi| < 1$, $\mu, \sigma > 0$ and $\eta_t \overset{iid}{\sim} \mathcal{N}(0,1)$. With $g_t := \log(\lambda_t) - \mu$ and $\beta := \exp(\mu)$ we can write:

$$\begin{aligned} \lambda_t &= \beta \exp(g_t)\,, \\ \text{where} \quad g_t &= \phi g_{t-1} + \sigma\eta_t \,. \end{aligned} \tag{3.14}$$

This model has been previously considered, *inter alia*, by Zeger (1988) and Chan and Ledolter (1995). To write the model in state-space form, consider independent Poisson point processes $N_t(\cdot)$, $t = 1, \ldots, n$, of unit intensity. Then (3.14) can be replaced by

$$y_t = N_t\big((0, \beta \exp(g_t)]\big),$$

where $N_t\big((a,b]\big)$ denotes the number of events of $N_t(\cdot)$ in the time interval $(a,b]$ (cf. Fokianos *et al.* 2009).

To estimate the model parameters we apply the HMM method that was described in Section 3.1. That is, after discretization of the state space, the likelihood of the approximating model is that of an HMM which can be easily maximized numerically. We use the resolution $N = 200$ for the discretization and $-g_{min} = g_{max} = 2$ for the range of g_t-values. (The standard error of the fitted process g_t turns out to be approximately 0.3.) In this example the entries of the matrix $\mathbf{P}(y_t)$ in the approximated likelihood (3.3) are given by

$$f(y_t | g_t = b_i^*) = \exp(-\beta \exp(b_i^*)) \frac{(\beta \exp(b_i^*))^{y_t}}{y_t!} \,.$$

The transition probabilities γ_{ij} are given by (3.2), and we use the stationary distribution for $\boldsymbol{\delta}$. Maximizing (3.3) numerically yields the parameter estimates

$$\widehat{\phi} = 0.89 \,, \quad \widehat{\sigma} = 0.14 \quad \text{and} \quad \widehat{\beta} = 17.8 \,.$$

The minus log likelihood and AIC values obtained for this model are 332.27 and 670.54, respectively. In this example the AIC favours the Poisson SSM rather than the standard Poisson HMMs (the smallest AIC-value, 676.92, for Poisson HMMs is obtained when using three states; cf. Zucchini and MacDonald 2009). Figure 3.7 displays the decoded mean sequences of the Poisson SSM and of the three-state Poisson HMM; both were decoded using the Viterbi algorithm. In the HMM the earthquake rate changes between a finite number of levels. On the other hand in the SSM the transitions follow

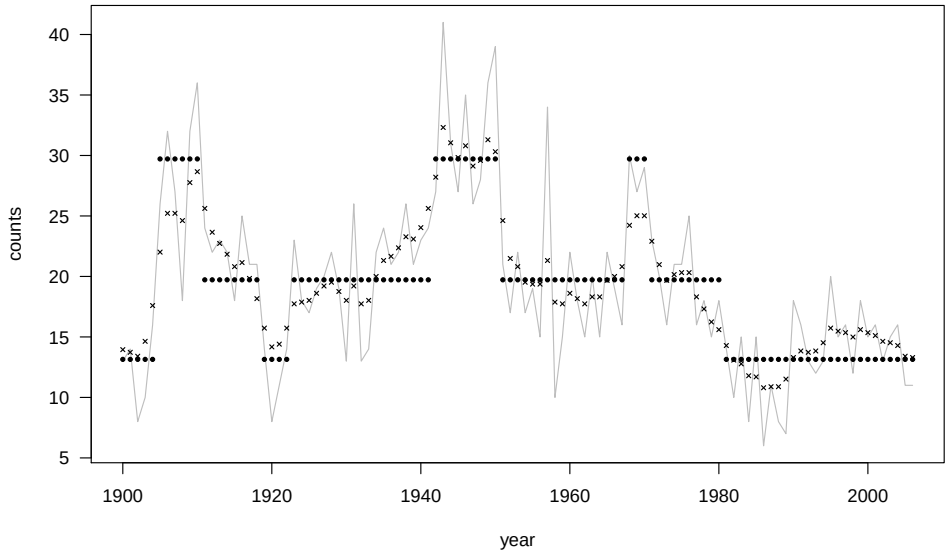

Figure 3.7: *Earthquake counts (solid grey line) and decoded mean sequences of the three-state Poisson HMM (filled circles) and the Poisson SSM (crosses).*

an autoregressive process and thus the rate is continuously distributed. There is no a priori reason to suppose that there is a finite number of rates for the occurrence of major earthquakes.

The proposed model is a basic example for a nonlinear and non-Gaussian SSM. It illustrates the two main ways in which SSMs can overcome possible drawbacks of conventional HMMs. Firstly, SSMs can be more parsimonious in terms of the number of parameters, especially in the hidden part of the model. In the given application the state process of the chosen HMM involves six parameters while that of the SSM involves only two parameters. Secondly, interpretation seems to be easier and more intuitive in case of the fitted SSM.

3.4 Application to polio counts

We now consider the time series of monthly counts of cases of poliomyelitis between January 1970 and December 1983, in the U.S.. The observations are displayed in Figure 3.8. After its initial appearance in Zeger (1988), the time series of polio counts has been analysed, *inter alia*, by Chan and Ledolter (1995), Le Strat and Carrat (1999) and Davis and Rodriguez-Yam (2005). A question of principal interest is whether the

data follow a decreasing time trend. We adopt the nonlinear seasonal Poisson SSM proposed by Zeger (1988), which, in contrast to the Poisson SSM considered in Section 3.3, additionally contains a trend and seasonal components:

$$\begin{aligned} y_t &= N_t\big((0, \beta_t \exp(g_t)]\big)\,, \\ \text{where} \quad g_t &= \phi g_{t-1} + \sigma\eta_t \\ \text{and} \quad \log(\beta_t) &= \mu_1 + \mu_2 \frac{t}{1000} + \mu_3 \cos\left(\frac{2\pi t}{12}\right) + \mu_4 \sin\left(\frac{2\pi t}{12}\right) \\ &\quad + \mu_5 \cos\left(\frac{2\pi t}{6}\right) + \mu_6 \sin\left(\frac{2\pi t}{6}\right)\,, \end{aligned}$$

with $\mu_i \in \mathbb{R}$, $i = 1, \ldots, 6$, and N_t and η_t defined as in Section 3.3. Zeger (1988) estimates the model parameters of this seasonal Poisson SSM using an estimating equation approach. Chan and Ledolter (1995) use a Monte Carlo EM algorithm. Davis and Rodriguez-Yam (2005) approximate the likelihood using a Taylor expansion. We again approximate the likelihood by that of an HMM and then apply numerical maximization. Here we use $N = 200$ as resolution for the discretization, and $-g_{min} = g_{max} = 4$ for the range of g_t-values. (The standard error of the fitted process g_t turns out to be approximately 0.7.) The likelihood components, $\boldsymbol{\delta}$, $\boldsymbol{\Gamma}$, and $\mathbf{P}(y_t)$, are computed as in the earthquake counts example, except that β is replaced by β_t. The main advantage of the HMM method, compared to the methods referred to above, is its simplicity.

Table 3.6: *Estimated model parameters and bootstrap standard errors for the seasonal Poisson SSM, obtained via the HMM approximation method.*

para.	estimate	s.e.
μ_1	0.24	0.29
μ_2	-3.75	3.05
μ_3	0.16	0.15
μ_4	-0.48	0.17
μ_5	0.41	0.13
μ_6	-0.01	0.13
ϕ	0.66	0.19
σ^2	0.27	0.11

Table 3.6 gives the estimated parameters for the seasonal Poisson SSM, as well as (parametric) bootstrap standard errors based on 400 replications. The results are very close to those of Davis and Rodriguez-Yam (2005). (Zeger (1988) as well as Chan and Ledolter (1995) consider different parameterizations of the model.) The estimated trend

is negative and, according to the fitted model, the poliomyelitis rate within each year peaks in November. The values of minus the log likelihood and the AIC are 248.25 and 512.5, respectively. Figure 3.8 displays the observations and the decoded sequence of means.

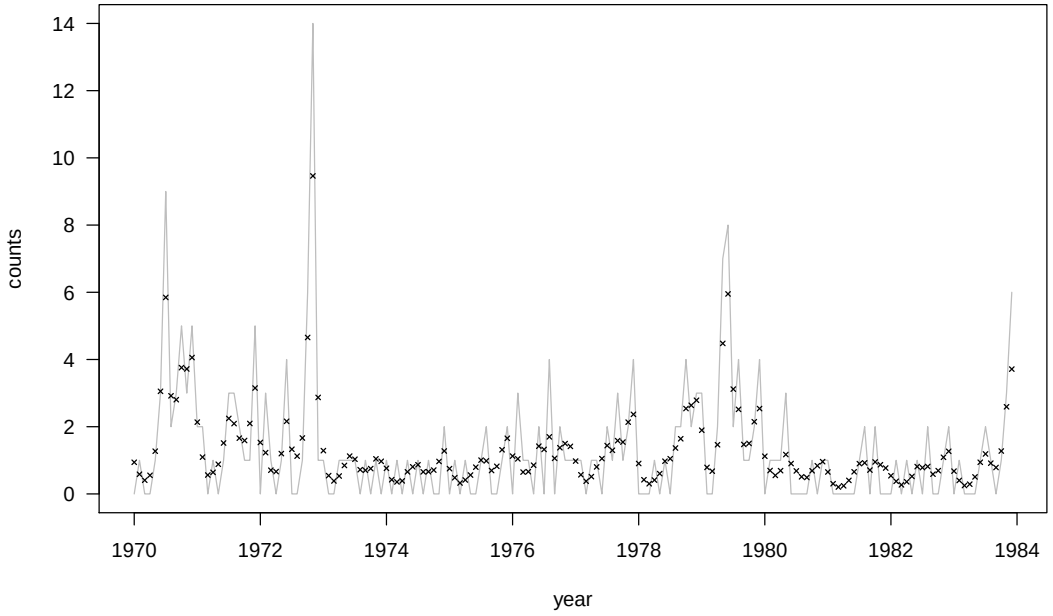

Figure 3.8: *Polio counts (solid line) in the U.S., January 1970 – December 1983, and decoded mean sequence (crosses) of the fitted seasonal Poisson SSM.*

Le Strat and Carrat (1999) fitted Poisson HMMs to the polio count data. For comparison purposes we also do so. We consider seasonal Poisson HMMs with trend: the state-dependent distribution at time t, given state k, is assumed to be a Poisson with parameter $\lambda_{t,k}$ being determined by

$$\log(\lambda_{t,k}) = \mu_{1,k} + \mu_2 \frac{t}{1000} + \mu_3 \cos\left(\frac{2\pi t}{12}\right) + \mu_4 \sin\left(\frac{2\pi t}{12}\right) + \mu_5 \cos\left(\frac{2\pi t}{6}\right) + \mu_6 \sin\left(\frac{2\pi t}{6}\right),$$

with $\mu_{1,k}$, $\mu_i \in \mathbb{R}$, $i = 2, \ldots, 6$. Note that the constant $\mu_{1,k}$ varies across states, whereas the trend and the seasonal components are fixed across states. In principle one could model seasonal components that also vary across states. However, in view of the relatively small sample size the large number of parameters in such a model cannot be estimated reliably.

Table 3.7: *Results of the seasonal Poisson HMM fits to the polio data.*

seasonal HMMs			
# states	*# para.*	*mllk*	*AIC*
1	6	287.27	586.55
2	9	250.01	518.01
3	14	240.46	**508.93**

Table 3.7 gives the resulting minus log likelihood and AIC values for such HMMs with different numbers of states. In terms of the AIC, the HMM with three states performed best. The AIC value of that model is also superior to that of the seasonal Poisson SSM. However, the SSM is determined by fewer parameters, which in view of the relatively small number of observations ($T = 168$) seems preferable.

In this section we demonstrated that the proposed model fitting strategy for SSMs is capable of adjusting for trend and seasonality. We also showed that the estimates produced by the HMM approximation method are in agreement with those produced by other methods. By means of the parametric bootstrap we obtained standard errors for the estimates. Bootstrap is feasible in this application as the analysed time series of polio counts is rather short; the computation of the standard errors took about 1.5 hours. For longer time series the increased effort required to compute standard errors can be a drawback of the HMM approximation method.

3.5 Application to daily rainfall occurrences

Consider again the time series of wet and dry days in Zlatograd that was analysed in Section 2.3. We now demonstrate how SSMs can be applied to this nonhomogeneous binary time series. Similar applications were previously discussed e.g. in Kitagawa (1987) and Czado and Song (2008). We assume that the observations were generated by a Bernoulli random variable whose parameter is driven nonlinearly by a continuous-valued state process, g_t, and a seasonal component, s_t. More precisely, we consider the following seasonal *Bernoulli SSM*:

$$\begin{aligned} y_t &\sim \text{Bern}(\pi_t)\,, \\ \text{where} \quad \log\left(\frac{\pi_t}{1-\pi_t}\right) &= g_t + \beta + s_t\,. \end{aligned}$$

The logit transform is applied to ensure that π_t is in $[0,1]$. The process g_t is assumed to be a Gaussian AR(1), while the seasonality is modelled by trigonometric functions involving the first two harmonics:

$$g_t = \phi g_{t-1} + \sigma\eta_t\,,$$
$$s_t = \mu_1 \cos\left(\frac{2\pi t}{365}\right) + \mu_2 \sin\left(\frac{2\pi t}{365}\right) + \mu_3 \cos\left(\frac{4\pi t}{365}\right) + \mu_4 \sin\left(\frac{4\pi t}{365}\right).$$

The seasonal Bernoulli SSM can be written in state-space form as follows:

$$y_t = \begin{cases} 1 & \text{if } U_t \in [0,\ \text{logit}^{-1}(g_t + \beta + s_t)] \\ 0 & \text{otherwise} \end{cases},$$

with U_t denoting i.i.d. uniformly distributed random variables on $[0,1]$.

To fit the model, we used the HMM method with resolution $N = 400$ and $-g_{min} = g_{max} = 75$. (The standard error of the fitted process g_t turns out to be approximately 20.) The components of the approximated likelihood are computed analogously to those in the previous examples. In particular,

$$f(y_t|g_t = b_i^*) = \left(\text{logit}^{-1}(b_i^* + \beta + s_t)\right)^{y_t} \left(1 - \text{logit}^{-1}(b_i^* + \beta + s_t)\right)^{1-y_t}.$$

The (numerical) maximum likelihood estimates are

$$\widehat{\phi} = 0.49\,, \quad \widehat{\sigma} = 18.1\,, \quad \widehat{\beta} = -11.1\,,$$
$$\widehat{\mu}_1 = -2.47\,, \quad \widehat{\mu}_2 = -3.85\,, \quad \widehat{\mu}_3 = -0.95 \quad \text{and} \quad \widehat{\mu}_4 = -4.10\,.$$

Figure 3.9 displays the fitted seasonal component s_t. It is in agreement with the sample proportions of rainy days per month as displayed in Table 2.4. The AIC value of the seasonal Bernoulli SSM is 18833.07. In terms of this criterion the model performs similarly well as the HMM considered in Section 2.3 (AIC = 18832.25). However, it is inferior to the HSMM selected in Section 2.3 (AIC = 18806.04).

The application to daily rainfall occurrences illustrates how SSMs can be applied to nonhomogeneous binary time series. The proposed model can be easily extended to binomial responses. In this application the estimated standard deviation of g_t is surprisingly high, which underlines the point that the "essential range" of g_t in the discretization manoeuvre has to be chosen with care. Unlike in the proposed SSM, the seasonal components in the HMMs and HSMMs considered in Section 2.3 varied across states (cf. Figure 2.3). As there is no reason to suppose that seasonal fluctuations of "dry" and "wet" states, or more sophisticated weather conditions, follow the same pattern, this may be an advantage of conventional HMMs (and HSMMs) over the proposed type of SSMs.

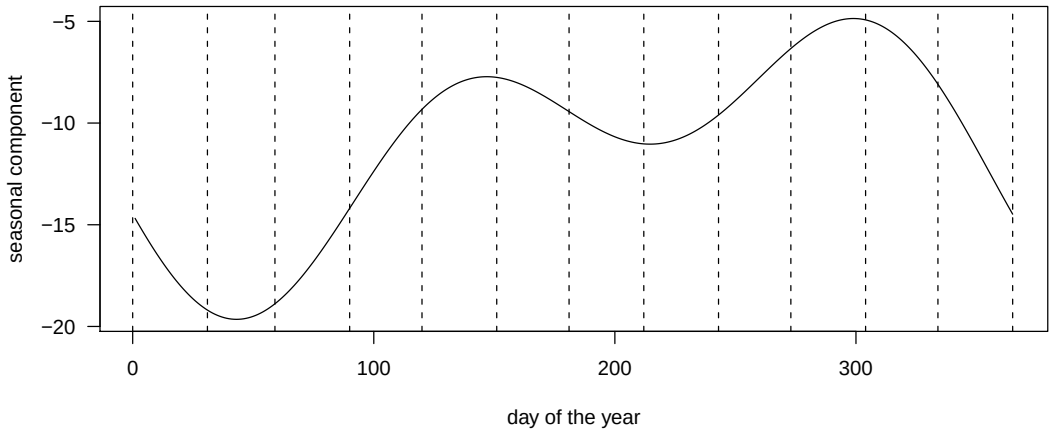

Figure 3.9: *Fitted seasonal component s_t of the seasonal Bernoulli SSM; the dashed lines separate the months.*

3.6 Application to glacial varve thicknesses

We consider the time series of glacial varve thicknesses that was analysed by Shumway and Stoffer (2006). Varves are layers of sediment that are deposited by melting glaciers. Each layer corresponds to the silt and sand deposited over a period of one year. Changes in varve thickness indicate yearly temperature changes, and thus records of varve thicknesses are potentially useful for long-term climate research (cf. Shumway and Stoffer 2006).

