Application of Hidden Markov Models and Hidden Semi-Markov Models to Financial Time Series

Dissertation
Presented for the Degree of Doctor of Philosophy
at the Faculty of Economics and Business Administration
of the Georg-August-University of Göttingen

by
Jan Bulla
from
Hannover, Germany

Göttingen, 2006

First Examiner:	Prof. Dr. Walter Zucchini
Second Examiner:	Prof. Dr. Heinrich Hering
Third Examiner:	Prof. Dr. Wolfgang Benner
Day of oral exams:	6.7.2006

... it's [still] shining blue! (J.-J. R.)

Contents

List of Figures

List of Tables

Acknowledgements

First of all, I would like to thank my thesis advisor, Prof. Walter Zucchini, for his academic support and the Center for Statistics as well as the Friedrich-Ebert Stiftung for their financial support.

Moreover, I would also like to render my thanks to the members of the Institute for Statistics and Econometrics and the Centre for Statistics at Göttingen. Thanks are also due especially to Daniel Adler, Oleg Nenadic, Richard Sachsenhausen and Karthinathan Thangavelu who provided constant help and support for my work and answered all my boring questions. I owe special thanks to my co-authors Andreas Berzel and Sascha Mergner who spent numerous hours on our joint research. I would like to sincerely thank Prof. Peter Thomson and Dr. Yann Guédon for their inspiring comments on my work with Hidden Semi-Markov Models.

I am also indebted to the office staff, in particular Hertha Zimmer from the Institute for Mathematical Stochastics for her exceptional help in all administrative matters. I greatly appreciate the financial support of Prof. Manfred Denker and Prof. Hartje Kriete who offered me the chance to participate in inspiring conferences and delicious wine-tasting sessions in Australia and New Zealand.

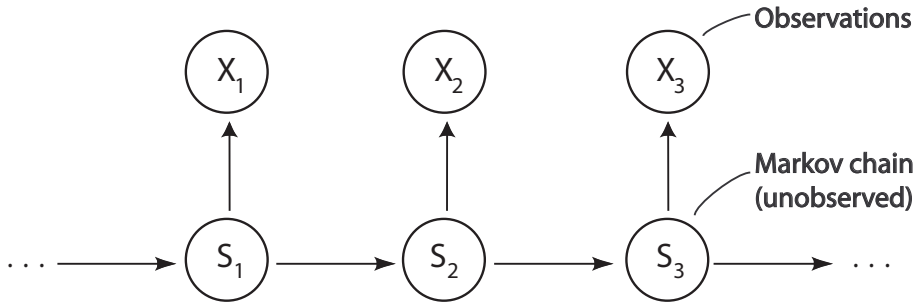
On a more personal note, I am grateful to Oana Serea and Ingo Bulla for their continual support. I would also like to thank my parents for supporting their son still studying at his babylonic age. My heartfelt thanks also go to Wolfgang Schwarzwäller for encouraging me to start the thesis and Prof. Heinrich Hering for his valuable advice over the many years.

Chapter 1

Introduction

Hidden Markov Models (HMMs) and Hidden Semi-Markov Models (HSMMs) provide flexible, general-purpose models for univariate and multivariate time series, especially for discrete-valued series, categorical series, circular-valued series and many other types of observations. They can be considered as a special class of mixture models. The common properties of HMMs and HSMMs are, first of all, that both are built from two stochastic processes: an observed process and an underlying 'hidden' (unobserved) process. The basic structure of HMMs and HSMMs is illustrated in Figures 1.1 and 1.2, respectively.

Figure 1.1: Basic structure of a Hidden Markov Model



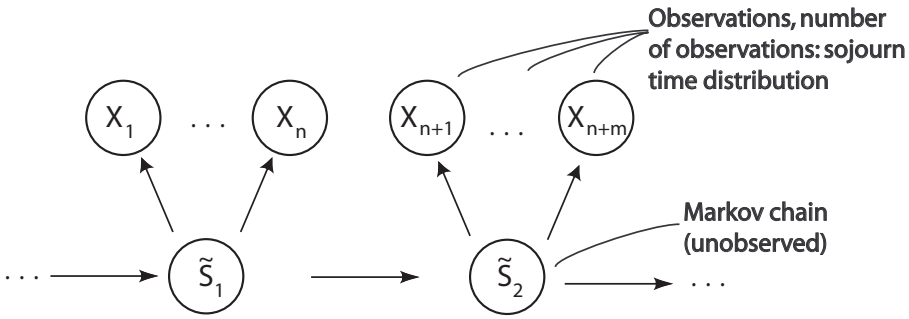
The models are a combination of the following two processes:

- a (semi-)Markov chain S_t which determines the state at time t, and
- a state-dependent process X_t which generates the observation depending

on the current state of S_t.

Moreover, they fulfill the so-called conditional independence property: Given the hidden state at time t, the distribution of the observation at this time is fully determined. A very important consequence of these assumptions is the correlation structure of the observed data. While the autocorrelation function of HMMs is of a particular shape due to the Markov-property of the hidden process, that of the HSMM is more flexible and offers a large variety of possible temporal dependence structures.

Figure 1.2: Basic structure of a Hidden Semi-Markov Model



HMMs and HSMMs have been used for more than two decades in signal-processing applications, especially in the context of automatic speech recognition (e.g. Ferguson 1980, Rabiner 1989). In this context, they allow one to make inferences about the unobserved process. In economic time series modeling, the regime-switching models based on the seminal works of Hamilton (1989, 1990) are a very well-known application of HMMs. Another application is described in the widely known article of Rydén et al. (1998) who analyzed the variation of a daily return series from the S&P 500 index by a HMM.

Though the study of HMMs began in the mid-sixties with the paper of Baum & Petrie (1966), the first application of HSMMs was analyzed in 1980 by Ferguson (1980). Subsequently, various aspects of the models have been considered, e.g., the estimation of the order of HMMs (Rydén 1995*b*) or asymptotic properties of maximum likelihood estimators for HMMs (Bickel et al. 1998, Douc & Matias 2001, Rydén 1995*a*) and HSMMs (Barbu & Limnios 2005). Although interest in HMMs and HSMMs has continuously increased during the past years, and

numerous articles on theoretical and practical aspects have been published, several gaps remain. This thesis addresses some of them, divided into three main topics:

1. Computational issues in parameter estimation of stationary hidden Markov models.

2. A Markov switching approach to model time-varying Beta risk of pan-European Industry portfolios.

3. Stylized facts of financial time series and HSMMs.

The decision to work on the first topic was motivated by the fact that the parameters of a HMM can be estimated by direct numerical maximization (DNM) of the log-likelihood function or, more popularly, using the expectation-maximization (EM) algorithm. Although neither of the algorithms is superior to the other in all respects, researchers and practitioners who work with HMMs tend to use only one of the two, and to ignore the other. We compared the two methods in terms of their speed of convergence, effect of different model parameterizations, how the fitted-log likelihood depends on the true parameter values and on the starting values of the algorithms. Further, it is desirable to fit a stationary HMM in many applications. However, the standard form of the EM algorithm is not designed to do this and therefore, in most cases, authors who use it fit homogeneous but non-stationary models instead. We show how the EM algorithm could be modified to fit stationary HMMs. We propose a hybrid algorithm that is designed to combine the advantageous features of the EM and DNM algorithms, and compare the performance of the three algorithms (EM, DNM and the hybrid) using simulated data from a designed experiment, and also a real data set. We then describe the results of an experiment to assess the true coverage probability of bootstrap-based confidence intervals for the parameters.
The results of the comparison of the EM algorithm and DNM clearly show the trade-off between stability and performance. The hybrid algorithm seems to provide an excellent compromise; it is as stable as the EM-algorithm but it converges faster. Further, we show that the true coverage probability for bootstrap-based confidence intervals, obtained by parametric bootstrap, may be unreliable for models whose state-dependent parameters lie close to each other.

The rationale to take up the second topic, a Markov switching approach to model time-varying beta risk, was the development of a joint model for many

financial time series. The modeling of daily return series with HMMs has been investigated by several authors. After the seminal work of Rydén et al. (1998) who showed that the temporal and distributional properties of daily returns series are well reproduced by the two- and three-state HMMs with normal components, several other authors followed their ideas (see, e.g., Cecchetti et al. 1990, Linne 2002, Bialkowski 2003).
For many applications it is desirable to model a portfolio comprising multiple assets, e.g., a portfolio of European shares selected from the Dow Jones EURO STOXX 600. Fitting a multivariate HMM with normal component distributions would require the estimation of the covariance matrix for each of the states. In the worst case, considering the portfolio of a professional investor which is composed of all 600 shares, the procedure would involve a matrix of dimension 600×600 yielding 180300 parameters to be estimated for each state. It is obvious that such a model would be grossly over-parameterized, resulting in unreliable estimates.
A possible solution to the quadratic increase of the number of parameters is based on the Capital Asset Pricing Model (CAPM). In this model, the return of each asset is linearly dependent to the market return (plus an error term):

$$R_{it} = \alpha_i + \beta_i R_{0t} + \epsilon_{it}, \quad \epsilon_{it} \sim N(0, \sigma_i^2),$$

where R_{it}, R_{0t} are the returns of the ith asset and the market, respectively. The error term is represented by ϵ_{it}; β_i is the market or systematic risk. In this setup, the number of parameters increases only linearly with the number of assets considered. The joint behavior of all assets is modeled by the common dependence on the market return.
We study the performance of two Markov switching models based on the approaches of Fridman (1994) and Huang (2000), and compare their forecast performances to three models, namely a bivariate t-GARCH(1,1) model, two Kalman filter based approaches and a bivariate stochastic volatility model. The main results of the comparisons indicate that the random walk process in connection with the Kalman filter is the preferred model to describe and forecast the time-varying behavior of sector betas in a European context, while the two proposed Markov switching models yielded unsatisfactory results.

The third and main topic addressed in this study is HSMMs, an extension of the well known class of HMMs. For HSMMs, the runlength distributions can be modeled explicitly instead of implicitly following the geometric distribution of a HMM. Ferguson (1980) considered HSMMs as an alternative approach to the classical HMMs for speech modeling because the latter were not flexible enough to describe the time spent in a given state. After this pioneering work,

several problems related to hidden semi-Markov chains were further investigated by different authors, e.g., Levinson (1986), Guédon & Cocozza-Thivent (1990), Guédon (2003), Sansom & Thomson (2001), Yu & Kobayashi (2003) and different parametric hypotheses were considered for the state occupancy, as well as for the distribution of the observations. We provide estimation procedures for a variety of HSMMs belonging to the recently introduced class of right-censored HSMMs. In contrast to the original model of Ferguson (1980), they do not require the assumption that the end of a sequence systematically coincides with the exit from a state. Such an assumption is unrealistic for many financial time series, daily return series in particular.
The ability of a HMM to reproduce several stylized facts of daily return series was illustrated by Rydén et al. (1998). However, they point out that one stylized fact cannot be reproduced by a HMM, namely the slowly decaying autocorrelation function of squared returns, which plays a key role in risk-measurement and the pricing of derivatives. The lack of flexibility of a HMM to model this temporal higher order dependence can be explained by the implicit geometric distributed sojourn time in the hidden states.
We present two alternative HSMM-based approaches to model eighteen series of daily sector returns with about 5000 observations. Our key result is that the slowly decaying autocorrelation function is significantly better described by a HSMM with negative binomial sojourn time and normal conditional distributions.

This thesis is structured as follows. An introduction to the basics of HMMs is provided in Chapter 2. The computational issues in parameter estimation of stationary HMMs are addressed in Chapter 3 and the Markov switching approach to model time-varying beta risk are subject of Chapter 4. Chapter 5 provides the theoretical framework for the estimation of HSMMs. The application of HSMMs to daily return series is presented in Chapter 6. The discussion in Chapter 7 recapitulates the main results and offers some suggestions for future research.

Chapter 2

Hidden Markov Models

Hidden Markov Models (HMMs) are a class of models in which the distribution that generates an observation depends on the state of an underlying but unobserved Markov process. In this chapter we provide a brief introduction to HMMs and explain the basics of the underlying theory.

HMMs have been applied in the field of signal-processing for more than two decades, especially in the context of automatic speech recognition. However, they also provide flexible, general-purpose models for univariate and multivariate time series, including discrete-valued series, categorical series, circular-valued series and many other types of observations. Consequently, the interest in the theory and applications of HMMs is rapidly expanding to other fields, e.g.:

- Various kinds of recognition: faces, speech, gesture, handwriting/signature.
- Bioinformatics: biological sequence analysis.
- Environment: wind direction, rainfall, earthquakes.
- Finance: daily return series.

The bibliography lists several articles and monographs that deal with the application of HMMs in these fields. Important references include Durbin et al. (1998), Elliott et al. (1995), Ephraim & Merhav (2002), Koski (2001), Rabiner (1989).

The application of HMMs in the above mentioned fields is mainly due to their versatility and mathematical tractability. In detail, they are characterized by the following properties (cf. MacDonald & Zucchini 1997):

- Availability of all moments: mean, variance, autocorrelations.
- The likelihood is easy to compute; the computation is linear in the number of observations.
- The marginal distributions are easy to determine and missing observations can be handled with minor effort.
- The conditional distributions are available, outlier identification is possible and forecast distributions can be calculated.

In addition, HMMs are interpretable in many cases and can easily accommodate additional covariates. Furthermore, they are moderately parsimonious; in many applications a simple two-state model provides a reasonable fit.

This chapter is organized as follows. In Section 2.1 we introduce independent mixture models and discrete Markov chains, the two main components of HMMs. Subsequently in Section 2.2, we present the construction of a HMM and show how the likelihood can be calculated.

2.1 Fundamentals

This section provides a brief introduction to two fundamental concepts that are necessary to understand the basic structure of Hidden Markov Models (HMMs). As the marginal distribution of a HMM is a discrete mixture model, we first provide a general outline of mixture distributions in Section 2.1.1. Then, we introduce Markov chains in Section 2.1.2 since the selection process of the parameters of a HMM is modeled by a Markov chain.

2.1.1 Independent Mixture Distributions

In general, an independent mixture distribution consists of a certain number of **component conditional distributions**. In some applications it isi reasonable to assume that the population heterogeneity is modeled by a continuous mixture. More details on continuous mixtures can be found, e.g., in Böhning (1999).
However, the focus of the subsequent work lies on discrete mixtures with a finite number of component distributions. These component distributions can be either discrete or continuous. In the case of two-component distributions,

the mixture distribution is characterized by the two random variables X_0 and X_1 along with their probability functions or probability density functions (pdf).

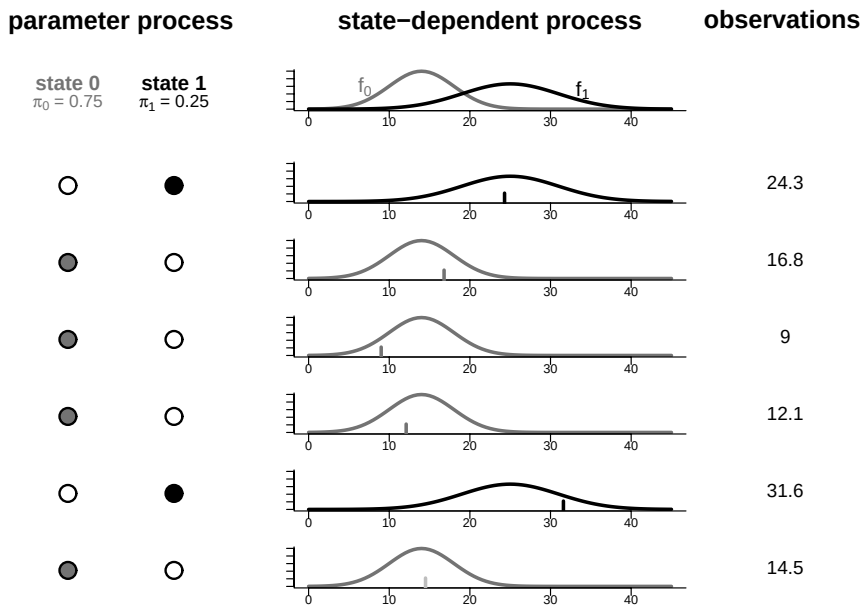
Random variable	Probability function	pdf
X_0	$p_0(x)$	$f_0(x)$
X_1	$p_1(x)$	$f_1(x)$

Moreover, for the **parameter process** a discrete random variable S is needed to perform the mixture:

$$S := \begin{cases} 0 \text{ with probability } \pi_0 \\ 1 \text{ with probability } \pi_1 = 1 - \pi_0 \end{cases}.$$

One may imagine S like tossing a coin: If S takes the value 0, then an observation is a realization of X_0; if S takes the value 1, then an observation is a realization of X_1. The structure of that process for the case of two continuous component distributions is shown in Figure 2.1.

Figure 2.1: Process structure of a two-component mixture distribution



Note that, in practice, we do not know which way the coin landed. Only the observations generated by either X_0 or X_1 can be observed and, in most cases,

they cannot be assigned to a distinct random variable.

Given the probability of each component and the respective probability distributions, the probability density function of the mixture can be computed easily. For ease of notation we only treat the continuous case. Let X denote the outcome of the mixture. Then, its probability density function is given by

$$f(x) = \pi_0 f_0(x) + \pi_1 f_1(x).$$

The extension to the J-component case is straightforward. Let $\pi_0, \ldots, \pi_{J-1}$ denote the weights assigned to the different components and $f_0, \ldots, f_{J-1}$ denote their corresponding probability density functions. Then, the distribution of the outcome, X, is a mixture and can be easily calculated as a linear combination of the component distributions:

$$f(x) = \sum_{i=0}^{J-1} \pi_i f_i(x).$$

Moreover, the calculation of the k-th moment $E(X^k)$ is simply a linear combination of the respective moments of its components:

$$E(X^k) = \sum_{i=0}^{J-1} \pi_i E(X_i^k), \quad k \in \{1, 2, ...\}.$$

Note that this does not hold for the central moments, e.g., the variance of a mixture:

$$Var(X) \neq \sum_{i=1}^{J-1} \pi_i Var(X_i).$$

The estimation of the parameters of a mixture distribution is usually performed by a maximum likelihood (ML) algorithm. The likelihood of a mixture model with J components is given by

$$L(\theta_0, \ldots, \theta_{J-1}, \pi_0, \ldots, \pi_{J-1}, x_0, \ldots, x_{\tau-1}) = \prod_{j=0}^{\tau-1} \sum_{i=0}^{J-1} \pi_i f_i(x_j, \theta_i).$$

where $\theta_0, \ldots, \theta_{J-1}$ are the parameter vectors of the component distributions, $\pi_0, \ldots, \pi_{J-1}$ are the mixing parameters, and $x_0, ..., x_{\tau-1}$ are the observations.

It is not possible to maximize this likelihood function analytically. Therefore, the parameter estimation has to be carried out by numerical maximization of the likelihood using special software. A very useful software package for the estimation of mixture models is C.A.MAN, developed by Böhning et al. (1992).[1]

One example in which mixture distributions with continuous components can be applied is the analysis of stock returns, as demonstrated in the following example. Figure 2.2 shows the daily percentage returns of the DAX 30, DJ STOXX, and FTSE 100 Index between 1$^{\text{st}}$ January 1994 and 31$^{\text{st}}$ December 2004.
It is visible that the variance of the returns is not constant over the whole trading period. Instead, there are some periods with low absolute returns and others with high absolute returns – there is "volatility clustering" observable for many financial time series. For that reason, a simple normal distribution does not provide an adequate description of the daily percentage return on the indices, as can be seen in Figure 2.3, which shows a histogram of the daily returns and a fitted normal distribution.

The fitted normal distribution underestimates the probability of extremely low and high absolute returns. The return series also shows excess kurtosis compared to the normal distribution. In contrast, a mixture of three normal distributions as shown in Figure 2.4 provides a better fit. The mixing weights correspond to those obtained by fitting a HMM.

[1]The software package can be downloaded from *http://www.personal.rdg.ac.uk/~sns05dab/Software.html.*

Figure 2.2: Percentage return of the DAX 30, DJ STOXX, and FTSE 100 Index

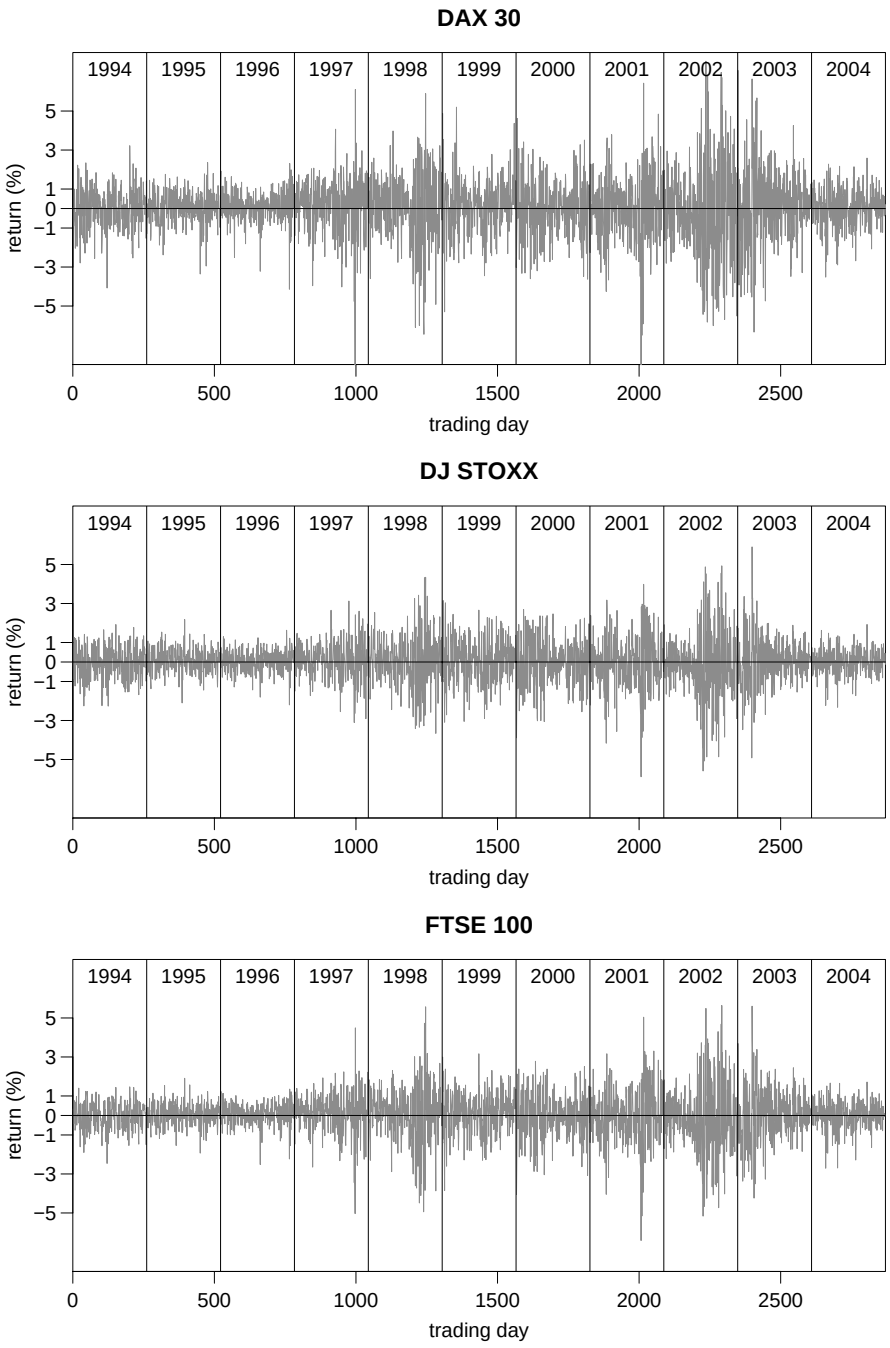


Figure 2.3: Histogram of daily returns of the DAX 30, DJ STOXX, and FTSE 100 Index with fitted normal distributions

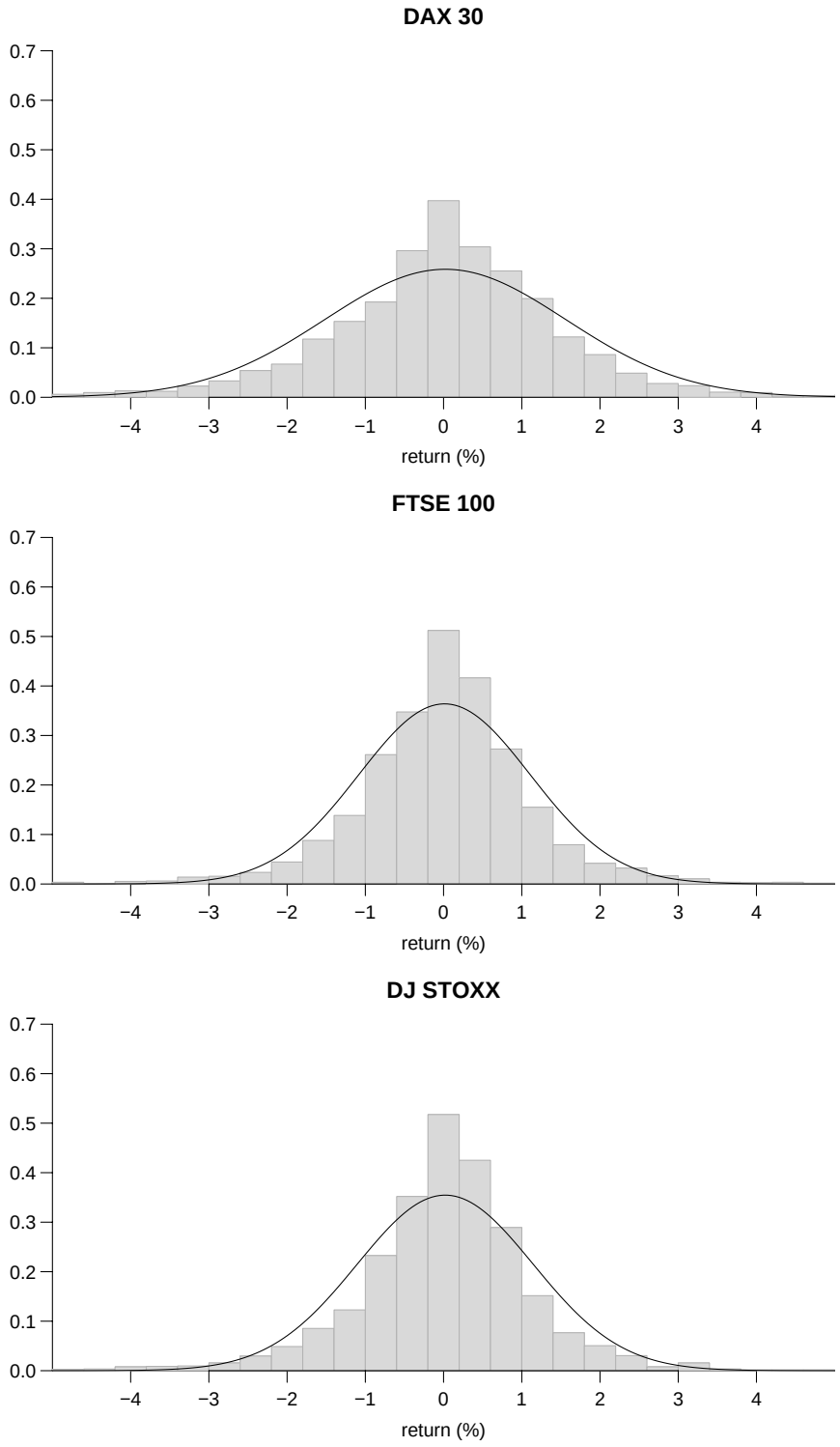
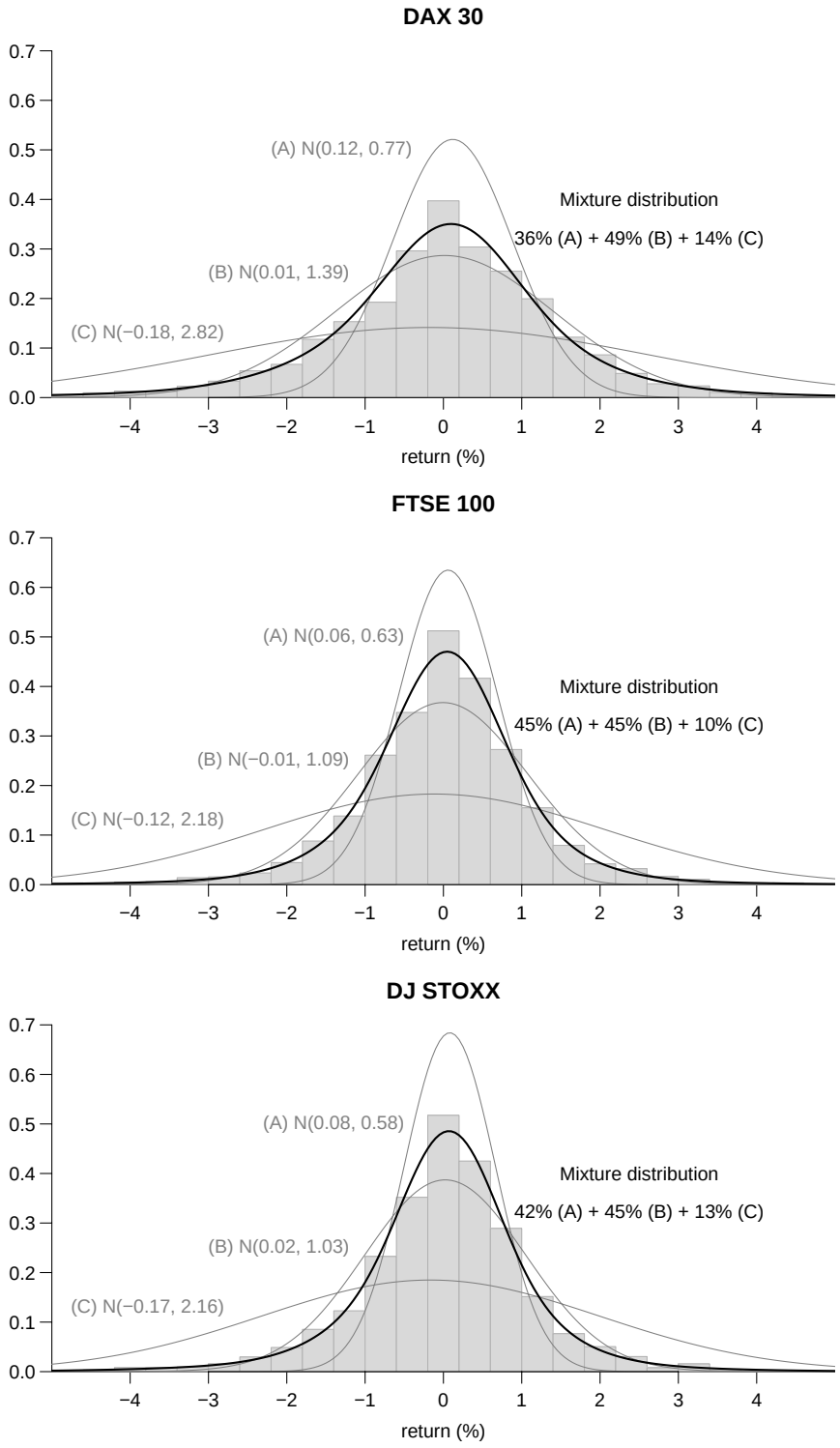


Figure 2.4: Histogram of daily returns of the DAX 30, DJ STOXX, and FTSE 100 Index with fitted mixtures of normal distributions



2.1.2 Markov Chains

As the theory of Markov chains is well documented, we present only a short introduction to the topic and some of their basic properties that are necessary for the construction of HMMs. For a detailed description of Markov chains see, e.g., Grimmett & Stirzaker (2001) or Parzen (1962).