The series gives the varve thicknesses (in millimeters) from a location in Massachusetts for 634 years, beginning 11 844 years before present. Figure 3.10 displays the observations. The sample autocorrelation function and a histogram of the observations are displayed in Figures 3.11 and 3.12, respectively. Without recourse to formal definitions, there is some indication of long memory in the sense of a persistent autocorrelation function.

As the observations are necessarily positive, the gamma distribution is a more natural choice for modelling purposes than is the normal distribution. We wish to fit SSMs with continuous state spaces to the varve series. However, as it is not clear whether the state process ideally should influence shape and/or scale parameter of the gamma state-dependent distribution, we first consider HMMs wherein *both* parameters are driven by the states. We begin by considering stationary gamma HMMs, introduced in Section

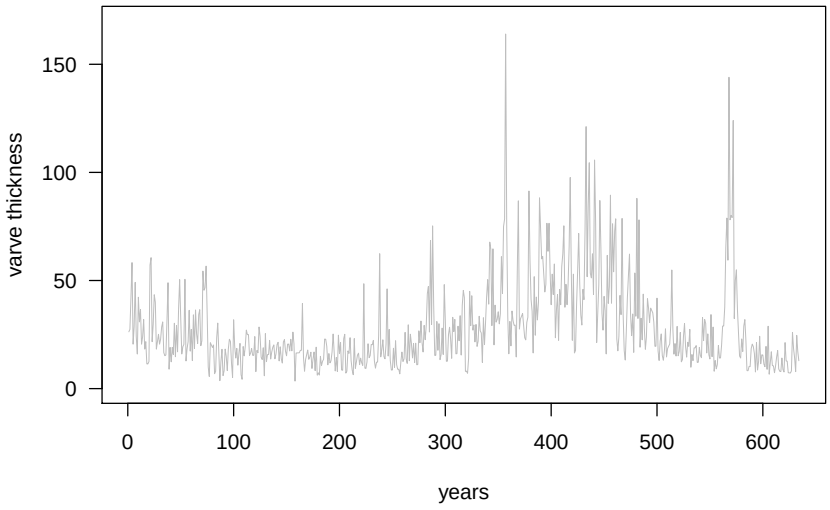

Figure 3.10: *Series of glacial varve thicknesses (in mm).*

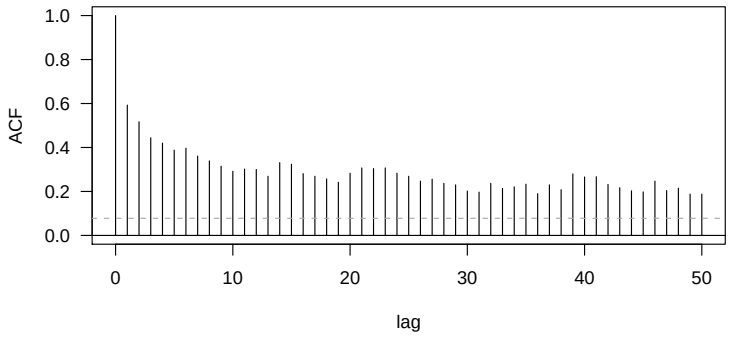

Figure 3.11: *Sample autocorrelation function of the varve time series.*

1.4.2. Table 3.8 summarizes the model fitting results for different numbers of states. The AIC selects the model with three states.

The fitted means, $\widehat{\mu}_n = \widehat{\kappa}_n\widehat{\theta}_n$, standard deviations, $\widehat{\sigma}_n = \widehat{\kappa}_n^{\frac{1}{2}}\widehat{\theta}_n$, and coefficients of variation, $\widehat{c}_n = \widehat{\sigma}_n/\widehat{\mu}_n = \widehat{\kappa}_n^{-\frac{1}{2}}$, for the individual states $n = 1, 2, 3$ in the fitted three-

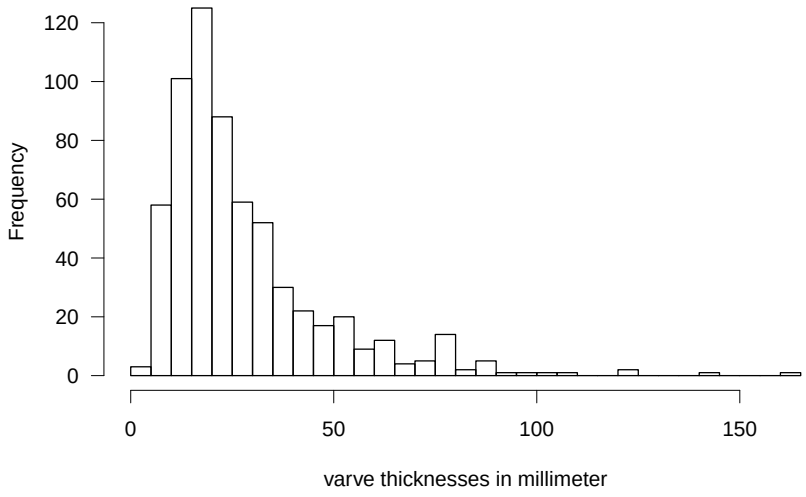

Figure 3.12: *Histogram of the varve thicknesses.*

Table 3.8: *Results of the gamma HMM fits to the varve data.*

# states	*# para.*	*mllk*	*AIC*
2	6	2448.22	4908.43
3	12	2409.48	**4842.96**
4	20	2405.11	4850.22

state gamma HMM are given by

$$\begin{aligned}
\widehat{\mu}_1 &= 16.02\,, \quad \widehat{\sigma}_1 = \ \ 6.72\,, \quad \widehat{c}_1 = 0.42\,,\\
\widehat{\mu}_2 &= 26.94\,, \quad \widehat{\sigma}_2 = 10.29\,, \quad \widehat{c}_2 = 0.38\,,\\
\widehat{\mu}_3 &= 57.99\,, \quad \widehat{\sigma}_3 = 23.09\,, \quad \widehat{c}_3 = 0.40\,.
\end{aligned}$$

Here, as in many similar applications, the standard deviation is approximately proportional to the mean, resulting in a coefficient of variation that is approximately constant. It therefore seems reasonable to reduce the number of parameters accordingly. We thus fit another gamma HMM in which the means, μ_n, vary across states and the coefficient of variation, c_v, is constant. (The shape and scale parameters are then given by $\kappa_n = c_v^{-2}$ and $\theta_n = \mu_n c_v^2$, respectively.) This model has 10 parameters, and the minus log likelihood and AIC values are given by 2410.16 and 4840.31, respectively.

Now that we have seen that the coefficient of variation can be assumed constant, the

following *gamma SSM* seems a natural choice:

$$\begin{aligned} & y_t = \epsilon_t \beta \exp(g_t)\,, \\ \text{where} \quad & g_t = \phi g_{t-1} + \sigma \eta_t \\ \text{and} \quad & \epsilon_t \overset{iid}{\sim} \Gamma\big(\kappa = c_v^{-2}, \theta = c_v^2\big)\,. \end{aligned}$$

(As in the previous applications $\eta_t \overset{iid}{\sim} \mathcal{N}(0,1)$.) Note that $\mathbb{E}\epsilon_t = \kappa\theta = 1$ and $\mathbb{E}(y_t|g_t) = \beta\exp(g_t)$, i.e. the mean of the gamma distributed random variable y_t is driven nonlinearly by the autoregressive process g_t. The coefficient of variation of y_t, given g_t, is constant:

$$\frac{\sqrt{\mathrm{var}(y_t|g_t)}}{\mathbb{E}(y_t|g_t)} = \sqrt{\mathrm{var}\,(\epsilon_t)} = \sqrt{\kappa\theta^2} = c_v\,.$$

In the approximate likelihood (3.3), $\boldsymbol{\delta}$ and $\boldsymbol{\Gamma}$ are computed as in the previous examples. In this example $f(y_t|g_t = b_i^*)$ is the p.d.f. of the gamma distribution with parameters $\kappa = c_v^{-2}$ (shape) and $\theta = \beta\exp(b_i^*)c_v^2$ (scale), evaluated at y_t. Fitting the gamma SSM with the HMM method, with $N = 200$ and $-g_{min} = g_{max} = 3$, yields the following parameter estimates:

$$\widehat{\beta} = 24.4\,, \quad \widehat{\sigma} = 0.15\,, \quad \widehat{\phi} = 0.95\,, \quad \widehat{c_v} = 0.40\,.$$

The minus log likelihood and AIC values of the gamma SSM are given by 2414.96 and 4837.93, respectively. In view of the relatively small number of parameters, the fit is surprisingly good. In particular, the AIC value is smaller than that of the selected HMM counterpart. However, it turns out the gamma SSM does not fully capture the autocorrelation structure of the series. Figure 3.13 displays the sample autocorrelation function for a series of length 1 000 000 that was simulated from the fitted gamma SSM; compared to the original varve series (Figure 3.11), the decay of the autocorrelation is too fast.

Allowing for more flexibility in the state process may be one way to overcome this limitation. Thus, consider a gamma SSM as above but wherein the innovations in the state process are mixtures of two normal distributions:

$$g_t = \phi g_{t-1} + (Z_t\sigma_1 + (1 - Z_t)\sigma_2)\eta_t\,,$$

where Z_t denotes a sequence of i.i.d. Bernoulli(α) random variables. This model, with y_t and ϵ_t as above, will be termed *gamma mixture SSM*.

The parameter estimates for the gamma mixture SSM, obtained via the HMM method with $N = 200$ and $-g_{min} = g_{max} = 3$, are given by

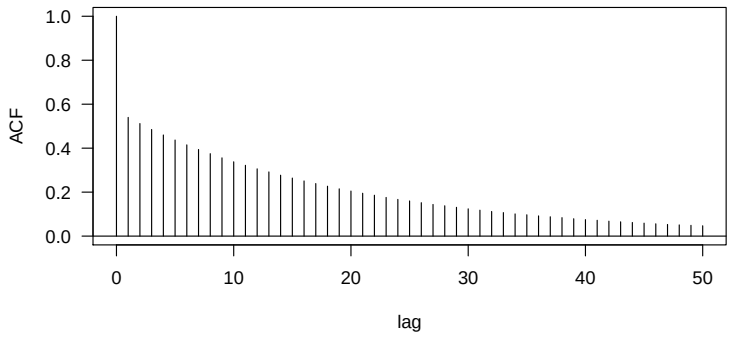

Figure 3.13: *Large sample autocorrelation function of the fitted gamma SSM.*

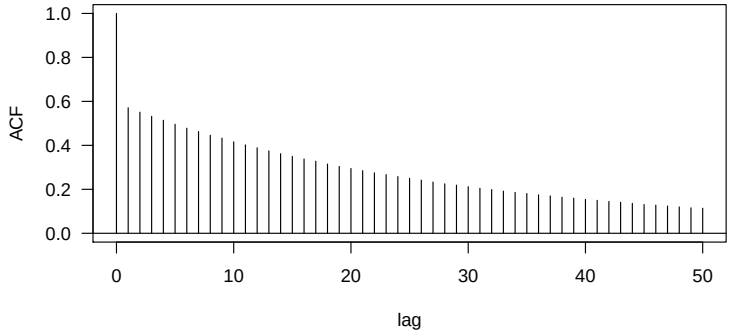

Figure 3.14: *Large sample autocorrelation function of the fitted gamma mixture SSM.*

$$\widehat{\beta} = 27.0\,, \quad \widehat{\boldsymbol{\sigma}} = (0.00, 0.38)\,, \quad \widehat{\phi} = 0.97\,, \quad \widehat{c_v} = 0.41\,, \quad \widehat{\alpha} = 0.90\,.$$

Minus log likelihood and AIC are given by 2409.04 and 4830.09 respectively. In terms of the AIC, the gamma mixture SSM is preferable to both the gamma SSM and the selected gamma HMM. Figure 3.14 shows the autocorrelation function of the fitted gamma mixture SSM as obtained from a simulated series of length 1 000 000. Evidently the model accounts better for the long memory of the series than does the gamma SSM considered before.

While the models appear very similar at first glance, and even though the gamma SSM is nested in the more flexible gamma mixture SSM, the two models require very different interpretations. This is due to the fact that σ_1 was estimated as (approximately) zero in the gamma mixture SSM. As a consequence the decoded underlying state sequences

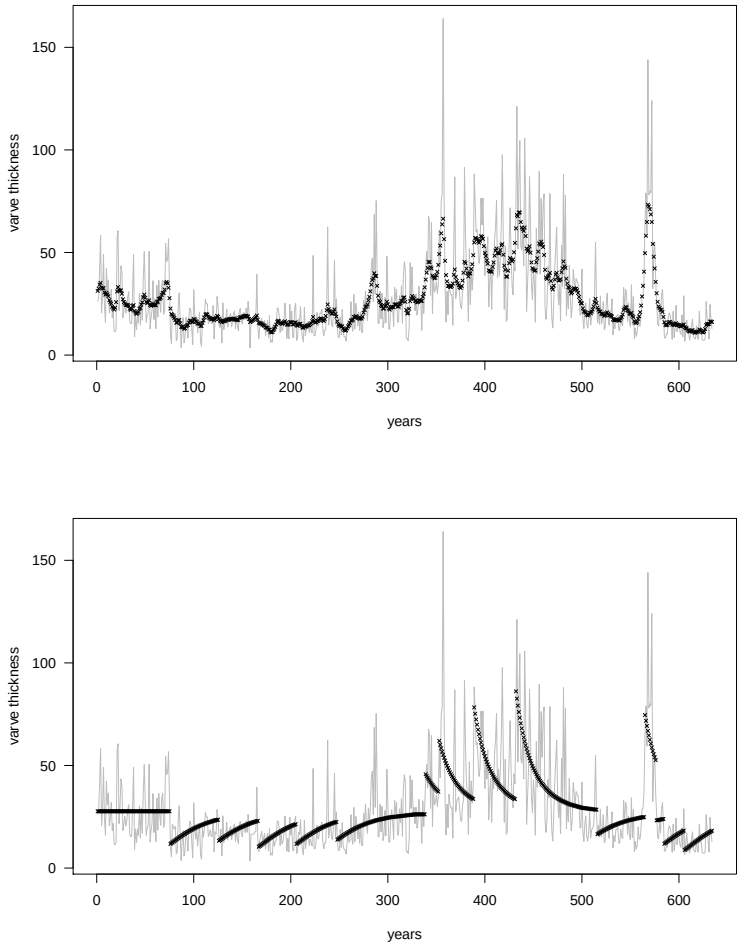

Figure 3.15: *Series of glacial varve thicknesses (solid grey lines), decoded mean sequence of the fitted gamma SSM (crosses in the top plot), and decoded mean sequence of the fitted gamma mixture SSM (crosses in the bottom plot).*

have very different patterns; see Figure 3.15.

The application to glacial varve thicknesses involves continuous observations that are strictly positive. It illustrates the flexibility of state-space modelling, and of the proposed estimation method via structured HMMs, in several ways. Firstly, distributions other than the normal can be fitted without difficulty. Secondly, the quite general structure of SSMs allows for adjustments of the model to peculiarities of the data; in the given application the assumption of a constant coefficient of variation seemed reasonable. Lastly, it was illustrated that models other than the standard AR(1) may be

worthwhile to consider for the state process.

3.7 Concluding remarks

Conventional HMMs have a *finite* number of states. The number of parameters in the state process increases quadratically in the number of states. In practice this limits the number of states one can use. In many applications, e.g. the ones considered in this chapter, there is no a priori reason to suppose that the number of states should be small, or even finite. On the other hand SSMs with continuous state spaces usually depend on only few parameters, and thus overcome this problem. However, they are demanding to fit, both in terms of the effort needed to code the software and in terms of computing time.

In this chapter we have demonstrated how it is possible to approximate the likelihood of an SSM by that of a suitably structured HMM. The proposed approximation method has the important advantage that it is easy to implement. Unlike in the case of SSMs, the likelihood of structured HMMs is easy to compute; numerical maximization thus is feasible. That makes it possible to experiment with variations of models with relatively little programming effort. The applications in this chapter illustrate the flexibility of the models and the methodology, as well as the wide range of potential fields of application.

4 Population hidden Markov models for sleep EEG data[1]

In this chapter we introduce extensions of hidden Markov models for the analysis of the Fourier power spectrum of the electroencephalogram (EEG) during sleep. The two key accomplishments are as follows: first, an easily implemented extension of HMMs for populations of time series and second, a detailed HMM analysis of EEG data recorded during a full-montage sleep study. In this analysis, parameters from the population model for a well-matched subset of subjects with and without sleep disordered breathing (SDB) are compared. SDB is a chronic condition whereby subjects have repeated either complete (apneas) or partial (hypopneas) collapses of the upper airway during sleep. SDB has been shown to have a number of health consequences such as daytime sleepiness, increased risk for motor vehicle accidents, incident hypertension, cardiovascular disease, stroke, all of which can cause mortality (cf. Punjabi *et al.* 2009). We develop a benchmark method for the application of HMMs in complex epidemiological studies and illustrate the method on a unique data set created to study an important public health issue.

The application under study involves the potential correlation of SDB with cortical brain activity. Human sleep and its physiological and health correlates comprise extremely complex biological phenomena. Rather than being simply inert, sleep is highly dynamic, involving changes in physical and neural activity. In children, sleep has been shown to be instrumental in physical and cognitive development. Sleep is also crucial for memory consolidation and immune system repair. Research in sleep continues to unravel the crucial role that sleep plays in health and well being (cf. Gami *et al.* 2005). Electrophysiological measures are an objective means to characterize the electrical activity of the brain during sleep. The electroencephalogram, along with a battery of other biological signals, is collected as part of the overnight polysomnogram (PSG). In clinical and research settings the PSG is used to characterize sleep quality and to assess the presence of various disorders, such as SDB. We focus entirely on the EEG signal, and particularly on banded frequency components of its power spectrum. Such bands

[1]This chapter is based on Langrock, Swihart, Caffo, Crainiceanu and Punjabi (2010).

are crucial for understanding the overnight dynamics of sleep brain activity and any possible alternation due to disease or behaviour. Analyses investigating correlates of sleep-EEG spectra include Crainiceanu *et al.* (2009), Di *et al.* (2009) and Zhang *et al.* (2008).

The application of HMMs to EEG spectrum data is natural, as sleep in humans and many other species is often characterized by sleep states. In humans, the biological signals of the PSG are used to visually classify sleep into light sleep (Stage I and II), deep sleep (slow wave sleep) and rapid eye movement (REM) sleep. The resulting sleep hypnograms, which are single discrete-time, discrete-state processes, have been well studied in the clinical/medical and statistical literature (see e.g. Yassouridis *et al.* 1999, Kneib and Hennerfeind 2008). The present investigation does not focus on visually classified sleep stage data (i.e. hypnograms), other than as partial motivation for using latent states via HMMs to study sleep EEGs. Hence our use of the term "state" always refers to latent nominal classifications estimated via HMMs, not realizations of sleep stages as used in the clinical/medical literature and hypnogram data. Further motivation for HMMs in this setting is given by the fact that EEG spectra show high autocorrelation, which HMMs elegantly address. The general benefits of using HMMs in the context of EEG classification have been discussed by Penny and Roberts (1998) and Zhong and Ghosh (2002). In the context of sleep staging, such an approach has been taken previously in Flexer *et al.* (2005) and Doroshenkov *et al.* (2007), however, with different objectives than those of the current investigation. Those authors essentially try to replicate the hypnogram via automated scoring, which is not at all what we are aiming at. Instead, our model represents an alternative way for summarizing sleep dynamics, in particular for populations of EEG time series.

We analyse the Fourier power spectrum of the sleep EEG signal. As brain activity during sleep is highly non-stationary, we consider the Fourier transform in thirty-second bins within which stationarity can be reasonably assumed. We further summarize the Fourier transform by considering the power in bands of the spectrum. We employ the standard four bands that are typically used in EEG research. When training the population HMMs, we face large amounts of data and numbers of model parameters that are to be estimated. For this reason, unlike other approaches to analysing the raw EEG data, we achieve a great deal of data reduction via the preprocessing of the raw EEG signal into spectral bands, thereby simultaneously focusing on the core components of the signal of interest and greatly alleviating the computational burden.

The considered sleep EEG data set is described in detail in Section 4.1. In Section 4.2, the model is introduced and its estimation is discussed. The results of fitting the model to the EEG data are given in Section 4.3. Concluding remarks are given in Section 4.4.

4.1 Description of the sleep EEG data[2]

The Sleep Heart Health Study (SHHS) is a landmark study of sleep, sleep disorders and their cardiovascular correlates (Quan *et al.* 1997). In this study, over six thousand subjects underwent in-home polysomnography with measurements of the EEG during sleep. Approximately four thousand subjects had a repeat polysomnogram four years after the baseline sleep study. In this analysis, we restrict ourselves to 102 carefully matched subjects with and without SDB.

Matching is appealing, as the data are observational and epidemiologic confounding of the disease effect is of concern. The number of subjects in the SHHS dataset allow for well populated, well selected sub-groups for the desired comparisons. To assess the independent effects of SDB on sleep structure, strict exclusion criteria were employed and included prevalent cardiovascular disease, hypertension, chronic obstructive pulmonary disease, asthma, coronary heart disease, history of stroke, and smoking. For the purpose of this analysis we examine subjects with moderate to severe SDB as assessed by a respiratory disturbance index (RDI) of at least 30 events/hour. Subjects without SDB were identified as those with an RDI < 5 events/hour. Propensity score matching was utilized to balance the SDB and non-SDB groups on demographic factors and to minimize confounding (Rosenbaum and Rubin 1983). Subjects with SDB were matched with those without SDB on the factors of age, body mass index (BMI), race and sex. Race and sex were exactly matched, while age and BMI were matched using the nearest neighbour Mahalanobis technique, so that matches had to be within a Mahalanobis distance (caliper) of 0.1, with multiple matches within the caliper being settled by random selection (Ho *et al.*, forthcoming).

The resultant match was 51 pairs that met the strict inclusion criteria outlined above and exhibiting very low standardized biases. Table 4.1 gives the summary statistics for the group of individuals with SDB and the control group. All measures are not significantly different (RDI is different by design).

The sleep EEG was processed in Matlab (Mathworks) as follows. Separately, for each of two nodes per subject, the signal was partitioned into non-overlapping 30 second bins. The fast Fourier transform was applied to each bin. Band pass filters were applied to separate the signal into four bands: δ (up to 4 Hz), θ (4 – 7 Hz), α (8 – 12 Hz) and β (12 – 30 Hz). The Fourier coefficients were squared and summed to obtain the spectral power within each band. These terms were normalized by dividing by the total power, resulting in a proportion of the total power represented in each band.

[2]This section was written mainly by B. S. Caffo and B. J. Swihart.

Table 4.1: Demographic covariates and sleep variables, means of the two groups.

Variable	SDB	no-SDB	p-value
RDI ($events/hour$)	40.5	2.1	0.000
BMI (kg/m^2)	30.3	30.2	0.972
Age ($years$)	61.8	61.8	1.000
Race (% $white$)	92.2	92.2	1.000
Sex (% $male$)	66.7	66.7	1.000
Total Sleep Time ($min.$)	351	357	0.593
% Total Sleep Time asleep	81.9	83.4	0.743

Thus each observation is a point on the simplex for each 30 second epoch. Normalizing the spectrum was performed for a variety of reasons, including alleviating inter-subject variability.

4.2 Model description

4.2.1 Introducing the population HMM

For each time instant t, the vector of observations is an element of the unit 4-simplex

$$\Delta_4 = \{(x_1, x_2, x_3, x_4) \mid x_i \geq 0, \sum_i x_i = 1\} \subset \mathbb{R}^4 \,.$$

Here x_1, x_2, x_3 and x_4 represent the proportions of the δ-, θ-, α- and β-waves, respectively, as obtained from the fast Fourier transforms of the EEG data. They sum to one as the raw power was normalized by dividing individual band power by the sum of power over the δ, θ, α and β power bands. The Dirichlet distribution $\mathcal{D}(\boldsymbol{\lambda})$, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3, \lambda_4) \in \mathbb{R}^4_{\geq 0}$, with density

$$f_{\boldsymbol{\lambda}}(\boldsymbol{x}) = f_{\boldsymbol{\lambda}}(x_1, x_2, x_3, x_4) = \frac{\Gamma(\sum_{i=1}^4 \lambda_i)}{\prod_{i=1}^4 \Gamma(\lambda_i)} x_1^{\lambda_1 - 1} x_2^{\lambda_2 - 1} x_3^{\lambda_3 - 1} x_4^{\lambda_4 - 1} \,,$$

is a convenient and flexible model to describe random samples from Δ_4. In particular, $\mathcal{D}(\boldsymbol{\lambda})$ is a member of the exponential family, has finite dimensional sufficient statistics and is conjugate prior to the multinomial distribution (Blei *et al.* 2003). Expectation and standard deviations of the marginal distributions of $\mathcal{D}(\boldsymbol{\lambda})$ are $\mu_i := \mathbb{E}(x_i) = \frac{\lambda_i}{s}$ and $\sigma_i^2 := \mathrm{Var}(x_i) = \frac{\mu_i(1-\mu_i)}{s+1}$, where $s := \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4$. Thus a Dirichlet distribution with a fixed mean $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4)$ can account for different levels of variability using

a parameterization of the type $\mathcal{D}(c \cdot \boldsymbol{\lambda})$ and varying the positive scalar c. The parameter s — sometimes called *concentration parameter* — is a measure of how concentrated the distribution $\mathcal{D}(\boldsymbol{\lambda})$ is around its mean; the larger s, the less dispersed are the observed values.

To model the EEG spectral data using Dirichlet distributions, consider a *Dirichlet HMM* with N-state homogeneous Markov chain $\{S_t\}_{t=1,2,\ldots}$ and state-dependent process $\{\boldsymbol{X}_t\}_{t=1,2,\ldots}$. As before, we summarize the probabilities of state switches in the $(N \times N)$-transition probability matrix given by $\boldsymbol{\Gamma} = \{\gamma_{ij}\}$, $i, j = 1, \ldots, N$, where $\gamma_{ij} = \mathbb{P}(S_{t+1} = j \mid S_t = i)$. For the given time series of spectral band powers we use $\{\mathcal{D}(\boldsymbol{\lambda}) \,|\, \boldsymbol{\lambda} \in \mathbb{R}^4_{\geq 0}\}$ as approximating family of distributions for the state-dependent process $\{\boldsymbol{X}_t\}_{t=1,2,\ldots}$. We then have N different Dirichlet distributions $\mathcal{D}(\boldsymbol{\lambda}^{(n)})$, $n = 1, \ldots, N$ — one for each state of the Markov chain — and the current state of the Markov chain determines which of these distributions is selected:

$$\boldsymbol{X}_t \,|\, S_t = n \;\sim\; \mathcal{D}(\boldsymbol{\lambda}^{(n)})$$

We want to employ HMMs to analyse and quantify the stochastic properties of the trajectory of the EEG spectral data during sleep, in particular with regard to underlying state processes. Furthermore, we want to compare these properties between groups of subjects with and without SDB. Fitting a separate HMM to each individual would substantially limit our ability to compare results across subjects or groups of subjects because: 1) the estimated HMMs may not be synchronized across subjects (if both state-dependent distributions and underlying Markov chains vary across subjects, then the HMMs essentially are incommensurable) and 2) the Dirichlet distribution parameters, and thus the EEG states, may have different interpretations. Thus our methods differ substantially from standard HMM methods where interest usually centres on estimating and quantifying the HMM underlying an *individual* time series. Here we are concerned with populations of time series and differences across individuals. To achieve this we introduce the *population HMM* (PoHMM) which assumes that the Dirichlet parameters are variable across states but *fixed across subjects*, and that the transition probabilities across states are subject-specific. Fixing the state-dependent parameters across individuals enables us to compare different individuals in terms of the different Markov chains that result from the HMM fits. Conditional on a particular state n, the distribution of $\boldsymbol{X}_t$ for different individuals is identical; what differs is the stochastic structure of the succession of states. The main questions of interest regarding the state sequence of an individual include:

i) How much time, on average, is spent in a particular latent EEG state?

ii) What is the expected total number of state changes per hour?

iii) What is the expected number of transitions between particular states?

At the population level these questions focus on differences between individuals and between sub-populations (diseased and non-diseased groups). We return to these questions in the course of the evaluation of the results in Section 4.3.

4.2.2 Parameter estimation for the population HMM

Fitting the PoHMM using numerical maximization of the joint likelihood is infeasible. Indeed, for the 102 subjects selected from the SHHS, the number of parameters in case of stationarity would be $N \cdot (N-1) \cdot 102 + 4 \cdot N$ (the first summand corresponds to the Markov chain parameters, the second summand to the Dirichlet distribution parameters), e.g. 1240 for a basic model with $N = 4$ states. An additional difficulty is the large size of the data set, which contains roughly 170 000 observations, rendering standard fitting approaches infeasible. We also anticipate that the size and complexity of data sets with similar structure will increase dramatically in the future.

To circumvent these problems we consider the following pragmatic two-stage approach to model fitting: in Stage I we calibrate the Dirichlet parameters which will be fixed in Stage II to fit HMMs to all individuals. This strategy partitions the infeasible maximization problem into several relatively simple maximization problems, each of them involving a small number of parameters. The calibration of the Dirichlet parameters in Stage I is carried out by fitting an independent mixture of N Dirichlet distributions to all individuals ($5 \cdot N - 1$ parameters). By first considering independent, rather than dependent mixtures (i.e. HMMs), the complexity of the problem is reduced substantially. Nevertheless, this approach is likely to yield estimators that are not considerably different from those that would have been obtained from the HMM fit considering all parameters simultaneously (cf. Section 4.3.1). In Stage II the Dirichlet parameters are fixed and only the Markov chain parameters, i.e. the entries of the t.p.m. $\mathbf{\Gamma}$, are estimated for each individual ($N \cdot (N-1)$ parameters for each individual). This can now be realized for each individual separately, which drastically decreases the computational complexity.

In Stage I the likelihood to be maximized is a product of probability density functions of mixture distributions:

$$\mathcal{L}(\boldsymbol{\lambda}^{(1)}, \ldots, \boldsymbol{\lambda}^{(N)}, \gamma_1, \ldots, \gamma_N) = \prod_{m=1}^{M} \prod_{i=1}^{2} \prod_{t=1}^{T_{m,i}} \sum_{n=1}^{N} \gamma_n f_{\boldsymbol{\lambda}^{(n)}}(\boldsymbol{x}_t^{(m,i)}) \,, \tag{4.1}$$

where $\boldsymbol{x}_t^{(m,i)}$ denotes the vector of observed proportions of δ-, θ-, α- and β-waves for individual m at time t in night i, and γ_n denotes the mixing probability of the Dirichlet

distribution associated with state n. The number of individuals used to calibrate the Dirichlet parameters is M and, for individual m, the number of observations available in night i is $T_{m,i}$. The observations originate from a set of individuals (see 4.3.2.1 for more details), and for each individual from two separate overnight recordings of the EEG. Clearly the ordering in which the observations appear in the likelihood computation does not play any role. The likelihood given by (4.1) is maximized over the mixing parameters γ_i (with the constraint $\gamma_1 + \ldots + \gamma_N = 1$) and the Dirichlet parameter vectors $\boldsymbol{\lambda}^{(i)} \in \mathbb{R}^4_{\geq 0}$, $i = 1, \ldots, N$.

In Stage II the likelihood to be maximized for individual m is

$$\begin{aligned}\mathcal{L}(\boldsymbol{\Gamma}^{(m)}) =& \boldsymbol{\pi}^{(m)}\mathbf{P}(\boldsymbol{x}_1^{(m,1)})\boldsymbol{\Gamma}^{(m)}\mathbf{P}(\boldsymbol{x}_2^{(m,1)})\boldsymbol{\Gamma}^{(m)} \cdot \ldots \cdot \boldsymbol{\Gamma}^{(m)}\mathbf{P}(\boldsymbol{x}_{T_{m,1}}^{(m,1)})\mathbf{1}^t \\ & \cdot \boldsymbol{\pi}^{(m)}\mathbf{P}(\boldsymbol{x}_1^{(m,2)})\boldsymbol{\Gamma}^{(m)}\mathbf{P}(\boldsymbol{x}_2^{(m,2)})\boldsymbol{\Gamma}^{(m)} \cdot \ldots \cdot \boldsymbol{\Gamma}^{(m)}\mathbf{P}(\boldsymbol{x}_{T_{m,2}}^{(m,2)})\mathbf{1}^t \,, \end{aligned} \tag{4.2}$$

where

$$\mathbf{P}\big(\boldsymbol{x}_t^{(m,i)}\big) := \operatorname{diag}\left(f_{\boldsymbol{\lambda}^{(1)}}(\boldsymbol{x}_t^{(m,i)}), \ldots, f_{\boldsymbol{\lambda}^{(N)}}(\boldsymbol{x}_t^{(m,i)})\right),$$

$\boldsymbol{\Gamma}^{(m)} = (\gamma_{ij}^{(m)})$, $i, j = 1, \ldots, N$, denotes the t.p.m. of the Markov chain for individual m, $\mathbf{1}$ is a row vector of ones and $\boldsymbol{\pi}^{(m)} = ({\pi_1}^{(m)}, \ldots, {\pi_N}^{(m)})$ is the solution to the linear system $\boldsymbol{\pi}^{(m)}\boldsymbol{\Gamma}^{(m)} = \boldsymbol{\pi}^{(m)}$ subject to $\sum_i \pi_i^{(m)} = 1$, i.e. the stationary distribution of the fitted Markov chain[3], associated with individual m. Note that the likelihood given by (4.2) is maximized only over the parameters of the underlying hidden Markov chain of the model; the parameters at the observation level, i.e. the Dirichlet parameters $\boldsymbol{\lambda}^{(i)}$, are fixed at the values obtained in Stage I.

Likelihood maximization in Stage I and II cannot be carried out analytically and hence a numerical maximization algorithm is used instead (cf. Section 1.2 for more details). A question of interest concerns the choice of the number of states N. We discuss this in detail in Section 4.3.2.3.