Consider a stochastic process, i.e. a sequence of random variables $\{S_t : t \in 0, 1, \dots\}$ taking values in the state space $\{0, \dots, J-1\}$. For more general Markov processes, the time and state space may also be continuous. However, for this work we deal only with discrete-time Markov processes with discrete state space. Such processes are called **Markov chains**.
A stochastic process $\{S_t\}$ is a Markov process if, roughly speaking, given the current state of the process S_t, the future S_{t+1} is independent of its past S_{t-1}, S_{t-2},..., S_0. More precisely, let $s_0, \dots, s_t, s_{t+1}$ denote a sequence of observations of a stochastic process $\{S_t, t = 0, 1, \dots\}$. $\{S_t\}$ is a Markov process if it has the **Markov property**, namely

$$P(S_{t+1} = s_{t+1} | \underbrace{S_t = s_t, S_{t-1} = s_{t-1}, ..., S_0 = s_0}_{\text{"entire history"}}) = P(S_{t+1} = s_{t+1} | S_t = s_t)$$

for all $t \in \{0, 1, \dots\}$.

A Markov chain is called **homogeneous**, or Markov chain with stationary transition probabilities $p_{ij} := P(S_{t+1} = j \,|\, S_t = i)$, if the transition probabilities are independent of t. The transition probabilities of a homogeneous J-state Markov chain can be summarized in a $J \times J$ **transition probability matrix** (TPM) and can be presented as

$$\boldsymbol{T} := \begin{pmatrix} p_{0\,0} & \cdots & p_{0\,J-1} \\ \vdots & \ddots & \vdots \\ p_{J-1\,0} & \cdots & p_{J-1\,J-1} \end{pmatrix},$$

$$\text{with} \quad p_{ij} = P(S_{t+1} = j \,|\, S_t = i) \;\text{ and }\; \sum_{j=0}^{J-1} p_{ij} = 1,\; i \in \{0, \dots, J-1\}.$$

The TPM $\boldsymbol{T}$ contains the one-step transition probabilities and thus, describes the short-term behavior of the Markov chain. For describing the long-term behavior of a Markov chain, one can define the k-step transition probabilities $p_{ij}(k) := P(S_{t+k} = j \,|\, S_t = i)$. It can be shown that the matrix $\boldsymbol{T}(k)$, which

contains the k-step transition probabilities can be calculated as the k^{th} power of the TPM $\boldsymbol{T}$. That is,

$$\boldsymbol{T}(k) := \begin{pmatrix} p_{0\,0}(k) & \cdots & p_{0\,J-1}(k) \\ \vdots & \ddots & \vdots \\ p_{J-1\,0}(k) & \cdots & p_{J-1\,J-1}(k) \end{pmatrix} = \boldsymbol{T}^k \,.$$

For a proof, see Grimmett & Stirzaker (2001).

In this context one says that state j is accessible from state i, written $i \to j$, if the chain may ever reach state j with positive probability, starting from state i. That is, $i \to j$ if there exists some $k \in \{1, 2, \ldots\}$ with $p_{ij}(k) > 0$. Furthermore, states i and j communicate with each other, which is written as $i \leftrightarrow j$, if $i \to j$ and $j \to i$. We can then call a Markov chain to be **irreducible** if $i \leftrightarrow j$ for all $i, j \in \{0, \ldots, J-1\}$. In the following, as in most applications, we assume the Markov chain to be irreducible.

The k-step transition probabilities provide the conditional probabilities to be in state j at time $t+k$, given that the Markov chain is in state i at time t. However, in general, the marginal probability of the Markov chain to be in state i at a given time t is also of interest. Given the probability distribution for the initial state[2], $\boldsymbol{\pi} := (P(S_1 = 1), \ldots, P(S_1 = m))$ with $\sum_{i=1}^{m} \pi_i = 1$, the distribution of the state at time t can be computed as

$$(P(S_t = 0), \ldots, P(S_t = J-1)) = \boldsymbol{\pi}\boldsymbol{T}^{k-1}.$$

If the Markov chain is homogeneous and irreducible, one can show that $\boldsymbol{\pi}\boldsymbol{T}^{k-1}$ converges to a fixed vector, say $\boldsymbol{\pi}_s$, for large t. This unique vector is called the **stationary distribution** and can be determined by solving

$$\boldsymbol{\pi}_s = \boldsymbol{\pi}_s\boldsymbol{T} \quad \text{subject to} \quad \boldsymbol{\pi}_s\mathbf{1}' = 1.$$

For a proof of this result, see Seneta (1981). A Markov chain is said to be stationary, if the stationary distribution $\boldsymbol{\pi}_s$ exists and if it describes the marginal distribution of the states for all $t \in \{0, 1, \ldots\}$. In particular, for the distribution of the initial state one has that $\boldsymbol{\pi} = \boldsymbol{\pi}_s$. In practice, depending on the application, one has to decide whether it is sensible to assume that the underlying Markov chain of a HMM is stationary or not.

[2]We use the convention that vectors are row vectors

2.2 Hidden Markov Models

In this section we give a brief introduction to HMMs and their basic properties. For further reading, see, e.g., Ephraim & Merhav (2002) or MacDonald & Zucchini (1997). If not indicated otherwise, we also refer to the latter as standard reference for this section.

In an independent mixture model, the sequence of hidden states as well the sequence of observations is independent by definition. If there is correlation between the states, the independent mixture is not an appropriate model anymore as it does not take account of all the information contained in the data. One way of modeling data series with serial correlation is to let the parameter selection process be driven by an unobserved (i.e. hidden) Markov chain. This approach yields the HMM, which is a special case of a dependent mixture. Different underlying processes can also be treated. For example, in Chapter 5 we generalize the parameter selection process to a semi-Markov chain, which yields the HSMMs.

2.2.1 The basic Hidden Markov Model

Let $\{X_t\} = \{X_t, t = 0, 1, \dots\}$ denote a sequence of observations and $\{S_t\} = \{S_t, t = 0, 1, \dots\}$ a Markov chain defined on the state space $\{0, \dots, J-1\}$. For better readability, we introduce the notation

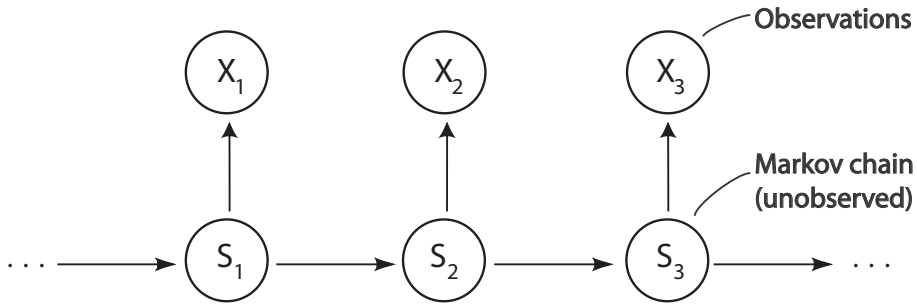
$$X_{t_0}^{t_1} := \{X_{t_0}, \dots, X_{t_1}\}$$

with $t_0 < t_1$; $S_{t_0}^{t_1}$ is defined similarly.
Consider a stochastic process consisting of two parts: Firstly the underlying but unobserved parameter process $\{S_t\}$, which fulfills the Markov property $P(S_t = s_t \,|\, S_1^{t-1} = s_1^{t-1}) = P(S_t = s_t \,|\, S_{t-1} = s_{t-1})$, and secondly the state-dependent observation process $\{X_t\}$, for which the **conditional independence property**

$$P(X_t = x_t \,|\, X_0^{t-1} = x_0^{t-1}, S_0^t = s_0^t) = P(X_t = x_t \,|\, S_t = s_t) \tag{2.1}$$

holds. Then, the pair of stochastic processes $\{(S_t, X_t)\}$ is called a J-state **Hidden Markov Model**. Equation (2.1) means that, if S_t is known, X_t depends only on S_t and not on any previous states or observations. The basic structure of the HMM is illustrated in Figure 2.5.

Figure 2.5: Basic structure of a Hidden Markov Model

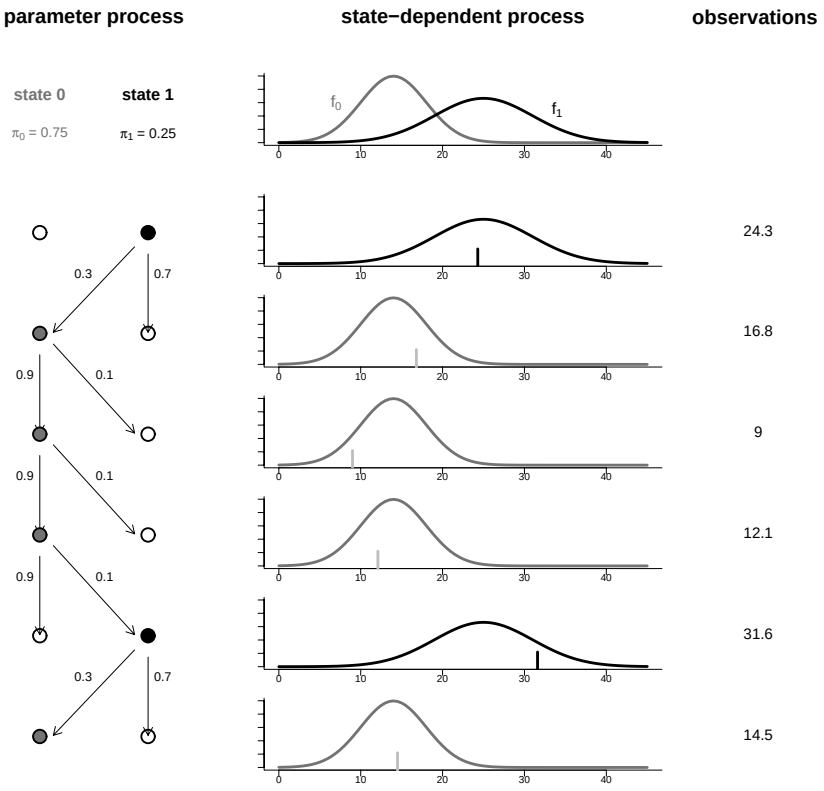


Thus a HMM is a combination of two processes, namely a Markov chain which determines the state at time t, $S_t = s_t$, and a state-dependent process which generates the observation $X_t = x_t$ depending on the current state s_t. In most cases a different distribution is imposed for each possible state of the state space. The Markov chain is assumed to be homogeneous and irreducible with transition probability matrix $\boldsymbol{T}$. By the irreducibility of $\{S_t\}$, there exists a unique stationary distribution of the Markov chain, $\boldsymbol{\pi} = \boldsymbol{\pi}_s$ (cf. Section 2.1.2).

A HMM is rather a theoretical construction. In reality, only the state-dependent process $\{X_t\}$ is observed while the underlying state process $\{S_t\}$ remains unknown. However, in many applications there is a reasonable interpretation for the underlying states. Suppose, for example, that the daily return series introduced in Section 2.1.1 is modeled with a two-state HMM. Then the states of the underlying Markov chain may be interpreted as condition of the financial market, namely a state with high volatility and a state with low volatility representing nervous and calm periods, respectively.

The process generating the observations of a stationary two-state HMM is demonstrated in Figure 2.6. Here the observed sequence equals $(24.3, 16.8, 9, 12.1, 31.6, 14.5)$ and $\text{diag}(\boldsymbol{T}) = (0.9, 0.7)$. In contrast to Figure 2.1, which shows the process structure of a two-component independent mixture model, the probabilities for the state S_{t+1} depend on the state S_t.

Figure 2.6: Process structure of a two-state Hidden Markov Model



For further details on the HMM, including a derivation of the moments and marginal distributions, the treatment of outliers and missing data, forecasting, decoding and smoothing procedures we refer to the manuscript of MacDonald & Zucchini (1997).

2.2.2 The Likelihood of a Hidden Markov Model

The likelihood of a HMM can be expressed in a closed formula, even in a relatively general framework. Let θ be set of all model parameters and let $\boldsymbol{P}(x_t)$ denote a diagonal matrix with the conditional probabilities $b_j(x_t) := P(X_t = x_t \,|\, S_t = j)$, $j = 1, \ldots, m$ on the main diagonal. Then, the likelihood of a HMM can be written as

$$\begin{aligned} L(\theta) &= P(\{X_0 = x_0, \ldots, X_{\tau-1} = x_{\tau-1}\}) \\ &= \boldsymbol{\pi}\boldsymbol{P}(x_0)\boldsymbol{T}\boldsymbol{P}(x_1)\boldsymbol{T}\ldots\boldsymbol{T}\boldsymbol{P}(x_{\tau-1})\mathbf{1}^t, \end{aligned} \tag{2.2}$$

where $\mathbf{1} := (1, \ldots, 1)$.

This form of the likelihood has several appealing properties. For example, stationary as well as non-stationary models can be handled and a (local) maximum can be found by numerical procedures such as Newton-type algorithms or via the so called EM-algorithm.
However, the evaluation of the likelihood is not completely straightforward, as it involves a large number of multiplications of matrices with elements between zero and one, numerical underflow occurs even on modern personal computers. The easiest way to overcome this difficulty is by applying rescaling techniques (see, e.g., Rabiner 1989). These rescaling techniques have to be applied for two of the most commonly utilized methods to maximize the likelihood of a HMM, namely direct numerical maximization and the EM-algorithm. Although other estimation procedures exist (e.g., Particle filters), most researchers prefer either the direct numerical maximization of the likelihood function or the EM-algorithm. Both methods have their own advantages and disadvantages; a comparison of the two approaches is presented in the following Chapter 3.

Chapter 3

Parameter Estimation for Hidden Markov Models

Maximum-likelihood (ML) parameter estimation in Hidden Markov Models (HMMs) can be carried out using either direct numerical maximization or the expectation maximization (EM) algorithm (Baum et al. 1970, Dempster et al. 1977). Although neither of the algorithms is superior to the other in all respects, researchers and practitioners who work with HMMs prefer to use only one of the two algorithms, and tend to ignore the other. The aim of this section is to explore the advantages and disadvantages of both estimation procedures for HMMs.

In many applications, it is desirable to fit a stationary HMM. The EM algorithm is not designed to do this and therefore, in most cases, authors who use the standard form of this algorithm fit homogeneous but non-stationary models instead. We show how the EM algorithm can be modified to fit stationary HMMs.

Direct numerical maximization of the likelihood using Newton-type algorithms generally converges faster than the EM algorithm, especially in the neighborhood of a maximum. However, it requires more accurate initial values than the EM to converge at all.

We implement both the new EM algorithm as well as direct numerical maximization using the software package `R` and assess their performances in terms of flexibility and stability using both simulated and real data sets. In particular, we analyze the speed of convergence, the effect of different model parameterizations and how the fitted-log likelihood depends on the true parameter values and on the initial values of the algorithms.

We suggest that it is possible to take advantage of the desirable properties

of each of the two methods by using a hybrid algorithm, and compare the performance of the three algorithms using simulated data from a designed experiment, and then with a real data set. Such algorithms have been proposed by some authors (e.g., Lange & Weeks 1989, Redner & Walker 1984), but the efficiency of such an algorithm has not yet been reported in the context of HMMs. We fill this gap and, as a by-product of the above simulation experiments, we also investigate the coverage probability of bootstrap interval estimates of the parameters.

This chapter is organized as follows. In Section 3.1 we give a brief description of the two most common methods for estimating the parameters of a HMM. Furthermore, we introduce the new EM algorithm for stationary time series and the hybrid algorithm. Section 3.2 describes the design of the simulation study, the results relating to the performance of direct maximization and the EM algorithm and then of the hybrid algorithm. The coverage probability of bootstrap-based confidence intervals is also addressed. In Section 3.3 we demonstrate the advantages of the hybrid algorithm, by fitting a set of real data. Section 3.4 summarizes the main findings of the chapter and offers some concluding remarks. To keep this section short, only the main results are presented. The entire analysis of this joint work with A. Berzel can be found in Bulla & Berzel (2006).

3.1 Estimation Algorithms for Stationary Hidden Markov Models

The parameters of HMMs are generally estimated using the method of maximum-likelihood (ML). Equation (2.2) shows that the likelihood equations have a highly nonlinear structure and there is no analytical solution for the ML estimates. The two most common approaches to estimate the parameters of a HMM are the EM algorithm and direct numerical maximization (DNM) of likelihood. In this section we present their strengths and weaknesses and introduce a hybrid algorithm, a combination of both. For alternative approaches including variations on ML estimation see, e.g., Archer & Titterington (2002).

3.1.1 Direct Numerical Maximization

We give only a brief account of parameter estimation of HMMs by direct numerical maximization (DNM) methods. For further details we refer to Mac-

Donald & Zucchini (1997). Recalling Equation (2.2), there exists a convenient explicit expression for the log-likelihood of a HMM that can be easily evaluated even for very long sequences of observations. This makes it possible to estimate the parameters by DNM of the log-likelihood function. DNM has appealing properties, especially concerning the treatment of missing observations, flexibility in fitting complex models and the speed of convergence in the neighborhood of a maximum. The main disadvantage of this method is its relatively small circle of convergence.
We use the open source statistical software `R` (R Development Core Team 2005), version 1.9.1, which allows the integrated functions `nlm()` and `optim()` to perform DNM of the negative log-likelihood. The function `nlm()` carries out minimization of a function using a Newton-type algorithm (Dennis & Moré 1977, Schnabel et al. 1985). The function `optim()` offers the Nelder-Mead simplex algorithm (Nelder & Mead 1965), a popular adaptive downhill simplex method for multidimensional unconstrained minimization, which does not require the computation of derivatives. In general, the Nelder-Mead algorithm is more stable; however, it may also get stuck in local minima and is rather slow when compared to the Newton-type minimization. In our study, we use the values of the scaling parameters proposed by Nelder & Mead (1965) and implemented those as default values in the `optim()` function.
Since both the functions `nlm()` and the Nelder-Mead algorithm can only perform unconstrained numerical minimization, the parameter constraints need to be taken into account by different transformation procedures. For the transition probability matrix (TPM), we apply the TR-transformation described in Zucchini & MacDonald (1998). In order to meet the non-negativity constraint of some of the parameters of the state-dependent distributions, we use different transformations and compare their performance.

For simplicity we consider a Poisson HMM; the extension to other models is straightforward. Let λ_i, $i = 0, \ldots, J-1$ denote the state-dependent parameters to be transformed. The simplest transformation is the natural logarithm $\log(\lambda_i)$. A second option is to make use of the fact that the ML estimates of the parameters of the state-dependent distributions, $\hat{\lambda}_i$, can only have support points in the interval $[x_{\min}, x_{\max}]$ where $x_{\min} := \min\{x_0, \ldots, x_{\tau-1}\}$ and $x_{\max} := \max\{x_0, \ldots, x_{\tau-1}\}$ (Böhning 1999). We can restrict the possible range of parameter estimates to that interval by applying a logit-type transformation, $\log\left((\lambda_i - x_{\min})/(x_{\max} - \lambda_i)\right)$.
Following the ideas of Robert & Titterington (1998) and Mengersen & Robert (1996, 1999) introduced for the case of a normal mixture model with two components, in some cases it might be convenient to "order" the states by modeling the differences between the state-dependent parameters instead of the state-

dependent parameters themselves: $(\tau_0, \tau_1, \ldots, \tau_{J-1}) := (\lambda_0, \lambda_1 - \lambda_0, \ldots, \lambda_{J-1} - \lambda_{J-2})$.
Since the ordering of the states can be used for both the log- and the logit-transformations, four different parameterizations have to be taken into account. In the case of the ordered logit parameterization, the range of the logit-transformation has to be adopted, i.e., $\log\left(\tau_i/(x_{\max} - \sum_{j=0}^{i} \tau_j)\right)$.
In the simulation study outlined in Section 3.2.2 we study the performance of these four parameterizations using both a Newton-type and the Nelder-Mead algorithm.

3.1.2 The Stationary EM Algorithm

A popular and routinely used alternative to DNM is the Baum-Welch algorithm, a special case of what subsequently became known as the EM algorithm. An introduction to the EM algorithm can be found in Appendix A. There exists a large literature on the EM algorithm and its application to HMMs. We do not provide any details on this well-established theory and refer to Baum et al. (1970), Dempster et al. (1977), Rabiner (1989), Liporace (1982), Wu (1983).

At this stage, we wish to note that the EM algorithm, in its original implementation in the context of HMMs, can be used to fit a homogeneous, but not a stationary HMM. Thus authors who apply this method of estimation are unable to maximize the likelihood under the assumption that the model is stationary, despite the fact that such an assumption is both natural and desirable in many applications. We show that the EM algorithm can be modified, at modest computational cost, so that it is able to fit a stationary HMM.

After assigning initial values to the parameters, the EM algorithm is implemented by successively iterating the E-step and the M-step until convergence is achieved.

E-step: Compute the Q-function

$$Q(\theta, \theta^{(k)}) = \mathbf{E}\left[\log P(X_0^{\tau-1} = x_0^{\tau-1}, S_0^{\tau-1} = s_0^{\tau-1} \,|\, \theta) \,|\, X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right],$$

where $\theta^{(k)}$ is the current estimate of the parameter vector θ.

M-step: Compute $\theta^{(k+1)}$, the parameter values that maximize the function Q w.r.t. θ:

$$\theta^{(k+1)} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta, \theta^{(k)}).$$

The feasibility of the computation of the M-step depends strongly on the conditional distributions of the observations. If the solution for this maximization problem cannot be obtained analytically then the maximization has to be carried out numerically (see, e.g., Wang & Puterman 2001). The maximization has to be executed for each M-step at considerable computational cost. Furthermore the rate of convergence of the EM can be very slow, namely linear in the neighborhood of a maximum (Dempster et al. 1977).
An important advantage of the EM algorithm is that (under mild conditions) the likelihood increases at each iteration, except at a stationary point (Wu 1983). Of course the increase may take one to only a local, rather than the global, maximum and thus the results do depend on the initial values of the parameters (Dunmur & Titterington 1998). Nevertheless the circle of convergence is relatively large compared to competing algorithms, which leads to high numerical stability in the form of robustness against poor initial values (Hathaway 1986). A major disadvantage of the EM algorithm in the context of HMMs is the lack of flexibility to fit complex models, as the E-step of the algorithm needs to be derived for each new model (Lange & Weeks 1989).

The EM algorithm for HMMs given in the literature works as follows. The three additive parts of the Q-function of a HMM given by

$$Q(\theta, \theta^{(k)}) = \sum_{i=0}^{J-1} \left[\underbrace{\log\, \pi_i\, \psi_1(i) + \left(\sum_{j=0}^{J-1} \sum_{t=0}^{\tau-2} \log\, p_{ij} \xi_t(i,j) \right)}_{(\star)} + \sum_{t=0}^{\tau-1} \log\, b_i(x_t)\, \psi_t(i) \right]$$

$$\begin{aligned} \text{with } \psi_t(i) &:= P(\{S_t = i | X_0^{\tau-1} = x_0^{\tau-1}, \theta\}) \\ \text{and } \xi_t(i,j) &:= P(\{S_t = i, S_{t+1} = j | X_0^{\tau-1} = x_0^{\tau-1}, \theta\} \end{aligned} \tag{3.1}$$

are split up in parts and maximized separately. Clearly, this procedure fits a homogeneous, but non-stationary, HMM because the individual treatment of the summands leads to an estimate $\hat{\boldsymbol{\pi}}$ which is not the stationary distribution of $\hat{\boldsymbol{T}}$. A popular way to impose stationarity is to simply neglect the first term, calculate $\hat{\boldsymbol{T}}$ and then set $\hat{\boldsymbol{\pi}}$ equal to the stationary distribution of $\hat{\boldsymbol{T}}$. However, this approach does not lead to the exact ML estimates of the parameters, except asymptotically.
In order to estimate a stationary Markov chain, the first two summands of (3.1) marked by $(\star)$ have to be treated simultaneously with the stationarity constraint

$$\boldsymbol{\pi} \tilde{\boldsymbol{T}} = (0, 0, \ldots, 0, 1), \tag{3.2}$$

at the M-Step of each iteration. $\tilde{\boldsymbol{T}}$ denotes the matrix obtained by replacing the last column of $\mathbf{1} - \boldsymbol{T}$ by the vector $(1, \ldots, 1)^T$ of length J.
The explicit calculation of a maximizing solution of the system of equations defined by $(\star)$ in (3.1), and (3.2) is more difficult than it appears. Even for the simplest non-trivial HMM with two states, the system becomes intractable.
To fit a stationary initial distribution, we carry out the modified M-step by embedding a numerical maximization procedure at each iteration of the EM algorithm. We note that this procedure is much more efficient than that of carrying out the entire M-step numerically. By taking the values of $\boldsymbol{T}$ of the preceding step as initial values for the maximization procedure, the EM algorithm is not slowed down significantly.