4.3 Fitting the population HMM to the sleep EEG data

We now fit the PoHMM to sleep EEG data acquired at the SHHS. We begin by looking at a simple three-subject example that is supposed to illustrate the data and to compare the proposed estimation method to the conventional maximum likelihood approach (Section 4.3.1). Subsequently, Section 4.3.2 discusses the model fitting results for the whole set of matched pairs as described in Section 4.1 (i.e. 102 subjects).

[3] In comparison to the previous chapters the notation was slightly modified in order to avoid confusion between stationary distribution and δ-waves.

4.3.1 An illustrative and method-comparative example

Figure 4.1 displays the observed EEG spectral powers of the δ-, θ-, α- and β-bands that were made in the SHHS for three subjects (in each case for two nights). Each time instant t refers to an interval of 30 seconds.

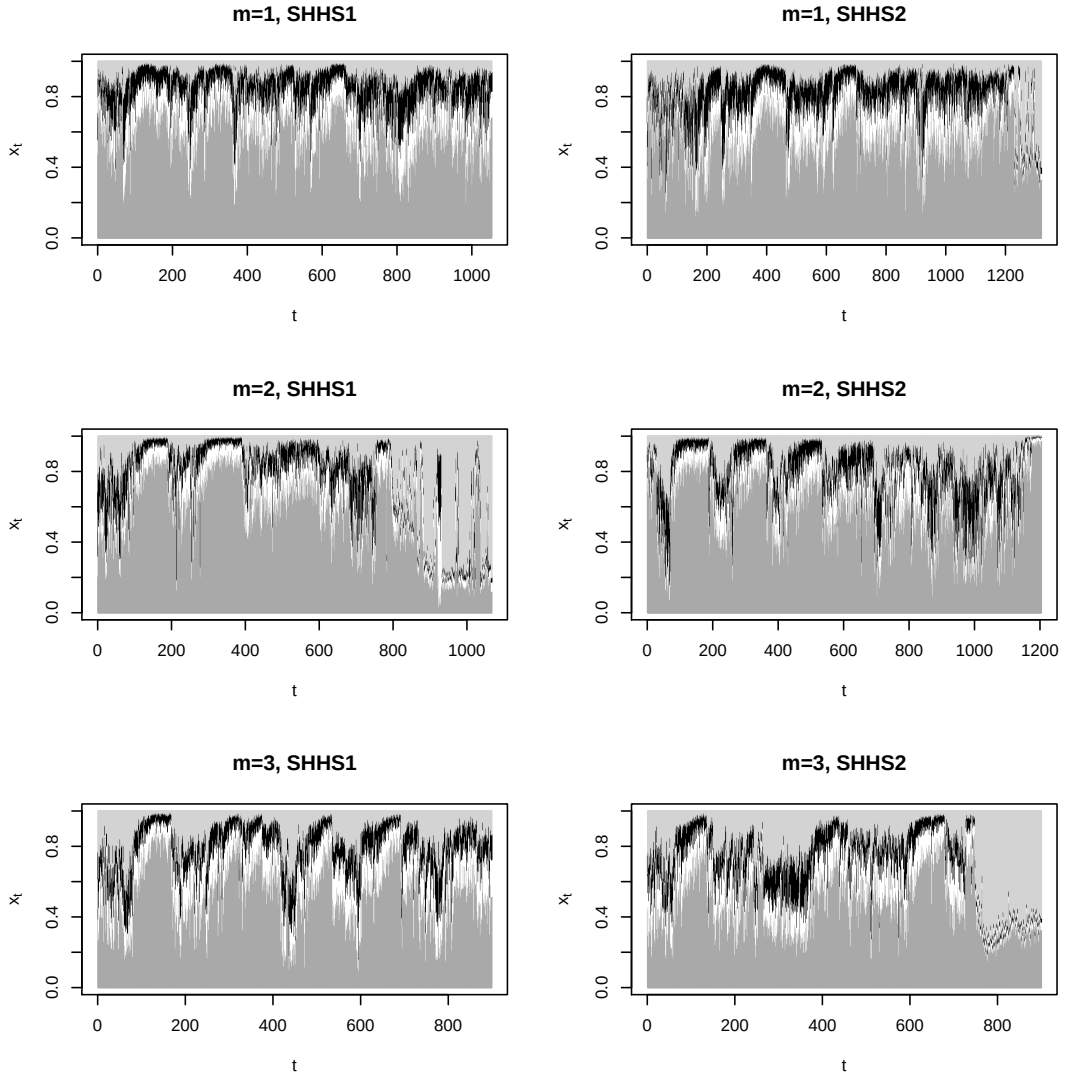

Figure 4.1: *Observations of three subjects acquired at SHHS1 (first night) and SHHS2 (second night); dark grey segments: δ-band spectral power, white segments: θ-band spectral power, black segments: α-band spectral power, light grey segments: β-band spectral power.*

The PoHMM was fit to these data in two different ways, first by maximizing the joint

likelihood and secondly by using the two-stage approach as described in Section 4.2.2. Table 4.2 compares the following: *i)* the computational time that was needed to perform the model fits, *ii)* the estimated Dirichlet parameters $\boldsymbol{\lambda}^{(n)}$ (with n denoting the associated state of the Markov chain), *iii)* the associated expected spectral band powers $\boldsymbol{\mu}^{(n)}$ and *iv)* the stationary Markov chain distributions for the three subjects, i.e. $\boldsymbol{\pi}^{(m)}$, $m = 1, 2, 3$ (where m refers to the subjects).

Table 4.2: *Four-state PoHMM fitted 1) via maximization of the joint likelihood and 2) via the two-stage approach for three subjects.*

	Joint likelihood				Two-stage			
Comp. time (Hrs.)	60.2				5.7			
	Dirichlet parameters				Dirichlet parameters			
	$\lambda_1^{(n)}$	$\lambda_2^{(n)}$	$\lambda_3^{(n)}$	$\lambda_4^{(n)}$	$\lambda_1^{(n)}$	$\lambda_2^{(n)}$	$\lambda_3^{(n)}$	$\lambda_4^{(n)}$
state $n = 1$	44.0	6.4	4.0	2.0	47.2	6.5	4.2	2.2
state $n = 2$	7.8	3.4	3.6	4.4	8.0	3.7	4.0	4.6
state $n = 3$	34.8	11.0	8.4	5.2	38.6	11.8	8.8	5.7
state $n = 4$	12.4	2.5	1.7	31.0	13.2	2.6	1.8	33.3
	Expected band power				Expected band power			
	δ	θ	α	β	δ	θ	α	β
state $n = 1$	0.78	0.11	0.07	0.04	0.79	0.11	0.07	0.04
state $n = 2$	0.41	0.18	0.19	0.23	0.39	0.18	0.20	0.23
state $n = 3$	0.59	0.19	0.14	0.09	0.59	0.18	0.14	0.09
state $n = 4$	0.26	0.05	0.04	0.65	0.26	0.05	0.03	0.65
	Stationary state prob.				Stationary state prob.			
	$\pi_1^{(m)}$	$\pi_2^{(m)}$	$\pi_3^{(m)}$	$\pi_4^{(m)}$	$\pi_1^{(m)}$	$\pi_2^{(m)}$	$\pi_3^{(m)}$	$\pi_4^{(m)}$
subject $m = 1$	0.23	0.22	0.53	0.02	0.22	0.24	0.52	0.02
subject $m = 2$	0.33	0.34	0.26	0.07	0.32	0.34	0.27	0.07
subject $m = 3$	0.14	0.42	0.27	0.16	0.14	0.42	0.28	0.16

As pointed out above, method 1) might theoretically be preferable but is infeasible for large populations. On the other hand, the two-stage method is feasible even for very large populations and, in any case, is substantially faster. What is more, the t.p.m.'s, $\boldsymbol{\Gamma}^{(1)}$, $\boldsymbol{\Gamma}^{(2)}$ and $\boldsymbol{\Gamma}^{(3)}$, of the three subjects were estimated sequentially, but they could have been estimated in parallel on three different processors, which would have substantially reduced the computing time. The fact that the two-stage method can be

implemented in parallel is a potentially enormous advantage in this context. As can be seen in Table 4.2, the two-stage method yields reasonable results in the sense that they are very close to those obtained by joint maximization of the likelihood. In particular, the expected spectral band powers differ by at most 0.02 (δ-waves in state 2). These findings verify that the two-stage-method can result in plausible estimates. It agrees well with joint maximum likelihood in the scenarios where the latter is possible and scales to large epidemiological studies of sleep.

The fit to this small sample illustrates that the fitted hidden Markov chains have significantly different characteristics across subjects. According to the fit, subject 1 in this example spends about 52% of the night in state 3 of the Markov chain, on average, whereas state 2 is the most frequented state by subjects 2 and 3. The subsequent section concentrates on the quantitative analysis of such differences for a large population ($M = 102$) containing 51 healthy subjects paired to 51 sleep apneics.

4.3.2 Results for the whole population

In this section we use the proposed two-stage method to fit the PoHMM to the whole population of matched pairs (i.e. to $M = 102$ subjects). For each individual, the EEG measurements of two nights as given in the SHHS are incorporated. We give the results and interpretations using $N = 5$ states for the Markov chain. (Section 4.3.2.3 discusses this choice and the consequences.) According to the proposed two-stage estimation method the model fitting exercise is split: in Section 4.3.2.1 the Dirichlet parameters are calibrated, and in Section 4.3.2.2 the Markov chain parameters are estimated.

4.3.2.1 Stage I — Calibrating the state-dependent distributions

In this first stage the state-dependent parameters are estimated by fitting an independent mixture of $N = 5$ Dirichlet distributions to the EEG data. The first task is to choose the calibration sample, i.e. the set of individuals to which the mixture is to be fitted. Two natural candidates are the set of all healthy individuals or the set of all individuals, regardless of their disease status. In our opinion the former is a better option because it would seem more relevant to regard the properties of the sleep states of healthy individuals as normative. For the diseased individuals the parameters might differ, and it is of interest to quantify and investigate any such deviance; we thus repeated the model fitting exercise also for the set of all diseased individuals. The EEG recordings from both nights made in the SHHS were taken into account. In Table 4.3 the estimated Dirichlet parameters and the associated expected spectral band powers for the diseased and non-diseased subgroups are displayed.

Table 4.3: *Estimated Dirichlet parameters, associated expected spectral band powers of δ-, θ-, α- and β-waves and concentration parameters for healthy and diseased subgroups.*

Healthy subgroup

state	Diri. parameter vector $\boldsymbol{\lambda}^{(n)}$				δ	θ	α	β	$s^{(n)}$
$n=1$	56.72	7.51	4.37	2.30	0.80	0.11	0.06	0.03	70.90
$n=2$	32.59	10.08	6.95	3.78	0.61	0.19	0.13	0.07	53.40
$n=3$	0.49	0.43	0.41	0.62	0.25	0.22	0.21	0.32	1.95
$n=4$	6.75	3.78	4.22	2.94	0.38	0.21	0.24	0.17	17.69
$n=5$	4.16	1.81	1.32	19.32	0.16	0.07	0.05	0.73	26.62

Diseased subgroup

state	Diri. parameter vector $\boldsymbol{\lambda}^{(n)}$				δ	θ	α	β	$s^{(n)}$
$n=1$	58.00	7.42	4.59	2.43	0.80	0.10	0.06	0.03	72.44
$n=2$	36.53	10.23	7.53	4.18	0.62	0.17	0.13	0.07	58.47
$n=3$	1.33	0.89	1.01	1.00	0.32	0.21	0.24	0.24	4.23
$n=4$	12.89	6.38	6.77	4.43	0.42	0.21	0.22	0.15	30.48
$n=5$	3.32	1.50	1.26	25.81	0.10	0.05	0.04	0.81	31.89

The two groups led to very similar results. The most striking differences between the two fitted models arise in states 3 and 5, which is explicable insofar as these are the least frequented states (see the discussion of the sleep architecture below) and hence are expected to have less stable estimates. It is also interesting to note that the variances themselves are variable for the different states. The lowest concentration parameter $s^{(n)}$ was estimated for state $n = 3$ (1.95 for the healthy and 4.23 for the diseased subgroup). This could be an indication that the make-up of this sleep state differs largely across individuals; hence a model with fixed Dirichlet parameters across individuals would try to capture the heterogeneity of this sleep state by a small concentration parameter s. As the two fits led to similar results it seems reasonable to adopt the parameters estimated from the set of healthy individuals to the whole population, which we do in the subsequent sections.

4.3.2.2 Stage II — Individual state switching probabilities

After fixing the Dirichlet parameters at the values obtained from the calibration fit performed above, we fitted a five-state Dirichlet HMM for each of the 102 individuals.

(Recall that the set of these 102 Dirichlet HMMs with identical state-dependent distributions across individuals is what we call the population HMM.) For each individual, the EEG measurements made in two separate nights are taken into account. Of interest is the stochastic structure of the resulting Markov chains. In what follows the fitted Markov chains are analysed in two different ways. We first discuss the sleep architecture by looking at the stationary distributions of the Markov chains. Subsequently, we analyse the estimated transition probabilities in terms of the resulting expected frequencies of transitions.

Sleep architecture

To gain insight in the results we start by looking at the stationary distributions $\boldsymbol{\pi}^{(m)}$, $m = 1, \ldots, 102$, that give the average proportions of time that the individuals spend in the different states according to the fitted model. In the following the indices $m = 1, \ldots, 51$ correspond to the healthy individuals while those with the indices $m = 52, \ldots, 102$ correspond to the diseased ones (and the matched pairs are $(1, 52)$, $(2, 53)$, $\ldots$, $(51, 102)$). We obtain

$$\bar{\boldsymbol{\pi}}_{healthy} := \frac{1}{51} \sum_{m=1}^{51} \boldsymbol{\pi}^{(m)} \approx (0.241, 0.407, 0.050, 0.270, 0.032)$$

and

$$\bar{\boldsymbol{\pi}}_{diseased} := \frac{1}{51} \sum_{m=52}^{102} \boldsymbol{\pi}^{(m)} \approx (0.223, 0.418, 0.054, 0.279, 0.026)\,.$$

Thus, according to the fit of the PoHMM, the EEG-derived sleep architecture, i.e. the average proportion of time the individuals spend in the different HMM states, is similar for healthy and diseased subjects. This confirms findings of papers dealing with sleep architecture analyses based on the hypnogram (e.g. Swihart *et al.* 2008). The most frequented HMM state is state 2 in which about $40 - 42\%$ of the night is spent. The least frequented HMM state is state 5 in which about 3% of the night is spent.

Apart from considering the population level averages it is interesting to note that the stationary distributions of the individuals show quite high variation. In states 1, 2 and 4 all individuals have stationary probabilities significantly larger than zero (i.e. > 0.01). States 3 and 5 on the other hand are not frequented by all individuals. According to the fit for four individuals (two healthy and two diseased) the stationary probability of being in state 3 is smaller than 10^{-5}. For state 5 this is the case even for 34 individuals (16 healthy and 18 diseased). We emphasize that this does not necessarily mean that these individuals never switch to the corresponding states — they might simply not

have done so on the two nights the observations were recorded.

Expected numbers of transitions

Another way of comparing the Markov chains is to analyse the state transition probabilities. The expected number of transitions of individual m from state i to state j in a series of T observations is obtained as

$$\mathbb{E}\, t_{ij}^{(m)}(T) = (T-1)\pi_i^{(m)}\gamma_{ij}^{(m)}$$

(see Zucchini and MacDonald 2009). Table 4.4 displays the averaged values of the expected numbers of transitions per hour, from state i to state j, for the two groups of interest (healthy and diseased individuals), i.e.

$$\frac{1}{51}\sum_{m=1}^{51}\mathbb{E}\, t_{ij}^{(m)}(120) \quad \text{and} \quad \frac{1}{51}\sum_{m=52}^{102}\mathbb{E}\, t_{ij}^{(m)}(120)\,,$$

for $i, j = 1, 2, 3, 4, 5$.

Table 4.4: *Averaged expected numbers of transitions per hour for healthy and diseased individuals.*

Healthy subgroup

	to state				
from state	1	2	3	4	5
1	23.97	3.44	0.12	1.08	0.01
2	3.84	41.71	0.04	2.89	0.00
3	0.09	0.01	5.51	0.18	0.16
4	0.70	3.32	0.14	27.92	0.01
5	0.02	0.00	0.14	0.02	3.68

Diseased subgroup

	to state				
from state	1	2	3	4	5
1	20.06	5.02	0.11	1.37	0.01
2	5.55	41.23	0.03	2.88	0.00
3	0.08	0.01	6.02	0.28	0.07
4	0.86	3.44	0.23	28.62	0.02
5	0.01	0.00	0.07	0.02	3.02

For transitions from state i to state j, $i \neq j$, we applied the following two-sided t-test to the differences between expected transition numbers for matched pairs:

$$H_0 : \mu_{ij} = 0\,,$$
$$H_A : \mu_{ij} \neq 0\,,$$

where μ_{ij} denotes the expectation (in the sense that the chosen individuals constitute a random sample of persons) of the pairwise differences

$$\mathbb{E}\, t_{ij}^{(m)}(120) - \mathbb{E}\, t_{ij}^{(m+51)}(120),\ m = 1, \ldots, 51\,.$$

At the 5% significance level the null hypothesis is rejected for transitions between states 1 and 2. In this case the expected number of transitions — in both directions — is significantly higher for diseased subjects. It is also rejected for transitions between states 3 and 5. In this case the expected number of transitions — again in both directions — is significantly higher for healthy subjects.

Summing up the off-diagonal elements from the tables yields the averaged expected total numbers of cross-state transitions:

$$\sum_{i,j\in\{1,2,3,4,5\},\, i\neq j} \frac{1}{51} \sum_{m=1}^{51} \mathbb{E}\, t_{ij}^{(m)}(120) = 16.21 \quad \text{(healthy subgroup)}$$

and

$$\sum_{i,j\in\{1,2,3,4,5\},\, i\neq j} \frac{1}{51} \sum_{m=52}^{102} \mathbb{E}\, t_{ij}^{(m)}(120) = 20.05 \quad \text{(diseased subgroup)}.$$

Not considering the group averages, and instead applying a two-sided t-test to the pairwise differences, yields a p-value of 0.014, meaning that the null hypothesis of a zero mean must be rejected at the 5% significance level. Thus the expected total number of cross-state transitions is significantly higher for diseased individuals.

In summary, diseased individuals tend to switch significantly more often between various states. Most of the switches occur between states 1 and 2, followed by the switches between states 2 and 4. The most striking difference between the groups lies in the former case: switches between states 1 and 2 occur about 50% more often in the group of diseased individuals.

Not captured by this analysis of the group averages is the heterogeneity within the groups. Although the difference between the average expected numbers of cross-state transitions is substantial, there are large fluctuations within the groups. This can be seen in Figure 4.2, where histograms and kernel density estimators of the expected

numbers of cross-state transitions per individual, i.e. the values

$$\sum_{i,j\in\{1,2,3,4,5\},\, i\neq j} \mathbb{E}\, t_{ij}^{(m)}(120)\,, \quad m = 1, \ldots, 102,$$

separated in the groups of healthy and diseased individuals, are displayed.

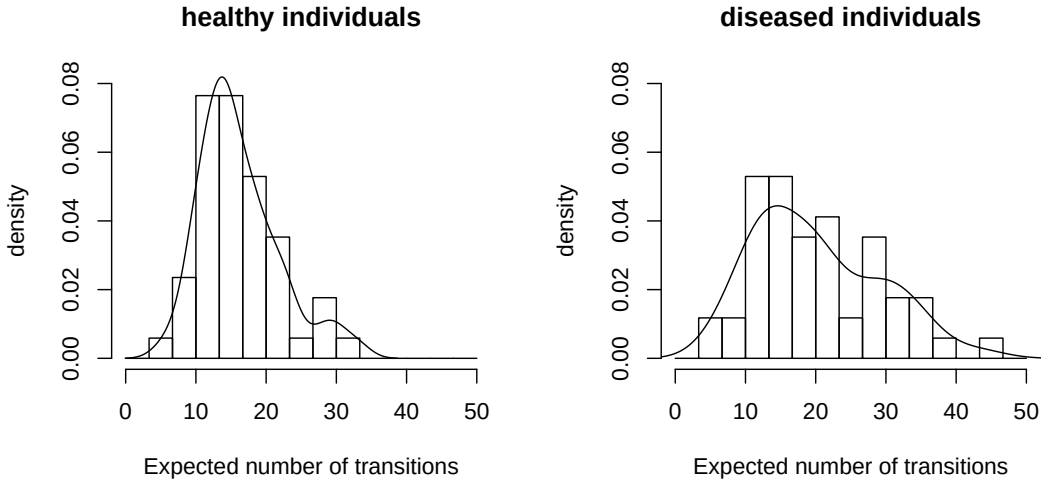

Figure 4.2: *Histogram and kernel density estimator for the expected total number of cross-state transitions.*

The plot does not suggest that the distribution for the set of diseased individuals is simply shifted. Instead, according to this plot about four-fifths of the diseased individuals do not show an anomalously high number of cross-state transitions. There seem to be merely a subgroup in the group of diseased individuals which is responsible for rendering the higher average expected number of cross-state transitions.

4.3.2.3 Choosing the number of states

We have given the results obtained for a PoHMM with five states. This number is in agreement with the sleep states stipulated by the American Academy of Sleep Medicine (AASM) manual from 2007 (REM, wake and Non-REM stages I-III). However, our approach is entirely empirical, and based only on banded spectral properties of one of the EEG nodes. Hence, an exploration into the robustness of conclusions to the number of states is warranted.

In order to choose an appropriate and statistically founded number of states, the PoHMM was fitted for numbers of states $N = 3, 4, 5, 6$. The results of the PoHMM fits

and the results of the independent Dirichlet mixture fits, i.e. the models fitted in Stage I, are given in Table 4.5. Apparently and unsurprisingly, the PoHMMs are superior to the independent mixtures since the latter do not take into account the significant autocorrelation of the series.

Table 4.5: *Log likelihood, number of parameters and BIC of the PoHMMs and the independent Dirichlet mixtures for different numbers of states.*

model	N	$\log \mathcal{L}$	# para.	BIC
Ind. mix.	3	661154.6	14	-1322140
	4	684817.9	19	-1369407
	5	692949.1	24	-1385609
	6	698586.3	29	-1396823
PoHMM	3	740792.2	624	-1473865
	4	775518.6	1240	-1536081
	5	788041.6	2060	-1551236
	6	793161.3	3084	-1549124

From the gaps between the likelihoods it is evident that employing less than five states leads to an underfitting. On the other hand, the models with five and six states yield similarly well fitting results. In the light of the possibility that classical sleep staging is playing a key role in the determination of the HMM states, both models seem reasonable. One might conjecture that the simpler layout possibly corresponds to the decomposition in three Non-REM states plus waking and REM, while the six-state model considers a fourth Non-REM state, as stipulated in the standard manual for scoring of sleep stages (Rechtschaffen and Kales 1968).

For parsimony, and as the improvement in the fit obtained by employing six states arguably is negligible (and in fact BIC chooses the five-state model), we presented our analysis for five states in Sections 4.3.2.1 and 4.3.2.2. Because this choice is, to some extent, arbitrary, it is interesting to investigate the differences in results between the five- and the six-state model. Table 4.6 gives the expected band powers associated with the states obtained in the PoHMMs with five and six states respectively.

The state labelled by $n = 2$ in the five-state PoHMM appears to have split into two different states, namely those labelled $n = 2, 3$ in the right table, when moving to the six-state model. Due to this split, two other states are slightly altered (states 1 and 4 in the left table correspond to states 1 and 5 in the right one), and two states are hardly affected (states 3 and 5 in the left table correspond to states 4 and 6 in the right one), when moving to the six-state PoHMM.

Table 4.6: *Expected band powers in the PoHMMs with five and six states.*

	5-state model			
state	expected band power			
n	δ	θ	α	β
1	0.80	0.11	0.06	0.03
2	0.61	0.19	0.13	0.07
3	0.25	0.22	0.21	0.32
4	0.38	0.21	0.24	0.17
5	0.16	0.07	0.05	0.72

	6-state model			
state	expected band power			
n	δ	θ	α	β
1	0.82	0.10	0.05	0.03
2	0.72	0.15	0.09	0.04
3	0.57	0.20	0.15	0.09
4	0.25	0.22	0.20	0.33
5	0.35	0.21	0.26	0.18
6	0.16	0.07	0.05	0.72

4.4 Concluding remarks

The HMM developed compares time series dynamics in a longitudinal data setting involving two EEG time series for each of 102 individuals. The nature of the data motivated the use of Dirichlet distributions for the state-dependent process. Fitting a separate Dirichlet HMM to each individual would have substantially limited the possibility to compare results across subjects. Thus the primary aim of the population HMM was to account for the heterogeneity across individuals whilst enabling for inter-subject comparisons. The proposed two-stage fitting process is easily carried out. It scales to large studies and integrates well with cluster computing. A potential blemish of the proposed model is that it does not account for the apparent nonhomogeneity of the data. The model thus represents a rather basic first approach to modelling of EEG series via HMMs.

The given application to sleep EEG data revealed that the time spent in the model-derived sleep states is equivalent across carefully matched diseased (sleep apnea) and non-diseased subgroups. Our analysis confirms results from studies on hypnograms, i.e. sleep stage time series obtained by visual classification (cf. Swihart *et al.* 2008). We do note differences, between diseased and non-diseased subjects, in the model-derived state transition rates. Individuals suffering SDB tend to switch more often between states than do healthy individuals. This also confirms results obtained when investigating hypnograms (cf. Swihart *et al.* 2008). However, unlike data analyses using the hypnogram, our approach is entirely automated, being directly applied to the processed EEG signal.

Summary and outlook

We conclude with a summary of the main results along with brief discussions of possible future research related to the individual topics. The three main parts of the thesis, Chapters 2 – 4, deal with HMMs that address special needs. The first part, Chapter 2, is concerned with HMMs that have dwell-time distributions other than geometric. In Chapter 3 it is demonstrated that HMMs are useful tools for fitting nonlinear and non-Gaussian state-space models. Lastly, in Chapter 4, an HMM for populations of sleep EEG time series is developed. Details of the findings are as follows.

A restrictive feature of standard HMMs is that the state dwell-time distributions are necessarily geometric. In Chapter 2 it is shown how this restriction can be relaxed to allow for arbitrary dwell-time distributions while preserving the Markov property of the latent process. This is done by implementing an existing idea, the use of state aggregates, in a new way. The resulting class of HMMs can represent any given dwell-time distribution, either exact or approximately, where, in general, the approximation can be made arbitrarily accurate. The models described in Chapter 2 can either be regarded as approximations to HSMMs or as extensions of ordinary HMMs that offer additional flexibility for the state dwell-time distributions.

The range of methodology that is currently available for HMMs is much more extensive than that for HSMMs. HMMs are easier to apply and to adapt to meet the needs of applications having special features. In particular one can easily incorporate covariates, in the state-dependent process as well as in the latent process. Furthermore, unlike in the case of HSMMs, it is simple to fit stationary models of the proposed type.

The literature on hidden semi-Markov modelling contains relatively few applications. Generally, in the author's view, HSMMs have not yet attracted the attention they deserve. Perhaps the ease of the HMM approximation method can make a contribution in that it makes them more conveniently accessible to practitioners.

HMMs with finite state space are nested in the broader family of SSMs. However, in general the likelihood of SSMs can not be evaluated directly; statistical inference for nonlinear and non-Gaussian SSMs with infinite state space is usually much more challenging than for (standard) HMMs. The material in Chapter 3 illustrates that structured HMMs provide convenient and flexible devices for accurately approximating a diverse variety of SSMs. More precisely, it is shown that general-type SSMs can be approximated by suitably structured HMMs. The approximation can be made arbitrarily accurate at the cost of increasing numerical complexity. One of the main benefits compared to competing approaches, in particular to Monte Carlo methods, is that the programming effort involved in fitting the structured HMMs is very modest (cf. Appendix A4, which contains the functions used to compute and to maximize the likelihood for the SVt model; fitting other SSMs with the proposed method generally involves no more than straightforward changes to that code.). As in case of the HMM approximations to HSMMs, the proposed method enables one to apply all standard HMM techniques.
One of the most important applications of the approximation method via structured HMMs is stochastic volatility modelling. The evidence presented in Section 3.2 supports the claim that nonstandard SV models can outperform the standard SV models SV_0 and SVt in terms of the AIC, goodness of fit (as assessed by the behaviour of residuals) and also the type of backtesting that is applied by central banks and regulatory authorities in order to assess the accuracy of models in terms of the Basel Accords. Series of daily returns are sufficiently long for it to be worthwhile to "invest" in the few additional parameters required for such extensions, in the hope of improving the fit and the forecasting performance.
Future research could involve the exploration of other nonstandard state-space models, for SV modelling as well as in other scenarios. Due to the high flexibility of HMMs there are countless possible applications. One could also attempt to apply structured HMMs to estimate SSMs with higher-dimensional state spaces. The two-and three-dimensional cases are likely to be of most interest — numerous applications with states representing locations are imaginable. As the method becomes more involved in higher dimensions, alternative ways of choosing appropriate grids would need to be explored.

The purpose of population HMMs, discussed in Chapter 4, is to enable comparisons between a number of HMMs fitted to longitudinal data. The proposed two-stage fitting process is easy to use and, unlike joint maximum likelihood, it scales to large studies and integrates well with cluster computing. Numerical studies demonstrate good agreement between the proposed two-stage fitting method and full maximum likelihood, while also

demonstrating substantial decreases in computing time. The proposed model is applied to a novel study of sleep and its correlates. Despite being based entirely on the EEG signal, our results confirm established hypotheses derived from hypnograms that are obtained by visual classification of the polysomnogram data.
Generally speaking, HMMs prove useful to extract features and study sleep phenomena for epidemiological studies. On the other hand the proposed model represents a rather basic first approach to modelling of EEG series via HMMs. Important future research would include covariate adjusted and nonhomogeneous variations of the model. Furthermore, the population HMM should be compared to alternative modelling approaches, in particular to models that incorporate random effects to explain the heterogeneity across the time series (see e.g. Altman 2007). The main challenge here will be to overcome the computational problems. The models proposed by Maruotti (2007) might offer a way forward.

Appendix

A1: Parameter estimates for models fitted to the Old Faithful data

Parameter estimates for the two-state gamma HMM

$$\widehat{\boldsymbol{\kappa}} = (116,\ 206) \qquad \text{(shape parameters)}$$
$$\widehat{\boldsymbol{\theta}} = (0.56,\ 0.45) \qquad \text{(scale parameters)}$$
$$\widehat{\boldsymbol{\mu}} = (64.6,\ 92.6) \qquad \text{(state-dep. means)}$$
$$\widehat{\boldsymbol{\Gamma}} = \begin{pmatrix} 0.00 & 1.00 \\ 0.08 & 0.92 \end{pmatrix} \qquad \text{(t.p.m.)}$$
$$\widehat{\boldsymbol{\delta}} = (0.07,\ 0.93) \qquad \text{(stationary distribution)}$$

Parameter estimates for the three-state gamma HMM

$$\widehat{\boldsymbol{\kappa}} = (94,\ 358,\ 285)$$
$$\widehat{\boldsymbol{\theta}} = (0.70,\ 0.25,\ 0.34)$$
$$\widehat{\boldsymbol{\mu}} = (65.6,\ 89.6,\ 97.2)$$
$$\widehat{\boldsymbol{\Gamma}} = \begin{pmatrix} 0.00 & 0.00 & 1.00 \\ 0.05 & 0.46 & 0.50 \\ 0.14 & 0.79 & 0.07 \end{pmatrix}$$
$$\widehat{\boldsymbol{\delta}} = (0.08,\ 0.55,\ 0.38)$$

Parameter estimates for the four-state gamma HMM

$$\widehat{\boldsymbol{\kappa}} = (100, 592, 497, 282)$$
$$\widehat{\boldsymbol{\theta}} = (0.65, 0.15, 0.19, 0.35)$$
$$\widehat{\boldsymbol{\mu}} = (65.4, 86.2, 93.0, 99.9)$$
$$\widehat{\boldsymbol{\Gamma}} = \begin{pmatrix} 0.00 & 0.00 & 0.25 & 0.75 \\ 0.05 & 0.00 & 0.65 & 0.30 \\ 0.06 & 0.26 & 0.57 & 0.11 \\ 0.19 & 0.51 & 0.28 & 0.02 \end{pmatrix}$$
$$\widehat{\boldsymbol{\delta}} = (0.08, 0.23, 0.51, 0.19)$$

Parameter estimates for the five-state gamma HMM

$$\widehat{\kappa} = (100, 564, 585, 406, 270)$$
$$\widehat{\boldsymbol{\theta}} = (0.66, 0.15, 0.16, 0.23, 0.37)$$
$$\widehat{\boldsymbol{\mu}} = (65.4, 86.4, 93.0, 93.4, 99.8)$$
$$\widehat{\boldsymbol{\Gamma}} = \begin{pmatrix} 0.00 & 0.00 & 0.20 & 0.01 & 0.79 \\ 0.05 & 0.00 & 0.29 & 0.35 & 0.30 \\ 0.13 & 0.00 & 0.00 & 0.66 & 0.20 \\ 0.00 & 0.54 & 0.46 & 0.00 & 0.00 \\ 0.18 & 0.53 & 0.08 & 0.19 & 0.01 \end{pmatrix}$$
$$\widehat{\boldsymbol{\delta}} = (0.08, 0.24, 0.23, 0.27, 0.18)$$

Parameter estimates for the four-state second-order gamma HMM

$$\widehat{\boldsymbol{\kappa}} = (101,\ 585,\ 518,\ 270)$$
$$\widehat{\boldsymbol{\theta}} = (0.65,\ 0.15,\ 0.18,\ 0.37)$$
$$\widehat{\boldsymbol{\mu}} = (65.3,\ 86.0,\ 92.6,\ 98.7)$$

$$100 \cdot \widehat{\boldsymbol{\Gamma}} = \begin{pmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 64 & 36 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 26 & 34 & 33 & 7 \\
40 & 60 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 7 & 8 & 62 & 23 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 9 & 76 & 13 & 3 \\
5 & 95 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 3 & 53 & 44 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 7 & 26 & 51 & 17 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 12 & 38 & 50 & 0 \\
10 & 90 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 6 & 54 & 40 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 8 & 14 & 68 & 10 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 100 & 0 & 0
\end{pmatrix}$$

Here the state pairs are, in order, $(1,3)$, $(1,4)$, $(2,1)$, $(2,3)$, $(2,4)$, $(3,1)$, $(3,2)$, $(3,3)$, $(3,4)$, $(4,1)$, $(4,2)$, $(4,3)$ and $(4,4)$; the state pairs $(1,1)$, $(1,2)$ and $(2,2)$ almost surely do not occur. Structural zeros are displayed in grey font. Note that the row sums of $\widehat{\boldsymbol{\Gamma}}$ given above do not all equal one due to rounding.