3.1.3 The Hybrid Algorithm

The relative merits of the EM algorithm and DNM have also been discussed by Campillo & Le Gland (1989) in the context of HMMs. They concluded that the EM algorithm is an interesting approach despite slow convergence, slow E-step and complicated M-step. Modifications of the EM algorithm, such as the integration of Newton-type 'accelerators' have been suggested to improve the rate of convergence, but these usually lead to a loss of stability and increase in complexity (Jamshidian & Jennrich 1997, Lange 1995).
An alternative approach is to use hybrid algorithms, which are constructed by combining the EM algorithm with a rapid algorithm with strong local convergence, in our case the Newton-type algorithm, as follows: the estimation procedure starts with the EM algorithm and switches to a Newton-type algorithm when a certain stopping criterion is fulfilled (Redner & Walker 1984).
This leads to a new algorithm that yields the stability and large circle of convergence from the EM algorithm along with superlinear convergence of the Newton-type algorithm in the neighborhood of the maximum.

3.2 A simulation experiment

In this section, we consider three aspects of the estimation procedure for HMMs: (i) the effect of different parameterizations in an unconstrained DNM, (ii) the performance (relative to the EM algorithm and DNM) of the hybrid algorithm introduced in the previous section and finally (iii) the reliability of parametric bootstrap confidence intervals for the parameters.

3.2.1 Study design

The influence of these methods on the estimation results, particularly on the resulting value of the log-likelihood and the performance of the hybrid algorithm are studied mainly with simulated two-state Poisson HMMs. In Section 3.3 we also analyze the effects on fitting a three-state Poisson HMM to a time series of earthquake counts.

The simulated time series are of three lengths (50, 200 and 500). For each length, four different transition probability matrices, namely,

$$\boldsymbol{T}_1 = \begin{pmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{pmatrix}, \boldsymbol{T}_2 = \begin{pmatrix} 0.7 & 0.3 \\ 0.8 & 0.2 \end{pmatrix}, \boldsymbol{T}_3 = \begin{pmatrix} 0.2 & 0.8 \\ 0.8 & 0.2 \end{pmatrix}, \boldsymbol{T}_4 = \begin{pmatrix} 0.55 & 0.45 \\ 0.45 & 0.55 \end{pmatrix},$$

and two different state-dependent parameter vectors $\boldsymbol{\lambda}_1 = (1, 2)$, $\boldsymbol{\lambda}_2 = (2, 5)$ served as parameters for the generation of the observations. This $3 \times 4 \times 2$ experimental design yields 24 different two-state Poisson HMMs. The realizations of these series were generated using the default random number generator in the base-library of `R`, an implementation of the Mersenne-Twister (Matsumoto & Nishimura 1998).

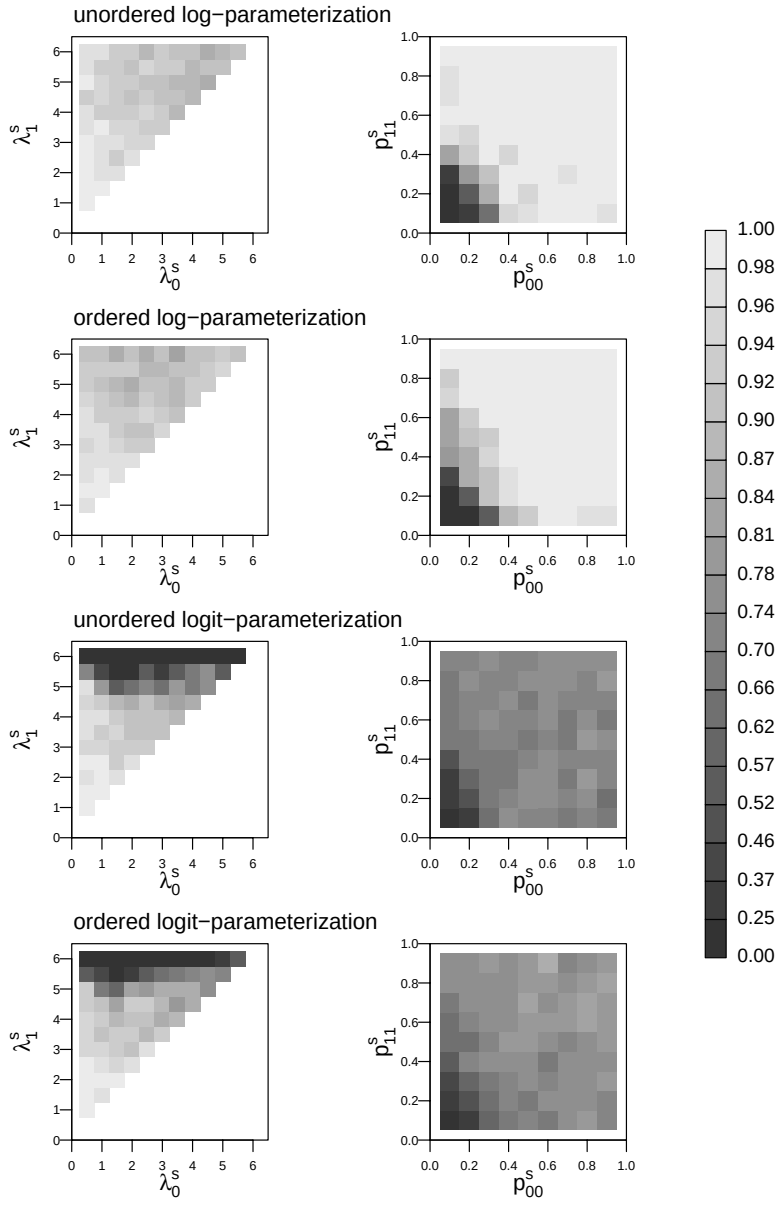
3.2.2 Results for different parameterizations

To test the effect of the different parameterizations from Section 3.1.1, we fit HMMs to each of the 24 generated time series where the initial values are combinations of $\lambda_0^s, \lambda_1^s \in \{0.5, 1, 1.5, \ldots, x_{\max}\}$, $\lambda_0^s < \lambda_1^s$, and $p_{00}^s, p_{11}^s \in \{0.1, 0.2, \ldots, 0.9\}$. Table 3.1 shows the percentage of failures, i.e. those cases in which the algorithm did not converge to a solution, and the percentage of successful convergence to the global maximum for the Newton-type and the Nelder-Mead algorithm, summed over all series.

Table 3.1: Performance of the Newton-type and Nelder-Mead algorithms with different parameterizations of the state-dependent parameters

	failures (%)		global maximum found (%)	
	Newton	Nelder-Mead	Newton	Nelder-Mead
unordered log	0.17	0.00	84.3	93.0
ordered log	0.22	0.00	81.4	93.0
unordered logit	1.04	0.00	78.1	83.9
ordered logit	0.64	0.00	77.2	85.5

Figure 3.1: Proportion of successful estimations for specific combinations of the parameter starting values using the Nelder-Mead algorithm and different parameterizations of the state-dependent parameters.



We find that, over all series, the simplest parameterization, i.e. the use of log-transformed state-dependent parameters, leads to the best results in terms

of the number of failures and the convergence to the global maximum, and we will therefore apply it to all further analysis. In general, this holds true for both the Newton-type and the Nelder-Mead algorithms. However, the Nelder-Mead algorithm provided better results than the Newton-type algorithm for all parameterizations, if not for each individual series.

As a typical example, Figure 3.1 demonstrates the results obtained using the Nelder-Mead algorithm and the four parameterizations for fitting a Poisson HMM to a time series with 200 observations that were simulated using the true parameters $\boldsymbol{\lambda}_1$ and $\boldsymbol{T}_1$ as defined above. The ML estimates obtained for this specific series are $\hat{\boldsymbol{\lambda}} = (0.55, 1.98)'$ and $\hat{\boldsymbol{T}}$ with diagonal $(0.89, 0.97)'$ leading to $\log L_{max} = -333.51$. Each graph in Figure 3.1 represents the proportion of successful estimations (i.e. estimations that led to the global maximum) for specific combinations of the initial values for the state-dependent parameters, λ_0^s, λ_1^s, or the diagonal elements of the initial TPM, p_{00}^s, p_{11}^s, given a specific parameterization of the state-dependent parameters. Light-colored areas indicate a high proportion of successful trials, while darker colors represent low proportions of success.
In this typical case the log-parameterization provides much more stable results than does the logit-parameterization. However, the performance of all four parameterizations improves as the initial values approach the values of the maximum likelihood estimates.
The general tendency that the unordered log-parameterization provides the most stable results, was found to hold true for all simulated series. Nevertheless, the stability of the estimation results depended on the properties of the true parameters used for the simulation. A detailed analysis of our experimental study provides the following results for both the Newton-type and the Nelder-Mead algorithms:

- The estimation results are clearly more stable for the true state-dependent parameter vector $\boldsymbol{\lambda}_2$, i.e. the case in which the true λ_i-values differ substantially.

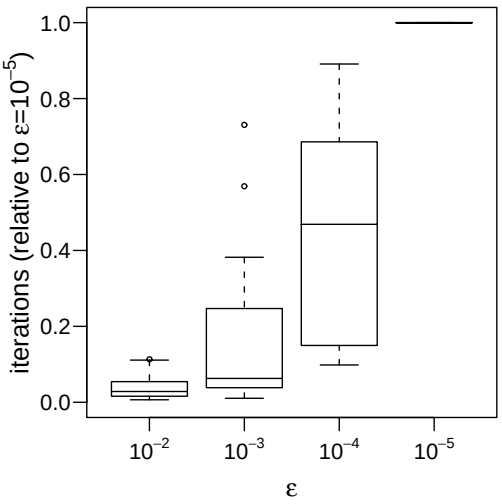
- The influence of the true TPM is not as straightforward as that of the true state-dependent parameters. However, we observed the tendency that one obtains the best results for $\boldsymbol{T}_2$, i.e. the case in which one state is dominant, and the worst results for $\boldsymbol{T}_3$, i.e. the case in which the series switches often between the states.

- The results for $\tau = 50$ are clearly less stable than those for $\tau = 200, 500$, while especially in those cases with true state-dependent parameter vector $\boldsymbol{\lambda}_1$ the results for $\tau = 500$ were worse than those for $\tau = 200$. This may be due to the fact that for longer series the algorithms are more likely to end up at local maxima.

3.2.3 Performance of the hybrid algorithm

The main parameter that influences the hybrid algorithm is the stopping criterion ϵ. The algorithm switches from the EM to the Newton-type algorithm as soon as the relative change in the log-likelihood of two subsequent steps falls below a predefined value ϵ.
Different choices of the stopping criterion ϵ have mainly two effects. Firstly, on the one hand, a small ϵ in general leads to a low failure rate of the algorithm (i.e., cases in which no maximum of the log-likelihood is found) and a high proportion of successful convergence to the global maximum. Secondly, on the other hand, small values of ϵ increase the computational time required by the algorithm. Figure 3.2 shows boxplots of the number of EM iterations of the hybrid algorithm for different values of ϵ, relative to the numbers of iterations obtained for the smallest value studied, i.e. $\epsilon = 10^{-5}$.

Figure 3.2: Effect of the stopping criterion ϵ in the hybrid algorithm on the number of EM iterations, relative to $\epsilon = 10^{-5}$



The relative number of EM iterations increases moderately when using $\epsilon = 10^{-3}$ instead of $\epsilon = 10^{-2}$, while it rises substantially when moving down to $\epsilon = 10^{-4}$ or $\epsilon = 10^{-5}$. Since the proportion of successful estimations improves only slightly for smaller values of ϵ, the choice of $\epsilon = 10^{-3}$ is a reasonable compromise to deal with the trade-off between speed and stability and is used in what follows.

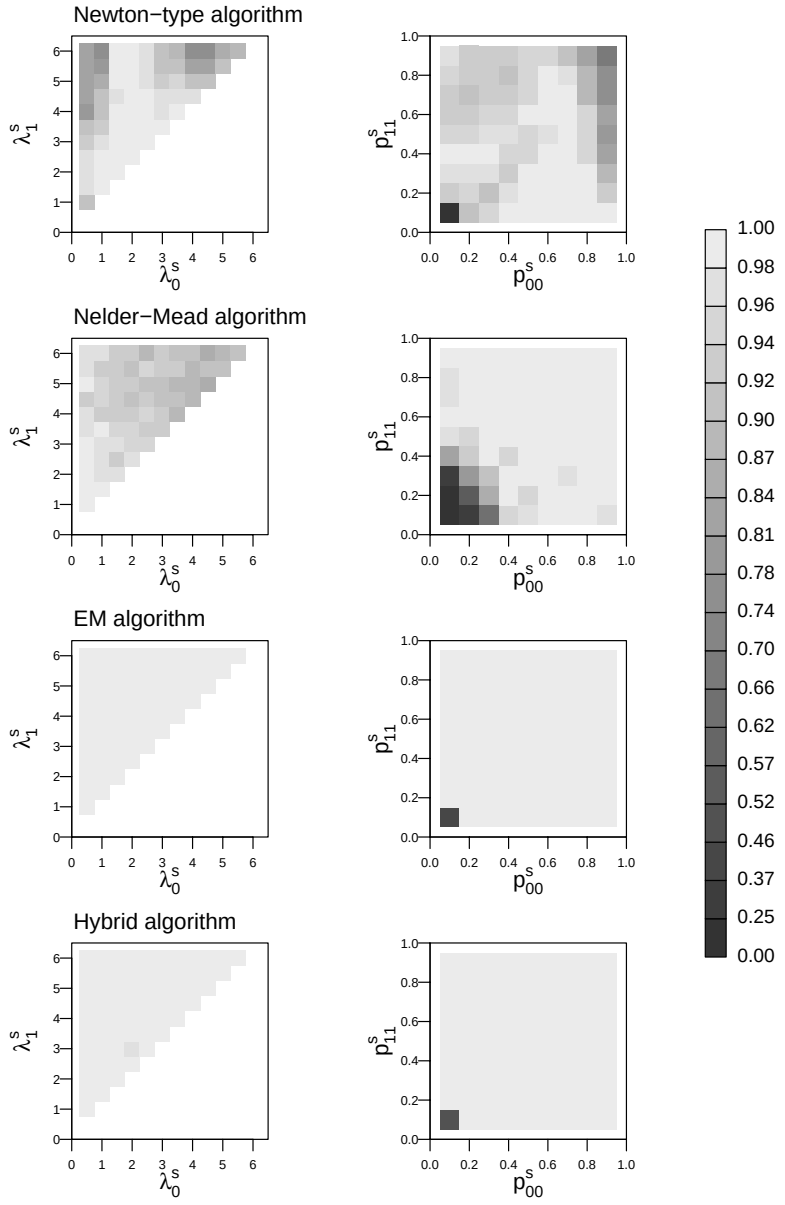
Fitting HMMs to the 24 time series with the same combinations of initial values mentioned above leads to the results displayed in Table 3.2, which clearly shows the high stability of the hybrid algorithm. The EM algorithm as well as the hybrid algorithm provide the most stable results (with their order changing from series to series). Not surprisingly, the Nelder-Mead algorithm is more stable than the Newton-type algorithm. However, there may be cases in which both the EM and the hybrid algorithm provide just a slight improvement over direct numerical maximization. It should also be mentioned that, as in the study of the parameterizations for direct numerical maximization, the results depend on the true parameters that were used to generate the observations. The results concerning the dependence of the stability on the true parameter settings described above also hold for the EM and the hybrid algorithm.

Table 3.2: Performance of the algorithms considered

	failures (%)	global maximum found (%)
Newton-type	0.17	84.3
Nelder-Mead	0.00	93.0
EM algorithm	0.00	95.6
Hybrid algorithm	0.00	95.4

The robustness of the EM and the hybrid algorithm is illustrated in Figure 3.3. The design of the figure corresponds to the design of Figure 3.1. It shows the proportion of successful estimations for specific combinations of the initial values for the state-dependent parameters, λ_0^s, λ_1^s, or the diagonal elements of the initial TPM, p_{00}^s, p_{11}^s. The algorithms for parameter estimation are applied for the same series as above.
While the Newton-type and Nelder-Mead algorithms provide stable estimation results for a few combinations of initial values, the EM and the hybrid algorithms reach the global maximum in almost all cases, except for some relatively extreme choices of the initial values for the TPM.

Figure 3.3: Proportion of successful estimations for specific combinations of the parameter starting values using different algorithms for parameter estimation



3.2.4 Coverage probability of confidence intervals

In this section we consider the estimation of confidence intervals for the parameters of HMMs and their properties. These may also be calculated based on the asymptotic properties of the estimators, but we will restrict our attention to those based on the parametric bootstrap as described, for example, in MacDonald & Zucchini (1997).

Bootstrap confidence intervals for the parameters of a HMM have already been studied by Visser et al. (2000), but we could find no detailed analysis of the true coverage probability of bootstrap confidence intervals in the context of HMMs.
We therefore apply a double parametric bootstrap method to analyze the coverage probability of bootstrap percentile confidence intervals (Efron & Tibshirani 1993, Chapter 13) using the hybrid algorithm introduced above, and at different levels of confidence (90%, 95% and 99%).
In a first step, for each of the 24 true parameter combinations listed in Section 3.2.1, we generated 1000 realizations of the respective Poisson HMM. In a second step, we computed the maximum likelihood estimates for these 1000 realizations and, in each case, simulated 200 new bootstrap realizations using the obtained parameter estimates. We then constructed confidence intervals applying the bootstrap percentile method to the new bootstrap samples, using different levels of confidence. Thus we obtained 1000 confidence intervals for each of the true parameters of the respective HMMs, and at each level of confidence. The coverage probabilities were then estimated as the proportion of those confidence intervals that cover the respective true parameter value.

The resulting coverage probabilities (and the respective confidence intervals) for the parameters λ_0 and p_{00}, for three different levels of confidence, depending on the length of the time series and the true TPM, are given in Figure 3.4 for a true state-dependent parameter vector $\boldsymbol{\lambda}$ with a relatively large difference between the state-dependent values λ_i, and in Figure 3.5 for a true $\boldsymbol{\lambda}$ vector with entries that lie rather close to each other.

Figure 3.4: Coverage probabilities of bootstrap confidence intervals with different levels of confidence for series simulated using $\boldsymbol{\lambda}_2$, i.e. a large difference between the state-dependent parameters

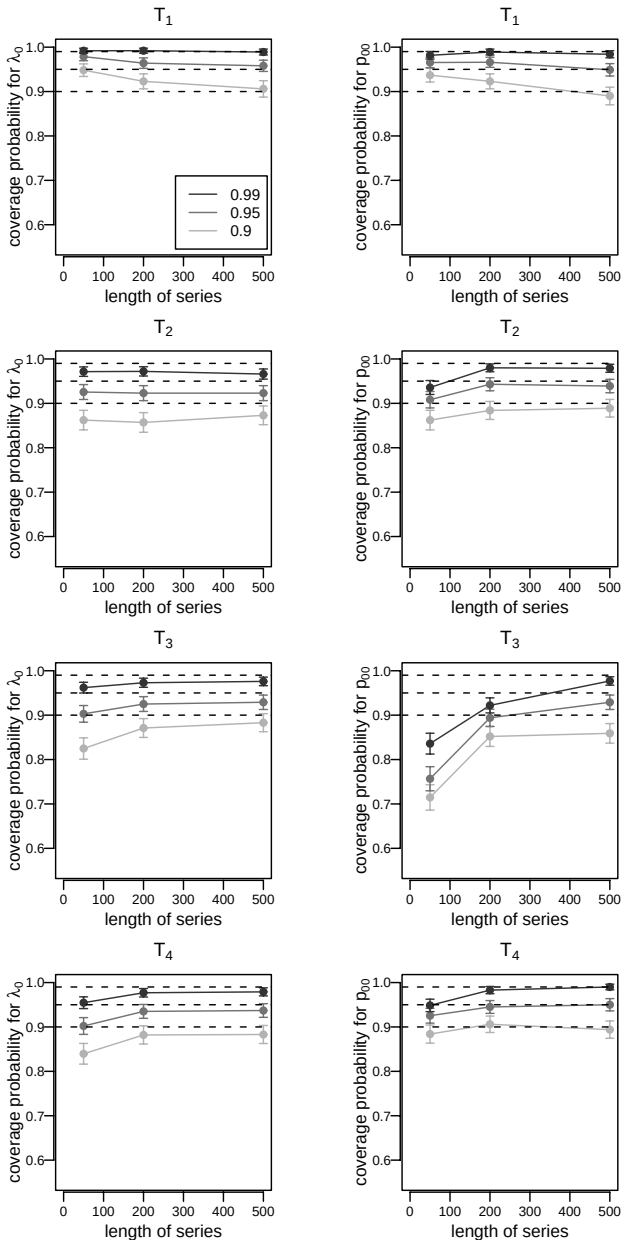
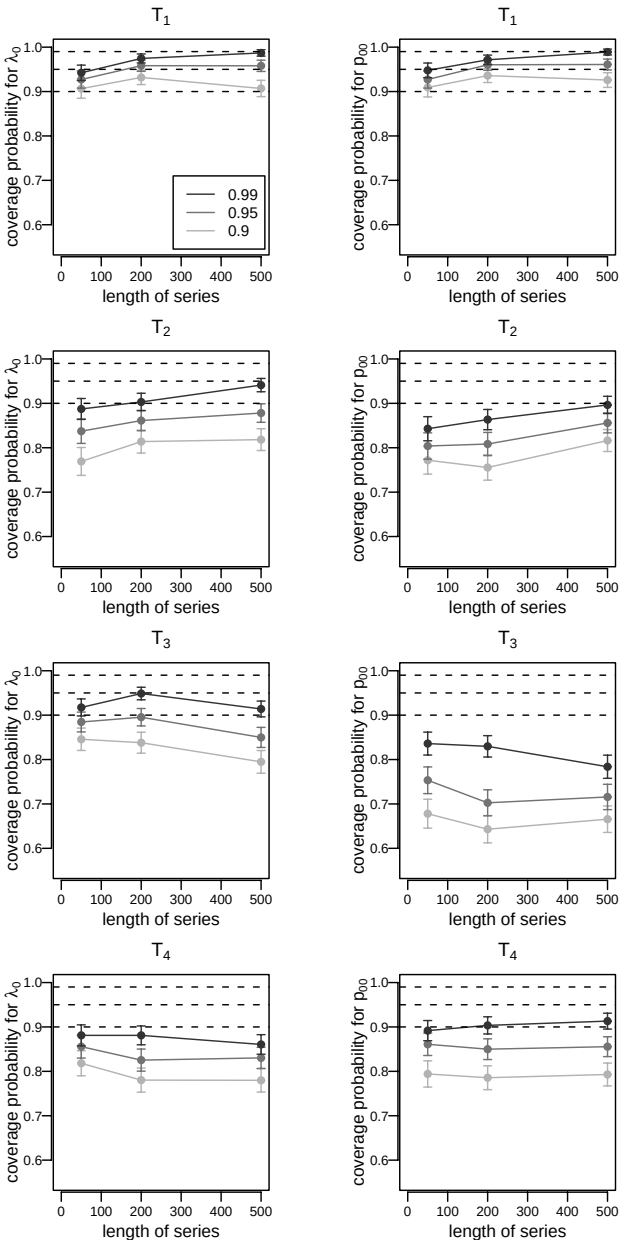


Figure 3.5: Coverage probabilities of bootstrap confidence intervals with different levels of confidence for series simulated using $\boldsymbol{\lambda}_1$, i.e. a small difference between the state-dependent parameters



It is observed that in most cases for $\tau = 50$ the true coverage probability deviates from the desired level of confidence. On the other hand, in the case of a relatively large difference between the true λ_i-values, the true coverage probabilities of the confidence intervals nearly correspond to the desired levels of confidence for series of length $\tau = 200$ or $\tau = 500$. The influence of the transition probability matrix on the coverage probability also becomes obvious. Especially in the case of the true TPM $\boldsymbol{T}_3$, which represents a HMM in which the states switch frequently, the coverage probability of the confidence interval for p_{00} deviates substantially from the nominal coverage probability even for $\tau = 200$.
In contrast, in the case where the λ_i-values differ less from each other, all coverage probabilities, except those for the transition probability matrix $\boldsymbol{T}_1$ (a case in which the states are highly persistent) are clearly smaller than the nominal level. This holds true even for relatively long time series of $\tau = 500$, indicating that the estimated confidence intervals are too narrow. Similar results hold for the other parameters, which are not shown in the figures. These results coincide with the findings of Nityasuddhi & Böhning (2003), who investigated normal mixtures and reported that the asymptotic properties of the EM algorithm are inaccurate when the means lie close to each other.

3.3 An application

In this section we report on the performance of the hybrid algorithm when this is applied to a set of real data. The time series consists of yearly counts of major world earthquakes, i.e. earthquakes of magnitude 7.0 or greater on the Richter scale, between the years 1900 to 2003[3].

This series has already been studied by Zucchini & MacDonald (1998), however with restriction to the period 1900-1997. Since these authors select the three-state HMM as the best model we restrict our attention here to the case of a three-state Poisson HMM.
We use an estimation grid similar to the one used in the previous section. The initial values $\lambda_0^s, \lambda_1^s, \lambda_2^s$ for the grid search are chosen from 10 equidistant points in $[x_{\min}, x_{\max}]$, where $\lambda_0^s < \lambda_1^s < \lambda_2^s$, and the starting values of the TPM are given by $p_{ii}^s \in \{0.2, 0.4, 0.6, 0.8\}$ for $i = 0, 1, 2$ and $p_{ij}^s = (1 - p_{ii}^s)/2$ for $i = 0, 1, 2, j \neq i$, yielding a total of 7680 grid points.

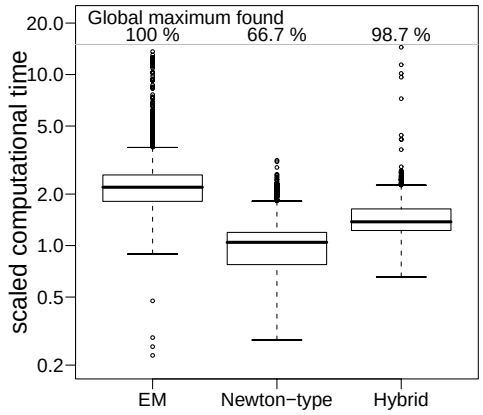
[3]The series can be downloaded from *http://wwwneic.cr.usgs.gov/neis/eqlists/7up.html*.

The ML estimates obtained for the earthquakes series are $\hat{\boldsymbol{\lambda}} = (11.4, 19.1, 29.4)$ and

$$\hat{\boldsymbol{T}} = \begin{pmatrix} 0.855 & 0.121 & 0.024 \\ 0.055 & 0.895 & 0.050 \\ 0.000 & 0.194 & 0.806 \end{pmatrix}$$

leading to $\log L_{max} = -324.79$. These results are close to the ones given by Zucchini & MacDonald (1998).
Figure 3.3 provides boxplots of the computational time needed when using a Newton-type algorithm, the EM algorithm and the hybrid algorithm. Since computational time depends on both the hardware and software used, the computational times in Figure 3.3 are normalized by the average computational time needed by the fastest method, the Newton-type algorithm as implemented in the R function `nlm()`. Thus the mean computational time is 1.00 for the Newton-type algorithm, 1.84 for the EM and 1.44 for the hybrid algorithm. All three algorithms considered never failed to provide a result, though the Newton-type algorithm succeeded only in 66.7% to attain the global maximum of the likelihood, while the EM and the hybrid algorithm led to the global maximum in 100% and 98.7% of all grid search trials, respectively. Hence, it is observed that the hybrid algorithm provides a reasonably efficient method to increase the stability when compared to the Newton-type algorithm and the speed when compared to the EM algorithm.

Figure 3.6: Scaled computational time and percentage of trials with convergence to the global maximum using different algorithms for parameter estimation of the quakes series



In this application, the reduction of computational time when using the hybrid algorithm instead of the EM algorithm seems to be relatively small. However, in the simulation experiment described above we found series of length $\tau = 200$ or $\tau = 500$ for which the mean computational time of hybrid algorithm could be reduced substantially compared to that of the EM algorithm.