Stationary distribution of the 13-state first-order Markov chain:

$$100 \cdot \widehat{\boldsymbol{\delta}} = (0.9,\ 6.8,\ 1.0,\ 11.7,\ 9.0,\ 3.2,\ 8.7,\ 26.1,\ 8.1,\ 3.5,\ 13.0,\ 7.5,\ 0.8)$$

Stationary distribution of the (equivalent) four-state second-order Markov chain:

$$100 \cdot \widehat{\boldsymbol{\delta}} = (7.6,\ 21.6,\ 46.1,\ 24.6)$$

A2: Some proofs related to HMMs with arbitrary dwell-time distributions

Proof of Proposition 1.

$$\begin{aligned}
\omega^\star_{ij} &= \mathbb{P}\big(S^\star_{t+1} \in I_j \,|\, S^\star_t \in I_i, S^\star_{t+1} \notin I_i\big) \\
&= \frac{\mathbb{P}\big(S^\star_{t+1} \in I_j, S^\star_t \in I_i\big)}{\mathbb{P}\big(S^\star_{t+1} \notin I_i, S^\star_t \in I_i\big)} \quad \text{(since, by assumption, } i \neq j\text{)} \\
&= \frac{\sum_{k\in I_i} \mathbb{P}\big(S^\star_{t+1} \in I_j \,|\, S^\star_t = k\big)\mathbb{P}\big(S^\star_t = k\big)}{\sum_{k\in I_i} \mathbb{P}\big(S^\star_{t+1} \notin I_i \,|\, S^\star_t = k\big)\mathbb{P}\big(S^\star_t = k\big)} \\
&= \frac{\sum_{k\in I_i} \omega_{ij} c_i(k - i_i^- + 1)\mathbb{P}\big(S^\star_t = k\big)}{\sum_{k\in I_i}\sum_{l\neq i} \omega_{il} c_i(k - i_i^- + 1)\mathbb{P}\big(S^\star_t = k\big)} \\
&= \omega_{ij}\,,
\end{aligned}$$

since $\sum_{l\neq i} \omega_{il} = 1$. □

Lemma 1. *Let $k \in \{1, \ldots, N\}$ and $n \in \mathbb{N}$. Then*

$$\prod_{i=1}^{n} (1 - c_k(i)) = 1 - F_k(n)\,. \tag{A.1}$$

Proof. By induction: First note that $1 - c_k(1) = 1 - p_k(1) = 1 - F_k(1)$. Now assume that (A.1) holds for some n. If $F_k(n) < 1$ then

$$\begin{aligned}
\prod_{i=1}^{n+1} (1 - c_k(i)) &= \Big(\prod_{i=1}^{n} (1 - c_k(i))\Big)\big(1 - c_k(n+1)\big) \\
&= \big(1 - F_k(n)\big)\left(1 - \frac{p_k(n+1)}{1 - F_k(n)}\right) \\
&= 1 - \big(F_k(n) + p_k(n+1)\big) = 1 - F_k(n+1)\,.
\end{aligned}$$

If $F_k(n) = 1$ then $F_k(n+1) = 1$ and

$$\prod_{i=1}^{n+1} \big(1 - c_k(i)\big) = \big(1 - F_k(n)\big)\big(1 - c_k(n+1)\big) = 0 = 1 - F_k(n+1)\,.$$

□

Proof of Proposition 2. The case $m_k = 1$ is trivial, so we consider the case $m_k \geq 2$. Since every sojourn in the state aggregate I_k starts in state i_k^-, and taking into account

the special structure of $\mathbf{\Gamma}$, it follows that

$$\begin{aligned} p_k^\star(1) &= \mathbb{P}\big(S_{t+1} \notin I_k \,|\, S_t \in I_k, S_{t-1} \notin I_k\big) \\ &= \sum_{1 \le s \le N, s \neq k} \omega_{ks} c_k(1) = c_k(1) = p_k(1). \end{aligned}$$

We now consider $p_k^\star(r)$ for $2 \le r \le m_k$. The structure of $\mathbf{\Gamma}$ is such that the dwell time in state aggregate I_k is of length r if and only if the state sequence successively runs through the states $i_k^-, i_k^- + 1, \ldots, i_k^- + r - 1$ and then immediately switches from state $i_k^- + r - 1$ to a different state aggregate.
If $F_k(r-1) < 1$ then, by (A.1), it follows that

$$\begin{aligned} p_k^\star(r) &= \prod_{i=1}^{r-1} \big(1 - c_k(i)\big) \sum_{1 \le s \le N, s \neq k} \omega_{ks} c_k(r) \\ &= \big(1 - F_k(r-1)\big) \frac{p_k(r)}{1 - F_k(r-1)} = p_k(r), \end{aligned}$$

and if $F_k(r-1) = 1$ then $p_k^\star(r) = 0 = p_k(r)$.

Finally we consider the case $r > m_k$. The dwell time in state aggregate I_k is of length $r > m_k$ if and only if the state sequence successively runs through the states $i_k^-, i_k^- + 1, \ldots, i_k^+ - 1$, then remains in state i_k^+ for $r - m_k + 1$ time units and finally switches to a different state aggregate.
If $F_k(m_k - 1) < 1$ then, again by (A.1), it follows that

$$\begin{aligned} p_k^\star(r) &= \prod_{i=1}^{m_k - 1} \big(1 - c_k(i)\big) \big(1 - c_k(m_k)\big)^{r - m_k} \sum_{1 \le s \le N, s \neq k} \omega_{ks} c_k(m_k) \\ &= \big(1 - F_k(m_k - 1)\big) \big(1 - c_k(m_k)\big)^{r - m_k} \frac{p_k(m_k)}{1 - F_k(m_k - 1)} \\ &= p_k(m_k) \big(1 - c_k(m_k)\big)^{r - m_k}, \end{aligned}$$

whereas if $F_k(m_k - 1) = 1$, then $c_k(m_k) = 1$, and so $p_k^\star(r) = 0 = p_k(m_k)\big(1 - c_k(m_k)\big)^{r - m_k}$. □

A3: Derivation of moments for the nonstandard SV models

Stationary moments of g_t in the $SVMt$ model

We start with a technical lemma:

Lemma 2. *If $\{g_t\}$ is second-order stationary, i.e. if the inequality (3.8) holds, then*

$$|\,\alpha\phi_1 + (1-\alpha)\phi_2\,| < 1\,, \tag{A.2}$$

i.e. $\{g_t\}$ is first-order stationary (cf. Wong and Li 2000).

Proof. Indirectly: we first show that

$$\alpha\phi_1 + (1-\alpha)\phi_2 \geq 1 \tag{A.3}$$

excludes second-order stationarity. If $|\phi_1|, |\phi_2| \geq 1$, then

$$\alpha\phi_1^2 + (1-\alpha)\phi_2^2 \geq \alpha + (1-\alpha) = 1\,,$$

which contradicts second-order stationarity (cf. the inequality (3.8)). If $\phi_1, \phi_2 < 1$, then

$$\alpha\phi_1 + (1-\alpha)\phi_2 < \alpha + (1-\alpha) = 1\,,$$

which is inconsistent with (A.3). Thus, assume without loss of generality that $|\phi_1| < 1$ and $\phi_2 \geq 1$. Then, from (A.3),

$$\begin{aligned}
&\alpha(\phi_1 - \phi_2) + \phi_2 \geq 1 \\
\Rightarrow\ &\alpha(\phi_1 - \phi_2)(\phi_1 + \phi_2) + \phi_2(\phi_1 + \phi_2) \geq \phi_1 + \phi_2 \\
\Rightarrow\ &\alpha(\phi_1 - \phi_2)(\phi_1 + \phi_2) + \phi_2^2 \geq \phi_1 + \phi_2 - \phi_1\phi_2 \\
\overset{(\star)}{\Longrightarrow}\ &\alpha(\phi_1 - \phi_2)(\phi_1 + \phi_2) + \phi_2^2 \geq 1 \\
\Rightarrow\ &\alpha(\phi_1^2 - \phi_2^2) + \phi_2^2 \geq 1 \\
\Rightarrow\ &\alpha\phi_1^2 + (1-\alpha)\phi_2^2 \geq 1\,,
\end{aligned}$$

which contradicts second-order stationarity. (Note: $(\star)$ holds as

$$\begin{aligned}
&\phi_1 < 1 \\
\Rightarrow\ &\phi_1(1-\phi_2) \geq 1 - \phi_2 \\
\Rightarrow\ &\phi_1 + \phi_2 - \phi_1\phi_2 \geq 1\,.)
\end{aligned}$$

Analogously one shows that

$$\alpha\phi_1 + (1-\alpha)\phi_2 \leq -1$$

excludes second-order stationarity. □

We now compute the mean of $\{g_t\}$ under second-order stationarity. First of all, we have

$$\mathrm{E}(g_{t+1} \,|\, g_t) = (\alpha\phi_1 + (1-\alpha)\phi_2)g_t \,.$$

Taking expectations on both sides yields

$$\mathrm{E}(g_{t+1}) = (\alpha\phi_1 + (1-\alpha)\phi_2)\mathrm{E}(g_t) \,.$$

Second-order stationarity implies first-order stationarity and thus $\mathrm{E}(g_t) = \mu_g \; \forall t$. Hence

$$\begin{aligned} \mu_g &= (\alpha\phi_1 + (1-\alpha)\phi_2)\mu_g \\ \Rightarrow \mu_g &= 0 \,, \end{aligned}$$

as $\alpha\phi_1 + (1-\alpha)\phi_2 \neq 1$ according to Lemma 2.

The stationary variance of $\{g_t\}$ can now be computed as follows:

$$\begin{aligned} \mathrm{var}(g_{t+1}) &= \mathrm{E}(\mathrm{var}(g_{t+1} \,|\, g_t)) + \mathrm{var}(\mathrm{E}(g_{t+1} \,|\, g_t)) \\ &= \mathrm{E}(\alpha\sigma_1^2 + (1-\alpha)\sigma_2^2 + \alpha(1-\alpha)(\phi_1 g_t - \phi_2 g_t)^2) \\ &\quad + \mathrm{var}(\alpha\phi_1 g_t + (1-\alpha)\phi_2 g_t) \\ &= \alpha\sigma_1^2 + (1-\alpha)\sigma_2^2 + \alpha(1-\alpha)(\phi_1 - \phi_2)^2(\mathrm{var}(g_t) + (\mathrm{E}(g_t))^2) \\ &\quad + (\alpha\phi_1 + (1-\alpha)\phi_2)^2 \mathrm{var}(g_t). \end{aligned}$$

Second-order stationarity implies that $\mathrm{E}(g_t) = \mu_g = 0 \; \forall t$ (see above), and that $\mathrm{var}(g_t) = \sigma_g^2 \; \forall t$, and thus

$$\begin{aligned} \sigma_g^2 &= \alpha\sigma_1^2 + (1-\alpha)\sigma_2^2 + \left(\alpha(1-\alpha)(\phi_1 - \phi_2)^2 + (\alpha\phi_1 + (1-\alpha)\phi_2)^2\right)\sigma_g^2 \\ &= \alpha\sigma_1^2 + (1-\alpha)\sigma_2^2 + \left(\alpha\phi_1^2 + (1-\alpha)\phi_2^2\right)\sigma_g^2 \\ \Leftrightarrow \sigma_g^2 &= \frac{\alpha\sigma_1^2 + (1-\alpha)\sigma_2^2}{1 - \left(\alpha\phi_1^2 + (1-\alpha)\phi_2^2\right)} \,. \end{aligned}$$

Stationary moments of g_t in the $MSSVt$ model

We have

$$\mathrm{E}(g_{t+1} \,|\, g_t, \alpha_t) = \phi g_t \,.$$

Taking expectations on both sides yields

$$\mathrm{E}(g_{t+1}) = \phi\mathrm{E}(g_t) \,.$$

Under stationarity we have $\mathrm{E}(g_t) = \mu_g \; \forall t$, and thus

$$\begin{aligned} \mu_g &= \phi \mu_g \\ \Rightarrow \mu_g &= 0 \,, \end{aligned}$$

as $|\phi| < 1$.

The stationary variance of $\{g_t\}$ can now be computed as follows:

$$\begin{aligned} \mathrm{var}(g_{t+1}) &= \mathrm{E}(\mathrm{var}(g_{t+1} \,|\, g_t, \alpha_t)) + \mathrm{var}(\mathrm{E}(g_{t+1} \,|\, g_t, \alpha_t)) \\ &= \mathrm{E}(\sigma^2_{\alpha_t}) + \phi^2 \mathrm{var}(g_t) \,. \end{aligned}$$

Stationarity of $\{\alpha_t\}$ and $\{g_t\}$ implies that $\mathrm{E}(\sigma^2_{\alpha_t}) = \delta_1^{(\alpha)} \sigma_1^2 + \delta_2^{(\alpha)} \sigma_2^2 \; \forall t$, and that $\mathrm{var}(g_t) = \sigma_g^2 \; \forall t$, and thus

$$\begin{aligned} \sigma_g^2 &= \delta_1^{(\alpha)} \sigma_1^2 + \delta_2^{(\alpha)} \sigma_2^2 + \phi^2 \sigma_g^2 \\ \Leftrightarrow \sigma_g^2 &= \frac{\delta_1^{(\alpha)} \sigma_1^2 + \delta_2^{(\alpha)} \sigma_2^2}{1 - \phi^2} \,. \end{aligned}$$

Stationary moments of g_t in the $SVVt$ model

Stationarity implies that $\mathrm{E}(g_t) = 0 \; \forall t$. (The proof is analogous as in case of the *MSSVt* model.) To obtain the stationary variance, consider

$$\begin{aligned} \mathrm{var}(g_{t+1}) &= \mathrm{E}(\mathrm{var}(g_{t+1} \,|\, g_t)) + \mathrm{var}(\mathrm{E}(g_{t+1} \,|\, g_t)) \\ &= \omega + \gamma \mathrm{E}\big(\exp(g_t)\big) + \phi^2 \mathrm{var}(g_t) \,. \end{aligned}$$

Using the Taylor approximation $\exp(x) \approx 1 + x + 0.5x^2$ and assuming second-order stationarity, i.e. $\mathrm{var}(g_t) = \sigma_g^2 \; \forall t$, yields

$$\begin{aligned} \sigma_g^2 &\approx \omega + \gamma\big(1 + \mathrm{E}(g_t) + 0.5\mathrm{E}(g_t^2)\big) + \phi^2 \sigma_g^2 \\ &\approx \omega + \gamma\big(1 + 0.5\sigma_g^2\big) + \phi^2 \sigma_g^2 \\ \Leftrightarrow \sigma_g^2 &\approx \frac{\omega + \gamma}{1 - 0.5\gamma - \phi^2} \,. \end{aligned}$$

Stationary moments of y_t in the models $SVMt$, $MSSVt$ and $SVVt$

In each of the models *SVMt*, *MSSVt* and *SVVt*, the observation equation is given by

$$y_t = \varepsilon_t(\beta \exp(0.5 g_t) + \xi) \,,$$

where ε_t is t_ν-distributed. Thus,

$$\mathrm{E}(y_t) = \mathrm{E}\big(\mathrm{E}(y_t \,|\, g_t)\big) = \mathrm{E}\big((\beta \exp(0.5 g_t) + \xi) \underbrace{\mathrm{E}\varepsilon_t}_{=0}\big) = 0 \;\forall t\,.$$

Furthermore, if $\{g_t\}$ is stationary, then

$$\begin{aligned}
\mathrm{var}(y_t) &= \mathrm{E}(y_t^2) \\
&= \mathrm{E}\big(\mathrm{E}(y_t^2 \,|\, g_t)\big) \\
&= \mathrm{E}\big((\beta \exp(0.5 g_t) + \xi)^2 \mathrm{E}(\varepsilon_t^2)\big) \\
&= \frac{\nu}{\nu - 2}\big(\beta^2 \mathrm{E}(\exp(g_t)) + 2\beta\xi \mathrm{E}(\exp(0.5 g_t)) + \xi^2\big) \\
&\approx \frac{\nu}{\nu - 2}\big(\beta^2 (1 + 0.5\sigma_g^2) + 2\beta\xi(1 + 0.125\sigma_g^2) + \xi^2\big) \\
&= \frac{\nu}{\nu - 2}\big((\beta + \xi)^2 + \sigma_g^2(0.5\beta^2 + 0.25\beta\xi)\big)\,,
\end{aligned}$$

$$\begin{aligned}
\mathrm{E}(y_t^4) &= \mathrm{E}\big(\mathrm{E}(y_t^4 \,|\, g_t)\big) \\
&= \mathrm{E}\big((\beta \exp(0.5 g_t) + \xi)^4 \mathrm{E}(\varepsilon_t^4)\big) \\
&= \frac{3\nu^2}{(\nu - 2)(\nu - 4)}\big(\beta^4 \mathrm{E}(\exp(2 g_t)) + 6\beta^2\xi^2 \mathrm{E}(\exp(g_t)) + \xi^4 \\
&\quad + 4\beta^3\xi \mathrm{E}(\exp(1.5 g_t)) + 4\beta\xi^3 \mathrm{E}(\exp(0.5 g_t))\big) \\
&\approx \frac{3\nu^2}{(\nu - 2)(\nu - 4)}\big(\beta^4 (1 + 2\sigma_g^2) + 6\beta^2\xi^2 (1 + 0.5\sigma_g^2) + \xi^4 \\
&\quad + 4\beta^3\xi(1 + 1.125\sigma_g^2) + 4\beta\xi^3(1 + 0.125\sigma_g^2)\big) \\
&= \frac{3\nu^2}{(\nu - 2)(\nu - 4)}\big((\beta + \xi)^4 + \sigma_g^2(2\beta^4 + 3\beta^2\xi^2 + 4.5\beta^3\xi + 0.5\beta\xi^3)\big)\,.
\end{aligned}$$

and, consequently,

$$\begin{aligned}
\mathrm{kurtosis}(y_t) &= \frac{\mathrm{E}(y_t^4)}{\big(\mathrm{var}(y_t)\big)^2} \\
&\approx 3\,\frac{\nu - 2}{\nu - 4}\,\frac{(\beta + \xi)^4 + c_2\sigma_g^2}{\big((\beta + \xi)^2 + c_1\sigma_g^2\big)^2}\,,
\end{aligned}$$

where $c_1 = 0.5\beta^2 + 0.25\beta\xi$ and $c_2 = 3\beta^2\xi^2 + 2\beta^4 + 4.5\beta^3\xi + 0.5\beta\xi^3$.