3.4 Conclusion

We presented different methods of parameter estimation for HMMs, among others an EM algorithm for stationary time series. In the cases investigated here, it turned out that the simplest parameterization for the direct numerical maximization provides the best results. Comparison of the EM algorithm and direct numerical maximization clearly showed the trade-off between stability and performance. The hybrid algorithm would seem to provide an excellent compromise, because it is not only as stable as the EM-algorithm, but also clearly faster. If a choice has to be made between the EM algorithm and DNM, the latter is preferable if one can provide accurate initial values, or if the estimation is time-critical. Clearly, if the formulae required for the EM algorithm are too difficult to derive, or if one wishes to avoid deriving these, then one has to use DNM. In all other situations, the EM algorithm is the preferred method due to its greater stability.
We also found that the true coverage probability for bootstrap-based confidence intervals, obtained by parametric bootstrap, can be unreliable for models whose state-dependent parameters lie close to each other.

Our analysis can easily be extended to cover other component distributions. A smaller investigation of Normal HMMs revealed tendencies similar to those obtained for the Poisson HMMs, although the results were not as clear-cut. A complicating factor in the important special case of Normal HMMs is that each state-dependent distribution depends on two parameters and, furthermore, the likelihood function is – as for independent normal mixtures – in fact unbounded (Nityasuddhi & Böhning 2003).
Unfortunately, the extension of our analysis to Hidden Semi-Markov Models failed due to computational complexity. However, smaller tests revealed similar tendencies and indicated a high robustness of the EM algorithm against poor initial guesses for the semi-Markovian case as well.

Chapter 4

Markov Switching Approaches to Model Time-Varying Betas

Modeling daily return series with HMMs has been investigated by several authors. After the seminal work of Rydén et al. (1998) who showed that two- and three-state HMMs with normal components reproduce well the temporal and distributional properties of daily returns from the S&P 500 index, several other authors followed their ideas (see, e.g., Cecchetti et al. 1990, Linne 2002, Bialkowski 2003).

The focus of this chapter lies on the development of a joint model for return series. Consider a portfolio consisting of multiple assets, e.g., a portfolio of European shares selected from the Dow Jones (DJ) EURO STOXX 600. Fitting a multivariate HMM with normal component distributions would require the estimation of the variance-covariance matrix for each of the states. In the worst case, the portfolio of a professional investor is composed of all 600 shares and thus the procedure would involve a matrix of dimension 600×600 yielding 180300 parameters to be estimated per state. It is obvious that such a model would be grossly over-parameterized resulting in very unstable estimates. Moreover, as shown in Chapter 3, the common estimation algorithms for HMMs depend on the choice of the initial values. Hence, choosing even a small number of different initial values for each parameter may yield an infeasible amount of estimations to be carried out.
A possible solution to the quadratic increase of the number of parameters is based on the Capital Asset Pricing Model (CAPM). In the CAPM, the return of every single asset is linearly dependent to the market return (plus an error term):

$$R_{it} = \alpha_i + \beta_i R_{0t} + \epsilon_{it}, \quad \epsilon_{it} \sim N(0, \sigma_i^2),$$

where R_{it}, R_{0t} are the returns of the i^{th} asset and the market, respectively. The error term is represented by ϵ_{it} and β_i is the market or systematic risk. Beta represents one of the most widely used concepts in finance. It is used by financial economists and practitioners

- to estimate a stock's sensitivity to the overall market,
- to identify mispricings of a stock,
- to calculate the cost of capital, and
- to evaluate the performance of asset managers.

In the context of the CAPM, beta is assumed to be constant over time and is estimated using the method of ordinary least squares (OLS). However, inspired by theoretical arguments, various studies have revealed that the systematic risk of an asset depends on microeconomic and macroeconomic factors and rejected the assumption of beta stability (e.g., Fabozzi & Francis 1978, Sunder 1980, Bos & Newbold 1984, Collins et al. 1987).

In our study we investigate the utility of two Markov switching models for the coefficients in the CAPM. As a consequence, the number of parameters increases only linearly with the number of assets; however, a joint behavior of all assets is modeled by the common dependence on the market return.
We investigate the time-varying behavior of systematic risk for eighteen pan-European sectors. Using weekly data over the period of 1987-2005, four different modeling techniques in addition to the standard constant coefficient model are employed:

- a bivariate t-GARCH(1,1) model,
- two Kalman filter based approaches,
- a bivariate stochastic volatility model estimated via the efficient Monte Carlo likelihood technique, and
- two Markov switching models.

In this thesis we will focus mainly on the two Markov switching models. This approach uses a Markov switching framework which belongs to the large class of Markov switching models introduced by Hamilton (1989, 1990). Although Markov switching regression models have been applied in many different settings, the literature dealing with time-varying betas is relatively scarce. Fridman (1994) considered monthly data from the years 1980 to 1991 to analyze

the excess returns of three oil corporation securities by fitting a two-state regression model. This resulted in an improved assessment of systematic risk associated with each security. He also noted two effects: beta increases whenever the process is in the more volatile state, and the state associated with higher volatility tends to be less persistent than the state associated with lower volatility. Huang (2000) also considered a Markov switching model with one high-risk and one low-risk state. Using monthly return data from April 1986 to December 1993, he performed several tests to check the consistency of different states with the CAPM and rejected the hypothesis that the data were from the same state.
The results presented in this chapter are aggregated from our joint work with S. Mergner. For a detailed analysis, we refer to the paper of Mergner & Bulla (2005), which is available on request.

This chapter is organized as follows. In Section 4.1 we present OLS, the most common method to estimate unconditional betas. Section 4.2 presents two Markov switching approaches to model time-varying betas. A description of the data series to be analyzed is given in Section 4.3 and Section 4.4 contains our empirical results. A short conclusion is drawn in Section 4.5 and Section 4.6 summarizes the estimation results.

4.1 The Unconditional Beta in the CAPM

As a starting point, market risk is treated as a constant. The benchmark for time-varying betas is the excess-return market model with constant coefficients where an asset's unconditional beta can be estimated via OLS:

$$R_{it} = \alpha_i + \beta_i R_{0t} + \epsilon_{it}, \quad \epsilon_{it} \sim N(0,\, \sigma_i^2),$$

with

$$\hat{\beta}_i = \frac{Cov(R_0, R_i)}{Var(R_0)},$$

where R_{0t} denotes the excess return of the market portfolio and R_{it} denotes the excess return to sector i for $i = 1, \ldots, I$, each for period $t = 0, \ldots, \tau - 1$. The error terms ϵ_{it} are assumed to be i.i.d. Normal with mean zero and constant variance σ_i^2. Following the version of the CAPM of Sharpe (1964) and Lintner (1965), investors can borrow and lend at a risk-free rate; all returns are in excess over a risk-free interest rate and α_i is expected to be zero. Table 4.2, at the end of this chapter, summarizes the OLS estimates of the excess market model. As expected the intercept is not different from zero at the 5% level of

significance for any sector. For further details on the CAPM, see Chapter 5 of Campbell et al. (1997).

4.2 The Markov Switching Approach

The Markov switching approach also belongs to the class of state space models. The implicit assumption of models switching between different regimes is that the data results from a process that undergoes abrupt changes which are induced, e.g., by political or environmental events.

In the Markov switching framework, the systematic risk of an asset is determined by the different regimes of beta, driven by an unobserved Markov chain. The switching behavior of beta is governed by the TPM. Under the assumption of a model with two states, the TPM is of the form

$$\boldsymbol{T} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix},$$

where the entries of each row describe the interaction of the two regimes from which beta is drawn: p_{00} is the probability of staying in state zero from period t to period $t+1$ and p_{01} is the probability of switching to state one. The second row of the $\boldsymbol{T}$ can be interpreted analogously.

In this study, two Markov switching models are employed. The first one is a simple **Markov Switching** (MS) model, i.e., a Markov switching regression. Let $\{s_0, \ldots, s_{\tau-1}\}$ denote the state sequence representing the different regimes; driven by the TPM of a stationary Markov chain, the states take values in $\{0, \ldots, J-1\}$. Following Huang (2000) the regime-switching CAPM is specified by

$$R_{it} = \alpha_{is_t} + \beta_{is_t} R_{0t} + \eta_{it}, \ \eta_{it} \sim N(0, \sigma^2_{is_t}), \tag{4.1}$$

which implies that the regression coefficients $(\alpha_{is_t}, \beta_{is_t})$ are selected according to the value of state s_t. Note that the model is designed to accommodate both the correlations across return series and the serial correlation of the individual series.

The second approach entails additional assumptions on the market returns to synchronize the switching times of beta with different market conditions and will be denoted as **Markov Switching Market** (MSM) model. Rydén et al. (1998) showed that the temporal and distributional properties of daily return series can be modeled by a HMM with normal or double-exponential state-dependent distributions. Following their approach, the dynamics of the assets'

returns follow the same regime-switching regression of Equation (4.1) with the distribution of the market returns being given by:

$$R_{0t} = \mu_{s_t} + \epsilon_{s_t}, \quad \epsilon_{s_t} \sim N(0, \sigma^2_{0s_t}).$$

This means that in the MSM model the regime of the market changes along with the regime of the regression setup because they depend on the same state sequence. This synchronous behavior offers an advantage of allowing for direct conclusions from the market conditions on the asset's risk represented by beta.

The estimation procedures for our Markov switching models are based on the maximum likelihood method for HMMs.[4] The likelihood of both models is available in an explicit form and hence the parameters of the models can be estimated directly by numerical maximization of the log-likelihood function. The EM algorithm was not our first choice for an estimation procedure. As we have knowledge on the range of the values that beta usually takes, and the fact that the states of financial models tend to be very persistent, the selection of reasonable initial values was a feasible task. This allowed us to take advantage of the greater speed of the direct numerical maximization procedures (cf. Chapter 3).
The estimates for the model parameters include, inter alia, the state-dependent betas for each asset i and state j denoted by $\hat{\beta}^{MS}_{ij}$ or $\hat{\beta}^{MSM}_{ij}$.

As mentioned above, the state sequence cannot be observed. Therefore information about the state-distribution at time t has to be derived in order to obtain the in-sample estimates as well as the out-of-sample forecasts of conditional betas. The desired probabilities of a sojourn in state j at time t can be computed by the so-called smoothing, filtering and state prediction algorithms (see e.g. Ephraim & Merhav 2002). Given the state-distribution at time t, estimates for the time-varying betas can be calculated by weighting the state-dependent $\hat{\beta}^{MS/MSM}_{ij}$ with the probability of a sojourn in the corresponding state:

$$\hat{\beta}^{MS/MSM}_{it} = \sum_{j=0}^{J-1} \Big[\beta_{ij} \cdot P(s_t = j | R_{01}, \ldots, R_{0T}, R_{i1}, \ldots, R_{iT}) \Big],$$

[4] All estimations procedures were carried out using the statistical software package R 2.1.1 (R Development Core Team 2005) which can be downloaded from *www.r-project.org*. The code for the estimation, decoding and forecasting algorithms are provided upon request.

with

$$P(S_t = j|R_{01}, \ldots, R_{0T}, R_{11}, \ldots, R_{1T}) = \begin{cases} \alpha_t(j)\beta_t(j)/L & \text{for } 0 \leq t \leq \tau - 1 \\ \alpha_t(j)(\boldsymbol{T}^{t-(\tau-1)})_{\bullet j}/L & \text{for } \tau - 1 < t \end{cases},$$

where $\alpha_t(j)$, $\beta_t(j)$ are the forward/backward probabilities from the forward-backward algorithm (Rabiner 1989) and $(\boldsymbol{T}^{t-(\tau-1)})_{\bullet j}$ denotes the jth column of the matrix $\boldsymbol{T}^{t-(\tau-1)}$.

4.3 Data and Preliminary Analysis

4.3.1 Data Series

The data used in this study are the weekly excess returns calculated from the total return indices for eighteen pan-European industry portfolios, covering the period from 2 December 1987 to 2 February 2005. All sector indices are from STO (2004), a joint venture of Deutsche Boerse AG, Dow Jones & Company and the SWX Group that develops a global free-float weighted index family, the DJ STOXX indices.
The DJ STOXX 600 return index, which includes the 600 largest stocks in Europe, serves as a proxy for the overall market. All indices are expressed in Euros as the common currency. Weekly excess returns between period $t-1$ and t for index i are computed continuously as

$$R_{it} = \ln(P_{it}) - \ln(P_{i,t-1}) - r_t^f,$$

where P_{it} is Wednesday's index closing price in week t, ln is the natural logarithm and r_t^f is the risk-free rate of return, calculated from the 3-month Frankfurt Interbank Offered Rate (FIBOR).[5] The FIBOR yields, fib_t, are represented as percentage per annum. They were converted to weekly rates r_t^f by the transformation

$$r_t^f = (1 + fib_t/100)^{1/52} - 1.$$

[5]All data were obtained from Thomson Financial Datastream.

4.3.2 Univariate Statistics

Descriptive statistics for the data are provided in Table 4.1. Over the entire sample, the Healthcare sector offered the highest mean excess return per week (0.17%), while the lowest was seen in Automobiles & Parts (0.02%). The risk, as measured by the standard deviation, ranges from 0.0203 for the defensive Utilities to 0.0422 for the high risk sector Technology. The market and all its segments are leptokurtic. Except for Healthcare and Travel & Leisure, all sectors and the market are negatively skewed. The Jarque-Bera statistic JB was used to test for normality. In the selected sample the null hypothesis of normality can be rejected at the 1% significance level for every sector, as well as for the overall market.

Table 4.1: Descriptive statistics of weekly excess returns

Weekly excess returns data of the eighteen DJ STOXX sector indices and the DJ STOXX Broad as European market portfolio, covering the period from 2 December 1987 to 2 February 2005.

Sector	N	Mean	Std. Dev.	Skew.	Kurt.	JB
Broad	897	0.0010	0.0231	−0.30	6.83	560.81
Automobiles	897	0.0002	0.0330	−0.56	6.30	452.55
Banks	897	0.0014	0.0270	−0.28	7.49	765.94
Basics	897	0.0012	0.0284	−0.24	5.13	177.41
Chemicals	897	0.0009	0.0257	−0.19	7.87	890.35
Construction	897	0.0008	0.0245	−0.32	4.97	159.58
Financials	897	0.0007	0.0259	−0.63	8.73	1286.90
Food	897	0.0010	0.0212	−0.27	5.86	317.60
Healthcare	897	0.0017	0.0253	0.18	5.52	242.96
Industrials	897	0.0007	0.0248	−0.47	5.69	303.08
Insurance	897	0.0004	0.0334	−0.85	13.97	4606.70
Media	897	0.0007	0.0342	−0.62	9.89	1832.40
Oil & Gas	897	0.0015	0.0267	−0.02	5.56	245.73
Personal	683	0.0009	0.0257	−0.22	4.95	113.83
Retail	683	0.0006	0.0298	−0.78	10.32	1594.50
Technology	897	0.0007	0.0422	−0.55	6.68	553.00
Telecom	897	0.0013	0.0344	−0.18	5.36	212.89
Travel	683	0.0007	0.0234	0.10	6.36	321.69
Utilities	897	0.0015	0.0203	−0.45	5.15	203.02

Remark: In September 2004, STOXX Ltd. switched its sector definitions from the DJ Global Classification Standard to the Industry Classification Benchmark and replaced the sectors Cyclical Goods & Services, Non-Cyclical Goods & Services and Retail (old) by the new sectors Travel & Leisure, Personal & Household Goods and Retail (new), respectively. As the history for the newly formed sectors is available only since 31 December 1991, for these three sectors only 683 weekly observations are available instead of 897.

4.4 Empirical Results

In the following we present the estimation results for the OLS model, the two KF, and the two Markov switching models. The KF models represent the best models while the OLS model is the most common model. To consider all the models described by Mergner & Bulla (2005) would go beyond the scope of this chapter. However, for the sake of completeness different approaches, namely a GARCH model, a stochastic volatility (SV) model, and two Kalman filter (KF) models (one mean reverting (MR) and another random walk (RW) model) were analyzed in the complete study.

4.4.1 Unconditional Beta Estimates

The estimated parameters of the OLS model are reported in Table 4.2 at the end of this chapter. According to the efficient market hypothesis and the implications of the Sharpe-Lintner version of the CAPM, all alphas should be zero. It can be seen from the first column that none of the estimated alphas is different from zero at an acceptable level of significance. In comparison, the estimated betas are all significant at the 1% level of significance. The lowest beta was estimated for Food & Beverages (0.65); the beta for Technology (1.49) was the highest, confirming the sector's high-risk profile. From the reported coefficients of determination (R^2), it can be seen that depending on the respective sector, between 43% (Oil & Gas) and 83% (Industrial Goods & Services) of the total return variation can be explained by movements of the overall market.

The last two columns of Table 4.2 provide the results of the classical Lagrange multiplier (LM) ARCH test for heteroskedasticity, as proposed by Engle (1982). ARCH(p) is the LM statistic of Engle's ARCH test for lag order p. With the exception of Retail, the null hypothesis of homoskedastic disturbances can be rejected at the 5% level of significance for all sectors and for

both lag orders tested.

4.4.2 Modeling Conditional Betas

The fit of the regime switching MS and MSM models to the data was tested with a different number of regimes. According to the AIC, two states turned out to be sufficient and therefore the results that are summarized in the Tables 4.3 and 4.4 concern the two-state models. As expected, all alphas are very close to zero and, for almost all sectors, the high- and the low-risk states could be well identified. While in the case of the MS model, the two state-dependent betas lie close together for the sectors Industrials and Retail, for the MSM model this occurs only for the Industrials sector. Generally it can be observed that the MSM model is characterized by a less clear separation of the two regimes; the state-dependent betas lie closer to each other than the betas of the corresponding MS model. This phenomenon can be explained by the lack of flexibility of the former model due to the enforced synchronous switching with the market regimes.

It should be also mentioned that the estimates for the expected market returns μ_0 and μ_1 of the MSM model are very close to zero, which supports Rydén et al. (1998) who proposed means equal to zero for daily return series. The estimates for p_{00} and p_{11}, mostly taking values between 95% and 99%, show a high persistence for both the high- and the low-risk states. Our results do not confirm the observations made by Fridman (1994) who reported lower persistence of the high-risk state.

4.4.3 Comparison of Conditional Beta Estimates

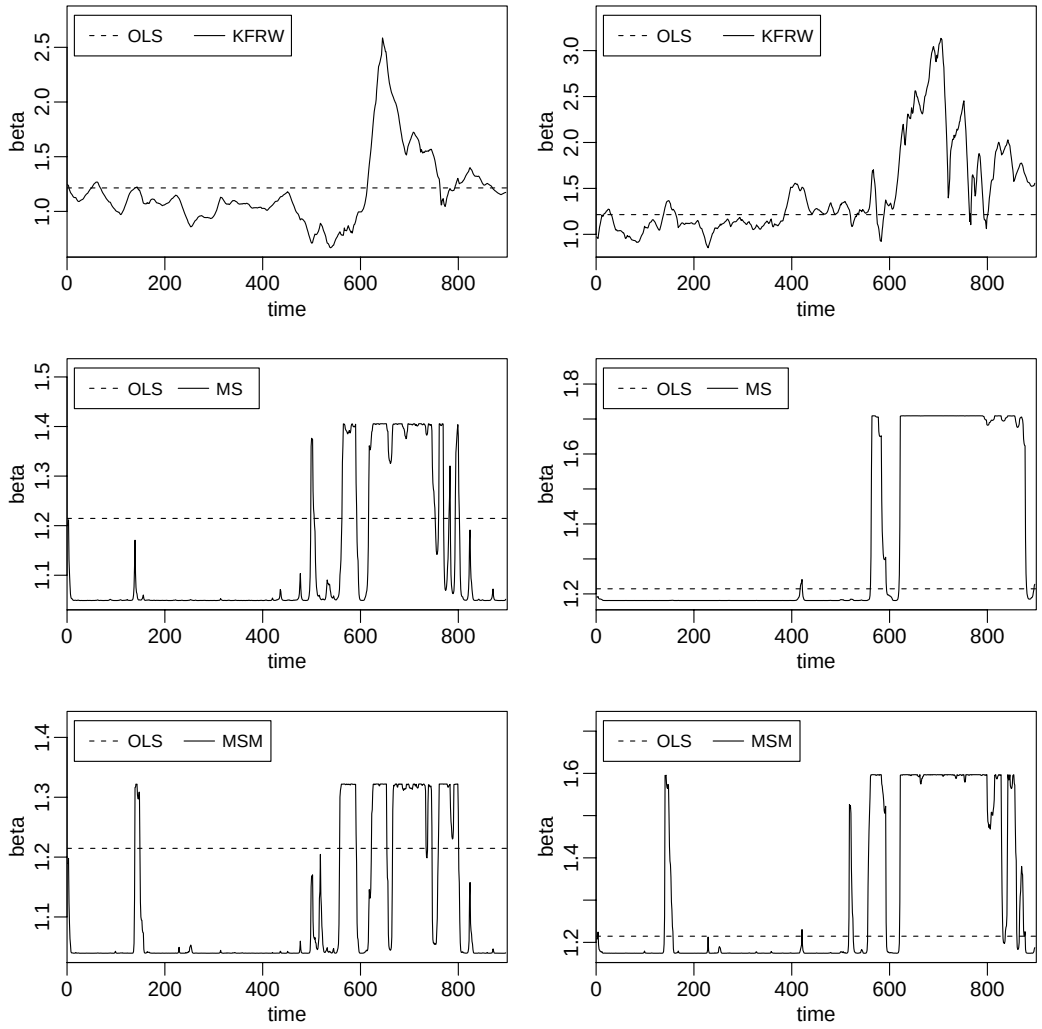
The conditional beta series of the MS, the MSM model, and the two Kalman filter models are summarized by their respective means and ranges in Tables 4.5 and 4.6 at the end of the chapter. While the widest range of beta across the sectors is observed for the Kalman MR model, the minimum and maximum of the conditional betas estimated by the two regime switching approaches do not deviate far from their respective means.

Figure 4.1 illustrates general similarities and differences between the alternative conditional beta series for the Media (left hand side) and the Technology sector (right hand side). The KF-based techniques display the greatest variation. The evolution of the betas during the TMT bubble and its aftermath is described in a convenient way by the KF models, while the Markov switching

framework displays a lack of flexibility. The latter models are not able to reflect the developments and dramatic shifts in terms of market risk in the course of the TMT bubble. In particular, the MSM model switches back and forth between the different states without giving a clear direction of the sectors' sensitivity to the overall market.

Figure 4.1: Various conditional betas for the Media and the Technology sector

The conditional betas for the Media sector are displayed in the three figures on the left hand side. The figures on the right hand side correspond to the Technology sector.



4.4.4 In-Sample and Out-Of-Sample Forecasting Accuracy

To determine the performance of the Markov switching models in generating a relatively best measure of time-varying systematic risk, the different techniques are formally ranked based on their in-sample performance. The first two criteria used to evaluate and compare the respective in-sample forecasts are the mean absolute error (MAE) and the mean squared error (MSE):

$$\begin{aligned} MAE_i &= \frac{1}{\tau}\sum_{t=0}^{\tau-1}\frac{|\hat{R}_{it}-R_{it}|}{\tau}, \\ MSE_i &= \frac{1}{\tau}\sum_{t=0}^{\tau-1}\frac{(\hat{R}_{it}-R_{it})^2}{\tau}, \end{aligned}$$

where τ is the number of forecast observations and $\hat{R}_{it}=\hat{\beta}_{it}R_{0t}$ denotes the series of return forecasts for sector i, calculated as the product of the conditional beta series estimated over the entire sample and the series of market returns which is assumed to be known in advance. The forecast quality is inversely related to the size of these two error measures.
While the mean error criteria can be used to evaluate the average forecast performance over a specified period of time for each model, and each sector individually, they do not allow for an analysis of forecast performances across sectors. From a practical perspective, it is interesting to observe how closely the rank order of forecasted sector returns corresponds to the order of realized sector returns at any time. Spearman's rank correlation coefficient (ρ_t^S), a non-parametric measure of correlation that can be used for ordinal variables in a cross-sectional context, is applied as the third evaluation criteria. After ranking the forecasted and observed sector returns separately for each point of time, where the sector with the highest return ranks first, ρ_t^S can be computed as

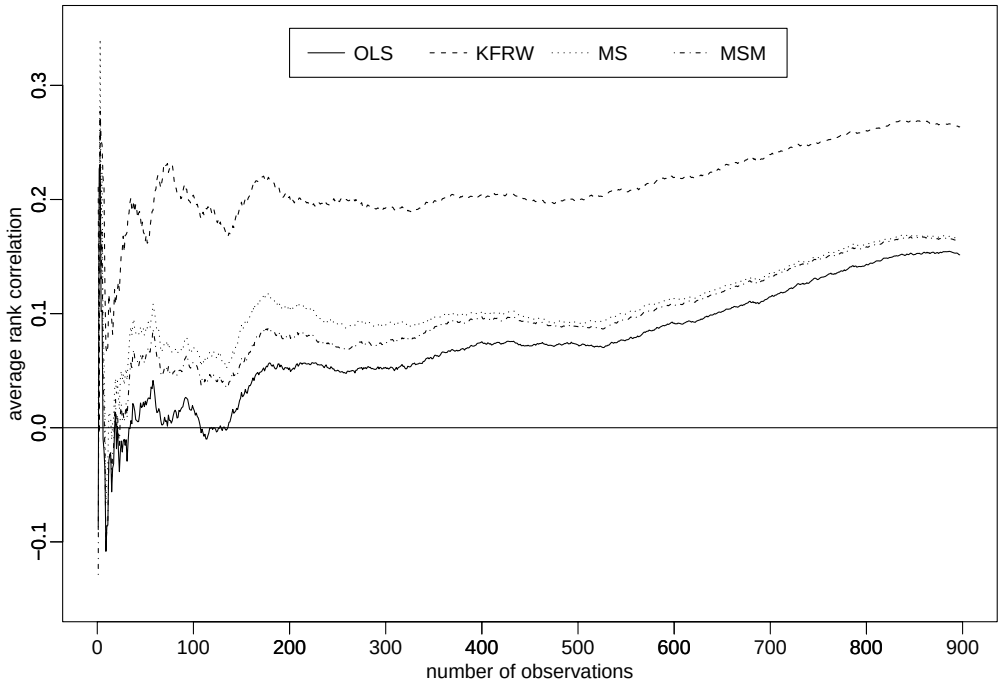
$$\rho_t^S = 1-\frac{6\sum_{i=1}^{I_t}D_{it}^2}{I_t(I_t^2-1)},$$

with D_{it} being the difference between the corresponding ranks for each sector and I_t being the number of pairs of sector ranks, each at time t.

The first step is the analysis of the in-sample forecasts using the first two criteria MAE and MSE. Compared to the Markov switching approaches, the degree of inferiority of OLS is remarkably low while the two KF techniques

clearly outperform their competitors. Within the Markov switching framework, the MS betas led to lower average errors than the MSM technique. Considering the third criterion, while the highest in-sample rank correlations are observed for the MR ($\rho^S = 0.46$) and the RW model (0.26), the MSM model (0.16) and the the MS (0.17) do only slightly better than OLS (0.15). Figure 4.2 illustrates how the average in-sample rank correlations develop over time for the various modeling techniques.

Figure 4.2: In-sample rank correlation coefficients



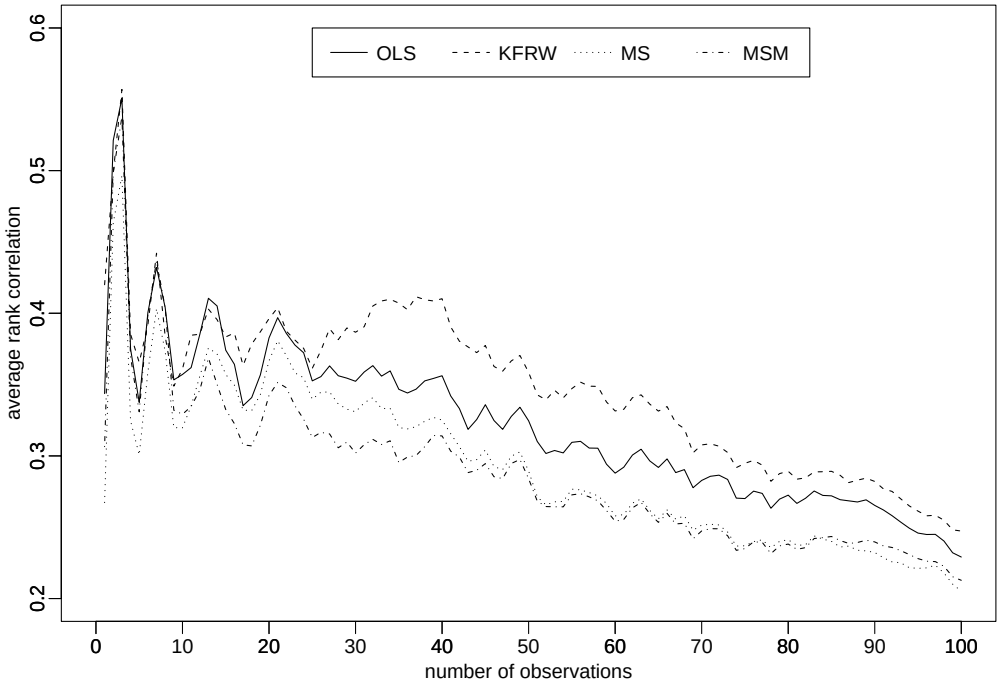
To sum up, the in-sample comparison suggests that the two Markov switching models are outperformed by their competitors.

While the in-sample analysis is useful to assess the various techniques' ability to fit the data, their out-of-sample forecast performance is more relevant from a practical point of view. For that purpose, 100 beta and return forecasts based on 100 samples of 520 weekly observations are estimated for each technique. Within this rolling window forecast procedure, the sample is rolled forward by one week while the sample size is kept constant at 520. The first sample, starting on 24 March 1993 and ending on 5 March 2003, is used to calculate

the out-of-sample conditional beta forecasts on 12 March 2003 based on the chosen modeling technique. The 100th beta forecast is then generated based on the last sample starting on 15 February 1995 and ending on 26 January 2005.

Without going into all the details, it can be observed that the KF approaches again offer the best forecast performance, while the two Markov switching approaches yield the worst results. These findings are broadly confirmed in a cross-sectional setting as shown in Figure 4.3, where the Markov switching techniques (each 0.21) produce the worst forecasts.

Figure 4.3: Out-of-sample rank correlation coefficients



4.5 Conclusion

The results of this study indicate that the out-of-sample forecast performances of the two proposed Markov switching models is inferior to that of any time-varying alternative and also to OLS. One reasonable explanation is that the Markov switching models were limited to two-state models, a very common

assumption in the switching CAPM framework. The undesirable consequence of this limitation, which leads to reduced flexibility compared to other approaches, results in an insufficient fit to the evolution of beta during the TMT bubble and the subsequent crash at the stock markets.

One possible solution is to significantly increase the number of states. However, as the number of parameters of the TPM increases quadratically with the number of states, this approach complicates the estimation procedure significantly. One possibility to reduce the number of variables could be a structuring of the TPM. For many models, the main contribution to the likelihood is obtained by the elements on the TPM's main diagonal. The off-diagonal elements could, for instance, be expressed in polynomial dependence on the main-diagonal element of the respective row. However, this approach is subject to future research.

4.6 Estimation Results

Table 4.2: OLS estimates of excess market model

Figures in parentheses denote p-values.

Sector	α	β	R^2	$ARCH(1)$	$ARCH(6)$
Automobiles	−0.001	1.148	0.64	26.14	53.38
	(0.150)	(0.000)		(0.000)	(0.000)
Banks	0.000	1.062	0.82	14.95	44.81
	(0.409)	(0.000)		(0.000)	(0.000)
Basics	0.000	0.902	0.54	61.08	171.64
	(0.610)	(0.000)		(0.000)	(0.000)
Chemicals	0.000	0.907	0.66	39.15	86.82
	(0.989)	(0.000)		(0.000)	(0.000)
Construction	0.000	0.886	0.69	25.91	47.81
	(0.776)	(0.000)		(0.000)	(0.000)
Financials	0.000	0.997	0.79	8.45	79.92
	(0.470)	(0.000)		(0.004)	(0.000)
Food	0.000	0.648	0.50	17.74	184.76
	(0.443)	(0.000)		(0.000)	(0.000)
Healthcare	0.001	0.777	0.50	5.00	58.89
	(0.121)	(0.000)		(0.025)	(0.000)
Industrials	0.000	0.977	0.83	12.44	58.00
	(0.391)	(0.000)		(0.000)	(0.000)
Insurance	−0.001	1.268	0.77	16.03	74.91
	(0.106)	(0.000)		(0.000)	(0.000)
Media	−0.001	1.215	0.67	21.26	74.82
	(0.423)	(0.000)		(0.000)	(0.000)
Oil & Gas	0.001	0.758	0.43	28.53	114.59
	(0.268)	(0.000)		(0.000)	(0.000)
Personal	0.000	0.907	0.74	84.39	95.63
	(0.924)	(0.000)		(0.000)	(0.000)
Retail	0.000	0.949	0.61	1.61	7.26
	(0.519)	(0.000)		(0.204)	(0.297)
Technology	−0.001	1.489	0.66	15.38	98.73
	(0.337)	(0.000)		(0.000)	(0.000)
Telecom	0.000	1.194	0.64	31.76	65.18
	(0.910)	(0.000)		(0.000)	(0.000)
Travel	0.000	0.770	0.65	7.15	38.80
	(0.863)	(0.000)		(0.008)	(0.000)
Utilities	0.001	0.694	0.62	11.44	36.20
	(0.068)	(0.000)		(0.001)	(0.000)

Table 4.3: Parameter estimates for MS models

Estimated parameters for the MS/MSM model for the first nine of eighteen DJ STOXX sectors.

Sector	Model	$\alpha_0, \alpha_1 \times 10^4$	β_0, β_1	$\sigma_{i0}^2, \sigma_{i1}^2 \times 10^2$	p_{00}	p_{11}	$\mu_0, \mu_1 \times 10^3$	$\sigma_{00}^2, \sigma_{01}^2 \times 10^2$
Automobiles	MS	-13.4; -4.87	1.26; 1.03	1.51; 2.89	0.993	0.980	–	–
	MSM	-3.05; -35.5	1.22; 1.07	1.47; 2.92	0.976	0.926	2.40; -2.97	1.56; 3.70
Banks	MS	1.57; 8.70	1.01; 1.11	0.70; 1.77	0.988	0.970	–	–
	MSM	2.02; 1.07	0.97; 1.10	0.67; 1.75	0.962	0.904	2.30; -2.02	1.58; 3.36
Basics	MS	-3.26; 8.62	1.05; 0.84	1.04; 2.64	0.998	0.998	–	–
	MSM	-3.90; 9.56	1.02; 0.82	1.08; 2.95	0.972	0.917	-2.58; -2.28	1.61; 3.31
Chemicals	MS	1.70; -4.37	1.00; 0.87	0.93; 2.05	0.998	0.996	–	–
	MSM	1.04; -12.6	0.98; 0.85	1.05; 2.38	0.978	0.901	2.65; -4.16	1.63; 3.70
Construction	MS	-7.04; 0.22	1.14; 0.77	0.92; 1.75	0.997	0.993	–	–
	MSM	-4.38; -4.11	1.07; 0.79	0.94; 1.89	0.982	0.959	2.55; -3.26	1.55; 3.42
Financials	MS	-0.61; -3.31	0.91; 1.08	0.84; 1.82	0.997	0.990	–	–
	MSM	-0.05; 3.33	0.91; 1.05	0.81; 1.89	0.969	0.883	2.52; -3.25	1.60; 3.62
Food	MS	-0.95; 1.96	0.91; 0.49	0.76; 2.23	0.993	0.984	–	–
	MSM	-1.48; -1.47	0.89; 0.52	0.78; 2.34	0.985	0.928	2.59; -2.51	1.60; 3.33
Healthcare	MS	14.5; -0.88	0.91; 0.68	1.13; 2.39	0.981	0.971	–	–
	MSM	10.0; 2.72	0.90; 0.73	1.16; 2.40	0.969	0.952	2.62; -1.31	1.51; 3.11
Industrials	MS	-0.59; -7.62	1.02; 0.96	0.62; 1.45	0.998	0.996	–	–
	MSM	1.52; -1.41	1.00; 0.97	0.64; 1.57	0.975	0.951	2.24; -1.45	1.59; 3.38

Table 4.4: Parameter estimates for MS models

Estimated parameters for the MS/MSM model for the last nine of eighteen DJ STOXX sectors.

Sector	Model	$\alpha_0, \alpha_1 \times 10^4$	β_0, β_1	$\sigma_{i0}^2, \sigma_{i1}^2 \times 10^2$	p_{00}	p_{11}	$\mu_0, \mu_1 \times 10^3$	$\sigma_{00}^2, \sigma_{01}^2 \times 10^2$
Insurance	MS	-5.81; -8.25	1.12; 1.37	1.01; 2.48	0.996	0.989	–	–
	MSM	-6.14; 1.58	1.11; 1.39	1.05; 2.64	0.971	0.901	2.44; -2.93	1.64; 3.72
Media	MS	-0.81; -12.4	1.05; 1.41	1.20; 3.40	0.989	0.960	–	–
	MSM	-2.57; -1.51	1.04; 1.32	1.21; 3.34	0.991	0.935	2.37; -4.21	1.65; 3.60
Oil & Gas	MS	6.02; 3.57	0.96; 0.58	1.37; 2.86	0.992	0.983	–	–
	MSM	6.12; 1.92	0.91; 0.71	1.38; 2.94	0.987	0.926	2.20; -1.68	1.62; 3.45
Personal	MS	-2.33; 6.06	0.95; 0.80	1.03; 1.87	0.995	0.980	–	–
	MSM	-1.76; -5.45	0.99; 0.85	1.72; 1.84	0.986	0.962	2.79; -4.19	1.72; 3.88
Retail	MS	2.29; -11.9	0.90; 0.97	1.12; 2.48	0.909	0.891	–	–
	MSM	2.50; -19.6	0.87; 0.99	1.45; 2.83	0.967	0.872	1.68; -5.45	1.68; 4.04
Technology	MS	2.45; -1.61	1.18; 1.71	1.25; 3.85	0.996	0.990	–	–
	MSM	3.67; -7.46	1.17; 1.60	1.25; 3.80	0.993	0.947	2.48; -2.35	1.62; 3.38
Telecom	MS	6.73; -8.51	1.09; 1.31	1.49; 3.24	0.997	0.989	–	–
	MSM	3.29; -3.24	1.12; 1.21	1.49; 3.22	0.992	0.974	2.76; -4.10	1.69; 3.60
Travel	MS	0.05; -0.20	0.82; 0.50	1.09; 2.71	0.967	0.733	–	–
	MSM	-1.36; -8.04	0.84; 0.73	1.05; 2.13	0.976	0.922	3.04; -5.01	1.67; 3.94
Utilities	MS	6.97; -8.05	0.82; 0.56	1.03; 1.63	0.994	0.977	–	–
	MSM	7.07; -3.89	0.81; 0.63	1.03; 1.67	0.989	0.960	2.90; -5.67	1.59; 3.69

Table 4.5: Comparison of OLS betas and various conditional beta series

Summary of various conditional beta series reporting the mean betas and their ranges (in brackets) for the first nine sectors.

Sector	β^{OLS}	β^{KFRW}	β^{KFMR}	β^{MSM}	β^{MS}
Automobiles	1.148	1.145	1.145	1.182	1.203
		(0.123; 1.609)	(0.025; 2.370)	(1.065; 1.225)	(1.029; 1.262)
Banks	1.062	1.019	1.034	1.014	1.041
		(0.367; 1.337)	(−0.156; 1.978)	(0.971; 1.103)	(1.011; 1.109)
Basics	0.902	0.956	0.945	0.955	0.950
		(−0.018; 1.489)	(−0.364; 1.616)	(0.815; 1.025)	(0.839; 1.047)
Chemicals	0.907	0.913	0.900	0.947	0.941
		(0.122; 1.299)	(0.031; 1.395)	(0.849; 0.980)	(0.865; 0.996)
Construction	0.886	0.964	0.933	0.980	0.992
		(0.617; 1.358)	(−0.036; 1.581)	(0.794; 1.070)	(0.766; 1.142)
Financials	0.997	0.937	0.947	0.948	0.956
		(0.552; 1.267)	(0.139; 2.081)	(0.911; 1.049)	(0.906; 1.083)
Food	0.648	0.710	0.708	0.773	0.763
		(−0.345; 1.115)	(−0.362; 1.116)	(0.519; 0.894)	(0.486; 0.910)
Healthcare	0.777	0.809	0.806	0.830	0.811
		(0.055; 1.142)	(0.010; 1.173)	(0.731; 0.903)	(0.678; 0.913)
Industrials	0.977	0.994	0.996	0.990	0.992
		(0.816; 1.233)	(−0.198; 1.836)	(0.974; 0.997)	(0.957; 1.017)

Table 4.6: Comparison of OLS betas and various conditional beta series

Summary of various conditional beta series reporting the mean betas and their ranges (in brackets) for the last nine sectors.

Sector	β^{OLS}	β^{KFRW}	β^{KFMR}	β^{MSM}	β^{MS}
Insurance	1.268	1.144	1.155	1.177	1.197
		(0.456; 1.929)	(0.032; 3.055)	(1.105; 1.392)	(1.117; 1.372)
Media	1.215	1.184	1.181	1.110	1.132
		(0.667; 2.586)	(−0.538; 3.820)	(1.039; 1.322)	(1.049; 1.406)
Oil & Gas	0.758	0.781	0.753	0.850	0.834
		(0.318; 1.056)	(−0.217; 1.372)	(0.713; 0.912)	(0.584; 0.958)
Personal	0.907	0.956	0.952	0.955	0.913
		(0.619; 1.186)	(0.576; 1.186)	(0.853; 0.992)	(0.802; 0.949)
Retail	0.949	0.907	0.898	0.903	0.934
		(0.264; 1.599)	(−0.470; 2.110)	(0.876; 0.994)	(0.903; 0.972)
Technology	1.489	1.460	1.488	1.313	1.356
		(0.853; 3.134)	(0.761; 3.438)	(1.174; 1.597)	(1.181; 1.709)
Telecom	1.194	1.246	1.266	1.146	1.145
		(0.738; 2.256)	(0.679; 2.290)	(1.122; 1.213)	(1.088; 1.314)
Travel	0.770	0.791	0.752	0.810	0.781
		(0.500; 0.981)	(−0.342; 1.453)	(0.728; 0.837)	(0.501; 0.814)
Utilities	0.694	0.753	0.742	0.762	0.744
		(0.239; 1.024)	(0.175; 1.018)	(0.626; 0.812)	(0.561; 0.819)

Chapter 5

Hidden Semi-Markov Models

Hidden semi-Markov models (HSMMs) are an extension of the well-known class of HMMs. While the runlength distribution of the HMM implicitly follows a geometric distribution, HSMMs allow for more general runlength distributions. In this chapter we study the estimation techniques for HSMMs from discrete – possibly multivariate – sequences which consider several runlength and conditional distributions.

Hidden semi-Markov chains with nonparametric state occupancy (or sojourn time, dwell time, runlength) distributions were first proposed in the field of speech recognition by Ferguson (1980). They were considered to be an alternative approach to classical HMMs for speech modeling because the latter are not flexible enough to describe the time spent in a given state, which follows a geometric distribution as a consequence of the Markov property of the underlying Markov chain. After this pioneering work, several problems related to hidden semi-Markov chains were further investigated by different authors, e.g., Levinson (1986), Guédon & Cocozza-Thivent (1990), Guédon (1999, 2003), Sansom & Thomson (2001, 2000), Yu & Kobayashi (2003) and different parametric hypotheses were considered for the state occupancy as well as for the observation distributions.

In the following, the state process of the HSMMs is assumed to be a semi-Markov chain with finite number of states. The conditional independence assumption for the observation process is similar to a simple hidden Markov chain. A semi-Markov chain can be constructed as follows: An embedded first-order Markov chain models the transitions between distinct states, while explicitly given discrete state occupancy distributions model the sojourn time for each of the states.

The first estimation procedure of Ferguson (1980), which has been applied by several authors, is based on the assumption that the end of a sequence systematically coincides with the exit from a state. This very specific assumption eases the notation of the likelihood functions but also has some disadvantages. One of the disadvantages is that the enforced exit from a state at the last observed data point may not be a realistic assumption in every case. The other is that the resulting models do not allow absorbing states and can therefore not be considered to be a true generalization of hidden Markov chains.
We focus on the theory for right-censored models introduced by Guédon (2003). His approach allows us to overcome the limitations of the classical HSMMs by defining HSMMs with an extended state sequence of the underlying semi-Markov chain. The last observation does not necessarily coincide with an exit from the last visited state. However, the estimation procedures become more complicated due to the inclusion of a right-censoring of the time spent in the last visited state.

This chapter is structured as follows. In Section 5.1 we introduce the concept of HSMMs. The derivation of the likelihood function of a HSMM is presented in Section 5.2. The maximization of the likelihood by the EM algorithm is the main subject of Section 5.3 which includes the derivation of the Q-function, the forward-backward algorithm and re-estimation formulae for various conditional and runlength distributions. Finally, the asymptotic properties and the implementation of stationary HSMMs are treated briefly in the Sections 5.4 and 5.5.

5.1 The Basic Definitions

A HSMM consists of a pair of discrete-time stochastic processes $\{S_t\}$ and $\{X_t\}$, $t \in \{0, \ldots, \tau - 1\}$. The observed process $\{X_t\}$ is linked to the hidden, i.e., unobserved state process $\{S_t\}$ by the conditional distribution depending on the state process. This construction is comparable to the HMM from Chapter 2. However, for a HSMM the state process is a finite-state semi-Markov chain. As for the HMMs, the support of the conditional distributions usually overlaps and so, in general, a specific observation can arise from more than one state. Thus the state process $\{S_t\}$ is not observable directly through the observation process $\{X_t\}$ but can only be estimated. The observation process $\{X_t\}$ itself may either be discrete or continuous, univariate or multivariate.

5.1.1 Semi-Markov Chains

We introduce semi-Markov chains briefly here; for a general reference about various semi-Markov models, see Kulkarni (1995). Since HSMMs with absorbing states are not appropriate for our applications in economics and also require a more difficult notation, we restrict ourselves to semi-Markov chains without absorbing states. The sojourn time in each of the states is a discrete non-negative random variable with an arbitrary distribution.

To construct a J-state semi-Markov chain we require an embedded first-order Markov chain. This J-state first-order Markov chain is defined by the following parameters:

- the **initial probabilities** $\pi_j := P(S_0 = j)$ with $\sum_j \pi_j = 1$, and
- the **transition probabilities** for the state i. For each $j \neq i$
  $p_{ij} := P(S_{t+1} = j | S_{t+1} \neq i, S_t = i)$ with $\sum_{j \neq i} p_{ij} = 1$ and $p_{ii} = 0$.

Note that the diagonal elements of the transition probability matrix have to be zero in contrast to ordinary HMMs (certainly, this is not true anymore if the assumption that all states are non-absorbing is relaxed). This embedded first-order Markov chain represents transitions between distinct states. To build a semi-Markov chain the **occupancy** (or **sojourn time**, **dwell time**) **distributions** $d_j(u)$ have to be assigned to each of the states by

$$d_j(u) := P(S_{t+u+1} \neq j, S_{t+u-v} = j, v = 0, \ldots, u-2 \,|\, S_{t+1} = j, S_t \neq j),$$

for $u \in \{1, \ldots, M_j\}$. The sojourn of the unobserved process of length u from $t+1$ until $t+u$ in the state j is denoted by $d_j(u)$. Before and after this sojourn in state j, the process has to be in some different state. The upper bound of the time spent in state j is denoted by M_j. We will assume that the state occupancy distribution is concentrated on the finite set of time points $\{1 \ldots, M_j\}$, where M_j may also increase up to the entire length of the observed sequence (in particular for parametric dwell time distributions).
For the particular case of the last visited state, we introduce the so-called **survivor function** of the sojourn time in state j,

$$D_j(u) := \sum_{v \geq u} d_j(v).$$

The survivor function sums up the individual probability masses of all possible sojourns of length $v \geq u$. It plays an important role in the Sections 5.2.1 and

5.2.2 for the extension from the classical HSMMs to right-censored models in which the last observation is no longer assumed to coincide with the exit from a state. The first-order Markov chain and the state occupancy distributions together constitute a semi-Markov chain.
If the process starts in state j at time $t = 0$, the following relation can be verified

$$P\left(S_t \neq j, S_{t-v} = j, v = 1, \ldots, t\right) = d_j\left(t\right)\pi_j. \tag{5.1}$$

Recall that conditional independence holds at each time step for a Markov chain, i.e., the current state at time t depends only on the last state at time $t-1$. Considering the definitions above, the difference between a Markov chain and a semi-Markov chain becomes clear. The Markov-property does not need to hold at each time t, but is transferred to the level of the embedded first-order Markov chain. To illustrate the transfer of the Markov property, we assume the classical assumption of an exit from the last visited state at the last observation to hold true for the rest of this section. This simplifies the notation, however the extension to the right-censored case would be straightforward. Let the number of different states visited consecutively by the semi-Markov chain be $R+1$. Then the number of sojourn times, denoted by $u_0, \ldots, u_R$, fulfills

$$u_0 + \cdots + u_R = \tau.$$

The relationship between the sojourn times $u_0, \ldots, u_R$ and the state sequence can be written in a simplified way by reducing the entire sequence of states $s_0, \ldots, s_{\tau-1}$ to the sequence of states $\tilde{s}_0, \ldots, \tilde{s}_R$ which have been visited:

$$\begin{aligned}
\tilde{s}_0 &:= \left\{s_0, s_1, \ldots, s_{u_0-1}\right\} \\
\tilde{s}_1 &:= \left\{s_{u_0}, s_{u_0+1}, \ldots, s_{u_0+u_1-1}\right\} \\
\tilde{s}_2 &:= \left\{s_{u_0+u_1}, s_{u_0+u_1+1}, \ldots, s_{u_0+u_1+u_2-1}\right\} \\
\vdots &:= \vdots \\
\tilde{s}_R &:= \left\{s_{u_0+\ldots+u_{R-1}}, s_{u_0+\ldots+u_{R-1}+1}, \ldots, s_{\tau-1}\right\}.
\end{aligned}$$

The Markov property is transferred to the sequence of visited states $\tilde{s}_r$, $r = 0, \ldots, R$ which constitute a hidden Markov chain. Focusing on hidden semi-Markov chains, the conditional independence between the past and the future holds true only when the process changes the state.

Remark: Relation (5.1) implies that the process enters a "new" state at time 0. This assumption makes the process non-stationary (in general), but we will

show how it may be relaxed in the context of stationary HSMMs in Section 5.5. To start with, we consider homogeneous but non-stationary HSMMs.

5.1.2 Hidden Semi-Markov Models

As in the case with discrete HMMs, a discrete HSMM can be seen as a pair of stochastic processes $\{S_t, X_t\}$. The discrete output process $\{X_t\}$ at $t \in \{0, \ldots, \tau - 1\}$ is associated with the state process $\{S_t\}$ by the J conditional or rather component distributions in the context of mixture distributions. The component distributions can be either discrete or continuous, but the state process is a finite-state semi-Markov chain.
In the discrete case the output process $\{X_t\}$ is related to the semi-Markov chain $\{S_t\}$ by the **observation** (or **emission**) **probabilities**

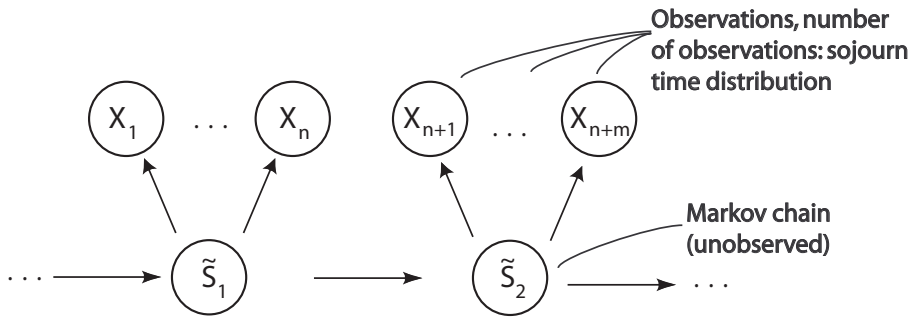
$$b_j(x_t) = P(X_t = x_t \mid S_t = j) \text{ with } \sum_{x_t} b_j(x_t) = 1.$$

Of course, in the case of continuous component distributions $\sum_{x_t}$ has to be replaced by $\int_{x_t}$. The observation process is characterized by the conditional independence property,

$$P(X_t = x_t \mid X_0^{\tau-1} = x_0^{\tau-1}, S_0^{\tau-1} = s_0^{\tau-1}) = P(X_t = x_t \mid S_t = s_t),$$

which implies the fact that the output process at time t depends only on the state of the underlying semi-Markov chain at time t similar to HMMs. The basic structure of the HSMM is illustrated in Figure 5.1, which is analogous to Figure 2.5 in Section 2.2.1.

Figure 5.1: Basic structure of a Hidden Semi-Markov Model



Without loss of generality the observation process $\{X_t\}$ is initially assumed to be one-dimensional for convenience of the reader and notational reasons. The extension to the multivariate case is straightforward and will be treated in Section 5.3.4, where both univariate and multivariate component distributions are considered for the observation distributions.

5.2 The Likelihood Function of a Hidden Semi-Markov Model

The crucial step for parameter estimation of HSMMs is the derivation of a tractable expression for the likelihood function in order to perform maximum likelihood estimation. The difficulty in deriving the likelihood lies in the fact that we are faced with a missing data problem because the state sequence remains unobserved. A very convenient approach to deal with this type of problem is the derivation of the likelihood of the complete data, which allows one to apply the expectation maximization (EM) algorithm in Section 5.3.
In the following, we consider the case of a single observed sequence which is relevant for later applications. The generalization to the case of a sample of sequences is straightforward, cf. Guédon & Cocozza-Thivent (1990), Guédon (1992). We use the notation introduced in Chapter 2. Recall that the observed as well as the state sequences of length $t_1 - t_0 + 1$ with $0 \leq t_0 < t_1 \leq \tau - 1$ are denoted by

$$\begin{aligned}\left\{X_{t_0}^{t_1} = x_{t_0}^{t_1}\right\} &:= \left\{X_{t_0} = x_{t_0}, \ldots, X_{t_1} = x_{t_1}\right\}, \text{ and} \\ \left\{S_{t_0}^{t_1} = s_{t_0}^{t_1}\right\} &:= \left\{S_{t_0} = s_{t_0}, \ldots, S_{t_1} = s_{t_1}\right\}, \text{ respectively.}\end{aligned}$$

As a first step we consider the classical form of the complete-data likelihood $\tilde{L}_c$ introduced by Ferguson (1980) which allows only for sequences in which the last observation coincides with an exit from the hidden state. For the complete-data formulation, both the outputs $x_0^{\tau-1}$ and the states $s_0^{\tau-1}$ of the underlying semi-Markov chain are known, and thus

$$\tilde{L}_c\left(s_0^{\tau-1}, x_0^{\tau-1} \,|\, \theta\right) = P\left(S_0^{\tau-1} = s_0^{\tau-1}, X_0^{\tau-1} = x_0^{\tau-1} \,|\, \theta\right),$$

where θ denotes the vector of all parameters. In this framework it can be further assumed that the number of states visited, $R+1$, is a fixed, but unknown number. This allows us to write the complete-data likelihood as

$$\begin{aligned}\tilde{L}_c\left(s_0^{\tau-1}, x_0^{\tau-1} \,|\, \theta\right) &= P\left(S_0^{\tau-1} = s_0^{\tau-1}, X_0^{\tau-1} = x_0^{\tau-1} \,|\, \theta\right) \\ &= P(X_0^{\tau-1} = x_0^{\tau-1} \,|\, S_0^{\tau-1} = s_0^{\tau-1}, \theta) P(S_0^{\tau-1} = s_0^{\tau-1} \,|\, \theta) \\ &= \prod_{t=0}^{\tau-1} b_{s_t}(x_t)\, \pi_{\tilde{s}_0} d_{\tilde{s}_0}(u_0) \prod_{r=1}^{R} p_{\tilde{s}_{r-1}\tilde{s}_r} d_{\tilde{s}_r}(u_r) \qquad (5.2)\end{aligned}$$

where $\tilde{s}_r$ is the $(r+1)^{\text{th}}$ visited state and u_r denotes the time spent in state $\tilde{s}_r$. Equation (5.2) is the most popular form of the complete-data likelihood for HSMM. See, e.g., Ferguson (1980), Levinson (1986), and Rabiner (1989).