Stationary moments in the $GSVt$ model

Since, conditional on g_t,

$$g_{t+1} \sim \Gamma(\phi g_t + \lambda, 1)\,,$$

we have

$$\mathrm{E}(g_{t+1} \,|\, g_t) = \phi g_t + \lambda \,,$$

and so

$$\mathrm{E}(g_{t+1}) = \phi \mathrm{E}(g_t) + \lambda \,.$$

If $\{g_t\}$ is stationary, then $\mathrm{E}(g_t) = \mu_g \; \forall t$, and

$$\begin{aligned} \mu_g &= \phi \mu_g + \lambda \\ \Leftrightarrow \mu_g &= \frac{\lambda}{1-\phi} \,. \end{aligned}$$

The details to derive the stationary variance are as follows:

$$\begin{aligned} \mathrm{var}(g_{t+1}) &= \mathrm{E}(\mathrm{var}(g_{t+1} \,|\, g_t)) + \mathrm{var}(\mathrm{E}(g_{t+1} \,|\, g_t)) \\ &= \mathrm{E}(\phi g_t + \lambda) + \mathrm{var}(\phi g_t + \lambda) \\ &= \phi \mathrm{E}(g_t) + \lambda + \phi^2 \mathrm{var}(g_t). \end{aligned}$$

Stationarity implies that $\mathrm{var}(g_t) = \sigma_g^2 \; \forall t$, and

$$\begin{aligned} \sigma_g^2 &= \big(\phi \frac{\lambda}{1-\phi} + \lambda\big)(1-\phi^2)^{-1} \\ &= \lambda\big(\frac{\phi}{1-\phi} + 1\big)(1-\phi^2)^{-1} \\ &= \frac{\lambda}{(1-\phi)(1-\phi^2)} \end{aligned}$$

Given stationarity of $\{g_t\}$ (i.e. if $\phi \in [0,1)$), one further obtains

$$\mathrm{E}(y_t) = \mathrm{E}\big(\mathrm{E}(y_t \,|\, g_t)\big) = \mathrm{E}\big(\beta \sqrt{g_t + \xi} \underbrace{\mathrm{E}\varepsilon_t}_{=0}\big) = 0 \; \forall t \,,$$

$$\begin{aligned} \mathrm{var}(y_t) &= \mathrm{E}(y_t^2) \\ &= \mathrm{E}\big(\mathrm{E}(y_t^2 \,|\, g_t)\big) \\ &= \mathrm{E}\big(\beta^2 (g_t + \xi) \mathrm{E}(\varepsilon_t^2)\big) \\ &= \frac{\nu}{\nu - 2} \beta^2 (\mu_g + \xi) \,, \end{aligned}$$

$$\begin{aligned} \mathrm{E}(y_t^4) &= \mathrm{E}\big(\mathrm{E}(y_t^4 \,|\, g_t)\big) \\ &= \mathrm{E}\big(\beta^4 (g_t + \xi)^2 \mathrm{E}(\varepsilon_t^4)\big) \\ &= \frac{3\nu^2}{(\nu-2)(\nu-4)} \beta^4 \big(\sigma_g^2 + \mu_g^2 + 2\xi\mu_g + \xi^2\big) \end{aligned}$$

and

$$\begin{aligned}\text{kurtosis}(y_t) &= \frac{\mathrm{E}(y_t^4)}{\big(\mathrm{var}(y_t)\big)^2} \\ &= 3\,\frac{\nu-2}{\nu-4}\cdot\frac{\sigma_g^2+(\mu_g+\xi)^2}{(\mu_g+\xi)^2} \\ &= 3\,\frac{\nu-2}{\nu-4}\left(1+\frac{\mu_g}{(\mu_g+\xi)^2(1-\phi^2)}\right).\end{aligned}$$

A4: R code for fitting an SVt model

This appendix gives the **R** code for fitting the stochastic volatility model SVt by means of structured HMMs. The computation uses four functions:

- `SV.HMM.pn2pw` transforms the (constrained) natural parameters to (unconstrained) working parameters; e.g. ϕ is mapped from $(-1, 1)$ to the whole real line;
- `SV.HMM.pw2pn` performs the inverse transformation to the working parameters;
- `SV.HMM.mllk` computes minus the log likelihood of the structured HMM;
- `SV.HMM.mle` performs the numerical minimization of the function `SV.HMM.mllk`.

Transform natural parameters to working parameters

```
SV.HMM.pn2pw<-function(phi,beta,sigma,nu)
{
 pvec <- c(log((1+phi)/(1-phi)),log(beta),log(sigma),log(nu))
 return(pvec)
}
```

This function bijectively maps each of the constrained natural parameters, i.e. the model parameters $\phi \in (-1, 1)$, $\beta > 0$, $\sigma > 0$ and $\nu > 0$, to $\mathbb{R}$, and returns the transformed values summarized in a vector `pvec`; cf. Section 1.2.

Transform working parameters to natural parameters

```
SV.HMM.pw2pn<-function(pvec)
{
 return(list(phi=(exp(pvec[1])-1)/(exp(pvec[1])+1),
        beta=exp(pvec[2]),sigma=exp(pvec[3]),nu=exp(pvec[4])))
}
```

This function performs the inverse transformation to the vector `pvec` of working parameters. It returns a list that comprises the natural model parameters ϕ, β, σ and ν.

Minus the log likelihood of structured HMM that approximates the SVt model

```
SV.HMM.mllk<-function(pvec,x,N,gbmax)
{
 p      <- SV.HMM.pw2pn(pvec)
 gb     <- seq(-gbmax,gbmax,length=N+1)
  # midpoints of the intervals used in the discretization:
 g      <- (gb[-1]+gb[-(N+1)])*0.5
 Gamma  <- matrix(0,N,N)
 for (i in 1:N)
   {
   goo       <- diff(pnorm(gb,p$phi*g[i],p$sigma))
   Gamma[i,] <- goo/sum(goo)
   }
 delta  <- diff(pnorm(gb,0,p$sigma/sqrt(1-p$phi^2)))
 beg    <- p$beta*exp(g/2)
 foo    <- delta*1/beg*dt(x[1]/beg,p$nu)
  # scaling:
 sumfoo <- sum(foo)
 lscale <- log(sumfoo)
 foo    <- foo/sumfoo
 for (t in 2:length(x))
   {
   foo    <- foo%*%Gamma*1/beg*dt(x[t]/beg,p$nu)
    # scaling:
   sumfoo <- sum(foo)
   lscale <- lscale+log(sumfoo)
   foo    <- foo/sumfoo
   }
 mllk   <- -lscale
 return(mllk)
}
```

This function computes minus the log likelihood for a given vector `pvec` of working parameters, a given vector `x` of observations, a given resolution `N` for the discretization and a given range $[-\texttt{gbmax}, \texttt{gbmax}]$ for the g_t-values to allow for; cf. Section 3.1.

Maximum likelihood estimation of the structured HMM

```
SV.HMM.mle<-function(x,N,gbmax,phi0,beta0,sigma0,nu0)
{
 pvec0 <- SV.HMM.pn2pw(phi0,beta0,sigma0,nu0)
 mod   <- nlm(SV.HMM.mllk,pvec0,x=x,N=N,gbmax=gbmax,print.level=2)
 p     <- SV.HMM.pw2pn(mod$estimate)
 return(list(phi=p$phi,beta=p$beta,sigma=p$sigma,nu=p$nu))
}
```

This function applies the minimization routine `nlm` in order to numerically minimize the function `SV.HMM.mllk`. Before applying the algorithm the initial values, `phi0`, `beta0`, `sigma0` and `nu0`, are transformed to the working parameter space. The resulting estimates for the (natural) parameters are returned in a list.

A5: Parameter estimates for the SV models

Table A.1: *Parameter estimates for the SV_0 model ($N = 100$, $-g_{min} = g_{max} = 5$).*

	$\widehat{\phi}$	$\widehat{\sigma}$	$100\,\widehat{\beta}$
Sony	0.960	0.238	1.947
Time Warner	0.995	0.125	2.288
Toyota	0.976	0.171	1.676
Trav. Comp.	0.969	0.239	1.625
BP	0.986	0.113	1.519
Roy. D. Sh.	0.987	0.118	1.571
Bank of Am.	0.993	0.167	1.658
Citigroup	0.991	0.179	1.919
Deu. Bank	0.988	0.150	2.034
Morgan St.	0.990	0.149	2.364

Table A.2: *Parameter estimates for the SVt model ($N = 100$, $-g_{min} = g_{max} = 5$).*

	$\widehat{\phi}$	$\widehat{\sigma}$	$100\,\widehat{\beta}$	$\widehat{\nu}$
Sony	0.983	0.141	1.759	8.5
Time Warner	0.997	0.085	2.138	11.3
Toyota	0.984	0.129	1.569	12.9
Trav. Comp.	0.987	0.140	1.456	8.0
BP	0.990	0.092	1.436	16.7
Roy. D. Sh.	0.990	0.101	1.494	19.5
Bank of Am.	0.996	0.119	1.588	11.0
Citigroup	0.995	0.122	1.822	10.0
Deu. Bank	0.994	0.101	1.817	8.5
Morgan St.	0.993	0.116	2.187	11.9

Table A.3: *Parameter estimates for the SVMt model ($N = 100$, $-g_{min} = g_{max} = 8$).*

	$\widehat{\boldsymbol{\phi}}$	$\widehat{\boldsymbol{\sigma}}$	$\widehat{\alpha}$	$100\,\widehat{\beta}$	$\widehat{\nu}$	$100\,\widehat{\xi}$
Sony	$\begin{pmatrix}0.738\\1.016\end{pmatrix}$	$\begin{pmatrix}0.004\\0.217\end{pmatrix}$	0.111	1.048	8.6	0.703
Time Warner	$\begin{pmatrix}0.974\\1.178\end{pmatrix}$	$\begin{pmatrix}0.031\\0.355\end{pmatrix}$	0.923	1.161	12.2	0.465
Toyota	$\begin{pmatrix}0.979\\1.336\end{pmatrix}$	$\begin{pmatrix}0.101\\1.187\end{pmatrix}$	0.974	0.849	13.7	0.650
Trav. Comp.	$\begin{pmatrix}0.958\\1.086\end{pmatrix}$	$\begin{pmatrix}0.021\\0.453\end{pmatrix}$	0.746	0.549	7.5	0.758
BP	$\begin{pmatrix}0.972\\1.063\end{pmatrix}$	$\begin{pmatrix}0.026\\0.441\end{pmatrix}$	0.756	0.206	16.5	0.969
Roy. D. Sh.	$\begin{pmatrix}0.975\\1.073\end{pmatrix}$	$\begin{pmatrix}0.025\\0.503\end{pmatrix}$	0.798	0.203	16.6	0.996
Bank of Am.	$\begin{pmatrix}0.970\\1.036\end{pmatrix}$	$\begin{pmatrix}0.246\\0.003\end{pmatrix}$	0.579	1.286	10.6	0.576
Citigroup	$\begin{pmatrix}0.945\\1.166\end{pmatrix}$	$\begin{pmatrix}0.035\\0.138\end{pmatrix}$	0.781	1.343	11.0	0.000
Deu. Bank	$\begin{pmatrix}0.987\\1.087\end{pmatrix}$	$\begin{pmatrix}0.174\\0.002\end{pmatrix}$	0.916	0.933	8.5	0.754
Morgan St.	$\begin{pmatrix}0.952\\1.185\end{pmatrix}$	$\begin{pmatrix}0.037\\0.220\end{pmatrix}$	0.825	1.265	13.0	0.485

Table A.4: *Parameter estimates for the MSSVt model ($N = 200$, $-g_{min} = g_{max} = 6$).*

	$\widehat{\phi}$	$\widehat{\boldsymbol{\sigma}}$	$\widehat{\boldsymbol{\Gamma}}^{(\alpha)}$	$100\,\widehat{\beta}$	$\widehat{\nu}$	$100\,\widehat{\xi}$
Sony	0.986	$\begin{pmatrix} 0.149 \\ 2.253 \end{pmatrix}$	$\begin{pmatrix} 0.994 & 0.006 \\ 0.357 & 0.643 \end{pmatrix}$	0.962	9.1	0.708
Time Warner	0.998	$\begin{pmatrix} 0.085 \\ 0.292 \end{pmatrix}$	$\begin{pmatrix} 0.986 & 0.014 \\ 0.089 & 0.911 \end{pmatrix}$	1.078	11.3	0.654
Toyota	0.986	$\begin{pmatrix} 0.092 \\ 0.448 \end{pmatrix}$	$\begin{pmatrix} 0.989 & 0.011 \\ 0.029 & 0.971 \end{pmatrix}$	0.843	14.6	0.692
Trav. Comp.	0.988	$\begin{pmatrix} 0.016 \\ 0.423 \end{pmatrix}$	$\begin{pmatrix} 0.991 & 0.009 \\ 0.013 & 0.987 \end{pmatrix}$	0.487	7.4	0.775
BP	0.987	$\begin{pmatrix} 0.025 \\ 0.419 \end{pmatrix}$	$\begin{pmatrix} 0.989 & 0.011 \\ 0.037 & 0.963 \end{pmatrix}$	0.448	16.9	0.831
Roy. D. Sh.	0.984	$\begin{pmatrix} 0.022 \\ 0.434 \end{pmatrix}$	$\begin{pmatrix} 0.983 & 0.017 \\ 0.055 & 0.945 \end{pmatrix}$	0.512	17.9	0.803
Bank of Am.	0.992	$\begin{pmatrix} 0.011 \\ 0.239 \end{pmatrix}$	$\begin{pmatrix} 0.997 & 0.003 \\ 0.004 & 0.996 \end{pmatrix}$	1.365	11.4	0.350
Citigroup	0.987	$\begin{pmatrix} 0.025 \\ 0.491 \end{pmatrix}$	$\begin{pmatrix} 0.968 & 0.032 \\ 0.162 & 0.838 \end{pmatrix}$	1.049	11.2	0.359
Deu. Bank	0.993	$\begin{pmatrix} 0.050 \\ 0.367 \end{pmatrix}$	$\begin{pmatrix} 0.981 & 0.019 \\ 0.049 & 0.951 \end{pmatrix}$	0.777	8.8	0.765
Morgan St.	0.994	$\begin{pmatrix} 0.110 \\ 0.352 \end{pmatrix}$	$\begin{pmatrix} 0.998 & 0.002 \\ 0.004 & 0.996 \end{pmatrix}$	1.096	13.2	0.842

Table A.5: *Parameter estimates for the SVVt model ($N = 100$, $-g_{min} = g_{max} = 6$).*

	$\widehat{\phi}$	$\widehat{\omega}$	$100\,\widehat{\gamma}$	$100\,\widehat{\beta}$	$\widehat{\nu}$	$100\,\widehat{\xi}$
Sony	0.985	0.026	0.382	1.281	8.5	0.449
Time Warner	0.998	0.010	0.427	2.281	11.3	0.619
Toyota	0.985	0.043	0.042	0.898	13.2	0.006
Trav. Comp.	0.994	0.062	0.064	0.478	7.4	0.797
BP	0.995	0.030	0.151	0.487	16.5	0.863
Roy. D. Sh.	0.996	0.038	0.119	0.487	17.0	0.916
Bank of Am.	0.998	0.030	0.088	0.852	10.6	0.595
Citigroup	0.998	0.007	0.450	1.710	10.0	0.167
Deu. Bank	0.996	0.037	0.017	0.563	8.5	0.839
Morgan St.	0.997	0.008	0.826	2.132	12.9	0.441

Table A.6: *Parameter estimates for the GSVt model ($N = 100$, $g_{min} = 0$, $g_{max} = 200$).*

	$\widehat{\phi}$	$\widehat{\lambda}$	$100\,\widehat{\beta}$	$\widehat{\nu}$	$\widehat{\xi}$
Sony	0.984	0.547	0.326	8.4	5.435
Time Warner	0.999	0.146	0.333	11.8	4.503
Toyota	0.984	0.502	0.296	12.2	6.475
Trav. Comp.	0.991	0.149	0.356	8.0	4.869
BP	0.985	0.247	0.318	18.8	8.307
Roy. D. Sh.	0.987	0.187	0.355	22.8	7.129
Bank of Am.	0.999	0.013	0.532	12.9	0.619
Citigroup	0.998	0.108	0.516	9.2	0.524
Deu. Bank	0.993	0.194	0.441	9.5	3.859
Morgan St.	0.994	0.188	0.505	9.9	2.994

Bibliography

Altman, R. (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), pp. 201–210.