In reality the underlying state sequence cannot be observed and the number of states visited is not available. Nevertheless the state sequence contributes to the likelihood by regarding all admissible paths from length one to length τ. That is, we consider state sequences of the form

$$\pi_{\tilde{s}_0} d_{\tilde{s}_0}(u_0) \prod_{r \geq 1} p_{\tilde{s}_{r-1}\tilde{s}_r} d_{\tilde{s}_r}(u_r)\, \mathbf{1}_{\{\sum_r u_r = \tau\}}(u_r)_{r \geq 1}. \qquad (5.3)$$

The indicator function $\mathbf{1}_{\{\sum_r u_r = \tau\}}(u_r)_{r \geq 1}$ guarantees that the lengths of the paths equals the length of the observations.

The likelihood of a HSMM with exit from the last visited state at $\tau - 1$ is obtained by enumeration of the complete-data likelihood (5.2) over all possible state sequences, which yields

$$\tilde{L}(\theta) = \sum_{s_0,\ldots,s_{\tau-1}} \tilde{L}_c(s_0^{\tau-1}, x_0^{\tau-1} \,|\, \theta). \tag{5.4}$$

In this representation, the difficulty of solving equation (5.4) explicitly is obvious: $\sum_{s_0,\ldots,s_{\tau-1}}$ includes all admissible paths of the form given by Equation (5.3), which effectively eliminates any chance of obtaining an analytic solution.

In the subsequent section we relax the assumption that the last observation coincides with an exit from the last visited state and introduce the partial and the complete likelihood estimator to generalize the classical approach.

5.2.1 The Partial Likelihood Estimator

The standard formulation (5.3) from the classical HSMM assumes that the end of the sequence of observations always coincides with the exit from a state because the sojourn times u_r sum up to τ. This very specific assumption has two main consequences. While on the one hand, only semi-Markov chains without absorbing states can be considered, on the other hand, the assumption does not seem to be realistic in most applications. For example, in the context of financial time series the states often represent different economic situations (e.g., bull and bear markets, or periods of low and high volatility). Obviously the economic situation cannot be assumed to end with the last observation.

As first step to generalize the classical approach, Guédon (2003) proposes to write the contribution of the state sequence to the complete-data likelihood $L_c(s_0^{\tau-1}, x_0^{\tau-1} \,|\, \theta)$ as

$$\pi_{\tilde{s}_0} d_{\tilde{s}_0}(u_0) \left\{ \prod_{r=1}^{R-1} p_{s_{r-1}\tilde{s}_r} d_{\tilde{s}_r}(u_r) \right\} p_{\tilde{s}_{R-1}\tilde{s}_R} D_{\tilde{s}_R}(u_R)\, \mathbf{1}_{\{\sum_{r=0}^{R} u_r = \tau\}}(u_0, \ldots, u_R). \tag{5.5}$$

The main difference to the classical approach presented in equation (5.3) lies in the substitution of the ordinary sojourn time probability by the survivor function $D_{\tilde{s}_R}(u_R)$ for the last visited state. Recall that the survivor function is defined by

$$D_j(u) := \sum_{v \geq u} d_j(v)$$

and represents the marginal sojourn time of u by summing over all admitted sojourn times $v \geq u$, i.e., limited by the upper bound M_j. The survivor function performs a right-censoring of the sojourn time in the last visited state.
The complete-data likelihood $L_c(s_0^{\tau-1}, x_0^{\tau-1} \mid \theta)$ based upon (5.5) considers only the state sequence $s_0^{\tau-1}$, and ignores the states after $s_{\tau-1}$.
Guédon (2003) explains the introduction of the modified probability distribution of the state sequences in Equation (5.5) and the resulting estimator by: "This assumption corresponds to a more general statement of the problem but also generates some difficulties regarding the final right-censored sojourn time interval which cannot be used in the estimation procedure. The rationale behind the corresponding estimator is related to Cox's partial likelihood idea Cox (1975) in the sense that it is derived by maximizing parts of the likelihood function [...]. Nevertheless, the aim underlying the factorization of the likelihood is clearly different from that emphasized by Cox."
In the following, this estimator is denoted by **partial likelihood estimator** because the information given by the state sequence is only considered in parts. The consequences on the estimation procedures is analyzed more precisely in Section 5.3.4.3 in conjunction with the M-step for the state occupancy distributions.

A remarkable consequence of the application of the partial likelihood estimator is a downward bias of the estimated state occupancy distribution. This bias results from the fact that the long sojourn times are more likely to include the final right-censored sojourn time. Consequently they are more often excluded from the estimation procedure than short sojourn times.
The theoretical justification for the bias lies in the fact that censoring at a time which is not a stopping time may introduce bias (cf. Aalen & Husebye 1991). The end of the last, complete sojourn time is not a stopping time because, roughly speaking, for a stopping time it has to be possible to decide whether or not the stopping event has occurred until time t on the basis of the information contained in the corresponding σ-algebra at time t. In our case the family of increasing σ-algebras are the filtration of the hidden semi-Markov process. The information whether the state sequence dwells in the last visited state until time t, or whether the last visited state is left at time t, is only available with the knowledge of the state sequence at time $t + 1$. Therefore the end of the last complete sojourn time is not a stopping time.

The full likelihood estimator introduced in the next section includes a correction for the downward bias of the estimated state occupancy distribution.

5.2.2 The Complete Likelihood Estimator

Guédon (2003) provides an extension of the partial likelihood estimator which requires a modified complete-data likelihood. In this setting the complete-data likelihood incorporates both the outputs $x_0^{\tau-1}$ and the state sequence $s_0^{\tau-1+u}$ of the underlying semi-Markov chain. The difference to the partial likelihood estimator is that, in this situation the final right-censored sojourn time interval contributes to the estimation procedure.
In detail, the state sequence remains in the last visited state $s_{\tau-1}$ from time $\tau-1$ to $\tau-1+u$, $u=0,1,\ldots$. The exit from the last visited state takes place at time $\tau-1+u$, which yields the complete-date likelihood

$$\begin{aligned} L_c\left(s_0^{\tau-1+u}, x_0^{\tau-1} \,|\, \theta\right) &= P\big(S_0^{\tau-1} = s_0^{\tau-1}, S_{\tau-1+v} = s_{\tau-1}, v = 1,\ldots,u-1, \\ &\qquad S_{\tau-1+u} \neq s_{\tau-1}, X_0^{\tau-1} = x_0^{\tau-1} \,|\, \theta\big). \end{aligned}$$

The estimator based on this specification of the complete-data problem is called **complete likelihood estimator**.
Compared to formula (5.5), the contribution of the state sequence to the complete-data likelihood has to be modified to

$$\pi_{\tilde{s}_0} d_{\tilde{s}_0}(u_0) \prod_{r=1}^{R} p_{\tilde{s}_{r-1}\tilde{s}_r} d_{\tilde{s}_r}(u_r)\, \mathbf{1}_{\{\sum_{r=0}^{R-1} u_r < \tau \leq \sum_{r=0}^{R} u_r\}}(u_0,\ldots,u_r). \tag{5.6}$$

The characteristic of the state sequences involved in (5.6) is that the sum of the first R sojourn times has to be smaller than the length of the observations τ, but the overall sojourn time, summed over $u_0,\ldots,u_{\tau-1}$, may exceed τ.

Compared to the original likelihood given by Equation (5.4), the completed state sequence complicates the likelihood function by an additional sum over all possible prolongations of the state sequence $s_0,\ldots,s_{\tau-1}$. It is given by

$$L(\theta) = \sum_{s_0,\ldots,s_{\tau-1}} \sum_{u_{\tau+}} L_c\left(s_0^{\tau-1+u}, x_0^{\tau-1} \,|\, \theta\right), \tag{5.7}$$

where $\sum_{s_0,\ldots,s_{\tau-1}}$ denotes the summation over every possible state sequence of length τ and and $\sum_{u_{\tau+}}$ denotes the sum on every supplementary duration from time τ spent in the state occupied at time $\tau - 1$.

Remark: Note that the results of an estimation based on either the complete or the partial likelihood estimator both depend on the contribution of the right-censored last visited state, which is taken into account or not, respectively. Hence none of the estimators yields the results of the original algorithms of Ferguson (1980) which consider the time spent in the last visited state as a typical (uncensored) sojourn time. For our applications in Chapter 6, we focus on the complete likelihood estimator because it makes use of the maximum amount of information available from the data.

The objective of the following Section 5.3 is the derivation of a suitable estimation procedure to maximize the likelihood of the observed sequence $x_0^{\tau-1}$. The exact time spent in the last visited state remains unknown and cannot be determined; only the minimum time is known. Therefore the survivor function plays a key role.

5.3 The EM Algorithm for Hidden Semi-Markov Models

In this section we present the derivation of the EM algorithm for right-censored HSMMs. Readers who are not familiar with the EM theory are referred to the brief introduction to the EM procedure in Appendix A.

The estimation of HSMMs is an incomplete-data problem because the observations are the only data which are accessible, the underlying path of the hidden semi-Markov chain remains inaccessible. Therefore the EM algorithm is the suitable way for maximum-likelihood estimation of HSMMs. However, the Q-function has to be derived in this specific framework. The EM algorithm maximizes $L(\theta)$ from Equation (5.7) by iteratively maximizing $Q(\theta|\theta^{(k)})$ over θ. The next value $\theta^{(k+1)}$ is chosen as

$$\theta^{(k+1)} = \arg\max_{\theta} Q(\theta \,|\, \theta^{(k)}).$$

Each iteration of the EM algorithm increases $L(\theta)$ and, generally, the sequence of re-estimated parameters $\theta^{(k)}$ converges to a local maximum of $L(\theta)$.

Our estimation procedure is based on the application of the EM algorithm introduced by Guédon (2003) which has the following two appealing properties.

- Compared to earlier algorithms, the complexity of the forward-backward algorithm is significantly reduced, which allows the treatment of longer time series. In particular, the E-step of the EM algorithm has a complexity that is similar to the complexity of the forward algorithm alone, or of the Viterbi algorithm. That is, the EM algorithm requires $O\left(J\tau\left(J+\tau\right)\right)$-time in the worst case and $O\left(J\tau\right)$-space.
- The calculation of the forward/backward probabilities requires the multiplication of many probabilities which leads to the problem of numerical underflow (see e.g. MacDonald & Zucchini (1997) where this point is discussed for HMMs, or Ferguson (1980) for the semi-Markovian case). The proposed forward-backward algorithm is auto-scaling, i.e., additional scaling procedures are not necessary because of its immunity to numerical underflow.

The observations are given by $\{X_0^{\tau-1}\}$. The hidden variable of the semi-Markov chain is $\{S_0^{\tau-1+u}\}$. Two estimation procedures are considered, building on the partial and the complete likelihood estimator, respectively. We begin with the latter case. The estimation procedure for the partial likelihood estimator can easily be obtained by a slight modification, which is presented in Section 5.3.3.

5.3.1 The Q-Function

Let $\theta^{(k)}$ denote the current value of θ at iteration k. The Q-function is defined by the conditional expectation of the complete-data log-likelihood which yields

$$Q(\theta \,|\, \theta^{(k)}) = E\left\{\log L_c\left(S_0^{\tau-1+u}, X_0^{\tau-1} \,|\, \theta\right) \,|\, X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right\}.$$

To obtain a mathematically tractable formulation of the Q-function, the conditional expectation has to be rewritten path-wise. The conditional distribution of the missing observations is given by $P(S_0^{\tau-1+u} = s_0^{\tau-1+u} \,|\, X_0^{\tau-1} = x_0^{\tau-1}, \theta)$ and the distribution of the complete-data is given by $P(S_0^{\tau-1+u} = s_0^{\tau-1+u}, X_0^{\tau-1} = x_0^{\tau-1} \,|\, \theta)$.

Hence the Q-function becomes

$$\begin{aligned} Q(\theta \mid \theta^{(k)}) &= \sum_{s_1,\ldots,s_{\tau-1}} \sum_{u_{\tau+}} \Big[\log L_c(s_0^{\tau-1+u}, x_0^{\tau-1} \mid \theta) \\ &\qquad P(S_0^{\tau-1+u} = s_0^{\tau-1+u} \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}) \Big]. \end{aligned} \tag{5.8}$$

The contribution of a specific path to the likelihood of a HSMM for given parameter θ is presented in formula (5.6). Considering also the contribution of the observed sequence yields the complete-data likelihood

$$L_c(s_0^{\tau-1+u}, x_0^{\tau-1} \mid \theta) = \pi_{\tilde{s}_0} d_{\tilde{s}_0}(u_0) \prod_{r=1}^{R} p_{\tilde{s}_{r-1}\tilde{s}_r} d_{\tilde{s}_r}(u_r) \prod_{t=0}^{\tau-1} b_{s_t}(x_t).$$

Substituting this representation into equation (5.8) yields the Q-function as a sum of four terms, each of which depends on an independent subset of set of parameters θ:

$$\begin{aligned} &Q(\theta \mid \theta^{(k)}) = \\ &\quad \sum_{s_1,\ldots,s_{\tau-1}} \sum_{u_{\tau+}} \log \pi_{\tilde{s}_0} P(S_0^{\tau-1+u} = s_0^{\tau-1+u} \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}) \\ &\quad + \sum_{s_1,\ldots,s_{\tau-1}} \sum_{u_{\tau+}} \left(\sum_{r=1}^{R} \log p_{\tilde{s}_{r-1}\tilde{s}_r} \right) P(S_0^{\tau-1+u} = s_0^{\tau-1+u} \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}) \\ &\quad + \sum_{s_1,\ldots,s_{\tau-1}} \sum_{u_{\tau+}} \left(\sum_{r=0}^{R} \log d_{\tilde{s}_r}(u_r) \right) P(S_0^{\tau-1+u} = s_0^{\tau-1+u} \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}) \\ &\quad + \sum_{s_1,\ldots,s_{\tau-1}} \sum_{u_{\tau+}} \left(\sum_{t=0}^{\tau-1} \log b_{s_t}(x_t) \right) P(S_0^{\tau-1+u} = s_0^{\tau-1+u} \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}). \end{aligned} \tag{5.9}$$

This form of the Q-function is advantageous for the maximization procedure because each of the terms can be treated individually.

The first term of Equation (5.9) can be written as

$$\sum_{s_1,\ldots,s_{\tau-1}} \sum_{u_{\tau+}} \log \pi_{\tilde{s}_0} P(S_0^{\tau-1+u} = s_0^{\tau-1+u} \,|\, X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}) = \sum_{j=0}^{J-1} P(S_0 = j \,|\, X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}) \log \pi_j, \tag{5.10}$$

because summing over all possible paths is equivalent to repeatedly selecting the different π_j $(j = 0, \ldots, J-1)$ and can therefore be marginalized to $t = 0$.

The second term in Equation (5.9) is transformed similarly by marginalizing the full paths to the transitions from i to j at time t for all $t \in \{0, \ldots, \tau-2\}$:

$$\sum_{s_1,\ldots,s_{\tau-1}} \sum_{u_{\tau+}} \left(\sum_{r=1}^{R} \log p_{\tilde{s}_{r-1}\tilde{s}_r} \right) P(S_0^{\tau-1+u} = s_0^{\tau-1+u} \,|\, X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}) = \sum_{i=0}^{J-1} \sum_{j \neq i} \sum_{t=0}^{\tau-2} P(S_{t+1} = j, S_t = i \,|\, X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}) \log p_{ij}. \tag{5.11}$$

The third term containing the sojourn time distribution is also marginalized to the different runlengths $d_j(u)$ of length u arising in state j and can be split up into the two summands

$$\begin{aligned} &\sum_{s_1,\ldots,s_{\tau-1}} \sum_{u_{\tau+}} \left(\sum_{r=0}^{R} \log d_{\tilde{s}_r}(u_r) \right) P(S_0^{\tau-1+u} = s_0^{\tau-1+u} \,|\, X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}) = \\ &\sum_u \Bigg\{ \sum_{t=0}^{\tau-2} P\left(S_{t+u+1} \neq j, S_{t+u-v} = j, v = 0, \ldots, u-1, \right. \\ &\qquad \left. S_t \neq j | X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right) + P\left(S_u \neq j, S_{u-v} = j, v = 1, \ldots, u \,| \right. \\ &\qquad \left. X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right) \Bigg\} \log d_j(u). \end{aligned} \tag{5.12}$$

The second term within the parentheses deals with the sojourn times of length u which start at the first state of the semi-Markov chain, the first term treats

all other sojourn times of length u occurring in $[1, \ldots, \tau - 1]$. In Section 5.3.3 we will give a more tractable expression obtained from the forward-backward algorithm.

The last term of equation (5.9) including the conditional distributions is also transformed to the sum of the marginal distributions of the observations at time t in state j by

$$\sum_{s_1,\ldots,s_{\tau-1}} \sum_{u_{\tau+}} \left(\sum_{t=0}^{\tau-1} \log b_{s_t}(x_t) \right) P(S_0^{\tau-1+u} = s_0^{\tau-1+u} \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}) = \sum_{j=0}^{J-1} \sum_{t=0}^{\tau-1} P\left(S_t = j \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right) \log b_j\left(x_t\right). \quad (5.13)$$

The representation of the conditional expectation $Q(\theta \mid \theta^{(k)})$ given in (5.10), (5.11), (5.12), and (5.13) of the different terms allows one to rewrite Equation (5.9) as

$$\begin{aligned} Q(\theta \mid \theta^{(k)}) &= Q_\pi\left(\{\pi_j\}_{j=0}^{J-1} \mid \theta^{(k)}\right) + \sum_{i=0}^{J-1} Q_p\left(\{p_{ij}\}_{j=0}^{J-1} \mid \theta^{(k)}\right) \\ &\quad + \sum_{j=0}^{J-1} Q_d\left(\{d_j(u)\} \mid \theta^{(k)}\right) I\left(p_{jj} = 0\right) \\ &\quad + \sum_{j=0}^{J-1} Q_b\left(\left\{b_j\left(x_0^{\tau-1}\right)\right\} \mid \theta^{(k)}\right), \end{aligned} \quad (5.14)$$

where the first term

$$Q_\pi\left(\{\pi_j\}_{j=0}^{J-1} \mid \theta^{(k)}\right) := \sum_j P\left(S_0 = j \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right) \log \pi_j \quad (5.15)$$

corresponds to the initial probabilities. The second term

$$Q_p\left(\{p_{ij}\}_{j=0}^{J-1} \mid \theta^{(k)}\right) := \sum_{j \neq i} \sum_{t=0}^{\tau-2} P\left(S_{t+1} = j, S_t = i \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right) \log p_{ij} \quad (5.16)$$

corresponds to the transition probabilities.

The term for the sojourn times is given by

$$
\begin{aligned}
&Q_d\left(\{d_j(u)\} \mid \theta^{(k)}\right) \\
&\quad := \sum_u \Bigg\{ \sum_{t=0}^{\tau-2} P\left(S_{t+u+1} \neq j, S_{t+u-v} = j, v = 0, \ldots, u-1,\right. \\
&\qquad\qquad S_t \neq j \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\Big) + P\left(S_u \neq j, S_{u-v} = j, v = 1, \ldots, u \mid\right. \\
&\qquad\qquad \left.\left. X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right)\right\} \log d_j(u)
\end{aligned} \tag{5.17}
$$

and the last term models the observation component

$$
Q_b\left(\left\{b_j\left(x_0^{\tau-1}\right)\right\} \mid \theta^{(k)}\right) := \sum_{t=0}^{\tau-1} P\left(S_t = j \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right) \log b_j(x_t). \tag{5.18}
$$

The re-estimation formulae for the quantities involved in (5.15), (5.16), (5.17) and (5.18) can be obtained by independently maximizing each of these terms. In the following, we refer to them as **re-estimation quantities**.

The implementation of the E-step of the EM algorithm is performed by the forward-backward algorithm. It computes all the re-estimation quantities for all times t and for all states j. The implementation of the forward-backward algorithm is presented in the subsequent Section 5.3.2.

Subsequently, the M-step maximizes each of the terms w.r.t. θ to obtain the next set of initial values for the E-step of the following iteration. The difficulty of the maximization varies with the component distributions chosen and may also involve numerical maximization methods when an explicit solution is not available, e.g. for the t distribution in Section 5.3.4.

5.3.2 The Forward-Backward Algorithm

The key algorithm for the estimation of HSMMs, as well as for ordinary HMMs, is the forward-backward algorithm. The basic idea of the algorithm is the decomposition of the probability

$$
L_j(t) := P(S_t = j \mid X_0^{\tau-1} = x_0^{\tau-1})
$$

into two components each of which can be calculated by a forward and a backward pass, respectively, through the observations. The two iterative procedures yield the so-called **forward** and **backward probabilities**, which are denoted by $B_{\text{HMC}_j}(t)$ and $F_{\text{HMC}_j}(t)$ for HMMs, and by $B_j(t)$ and $F_j(t)$ for HSMMs.
In case of an ordinary HMM, the forward-backward algorithm is based on the decomposition

$$\begin{aligned} L_j(t) &= P\left(S_t = j \mid X_0^{\tau-1} = x_0^{\tau-1}\right) \\ &= \frac{P\left(X_{t+1}^{\tau-1} = x_{t+1}^{\tau-1} \mid S_t = j\right)}{P\left(X_{t+1}^{\tau-1} = x_{t+1}^{\tau-1} \mid X_0^t = x_0^t\right)} P\left(S_t = j \mid X_0^t = x_0^t\right) \\ &= B_{\text{HMC}_j}(t)\, F_{\text{HMC}_j}(t), \end{aligned} \tag{5.19}$$

which expresses the conditional independence between the past and the future of the process at time t. Devijver (1985) showed that

- the quantities $F_{\text{HMC}_j}(t)$ can be computed by a forward pass through the observed sequence $x_0^{\tau-1}$, i.e., from 0 to $\tau - 1$ and
- the quantities $B_{\text{HMC}_j}(t)$ and $L_j(t)$ can be computed by a backward pass through $x_0^{\tau-1}$, i.e., from $\tau - 1$ to 0.

The resulting algorithm is of complexity $O(J^2\tau)$-time and immune to problems caused by numerical underflow.

In addition to the $L_j(t)$ decomposed in Equation (5.19), the forward-backward algorithm for HSMMs requires the calculation of

$$L1_j(t) := P\left(S_{t+1} \neq j, S_t = j \mid X_0^{\tau-1} = x_0^{\tau-1}\right)$$

with the decomposition

$$\begin{aligned} L1_j(t) &= P\left(S_{t+1} \neq j, S_t = j \mid X_0^{\tau-1} = x_0^{\tau-1}\right) \\ &= \frac{P\left(X_{t+1}^{\tau-1} = x_{t+1}^{\tau-1} \mid S_{t+1} \neq j, S_t = j\right)}{P\left(X_{t+1}^{\tau-1} = x_{t+1}^{\tau-1} \mid X_0^t = x_0^t\right)} P\left(S_{t+1} \neq j, S_t = j \mid X_0^t = x_0^t\right) \\ &= B_j(t)\, F_j(t). \end{aligned} \tag{5.20}$$

The decomposition provided by Equation (5.20) expresses the conditional independence between the past and the future of the process at the times of

a state change of the hidden semi-Markov chain. It forms the basis of the forward-backward algorithm for HSMMs.

While the decomposition (5.19) is convenient for the implementation of the EM algorithm for HMMs, this is not the case for the semi-Markovian case with the decomposition (5.20). In the next paragraph we provide a short overview over the development of the EM-algorithm and motivate our focus on the latest procedures derived by Guédon (2003) who succeeded to develop a powerful estimation algorithm based on the decomposition (5.20).

Since the beginning of the Eighties there has been a rapid progress in the field of the estimation procedures for HMMs and HSMMs. The main disadvantage of Ferguson's (1980) initially proposed forward-backward algorithm is the fact that it only allows the computation of $P\left(S_{t+1} \neq j,\ S_t = j \mid\ X_0^{\tau-1} = x_0^{\tau-1}\right)$ for each time t and each state j instead of the much more convenient smoothing probabilities $P\left(S_t = j \,|\, X_0^{\tau-1} = x_0^{\tau-1}\right)$. Thus, the resulting procedures are comparably complex.

The next significant advancement was developed by Guédon & Cocozza-Thivent (1990) who showed that while the forward probabilities $F_j\left(t\right)$ can be computed by a forward pass through the observed sequence $x_0^{\tau-1}$, a backward pass computes either the backward probabilities $B_j(t)$ or the probabilities $L1_j\left(t\right)$ (or both). The major improvement was the modification of the recursion, which allowed for the computation of the smoothing probabilities $L_j(t) = P\left(S_t = j \,|\, X_0^{\tau-1} = x_0^{\tau-1}\right)$ for each time t and each state j. One drawback of the estimation procedure of Guédon & Cocozza-Thivent (1990) is the high computational demands resulting from a backward recursion with cubic complexity in time (in the worst case) and a forward recursion with quadratic complexity. However, Guédon & Cocozza-Thivent (1990) indirectly fitted the conditional independence properties of a HSMM with the EM estimate requirements, which was a major development in the estimation of HSMMs.

The most recent significant advancement is presented by Guédon (2003) who succeeded in developing a new backward recursion. The recursion's complexity both in time and in space is similar to that of the forward recursion, which is $O\left(J\tau\left(J+\tau\right)\right)$-time (in the worst case) and $O\left(J\tau\right)$-space. This means that the computation of $L_j(t) = P(S_t = j \,|\, X_0^{\tau-1} = x_0^{\tau-1})$ as well as $L1_j\left(t\right) = P\left(S_{t+1} \neq j,\ S_t = j \mid\ X_0^{\tau-1} = x_0^{\tau-1}\right)$ require the same complexity. This allows one to fit even long sequences of observations in a reasonable amount of time. Moreover he relaxes the assumption that the last visited state terminates at the time of the last observation.

Remark: It can be noted that the forward-backward algorithm of Guédon (2003) basically computes the smoothed probabilities $L_j(t) = P(S_t = j \mid X_0^{\tau-1} = x_0^{\tau-1})$ for all $t \in \{0, \ldots, \tau - 1\}$. Hence this forward-backward algorithm can be recognized as smoothing algorithm. It is based on quantities comparable to $B_{\text{HMC}_j}(t)$ and $F_{\text{HMC}_j}(t)$ from the HMM decomposition (5.19).

We use the algorithm of Guédon (2003) with some extensions to various distributions for the component distributions and also for the runlength distributions. The key estimation procedures of the algorithm will be presented in the following sections.

5.3.2.1 The Forward Iteration

The forward iteration involves the computation of the forward-probabilities $F_j(t) = P(S_{t+1} \neq j, S_t = j \mid X_0^t = x_0^t)$ for each state j forward from time 0 to time $\tau - 1$ and can be presented as follows.

Start: The start of the loop at $t = 0$ can be simplified to

$$\begin{aligned} F_j(0) &= P(S_1 \neq j, S_0 = j \mid X_0 = x_0) \\ &= \pi_j d_j(1). \end{aligned}$$

Iteration:

$$\begin{aligned} F_j(t) &= P\left(S_{t+1} \neq j, S_t = j \mid X_0^t = x_0^t\right) \\ &= \frac{b_j(x_t)}{N_t}\left[\sum_{u=1}^{t}\left\{\prod_{v=1}^{u-1}\frac{b_j(x_{t-v})}{N_{t-v}}\right\} d_j(u) \sum_{i \neq j} p_{ij} F_i(t-u)\right. \\ &\qquad \left. + \left\{\prod_{v=1}^{t}\frac{b_j(x_{t-v})}{N_{t-v}}\right\} d_j(t+1)\,\pi_j\right] \end{aligned} \tag{5.21}$$

for all $t \in \{0, \ldots, \tau - 2\}$ and $j \in \{0, \ldots, J-1\}$. (For details of the derivations, see Appendix B). The quantity

$$N_t := P\left(X_t = x_t \mid X_0^{t-1} = x_0^{t-1}\right)$$

is the so-called **normalizing factor**.

The key difference of the algorithm presented by Guédon (2003) and earlier algorithms lies in the treatment of the sojourn time in the last visited state at $t = \tau - 1$ which is subject to a censoring with the survivor function. Using arguments similar to those for the derivation of (5.21), the last step of the iteration can be written as

$$\begin{aligned} F_j(\tau-1) &= P\left(S_{\tau-1} = j \mid X_0^{\tau-1} = x_0^{\tau-1}\right) \\ &= \frac{b_j(x_{\tau-1})}{N_{\tau-1}} \left[\sum_{u=1}^{\tau-1} \left\{ \prod_{v=1}^{u-1} \frac{b_j(x_{\tau-1-v})}{N_{\tau-1-v}} \right\} D_j(u) \sum_{i \neq j} p_{ij} F_i(\tau-1-u) \right. \\ &\qquad \left. + \left\{ \prod_{v=1}^{\tau-1} \frac{b_j(x_{\tau-1-v})}{N_{\tau-1-v}} \right\} D_j(\tau)\, \pi_j \right] \end{aligned} \tag{5.22}$$

for $t = \tau - 1$, $j \in \{0, \ldots, J-1\}$.
The exact time spent in this last state is unknown; however, the minimum time is known. Thus the probability mass functions $d_j(u)$ of the sojourn times in state j of the general forward iteration formula (5.21) is replaced by the corresponding survivor functions $D_j(u)$ in (5.22).

Remark: The normalizing factor N_t is directly obtained during the forward recursion. It can be derived as

$$\begin{aligned} N_t &= P\left(X_t = x_t | X_0^{t-1} = x_0^{t-1}\right) \\ &= \sum_j P\left(S_t = j, X_t = x_t | X_0^{t-1} = x_0^{t-1}\right) \\ &= \sum_j b_j(x_t) \left[\sum_{u=1}^{t} \left\{ \prod_{v=1}^{u-1} \frac{b_j(x_{t-v})}{N_{t-v}} \right\} D_j(u) \sum_{i \neq j} p_{ij} F_i(t-u) \right. \\ &\qquad \left. + \left\{ \prod_{v=1}^{t} \frac{b_j(x_{t-v})}{N_{t-v}} \right\} D_j(t+1)\, \pi_j \right] \end{aligned} \tag{5.23}$$

for $t \in \{0, \ldots, \tau-1\}$. Note that N_t may also serve for forecasting procedures. Setting $t = \tau$, the forecast distribution of $X_\tau = x$, conditioned on the entire sequence of observations, follows directly from the definition of N_t.

5.3.2.2 The Backward Iteration

The backward iteration performs the computation of the smoothing probabilities $L_j(t) = P\left(S_t = j \mid X_0^{\tau-1} = x_0^{\tau-1}\right)$ for each state j, backward from time $\tau - 1$ to time 0.

Start: The backward iteration starts at $t = \tau - 1$ with

$$L_j(\tau - 1) = P\left(S_{\tau-1} = j \mid X_0^{\tau-1} = x_0^{\tau-1}\right) = F_j(\tau - 1)$$

for $j \in \{0, \ldots, J-1\}$.

Iteration: The key point in this step lies in rewriting the quantity $L_j(t)$ as a sum of three terms.

$$\begin{aligned} L_j(t) &= P\left(S_t = j \mid X_0^{\tau-1} = x_0^{\tau-1}\right) \\ &= P\left(S_{t+1} \neq j, S_t = j \mid X_0^{\tau-1} = x_0^{\tau-1}\right) + P\left(S_{t+1} = j \mid X_0^{\tau-1} = x_0^{\tau-1}\right) \\ &\quad - P\left(S_{t+1} = j, S_t \neq j \mid X_0^{\tau-1} = x_0^{\tau-1}\right) \\ &= L1_j(t) + L_j(t+1) - P\left(S_{t+1} = j, S_t \neq j \mid X_0^{\tau-1} = x_0^{\tau-1}\right). \end{aligned} \tag{5.24}$$

The second term $L_j(t+1)$ is obtained directly from the previous iteration step. The first term $L1_j(t)$ and the third term, $P\left(S_{t+1} = j, S_t \neq j \mid X_0^{\tau-1} = x_0^{\tau-1}\right)$, which represents the entrance into state j, require a bit more attention.

The backward iteration is based on $L1_j(t)$. (For details of the derivation, see Appendix B). For $t \in \{0, \ldots, \tau - 2\}$ and $j \in \{0, \ldots, J-1\}$ it can be transformed to

$$\begin{aligned} L1_j(t) = \Bigg[\sum_{k \neq j} \Bigg[\sum_{u=1}^{\tau-2-t} \frac{L1_k(t+u)}{F_k(t+u)} \left\{ \prod_{v=0}^{u-1} \frac{b_k(x_{t+u-v})}{N_{t+u-v}} \right\} d_k(u) \\ + \left\{ \prod_{v=0}^{\tau-2-t} \frac{b_k(x_{\tau-1-v})}{N_{\tau-1-v}} \right\} D_k(\tau - 1 - t) \Bigg] p_{jk} \Bigg] F_j(t). \end{aligned} \tag{5.25}$$

The third term in (5.24) can also be transformed. (The details of the derivation are also given in Appendix B.) For $t \in \{0, \ldots, \tau - 2\}$ and $j \in \{0, \ldots, J-1\}$,

$$\begin{aligned} & P\left(S_{t+1} = j, S_t \neq j \,|\, X_0^{\tau-1} = x_0^{\tau-1}\right) \\ & = \left[\sum_{u=1}^{\tau-2-t} \frac{L1_j\,(t+u)}{F_j\,(t+u)} \left\{ \prod_{v=0}^{u-1} \frac{b_j\,(x_{t+u-v})}{N_{t+u-v}} \right\} d_j\,(u) \right. \\ & \quad \left. + \left\{ \prod_{v=0}^{\tau-2-t} \frac{b_j\,(x_{\tau-1-v})}{N_{\tau-1-v}} \right\} D_j\,(\tau - 1 - t) \right] \sum_{i \neq j} p_{ij} F_i\,(t)\,. \end{aligned} \tag{5.26}$$

At first glance, the backward iteration for the $L_j\,(t)$ appears complicated. However, the calculation of $L1_j\,(t) = P(S_{t+1} \neq j, S_t = j \,|\, X_0^{\tau-1} = x_0^{\tau-1})$ in (5.25) as well as $P(S_{t+1} = j, S_t \neq j \,|\, X_0^{\tau-1} = x_0^{\tau-1})$ in (5.26) may easily be performed by introducing the auxiliary quantities

$$G_j\,(t+1, u) := \frac{L1_j\,(t+u)}{F_j\,(t+u)} \left\{ \prod_{v=0}^{u-1} \frac{b_j\,(x_{t+u-v})}{N_{t+u-v}} \right\} d_j\,(u)\,, \quad u \in \{1, \ldots, \tau - 2 - t\},$$

$$G_j\,(t+1, \tau - 1 - t) := \left\{ \prod_{v=0}^{\tau-2-t} \frac{b_j\,(x_{\tau-1-v})}{N_{\tau-1-v}} \right\} D_j\,(\tau - 1 - t)$$

and

$$\begin{aligned} G_j\,(t+1) &:= \frac{P\left(X_{t+1}^{\tau-1} = x_{t+1}^{\tau-1} \,|\, S_{t+1} = j, S_t \neq j\right)}{P\left(X_{t+1}^{\tau-1} = x_{t+1}^{\tau-1} \,|\, X_0^t = x_0^t\right)} \\ &= \sum_{u=1}^{\tau-1-t} G_j\,(t+1, u)\,. \end{aligned}$$

These auxiliary quantities allow for a simplification of the backward iteration which is performed in two steps:

1. At each time t, $G_j\,(t+1, u)$, $G_j\,(t+1, \tau-1-t)$ and $G_j\,(t+1)$ are precomputed.

2. $L1_j\,(t)$ and $P\left(S_{t+1} = j, S_t \neq j \,|\, X_0^{\tau-1} = x_0^{\tau-1}\right)$ can be transformed to

$$L1_j(t) = \left\{ \sum_{k \neq j} G_k(t+1)\, p_{jk} \right\} F_j(t), \tag{5.27}$$

and

$$\begin{aligned} & P\left(S_{t+1} = j, S_t \neq j \,|\, X_0^{\tau-1} = x_0^{\tau-1}\right) \\ & = \frac{P\left(X_{t+1}^{\tau-1} = x_{t+1}^{\tau-1} \,|\, S_{t+1} = j, S_t \neq j\right)}{P\left(X_{t+1}^{\tau-1} = x_{t+1}^{\tau-1} \,|\, X_0^t = x_0^t\right)} P\left(S_{t+1} = j, S_t \neq j \,|\, X_0^t = x_0^t\right) \\ & = G_j(t+1) \sum_{i \neq j} p_{ij} F_i(t). \end{aligned}$$

Thus the quantities involved in the backward iteration can be calculated as sums and products of the auxiliary variables and the forward probabilities.

Remark: Guédon (2003) also presented a variant of the backward iteration presented above and which is built on $B_j(t)$. It works as follows. For each $t < \tau - 1$

$$L1_j(t) = B_j(t)\, F_j(t)$$

holds true. Hence the backward recursion, based on $B_j(t)$, is directly deduced from (5.25) as

$$\begin{aligned} B_j(t) = \sum_{k \neq j} \Bigg[& \sum_{u=1}^{\tau-2-t} B_k(t+u) \left\{ \prod_{v=0}^{u-1} \frac{b_k(x_{t+u-v})}{N_{t+u-v}} \right\} d_k(u) \\ & + \left\{ \prod_{v=0}^{\tau-2-t} \frac{b_k(x_{\tau-1-v})}{N_{\tau-1-v}} \right\} D_k(\tau-1-t) \Bigg] p_{jk}. \end{aligned}$$

The third term in (5.24) can be transformed to

$$
\begin{aligned}
&P\left(S_{t+1}=j, S_t \neq j \,|\, X_0^{\tau-1}=x_0^{\tau-1}\right) \\
&\quad= \left[\sum_{u=1}^{\tau-2-t} B_j\left(t+u\right)\left\{\prod_{v=0}^{u-1}\frac{b_j\left(x_{t+u-v}\right)}{N_{t+u-v}}\right\} d_j\left(u\right)\right. \\
&\qquad\left.+\left\{\prod_{v=0}^{\tau-2-t}\frac{b_j\left(x_{\tau-1-v}\right)}{N_{\tau-1-v}}\right\} D_j\left(\tau-1-t\right)\right]\sum_{i\neq j} p_{ij}F_i\left(t\right).
\end{aligned}
$$

This alternative way is presented only for the sake of completeness, but has not been used for our implementation of the estimation procedures.

The implementation of this forward-backward algorithm is proposed by Guédon (2003) in pseudo-code form which served as basis for this thesis.

5.3.3 The Sojourn Time Distribution

The aim of this section is to show how $Q(\{d_j(u)\})$ – the part of the Q-function dealing with the sojourn time given in equation (5.17) – can be calculated using the quantities derived in Section 5.3.2.
As long as we deal with non-stationary HSMMs, this is the only one part of the estimation procedure which is affected by the use of either the partial likelihood estimator or the complete likelihood estimator from the Sections 5.2.1 and 5.2.2.

5.3.3.1 The Q-Function based on the Full Likelihood Estimator

Recall from equation (5.17) that the the state occupancy distribution for each state j is given by

$$\begin{aligned} & Q_d\left(\{d_j(u)\} \mid \theta^{(k)}\right) \\ &= \sum_u \log d_j(u) \left\{ \sum_{t=0}^{\tau-2} P\left(S_{t+u+1} \neq j, S_{t+u-v} = j, v = 0, \ldots, u-1,\right.\right. \\ &\quad S_t \neq j \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\big) \\ &\quad \left.\left. + P\left(S_u \neq j, S_{u-v} = j, v = 1, \ldots, u \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right)\right\} \qquad (5.28) \\ &= \sum_u \eta_{ju}^{(k)} \log d_j(u) \qquad (5.29) \end{aligned}$$

with

$$\begin{aligned} \eta_{ju}^{(k)} &:= \sum_{t=0}^{\tau-2} P\left(S_{t+u+1} \neq j, S_{t+u-v} = j, v = 0, \ldots, u-1,\right. \\ &\quad S_t \neq j | X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\big) \\ &\quad \left. + P\left(S_u \neq j, S_{u-v} = j, v = 1, \ldots, u \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right)\right. . \end{aligned}$$

The computation of the two terms involved in (5.28) can be performed utilizing the quantities derived for the forward-backward algorithm. We start with the first term, and consider the following two possible cases.

$u \leq \tau - 2 - t$:

$$\begin{aligned} & P\left(S_{t+u+1} \neq j, S_{t+u-v} = j, v = 0, \ldots, u-1, S_t \neq j \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right) \\ &= G_j(t+1, u) \sum_{i \neq j} p_{ij} F_i(t), \end{aligned}$$

which can directly be extracted from the computation of $L_j(t)$.

$u > \tau - 2 - t$:

$$P\left(S_{t+u+1} \neq j, S_{t+u-v} = j, v = 0, \ldots, u-1, S_t \neq j \,|\, X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right)$$

$$= \left\{ \prod_{v=0}^{\tau-2-t} \frac{b_j\left(x_{\tau-1-v}\right)}{N_{\tau-1-v}} \right\} d_j\left(u\right) \sum_{i \neq j} p_{ij} F_i\left(t\right).$$

The computation of the terms arising for the latter case $u > \tau - 2 - t$ can be combined with the computation of

$$P\left(S_{\tau-1-v} = j, v = 0, \ldots, \tau-2-t, S_t \neq j \,|\, X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right)$$

$$= \left\{ \prod_{v=0}^{\tau-2-t} \frac{b_j\left(x_{\tau-1-v}\right)}{N_{\tau-1-v}} \right\} D_j\left(\tau-1-t\right) \sum_{i \neq j} p_{ij} F_i\left(t\right),$$

which is shown in Appendix B.

The second term in (5.28) corresponding to the time spent in the initial state from $t = 0$, one requires some supplementary computation and involves the already known quantities from the forward-backward algorithm. Again, we consider two separate cases.

$u \leq \tau - 1$:

$$P\left(S_u \neq j, S_{u-v} = j, v = 1, \ldots, u \,|\, X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right)$$

$$= \frac{L1_j\left(u-1\right)}{F_j\left(u-1\right)} \left\{ \prod_{v=1}^{u} \frac{b_j\left(x_{u-v}\right)}{N_{u-v}} \right\} d_j\left(u\right) \pi_j.$$

$u > \tau - 1$:

$$P\left(S_u \neq j, S_{u-v} = j, v = 1, \ldots, u \,|\, X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right)$$

$$= \left\{ \prod_{v=1}^{\tau} \frac{b_j\left(x_{\tau-v}\right)}{N_{\tau-v}} \right\} d_j\left(u\right) \pi_j.$$

The last point of this section shows how Equation (5.28) can be reformulated in such a way that the difference between the partial likelihood estimator and the full likelihood estimator can be directly deduced from the quantities involved in the calculation of the $\eta_{ju}^{(k)}$.

Equation (5.28) yields

$$\begin{aligned}
&\sum_u \Bigg\{ \sum_{t=0}^{\tau-2} P\left(S_{t+u+1} \neq j, S_{t+u-v} = j, v = 0, \ldots, u-1,\right. \\
&\qquad \left. S_t \neq j \,|\, X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right) \\
&\qquad + P\left(S_u \neq j, S_{u-v} = j, v = 1, \ldots, u \,|\, X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right) \Bigg\} \\
&= \sum_{t=0}^{\tau-2} P\left(S_{t+1} \neq j, S_t = j \,|\, X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right) \\
&\qquad + P\left(S_{\tau-1} = j \,|\, X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right) \\
&= \sum_{t=0}^{\tau-2} L1_j(t) + L_j(\tau-1).
\end{aligned}$$

Considering the above equation along with Equation (5.29), we obtain:

$$\begin{aligned}
Q_d\left(\{d_j(u)\} \,|\theta^{(k)}\right) &= \sum_u \eta_{ju}^{(k)} \log d_j(u), \\
\text{where} \quad \sum_u \eta_{ju}^{(k)} &= \sum_{t=0}^{\tau-2} L1_j(t) + L_j(\tau-1). \qquad (5.30)
\end{aligned}$$

5.3.3.2 The Q-Function based on the Partial Likelihood Estimator

As shown in the Sections 5.2.1 and 5.2.2, the difference between the complete likelihood estimator and the partial likelihood estimator lies in the re-estimation of the state occupancy distributions, where the latter ignores the contribution of the last visited state. That is, the information associated with the time spent in the last visited state is not used in the estimation procedure. This yields a new version of the Q-function given in Equation (5.17) which is similar to Equation (5.30). We denote the Q-function calculated from the partial likelihood estimator by $\widetilde{Q}_d(\{d_j(u)\} \,|\, \theta^{(k)})$.

The new term of the occupancy distribution for the state j involves a censoring term $\mathbf{1}_{\{u\leq\tau-1\}}(u)$. It is given by

$$\begin{aligned}
&\widetilde{Q}_d\left(\{d_j(u)\}\mid\theta^{(k)}\right)\\
&= \sum_u \log d_j(u)\Bigg\{\sum_{t=0}^{\tau-2-u} P\left(S_{t+u+1}\neq j, S_{t+u-v}=j, v=0,\ldots,u-1,\right.\\
&\qquad S_t\neq j\mid X_0^{\tau-1}=x_0^{\tau-1},\theta^{(k)}\big)\\
&\qquad \left.+P\left(S_u\neq j, S_{u-v}=j, v=1,\ldots,u\mid X_0^{\tau-1}=x_0^{\tau-1},\theta^{(k)}\right)\mathbf{1}_{\{u\leq\tau-1\}}(u)\right\}\\
&= \sum_u \widetilde{\eta}_{j,u}^{(k)}\log d_j(u),
\end{aligned}$$

where

$$\sum_u \widetilde{\eta}_{j,u}^{(k)} = \sum_{t=0}^{\tau-2} L1_j(t).$$

Note that the last visited state represented by $L_j(\tau-1)$ does not contribute to $\widetilde{Q}_d\left(\{d_j(u)\}\mid\theta^{(k)}\right)$. The computation of the $\widetilde{\eta}_{j,u}^{(k)}$ is similar to the computation of the $\eta_{j,u}^{(k)}$ except for the indicator which has to be inserted additionally.

5.3.4 Parameter Re-estimation

The second part of the EM algorithm consists of a re-estimation procedure, the M-step. This step determines the likelihood-increasing next set of parameters $\theta^{(k+1)}$ by

$$\theta^{(k+1)} = \arg\max_\theta Q(\theta\mid\theta^{(k)}).$$

In Equation (5.14) we showed that the Q-function $Q(\theta\mid\theta^{(k)})$ of a HSMM can be decomposed into four different terms, each depending on a given subset of θ. Hence, the re-estimation formulae for the parameters can be derived by maximizing each of the different terms separately.

In this section we derive the re-estimation formula for each parameter subset by maximizing the terms (5.15), (5.16), (5.17) and (5.18). We impose various distributional assumptions on the dwell time distributions (5.17) as well as the conditional observation distributions (5.18) to cover a wide area of applications. Some of the re-estimation formulae can be found in the literature; however,

our intention is to provide the reader with an overview of the existing material and introduce some extensions.

5.3.4.1 The Initial Parameters

We start with the parameters involved in the underlying hidden semi-Markov chain. In Equation (5.15), the term of $Q(\theta|\theta^{(k)})$ corresponding to the initial parameters is given by

$$Q_\pi\left(\{\pi_j\}_{j=0}^{J-1} \mid \theta^{(k)}\right) = \sum_j P\left(S_0 = j \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right) \log \pi_j.$$

Adding a Lagrange multiplier with the constraint $\sum_{m=0}^{J-1} \pi_m = 1$, differentiating w.r.t. π_j, and summing over all states $j \in \{0, \ldots, J-1\}$, we get

$$\begin{aligned}
& \frac{\partial}{\partial \pi_j}\left[\sum_{l=1}^{J-1} P(S_0 = l \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}) \log \pi_l - K(\sum_{m=0}^{J-1} \pi_m - 1)\right] = 0 \\
\Rightarrow \quad & \frac{1}{\pi_j} P(S_0 = j \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}) - K = 0 \\
\Rightarrow \quad & \sum_{j=0}^{J-1} P(S_0 = j \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}) = K \sum_{j=0}^{J-1} \pi_j \\
\Rightarrow \quad & K = 1 \\
\Rightarrow \quad & \pi_j = P(S_0 = j \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}).
\end{aligned}$$

The re-estimation formula for the initial parameters is given by

$$\pi_j^{(k+1)} = P\left(S_0 = j \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right) = L_j(0). \tag{5.31}$$

5.3.4.2 The Transition Probabilities

The second component of the semi-Markov chain comprises the transition probabilities of the embedded Markov chain. The corresponding term of $Q(\theta \mid \theta^{(k)})$ from Equation (5.15) is

$$Q_p\left(\{p_{ij}\}_{j=0}^{J-1} \mid \theta^{(k)}\right) = \sum_{j \neq i} \sum_{t=0}^{\tau-2} P\left(S_{t+1} = j, S_t = i \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right) \log p_{ij}.$$

Adding a Lagrange multiplier with the the constraint $\sum_{m=0}^{J-1} p_{lm} = 1$, differentiating w.r.t. p_{ij} and summing over all states $j \in \{0, \ldots, J-1\}$, we get

$$\frac{\partial}{\partial p_{ij}} \Bigg[\sum_{l=0}^{J-1} \sum_{m \neq l} \sum_{t=0}^{\tau-2} P(S_{t+1} = m, S_t = l \,|\, X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}) \log p_{lm} - K \Big(\sum_{m=0}^{J-1} p_{lm} - 1 \Big) \Bigg] = 0$$

$$\Rightarrow \quad \sum_{t=0}^{\tau-2} \frac{1}{p_{ij}} P(S_{t+1} = j, S_t = i \,|\, X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}) - K = 0$$

$$\Rightarrow \quad \sum_{t=0}^{\tau-2} \underbrace{\sum_{j=0, j \neq i}^{J-1} P(S_{t+1} = j, S_t = i \,|\, X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)})}_{= P(S_{t+1} \neq i, S_t = i \,|\, X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)})} = K \underbrace{\sum_{j=0, j \neq i}^{J-1} p_{ij}}_{=1}$$

$$\Rightarrow \quad K = \sum_{t=0}^{\tau-2} P(S_{t+1} \neq i, S_t = i \,|\, X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)})$$

$$\Rightarrow \quad p_{ij} = \frac{\sum_{t=0}^{\tau-2} P(S_{t+1} = j, S_t = i \,|\, X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)})}{\sum_{t=0}^{\tau-2} P(S_{t+1} \neq i, S_t = i \,|\, X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)})}.$$

The re-estimation formula can be written as

$$\begin{aligned} p_{ij}^{(k+1)} &= \frac{\sum_{t=0}^{\tau-2} P\left(S_{t+1} = j, S_t = i \,|\, X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right)}{\sum_{t=0}^{\tau-2} P\left(S_{t+1} \neq i, S_t = i \,|\, X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right)} \\ &= \frac{\sum_{t=0}^{\tau-2} G_j(t+1)\, p_{ij} F_i(t)}{\sum_{t=0}^{\tau-2} L1_i(t)}. \end{aligned} \tag{5.32}$$

Note that the quantity in the numerator of above Equation (5.32) does not require additional calculations because it can be extracted directly from the computation of $L1_i(t)$, cf. Equation (5.27).

5.3.4.3 The State Occupancy Distribution

The third term of $Q(\theta \,|\, \theta^{(k)})$ corresponds to the sojourn time distributions of the semi-Markov chain. Explicit solutions can be calculated for some distributions. However, this part of the maximization may also involve numerical

maximization methods if a closed solution cannot be derived.

Non-Parametric State Occupancy Distribution with the Complete Likelihood Estimator

The term of the Q-function given in Equation (5.17) and (5.30) treating the non-parametric state occupancy distribution is given by

$$Q_d\left(\{d_j(u)\} \mid \theta^{(k)}\right) = \sum_u \eta_{ju}^{(k)} \log d_j(u).$$

It has to be maximized under the constraint $\sum_u d_j(u) = 1$, which leads to

$$\begin{aligned}
&\frac{\partial}{\partial d_j(u)} \left[Q_d\left(\{d_j(u)\} \mid \theta^{(k)}\right) - K \left(\sum_{v=1}^{M_j} d_j(v) \right) \right] = 0 \\
\Rightarrow \quad &\frac{\eta_{ju}^{(k)}}{d_j(u)} - K = 0 \quad \Rightarrow \quad K = \sum_{v=1}^{M_j} \eta_{jv}^{(k)} \\
\Rightarrow \quad &d_j(u) = \frac{\eta_{ju}^{(k)}}{\sum_{v=1}^{M_j} \eta_{jv}^{(k)}}.
\end{aligned}$$

Hence, the re-estimation formula for the state occupancy probabilities can be written as

$$\begin{aligned}
d_j^{(k+1)}(u) &= \frac{\eta_{ju}^{(k)}}{\sum_{v=1}^{M_j} \eta_{jv}^{(k)}} \\
&= \frac{\eta_{ju}^{(k)}}{\sum_{t=0}^{\tau-2} L1_j(t) + L_j(\tau - 1)}.
\end{aligned} \tag{5.33}$$

Non-Parametric State Occupancy Distribution with the Partial Likelihood Estimator

According to the Sections 5.3.3.1 and 5.3.3.2 the difference to the complete likelihood estimator lies in a modified Q-function given by

$$\widetilde{Q}_d\left(\{d_j(u)\}\,|\theta^{(k)}\right) = \sum_u \widetilde{\eta}_{ju}^{(k)} \log d_j(u)$$

with

$$\sum_u \widetilde{\eta}_{ju}^{(k)} = \sum_{t=0}^{\tau-2} L1_j(t).$$

Hence the maximization is performed analogously to the previous case and yields

$$\begin{aligned} d_j^{(k+1)}(u) &= \frac{\tilde{\eta}_{ju}^{(k)}}{\sum_v \tilde{\eta}_{jv}^{(k)}} \\ &= \frac{\tilde{\eta}_{ju}^{(k)}}{\sum_{t=0}^{\tau-2} L1_j(t)}. \end{aligned} \tag{5.34}$$

Remark: The partial likelihood estimator leads to a downward bias of the sojourn times. This can be explained by the fact that longer state sequences are more likely to include the last visited state. Hence, they are more often ignored in the re-estimation procedure.

Geometric State Occupancy Distribution

The geometric state occupancy distribution reduces the HSMM to an ordinary HMM. However, in this case the distribution function of the sojourn times is given by

$$d_j(u) = (1-p_j)^{u-1} p_j$$

with $j \in \{0,\ldots,J-1\}$ and $u \in \{1,\ldots,M_j = \tau-1\}$. Then, for each j, the corresponding part of the Q-function becomes:

$$Q_d\left(\{d_j(u)\}\,|\,\theta^{(k)}\right) = \sum_{u=1}^{\tau-1} \eta_{ju}^{(k)} \Big[(u-1)\log(1-p_j) + \log p_j\Big].$$

For the maximization w.r.t. p_j follows

$$\frac{\partial}{\partial p_j} Q_d\left(\{d_j(u)\} \,|\, \theta^{(k)}\right) = 0 \quad \Rightarrow \quad \sum_{u=1}^{\tau-1} \eta_{ju}^{(k)} \left(-\frac{(u-1)}{(1-p_j)} + \frac{1}{p_j} \right) = 0$$

$$\Rightarrow \quad \sum_{u=1}^{\tau-1} \eta_{ju}^{(k)} \Big(-up_j + 1 \Big) = 0 \quad \Rightarrow \quad p_j = \frac{\sum_{u=1}^{\tau-1} \eta_{ju}^{(k)}}{\sum_{u=1}^{\tau-1} u\eta_{ju}^{(k)}}.$$

The re-estimation formula for the parameter p_j is thus given by

$$p_j^{(k+1)} = \frac{\sum_{u=1}^{\tau-1} \eta_{ju}^{(k)}}{\sum_{u=1}^{\tau-1} u\eta_{ju}^{(k)}}. \tag{5.35}$$

Note that the re-estimation formula (5.35) has been derived for the case of the full likelihood estimator. If the partial likelihood estimator is required, the quantities $\eta_{ju}^{(k)}$ have to be replaced by $\widetilde{\eta}_{ju}^{(k)}$ and the re-estimation formula remains the same. This argument applies to other runlength distributions in the sections to follow.