Aston, J. A. D. and Martin, D. E. K. (2007), Distributions associated with general runs and patterns in hidden Markov models. *Annals of Applied Statistics*, 1 (2), pp. 585–611.

Azzalini, A. and Bowman, A. W. (1990), A look at some data on the Old Faithful geyser. *Applied Statistics*, 39 (3), pp. 357–365.

Banachewicz, K., Lucas, A. and Vaart, A. (2008), Modelling portfolio defaults using hidden Markov models with covariates. *Econometrics Journal*, 11 (1), pp. 155–171.

Bartolucci, F. and De Luca, G. (2003), Likelihood-based inference for asymmetric stochastic volatility models. *Computational Statistics and Data Analysis*, 42 (3), pp. 445–449.

Bartolucci, F., Lupparelli, M. and Montanari, G. E. (2009), Latent Markov model for longitudinal binary data: An application to the performance evaluation of nursing homes. *Annals of Applied Statistics*, 3 (2), pp. 611–636.

Basel Committee on Banking Supervision (2006), *International Convergence of Capital Measurement and Capital Standards: A Revised Framework.* Bank for International Settlements, Basel.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003), Latent Dirichlet allocation. *The Journal of Machine Learning Research*, 3, pp. 993–1022.

Brémaud, P. (1999), *Markov Chains.* New York: Springer.

Broto, C. and Ruiz, E. (2004), Estimation methods for stochastic volatility models: a survey. *Journal of Economic Surveys*, 18 (5), pp. 613–649.

Bulla, J. (2006), *Application of Hidden Markov Models and Hidden Semi-Markov Models to Financial Time Series.* PhD thesis, University of Göttingen.

Bulla, J. and Berzel, A. (2006), Computational issues in parameter estimation for stationary hidden Markov models. *Computational Statistics*, 23 (1), pp. 1–18.

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

Bulla, J., Bulla, I. and Nenadić, O. (2009), An R package for analyzing hidden semi-Markov models. *Computational Statistics and Data Analysis*, 54 (3), pp. 611–619.

Cappé, O., Moulines, E. and Rydén, T. (2005), *Inference in Hidden Markov Models.* New York: Springer.

Carlin, B. P., Polson, N. G. and Stoffer, D. S. (1992), A Monte Carlo Approach to Nonnormal and Nonlinear State-Space Modeling. *Journal of the American Statistical Association*, 87 (418), pp. 439–500.

Chan, K. S. and Ledolter, J. (1995), Monte Carlo EM Estimation for Time Series Models Involving Counts. *Journal of the American Statistical Association*, 90 (429), pp. 242–252.

Chaubert–Pereira, F., Guédon, Y., Lavergne, C. and Trottier, C. (2008), Estimating Markov and semi-Markov switching linear mixed models with individual-wise random effects. *Computational Statistics, COMPSTAT'2008, 18th Symposium of IASC*, II, pp. 11–18.

Chib, S., Nardari, F. and Shephard, N. (2002), Markov chain Monte Carlo methods for stochastic volatility models. *Journal of Econometrics*, 108 (2), pp. 281–316.

Cox, D. R. and Miller, H.D. (1965), *The Theory of Stochastic Processes.* London: Chapman and Hall.

Crainiceanu, C. M., Caffo, B. S., Di, C. Z. and Punjabi, N. M. (2009), Nonparametric signal extraction and measurement error in the analysis of electroencephalographic activity during sleep. *Journal of the American Statistical Association*, 104 (486), pp. 541–555.

Czado, C. and Song, P. X.-K. (2008), State space mixed models for longitudinal observations with binary and binomial responses. *Statistical Papers*, 49 (4), pp. 691–714.

Danielsson, J. (1994), Stochastic volatility in asset prices: Estimation with simulated maximum likelihood. *Journal of Econometrics*, 64 (1–2), pp. 375–400.

Davis, R. and Rodriguez-Yam, G. (2005), Estimation for state-space models based on a likelihood approximation. *Statistica Sinica*, 15 (2), pp. 381–406.

Di, C. Z., Crainiceanu, C. M., Caffo, B. S. and Punjabi, N. M. (2009), Multilevel functional principal component analysis. *Annals of Applied Statistics*, 3 (1), pp. 458-488.

Doroshenkov, L. G., Konyshev, V. A. and Selishchev, S. V. (2007), Classification of human sleep stages based on EEG processing using hidden Markov models. *Biomedical Engineering*, 41 (1), pp. 25–28.

Durbin, J. and Koopman, S. J. (1997), Monte Carlo maximum likelihood estimation for non-Gaussian state space models. *Biometrika*, 84 (3), pp. 669–684.

Durbin, R., Eddy, S., Krogh, A. and Mitchison, G. J. (1998), *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids.* Cambridge: Cambridge University Press.

Durbin, J. and Koopman, S. J. (2001), *Time Series Analysis by State Space Methods.* Oxford: Oxford University Press.

Ephraim, Y. and Merhav, N. (2002), Hidden Markov processes. *IEEE Transactions on Information Theory*, 48 (6), pp. 1518–1569.

Ferguson, J. D. (1980), Variable duration models for speech. In: *Proceedings of the Symposium on the Applications of Hidden Markov Models to Text and Speech*, pp. 143–179. Princeton, New Jersey.

Flexer, A., Gruber, G. and Dorffner, G. (2005), A reliable probabilistic sleep stager based on a single EEG signal. *Artificial Intelligence in Medicine*, 33 (3), pp. 199–208.

Fokianos, K., Rahbek, A. and Tjøstheim, D. (2009), Poisson Autoregression. *Journal of the American Statistical Association*, 104 (488), pp. 1430–1439.

Franses, P. H. and van Dijk, D. (2000), *Non-linear Time Series Models in Empirical Finance.* Cambridge: Cambridge University Press.

Frühwirth-Schnatter, S. (2006), *Finite Mixture and Markov Switching Models.* New York: Springer.

Fridman, M. and Harris, L. (1998), A maximum likelihood approach for non-Gaussian stochastic volatility models. *Journal of Business & Economic Statistics*, 16 (3), pp. 284–291.

Gami, A. S., Howard, D. E., Olson, E. J. and Somers, V. K. (2005), Day-Night Pattern of Sudden Death in Obstructive Sleep Apnea. *New England Journal of Medicine*, 352 (12), pp. 1206–1214.

Grunwald, G. K., Hyndman, R. J., Tedesco, L. and Tweedie, R. L. (2000), Non-Gaussian conditional linear AR(1) models. *Australian and New Zealand Journal of Statistics*, 42 (4), pp. 479–495.

Guédon, Y. (2003), Estimating hidden semi-Markov chains from discrete sequences. *Journal of Computational and Graphical Statistics*, 12 (3), pp. 604–639.

Guédon, Y. (2005), Hidden hybrid Markov/semi-Markov chains. *Computational Statistics and Data Analysis*, 49 (3), pp. 663–688.

Hamilton, J. D. (1989), A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle. *Econometrica*, 57 (2), pp. 357–384.

Harvey, A. C. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge: Cambridge University Press.

Harvey, A. C., Ruiz, E. and Shephard, N. (1994), Multivariate stochastic volatility models. *Review of Economic Studies*, 61 (2), pp. 247–264.

Ho, D. E., Imai, K., King, G. and Stuart, E. A. (forthcoming), MatchIt: Nonparametric Preprocessing for Parametric Causal Inference. *Journal of Statistical Software*, http://gking.harvard.edu/matchit.

Hurwitz, S., Kumar, A., Taylor, R. and Heasler, H. (2008), Climate-induced variations of geyser periodicity in Yellowstone National Park, USA. *Geology*, 36 (6), pp. 451–454.

Ihaka, R. and Gentleman, R. (1996), R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5 (3), pp. 299–314.

Jacquier, E., Polson, N. G. and Rossi, P. E. (1994), Bayesian analysis of stochastic volatility models (with discussion). *Journal of Business & Economic Statistics*, 12 (4), pp. 371–417.

Johnson, M. T. (2005), Capacity and complexity of HMM duration modeling techniques. *IEEE Signal Processing Letters*, 12 (5), pp. 407–410.

Julier, S. J. and Uhlmann, J. K. (2004), Unscented filtering and nonlinear estimation. *Proceedings of the IEEE*, 92 (3). pp. 401–422.

Karlin, S. (1983), 11th R. A. Fisher Memorial Lecture, Royal Society, 20 April 1983.

Kim, S., Shephard, N. and Chib, S. (1998), Stochastic volatility: Likelihood inference and comparison with ARCH models. *The Review of Economic Studies*, 65 (3), pp. 361–393.

Kitagawa, G. (1987), Non-Gaussian state-space modeling of nonstationary time series (with discussion). *Journal of the American Statistical Association*, 82 (400), pp. 1032–1063.

Kitagawa, G. (1996), Monte Carlo Filter and Smoother for Non-Gaussian Nonlinear State Space Models. *Journal of Computational and Graphical Statistics*, 5 (1), pp. 1–25.

Kleinrock, L. (1975), *Queueing Systems, Vol. 1: Theory*. New York: Wiley.

Kneib, T. and Hennerfeind, A. (2008), Bayesian Semiparametric Multi-State Models. *Statistical Modelling*, 8 (2), pp. 169–198.

Koopman, S. J., Shephard, N. and Doornik, J. (1999), Statistical algorithms for models in state space using SsfPack 2.2. *Econometrics Journal*, 2 (1), pp. 113–166.

Krogh, A., Brown, M., Mian, I. S., Sjölander, K. and Haussler, D. (1994), Hidden Markov models in computational biology. Applications to protein modeling. *Journal of Molecular Biology*, 235 (5), pp. 1501–1531.

Kulkarni, V. G. (1995), *Modeling and Analysis of Stochastic Systems.* London: Chapman and Hall.

Langrock, R., MacDonald, I. L. and Zucchini, W. (2010), Estimating standard and nonstandard stochastic volatility models using structured hidden Markov models. In revision for *Journal of Empirical Finance*.

Langrock, R., Swihart, B. J., Caffo, B. S., Crainiceanu, C. M. and Punjabi, N. M. (2010), Population hidden Markov models with application to comparing dynamics in sleep electroencephalograms. Preprint submitted to *Annals of Applied Statistics*.

Langrock, R. (2010), Some applications of nonlinear and non-Gaussian state-space modelling by means of hidden Markov models. In revision for *Journal of Applied Statistics*.

Langrock, R. and Zucchini, W. (2011), Hidden Markov models with arbitrary dwell-time distributions. *Computational Statistics and Data Analysis*, 55 (1), pp. 715–724.

Le Strat, Y. and Carrat, F. (1999), Monitoring epidemiologic surveillance data using hidden Markov models. *Statistics in Medicine*, 18 (24), pp. 3463–3478.

MacDonald, I. L. and Zucchini, W. (1997), *Hidden Markov and other Models for Discrete-valued Time Series.* London: Chapman and Hall.

Maruotti, A. (2007), *Hidden Markov models for longitudinal data.* PhD thesis, University of Rome La Sapienza.

Melino, A. and Turnbull, S. M. (1990), Pricing foreign currency options with stochastic volatility. *Journal of Econometrics*, 45 (1–2), pp. 239–265.

Penny, W. and Roberts, S. (1998), Gaussian Observation Hidden Markov Models for EEG Analysis. *Technical report TR-98-12*, Imperial College, London.

Punjabi, N. M., Caffo, B. S., Goodwin, J. L., Gottlieb, D. J., Newman, A. B., O'Connor, G. T., Rapoport, D. M., Redline, S., Resnick, H. E., Robbins, J. A., Shahar, E., Unruh, M. L. and Samet, J. M. (2009), Sleep-Disordered Breathing and Mortality: A Prospective Cohort Study. *PLoS Med*, 6 (8), e1000132.

Quan, S. E., Howard, T. V., Iber, C., Kiley, J. P., Nieto, F. J., O'Connor, G. T., Rapoport, D. M., Redline, S., Robbins, J., Samet, J. M., and Wahl, P. W. (1997), The Sleep Heart Health study: Design, rationale, and methods. *Sleep*, 20 (12), pp. 1077–1085.

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

Rechtschaffen, A. and Kales, A. (1968), *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects.* National Institute of Health Publications 204, US Government Printing Office, Washington DC.

Robert, C. P. and Titterington, D. M. (1998), Reparameterization strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation. *Statistics and Computing*, 8 (2), pp. 145–158.

Rosenbaum, P. R. and Rubin, D. B. (1983), The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70 (1), pp. 41–55.

Rosenblatt, M. (1952), Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, 23 (3), pp. 470–472.

Russell, M. J. and Cook, A. E. (1987), Experimental evaluation of duration modelling techniques for automatic speech recognition. *IEEE Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, pp. 2376–2379.

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), pp. 217–244.

Sandmann, G. and Koopman, S. J. (1998), Estimation of stochastic volatility models via Monte Carlo maximum likelihood. *Journal of Econometrics*, 87 (2), pp. 271–301.

Sansom, J. and Thomson, P. (2001), Fitting hidden semi-Markov models to breakpoint rainfall data. *Journal of Applied Probability*, 38A, pp. 142–157.

Sansom, J. and Thomson, P. (2007), On rainfall seasonality using a hidden semi-Markov model. *Journal of Geophysical Research*, 112, D15105, doi:10.1029/2006JD008342.

Sansom, J. and Thomson, P. (2010), A hidden seasonal switching model for high-resolution breakpoint rainfall data. *Water Resources Research*, 46, W08510, doi:10.1029/2009WR008602.

Shephard, N. (1996), Statistical aspects of ARCH and stochastic volatility. In: *Time Series Models: In econometrics, finance and other fields*, Cox, D. R., Hinkley, D. V. and Barndorff-Nielsen, O. E. (eds.), pp. 1–67. Chapman & Hall, London.

Shephard, N. (Editor) (2005), *Stochastic Volatility: Selected Readings.* Oxford: Oxford University Press.

Shumway, R. H. and Stoffer, D. S. (2006), *Time Series Analysis and Its Applications: With R Examples.* 2nd Edition. Springer Texts in Statistics.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis.* London: Chapman & Hall.

So, M. K. P., Lam, K. and Li, W. K. (1998), A stochastic volatility model with Markov switching. *Journal of Business & Economic Statistics*, 16 (2), pp. 244–253.

Srikanthan, R. and McMahon, T. A. (2001), Stochastic generation of annual, monthly and daily climate data: A review. *Hydrology and Earth System Sciences*, 5 (4), pp. 653–670.

Swihart, B. J., Caffo, B. S., Bandeen-Roche, K. and Punjabi, N. M. (2008), Characterizing sleep structure using the hypnogram. *Journal of Clinical Sleep Medicine*, 4 (4), pp. 349–355.

Swiss National Bank (2008), *Financial Stability Report 2008.*

Taylor, S. J. (2005), *Asset Price Dynamics, Volatility, and Prediction.* Princeton: Princeton University Press.

Varin, C. and Vidoni, P. (2005), A note on composite likelihood inference and model selection. *Biometrika*, 92 (3), pp. 519–528.

Vogler, C. and Metaxas, D. (1997), Adapting hidden Markov models for ASL recognition by using three-dimensional computer vision methods. *IEEE International Conference On Systems, Man, and Cybernetics. Computational Cybernetics And Simulation*, 1, pp. 156–161.

Weisberg, S. (2005), *Applied Linear Regression.* Third Edition. New York: Wiley.

Welch, G. and Bishop, G. (1995), An Introduction to the Kalman filter. *UNC-CH Computer Science Technical Report 95–041.*

Wong, C. S. and Li, W. K. (2000), On a mixture autoregressive model. *Journal of the Royal Statistical Society B*, 62 (1), pp. 95–115.

Wong, W. K. (2010), Backtesting value-at-risk based on tail losses. *Journal of Empirical Finance*, 17 (3), pp. 526–538.

Woolhiser, D. A. (1992), Modeling daily precipitation – Progress and problems. In: *Statistics in the Environmental and Earth Sciences*, Walden, A. T. and Guttorp, P. (eds.), pp. 71–89, Edward Arnold.

Yassouridis, A., Steiger, A., Klinger, A. and Fahrmeir, L. (1999), Modelling and exploring human sleep with event history analysis. *Journal of Sleep Research*, 8 (1), pp. 25–36.

Yu, S.-Z. and Kobayashi, H. (2003), An efficient forward-backward algorithm for an explicit-duration hidden Markov model. *IEEE Signal Processing Letters*, 10 (1), pp. 11–14.

Zeger, S. L. (1988), A Regression Model for Time Series of Counts. *Biometrika*, 75 (4), pp. 621–629.

Zhang, L., Samet, J., Caffo, B. S., Bankman, I. and Punjabi, N. M. (2008), Power Spectral Analysis of EEG Activity During Sleep in Cigarette Smokers. *Chest*, 133 (2), pp. 2427–2432.

Zhong, S. and Ghosh, J. (2002), HMMs and coupled HMMs for multi-channel EEG classification. *IEEE Proceedings of the International Joint Conference on Neural Networks*, pp. 1154–1159.

Zucchini, W. and MacDonald, I. L. (2009), *Hidden Markov Models for Time Series: An Introduction Using R.* London: Chapman & Hall.