Remark: Depending on the parameters of the geometric distribution, the results of the estimation may not change, or change only a little, when reducing the upper bound of the dwell time M_j from $\tau - 1$ to a smaller value. This can be explained by the fact that $d_j(u) \to 0$ for $u \to \infty$. Depending on the rate of convergence, the $d_j(u)$ may become practically zero for all u greater than some u_0 and may therefore be ignored.
A smaller M_j reduces the computational time but it may also affect the estimates for other re-estimation quantities. In particular, the state-dependent observation probabilities for the t distribution turn out to react very sensitively to this acceleration method. Thus the procedure has to be applied with care. On the other hand, the Normal and Poisson distributions showed a high robustness and provided reasonable estimation results.

Negative Binomial State Occupancy Distribution

The negative binomial distribution is an extension of the geometric distribution. A minor problem that should be mentioned is that there are several different parameterizations for the negative binomial distribution and none of them can be considered as a standard. We choose the distribution function of the sojourn times be given by

$$\begin{aligned} d_j(u) &= \binom{u-2+r}{u-1} \pi^r (1-\pi)^{u-1} \\ &= \frac{\Gamma(u-1+r)}{\Gamma(r)\Gamma(u)} \pi^r (1-\pi)^{u-1}, \end{aligned}$$

where $\Gamma()$ denotes the Gamma-function; $r > 0$ and $\pi \in (0,1)$ are the parameters of the distribution. Note that it is convenient to rewrite the ratio of the Gamma-functions as

$$\exp\big[\log\Gamma(u-1+r) - \log\Gamma(r) - \log\Gamma(u)\big]$$

to increase numerical stability. Then, for each j the corresponding part of the Q-function becomes

$$\begin{aligned} Q_d\left(\{d_j(u)\} \,|\, \theta^{(k)}\right) &= \sum_{u=1}^{\tau-1} \eta_{ju}^{(k)} \Big[\log\Gamma(u-1+r) - \log\Gamma(r) - \log\Gamma(u) \\ &\qquad + r\log\pi + (u-1)\log(1-\pi)\Big]. \end{aligned}$$

The maximization w.r.t. to the parameters r and π is not straightforward; numerical methods have to be applied. Differentiating w.r.t. π yields

$$\begin{aligned} & \frac{\partial}{\partial\pi} Q_d\left(\{d_j(u)\} \,|\, \theta^{(k)}\right) = 0 \quad \Rightarrow \quad \sum_{u=1}^{\tau-1} \eta_{ju}^{(k)} \left(\frac{r}{\pi} - \frac{u-1}{1-\pi}\right) = 0 \\ \Rightarrow \quad & \sum_{u=1}^{\tau-1} \eta_{ju}^{(k)} \Big(r - \pi(r+u-1)\Big) = 0 \\ \Rightarrow \quad & \pi = \frac{r \sum_{u=1}^{\tau-1} \eta_{ju}^{(k)}}{\sum_{u=1}^{\tau-1} \eta_{ju}^{(k)} (r+u-1)}. \end{aligned} \tag{5.36}$$

The differentiation w.r.t. r involves terms of the form $\log\Gamma(s)$. Recall that the Digamma function is defined as $\psi(s) := \frac{\partial \log\Gamma(s)}{\partial s}$. Thus,

$$\frac{\partial}{\partial r} Q_d\left(\{d_j(u)\} \,|\, \theta^{(k)}\right) = 0 \quad \Rightarrow \quad \sum_{u=1}^{\tau-1} \eta_{ju}^{(k)} \Big(\psi(u-1+r) - \psi(r) + \log\pi\Big) = 0. \tag{5.37}$$

Substituting π from Equation (5.36) in Equation (5.37) yields the expression

$$\sum_{u=1}^{\tau-1} \eta_{ju}^{(k)} \left(\psi(u-1+r) - \psi(r) + \log \left[\frac{r \sum_{v=1}^{\tau-1} \eta_{jv}^{(k)}}{\sum_{v=1}^{\tau-1} \eta_{jv}^{(k)} (r+v-1)} \right] \right) = 0,$$

which has to be solved numerically, e.g. by a bisectioning algorithm. The estimation of π follows directly from Equation (5.36).

Remark: In some cases the implementation of root-finding algorithms may not be desirable. As an alternative the M-step can be modified in terms of the One-Step-Late algorithm of Green (1990). Instead of solving a system of equations, simply insert $r^{(k)}$ and $\pi^{(k)}$ into Equation (5.36) and (5.37), respectively. This technique may slow down the rate of convergence but, in general, the absence of root-finding algorithms compensates (somewhat) for the loss of speed. Moreover the calculation of analytic solutions for possibly large systems of equations can be avoided.

Mixtures of Distributions

It is sometimes advantageous to use mixtures of distributions as state occupancy distributions. A fully non-parametric sojourn time distribution may involve too many degrees of freedom, while the parametric alternative does not lead to the desired fit of the distribution. In such a situation a possible alternative is the combination of a non-parametric "head" with a parametric "tail" distribution.

Formally, the state occupancy distribution is defined piecewise on a set of disjoint intervals by

$$d_j(u) = \phi_k d_{jk}(u, \theta_k),$$

for $u \in \{D_1, \ldots, D_m\}$, and

$$D_k := \{d_{k-1}, d_{k-1}+1, d_{k-1}+2, \ldots, d_k - 1\}$$

with $0 = d_0 < d_1 < \ldots < d_{m-1} < \tau - 1$, $d_m = \infty$, $0 \leq \phi_k \leq 1$ and $\sum_k \phi_k = 1$. The quantities $d_{jk}(u, \theta_k)$ themselves are sojourn time distributions and, for each state j, the overall distribution is given by their weighted sum.

The case of an arbitrary number of mixture components is treated by Sansom & Thomson (2000). We present the case that the mixture consists of only two components because this setup suffices for many applications. However, the generalization is straightforward.

The basic idea of a two component dwell time distribution is that the first $D-1$ state occupancies follow the head distribution, given by $d_{j1}(u, \theta_1)$; the state occupancies from D to at most $\tau - 1$ follow the tail distribution $d_{j2}(u, \theta_2)$. The re-estimation formulae cannot be solved in a general framework, e.g., a mixture of two geometrical distributions already yields a system of equations that can only be solved numerically (see Sansom & Thomson 2000). We analyze a distribution with non-parametric head and a geometric tail. This mixture of distributions is relatively flexible but still has a comparatively small number of parameters. It can be written as

$$\begin{aligned} d_{j1}(u, \theta_1) &= d_{j1}(u) && \text{for } u \in \{1, \ldots, D-1\} \\ d_{j2}(u, \theta_2) &= (1-p_j)^{u-D} p_j && \text{for } u \in \{D, D+1, \ldots\} \end{aligned},$$

where, for the ease of notation, $d_{j1}(u)$ denotes the non-parametric state occupancy component. The entire state occupancy distribution can then be written as

$$d_j(u) = \begin{cases} \phi d_{j1}(u), & u \in \{1, \ldots, D-1\} \\ (1-\phi)(1-p_j)^{u-D} p_j, & u \in \{D, D+1, \ldots\} \end{cases},$$

which can also be expressed by

$$d_j(u) = \phi d_{j1}(u) \mathbf{1}_{\{u<D\}}(u) + (1-\phi)(1-p_j)^{u-D} p_j \mathbf{1}_{\{D \leq u\}}(u), \tag{5.38}$$

with $\mathbf{1}_{\{\cdot\}}(\cdot)$ denoting the indicator function.

The re-estimation formulae can be deduced by maximizing the corresponding part of the Q-function given in equation (5.17) which becomes

$$\begin{aligned} Q_d\left(\{d_j(u)\} \mid \theta^{(k)}\right) &= \sum_u \eta_{ju}^{(k)} \log d_j(u) \\ &= \sum_{k=1}^{2} \sum_{u \in D_k} \eta_{ju}^{(k)} \log(\phi_k d_{jk}(u)) \\ &= \sum_{u=1}^{D-1} \eta_{ju}^{(k)} \big[\log \phi + \log(d_{j1}(u)) \big] \\ &\quad + \sum_{u=D}^{\tau-1} \eta_{ju}^{(k)} \big[\log(1-\phi) + (u-D)\log(1-p_j) + \log p_j \big), \end{aligned}$$

for each $j \in \{0, \ldots, J-1\}$. To obtain the re-estimation formulae, we have to maximize the function above w.r.t. each of the parameters involved in Equation (5.38):

$$\begin{aligned} \frac{\partial}{\partial \phi} Q_d\left(\{d_j(u)\} \mid \theta^{(k)}\right) = 0 \quad &\Rightarrow \quad \frac{1}{\phi} \sum_{u=1}^{D-1} \eta_{ju}^{(k)} = \frac{1}{1-\phi} \sum_{u=D}^{\tau-1} \eta_{ju}^{(k)} \\ &\Rightarrow \quad \phi \left[\sum_{u=1}^{D-1} \eta_{ju}^{(k)} + \sum_{u=D}^{\tau-1} \eta_{ju}^{(k)} \right] = \sum_{u=1}^{D-1} \eta_{ju}^{(k)} \\ &\Rightarrow \quad \phi = \frac{\sum_{u=1}^{D-1} \eta_{ju}^{(k)}}{\sum_{u=1}^{\tau-1} \eta_{ju}^{(k)}}. \end{aligned}$$

Analogous to the case of a fully non-parametric state occupancy distribution, the maximization w.r.t. $d_{j1}(u)$ is performed under the constraint $\sum_{v=1}^{D-1} d_{j1}(v) = 1$. That is,

$$\begin{aligned} & \frac{\partial}{\partial d_{j1}(u)} \left[Q_d\left(\{d_j(u)\} \mid \theta^{(k)}\right) - K \left(\sum_{v=1}^{D-1} d_{j1}(v) \right) \right] = 0 \\ \Rightarrow \quad & \frac{\eta_{ju}^{(k)}}{d_{j1}(u)} - K = 0 \quad \Rightarrow \quad K = \sum_{v=1}^{D-1} \eta_{jv}^{(k)} \\ \Rightarrow \quad & d_{j1}(u) = \frac{\eta_{ju}^{(k)}}{\sum_{v=1}^{D-1} \eta_{jv}^{(k)}}. \end{aligned}$$

The maximization w.r.t. the parameter of the geometric tail is also similar to the full geometric case:

$$\begin{aligned}
&\frac{\partial}{\partial p_j} Q_d\left(\{d_j(u)\} \mid \theta^{(k)}\right) = 0 \quad \Rightarrow \quad \sum_{u=D}^{\tau-1} \eta_{ju}^{(k)} \left[-\frac{u-D}{1-p_j} + \frac{1}{p_j}\right] = 0 \\
\Rightarrow \quad &\sum_{u=D}^{\tau-1} \eta_{ju}^{(k)} \left[-(u-D)p_j + 1 - p_j\right] = 0 \Rightarrow \sum_{u=D}^{\tau-1} \eta_{ju}^{(k)} \left[p_j(u-D+1)\right] = \sum_{u=D}^{\tau-1} \eta_{ju}^{(k)} \\
\Rightarrow \quad &p_j = \frac{\sum_{u=D}^{\tau-1} \eta_{ju}^{(k)}}{\sum_{u=D}^{\tau-1} (u-D+1)\eta_{ju}^{(k)}}.
\end{aligned}$$

Thus, the re-estimation formulae are given by

$$d_{j1}^{(k+1)}(u) = \frac{\eta_{ju}^{(k)}}{\sum_{v=1}^{D-1} \eta_{jv}^{(k)}} \tag{5.39}$$

$$p_j^{(k+1)} = \frac{\sum_{u=D}^{\tau-1} \eta_{ju}^{(k)}}{\sum_{u=D}^{\tau-1} (u-D+1)\eta_{ju}^{(k)}} \tag{5.40}$$

$$\phi^{(k+1)} = \frac{\sum_{u=1}^{D-1} \eta_{ju}^{(k)}}{\sum_{u=1}^{\tau-1} \eta_{ju}^{(k)}}. \tag{5.41}$$

5.3.4.4 The Observation Component

The conditional observations can be modeled by a large variety of distributions. In the context of financial time series, mixtures of Normal distributions and t distributions are of particular interest for the modeling of phenomena following skewed or leptokurtic distributions. For each state j, the corresponding part of the Q-function in equation (5.18) is given by

$$\begin{aligned}
Q_b\left(\left\{b_j\left(x_0^{\tau-1}\right)\right\} \mid \theta^{(k)}\right) &= \sum_{t=0}^{\tau-1} P\left(S_t = j \mid X_0^{\tau-1} = x_0^{\tau-1}, \theta^{(k)}\right) \log b_j\left(x_t\right) \\
&= \sum_{t=0}^{\tau-1} L_j(t) \log b_j\left(x_t\right). \qquad (5.42)
\end{aligned}$$

Depending on the distributional assumptions imposed on $b_j(x_t)$, the maximization of the corresponding term of $Q(\theta \mid \theta^{(k)})$ may also involve numerical

methods. In the following we deal with some common distributions, e.g., Poisson, Bernoulli, Normal, and t distribution.

Bernoulli Distribution

The conditional distributions follow a Bernoulli distribution, i.e., for each $j \in 0, \ldots, \tau - 1$

$$b_j(x_t) = p_j^{x_t}(1 - p_j)^{1-x_t}$$

holds. In this section, p_j denotes the parameter of the Bernoulli conditional distribution. Equation (5.42) becomes

$$Q_b(\{b_j(x_0^{\tau-1})\} \mid \theta^{(k)}) = \sum_{t=0}^{\tau-1} L_j(t)\Big[x_t \log p_j + (1 - x_t)\log(1 - p_j)\Big],$$

which has to be maximized w.r.t. p_j to perform the M-step:

$$\frac{\partial}{\partial p_j} Q_b(\{b_j(x_0^{\tau-1})\} \mid \theta^{(k)}) = 0 \quad \Rightarrow \quad \sum_{t=0}^{\tau-1} L_j(t)\Big[x_t(1 - p_j) + (x_t - 1)p_j\Big] = 0$$

$$\Rightarrow \quad \sum_{t=0}^{\tau-1} L_j(t)(x_t - p_j) = 0 \quad \Rightarrow \quad p_j = \frac{\sum_{t=0}^{\tau-1} L_j(t)x_t}{\sum_{t=0}^{\tau-1} L_j(t)}.$$

Thus the re-estimation quantity is given by

$$p_j^{(k+1)} = \frac{\sum_{t=0}^{\tau-1} L_j(t)x_t}{\sum_{t=0}^{\tau-1} L_j(t)}. \tag{5.43}$$

Poisson Distribution

The component distributions are assumed to be univariate Poisson distributions with parameter λ_j:

$$b_j(x_t) = \frac{\lambda_j^{x_t} e^{-\lambda_j}}{x_t!}.$$

Equation (5.42) becomes

$$Q_b(\{b_j(x_0^{\tau-1})\} \mid \theta^{(k)}) = \sum_{t=0}^{\tau-1} L_j(t)\Big[x_t \log \lambda_j - \lambda_j - \log x_t!\Big].$$

The maximization w.r.t. λ_j yields

$$\begin{aligned} &\frac{\partial}{\partial \lambda_j} Q_b(\{b_j(x_0^{\tau-1})\} \mid \theta^{(k)}) = 0 \quad \Rightarrow \quad \sum_{t=0}^{\tau-1} L_j(t)\left[\frac{x_t}{\lambda_j} - 1\right] = 0 \\ &\Rightarrow \quad \lambda_j = \frac{\sum_{t=0}^{\tau-1} L_j(t) x_t}{\sum_{t=0}^{\tau-1} L_j(t)}, \end{aligned}$$

and the re-estimation quantity is

$$\lambda_j^{(k+1)} = \frac{\sum_{t=0}^{\tau-1} L_j(t) x_t}{\sum_{t=0}^{\tau-1} L_j(t)}. \tag{5.44}$$

Multivariate Normal Distribution

For the case of multivariate Normal component distributions we follow the derivations given by Bilmes (1998) for HMMs. The density functions are given by

$$b_j(\boldsymbol{x}_t) = \frac{1}{2\pi^{p/2}|\boldsymbol{\Sigma}_j|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_t - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu}_j)\right)$$

with mean $\boldsymbol{\mu_j}$ and positive definite covariance matrix $\boldsymbol{\Sigma_j}$. The dimension of the observations is denoted by p and all vectors are column vectors. Representing the constant terms by C, Equation (5.42) becomes

$$Q_b\left(\left\{b_j(x_0^{\tau-1})\right\} | \theta^{(k)}\right) = \sum_{t=0}^{\tau-1} L_j(t)\Big[C - \frac{1}{2}\log(|\boldsymbol{\Sigma}_j|) - \frac{1}{2}(\boldsymbol{x}_t - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu}_j)\Big]. \tag{5.45}$$

The maximization of the Q-function requires some matrix calculus, which is described in detail by Bilmes (1998). We first treat the maximization w.r.t. $\boldsymbol{\mu}$.

Taking the derivative of Equation (5.45) w.r.t. $\boldsymbol{\mu}$ and setting it equal to zero yields

$$\sum_{t=0}^{\tau-1} L_j(t)\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu}_j) = 0 \quad \Rightarrow \quad \boldsymbol{\mu}_j = \frac{\sum_{t=0}^{\tau-1} L_j(t)\boldsymbol{x}_t}{\sum_{t=0}^{\tau-1} L_j(t)}.$$

The first step of the maximization w.r.t. $\boldsymbol{\Sigma}$ consists in transforming equation (5.45) to

$$\begin{aligned}\frac{1}{2}\log(|\boldsymbol{\Sigma}_j^{-1}|)\sum_{t=0}^{\tau-1} L_j(t) - \frac{1}{2}\sum_{t=0}^{\tau-1} L_j(t) tr(\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu}_j)(\boldsymbol{x}_t - \boldsymbol{\mu}_j)^T) \\ = \frac{1}{2}\log(|\boldsymbol{\Sigma}_j^{-1}|)\sum_{t=0}^{\tau-1} L_j(t) - \frac{1}{2}\sum_{t=0}^{\tau-1} L_j(t) tr(\boldsymbol{\Sigma}_j^{-1} N_{jt})\end{aligned}$$

with $N_{jt} := (\boldsymbol{x}_t - \boldsymbol{\mu}_j)(\boldsymbol{x}_t - \boldsymbol{\mu}_j)^T$. Differentiating w.r.t. $\boldsymbol{\Sigma}$ yields

$$\begin{aligned}\frac{1}{2}\sum_{t=0}^{\tau-1} L_j(t)\left(2\boldsymbol{\Sigma}_j - \mathrm{diag}(\boldsymbol{\Sigma}_j)\right) - \frac{1}{2}\sum_{t=0}^{\tau-1} L_j(t)\left(2N_{jt} - \mathrm{diag}(N_{jt})\right) \\ = \frac{1}{2}\sum_{t=0}^{\tau-1} L_j(t)\left(2M_{jt} - \mathrm{diag}M_{jt}\right) \\ = 2S - \mathrm{diag}(S),\end{aligned}$$

where $M_{jt} := \boldsymbol{\Sigma}_j - N_{jt}$ and $S := \sum_{t=0}^{\tau-1} L_j(t) M_{jt}$. Setting the last line equal to zero yields

$$2S - \mathrm{diag}(S) = 0 \quad \Rightarrow \quad S = 0.$$

This is equivalent to $\sum_{t=0}^{\tau-1} L_j(t)\left(\boldsymbol{\Sigma}_j - N_{jt}\right) = 0$ and it follows

$$\boldsymbol{\Sigma}_j = \frac{\sum_{t=0}^{\tau-1} L_j(t) N_{jt}}{\sum_{t=0}^{\tau-1} L_j(t)} = \frac{\sum_{t=0}^{\tau-1} L_j(t)(\boldsymbol{x}_t - \boldsymbol{\mu}_j)(\boldsymbol{x}_t - \boldsymbol{\mu}_j)^T}{\sum_{t=0}^{\tau-1} L_j(t)}.$$

Hence the re-estimation quantities for the Normal component distributions are given by

$$\boldsymbol{\mu}_j^{(k+1)} = \frac{\sum_{t=0}^{\tau-1} L_j(t)\boldsymbol{x}_t}{\sum_{t=0}^{\tau-1} L_j(t)} \tag{5.46}$$

$$\boldsymbol{\Sigma}_j^{(k+1)} = \frac{\sum_{t=0}^{\tau-1} L_j(t)(\boldsymbol{x}_t - \boldsymbol{\mu}_j^{(k+1)})(\boldsymbol{x}_t - \boldsymbol{\mu}_j^{(k+1)})^T}{\sum_{t=0}^{\tau-1} L_j(t)}. \tag{5.47}$$

Mixtures of Normal Distributions

For the case of mixtures of Normal distributions as component distributions, we do not provide the entire calculus but give a short overview and the resulting re-estimation formulae. We refer to Sansom & Thomson (2000) for details. Let the density of the Normal distribution with mean $\boldsymbol{\mu}$ and positive definite covariance matrix $\boldsymbol{\Sigma}$ be given by

$$f(\boldsymbol{x_t}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{-1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x_t} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x_t} - \boldsymbol{\mu})\right),$$

where p is the dimension of the observations. Let $m \in \{0, \ldots, M-1\}$ denote the M mixture components which occur for each of the j state dependent component distributions. Thus for each state j, there are M means, denoted by $\boldsymbol{\mu}_{jm}$, and M covariance matrices, denoted by $\boldsymbol{\Sigma}_{jm}$. Then Equation (5.42) becomes

$$b_j(\boldsymbol{x_t}) = \sum_{m=0}^{M-1} \phi_{jm} f(\boldsymbol{x_t}, \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm})$$

with $\sum_{m=0}^{M-1} \phi_{jm} = 1$. To simplify the notation of the re-estimation formulae, it is helpful to introduce the auxiliary variable

$$L_{jm}(t) := \frac{L_j(t)}{\sum_{m=0}^{M-1} \phi_{jm} f(\boldsymbol{x_t}, \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm})} \phi_{jm} f(\boldsymbol{x_t}, \boldsymbol{\mu}_{jm}, \boldsymbol{\Sigma}_{jm}),$$

which can be interpreted as weighted probability of observing $\boldsymbol{x}_t$ in the mixing component m of state j. Then the re-estimation formulae for $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and ϕ can be written as

$$\phi_{jm}^{(k+1)} = \frac{\sum_{t=0}^{\tau-1} L_{jm}(t)}{\sum_{m=0}^{M-1} \sum_{t=0}^{\tau-1} L_{jm}(t)} = \frac{\sum_{t=0}^{\tau-1} L_{jm}(t)}{\sum_{t=0}^{\tau-1} L_j(t)}, \tag{5.48}$$

$$\boldsymbol{\mu}_{jm}^{(k+1)} = \frac{\sum_{t=0}^{\tau-1} L_{jm}(t) \boldsymbol{x}_t}{\sum_{t=0}^{\tau-1} L_{jm}(t)}, \tag{5.49}$$

$$\boldsymbol{\Sigma}_{jm}^{(k+1)} = \frac{\sum_{t=0}^{\tau-1} L_{jm}(t)(\boldsymbol{x}_t - \boldsymbol{\mu}_{jm}^{(k+1)})(\boldsymbol{x}_t - \boldsymbol{\mu}_{jm}^{(k+1)})^T}{\sum_{t=0}^{\tau-1} L_{jm}(t)}. \tag{5.50}$$

The derivations for (5.48), (5.49) and (5.50) are similar to that of Normal component distributions.

t Distribution

The t distribution falls into the class of the elliptically symmetric distributions. In contrast to the Normal distribution it has an additional parameter (the degrees of freedom), which allows one to fit longer tails to deal with more extreme observations.
The derivation and maximization of the Q-function for this distribution is not entirely straightforward. However, the techniques presented by Peel & McLachlan (2000) for the estimation of mixtures of t distributions can be adopted to the case of a HSMM and we follow their approach.

Recall that the t distribution is derived from a Normal mixture model of the form

$$\int g(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}/u)\, d\, U(u), \tag{5.51}$$

where $g()$ denotes the density of the Normal distribution. The random variable U follows a gamma distribution, i.e.,

$$U \sim \text{gamma}\left(\frac{1}{2}\nu, \frac{1}{2}\nu\right),$$

where the density function of the gamma distribution is parameterized as follows:

$$f(u, \alpha, \beta) = \frac{\beta^\alpha u^{\alpha-1}}{\Gamma(\alpha)} \exp(-\beta u) \mathbf{1}_{\{u>0\}}(u),$$

where $\Gamma()$ denotes the gamma function.

Evaluating the integral given in (5.51) yields the density of the t distribution with location parameter $\boldsymbol{\mu}$, ν degrees of freedom and positive definite inner product matrix $\boldsymbol{\Sigma}$. The density is given by

$$f(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) = \frac{\Gamma(\frac{\nu+p}{2})|\boldsymbol{\Sigma}|^{-\frac{1}{2}}}{(\pi\nu)^{\frac{1}{2}p}\Gamma(\frac{\nu}{2})\left\{1 + \delta(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})/\nu\right\}^{\frac{1}{2}(\nu+p)}},$$

where $\delta(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Mahalanobis distance

$$\delta(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (\boldsymbol{x} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}),$$

and p the dimension of the observations.

Note that f converges to the density function of a Normal distribution with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ as ν tends to infinity. The mean $\boldsymbol{\mu}$ of the t distribution exists for all $\nu > 1$, and the covariance matrix is given by $\nu/(\nu - 2)\boldsymbol{\Sigma}$ for $\nu > 2$.

In the case of conditional t distribution, the observation distribution from Equation (5.42) is

$$b_j(\boldsymbol{x_t}) = \frac{\Gamma(\frac{\nu_j+p}{2})|\boldsymbol{\Sigma}_j|^{-\frac{1}{2}}}{(\pi\nu)^{\frac{1}{2}p}\Gamma(\frac{\nu_j}{2})\left\{1 + \delta(\boldsymbol{x}_t, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)/\nu_j\right\}^{\frac{1}{2}(\nu_j+p)}}. \tag{5.52}$$

Unfortunately, the re-estimation formulae of the parameters involved in (5.52) cannot be derived directly, as was the case for observations following the Normal or the Poisson distribution. We adopt the derivation of the re-estimation formulae from Peel & McLachlan (2000), details of which can be found in their article.

In addition to the observations and the states of the semi-Markov chain, the complete-data log-likelihood has to be enriched by two more variables. Firstly, by the indicator function $z_{jt} = (z_t)_j$ which takes the value one if the observation $\boldsymbol{x}_t$ belongs to component j and zero otherwise. Secondly, the missing data from the gamma distributed random variable U, denoted by $u_0, \ldots, u_{\tau-1}$, has to be added to the complete-data with

$$X_t \,|\, u_t, z_{jt} = 1 \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j/u_t)$$

for $t \in \{0, \ldots, \tau - 1\}$, and

$$U_t \,|\, z_{jt} = 1 \sim \text{gamma}\left(\frac{1}{2}\nu_j, \frac{1}{2}\nu_j\right).$$

This "enriched" complete-data allows for a modified formulation of the complete-data likelihood. Given $z_0, \ldots, z_{\tau-1}$, the quantities $U_0, \ldots, U_{\tau-1}$ are conditionally independent, and thus the complete data likelihood can be factored into

the product of the marginal densities of Z_t, the conditional densities of U_t given z_t, and the conditional densities of X_t given u_t and z_t.
The complete-data log-likelihood of the observations of component j then becomes

$$\log L_c(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \nu_j) = \log L_{c1}(\nu_j) + \log L_{c2}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$$

with

$$\begin{aligned}
\log L_{c1}(\nu_j) &= \sum_{t=0}^{\tau-1} z_{jt} \Bigg\{ -\log \Gamma\left(\frac{1}{2}\nu_j\right) + \frac{1}{2}\nu_j \log\left(\frac{1}{2}\nu_j\right) \\
&\qquad + \frac{1}{2}\nu_j(\log u_t - u_t) - \log u_t \Bigg\} \qquad (5.53) \\
\log L_{c2}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) &= \sum_{t=0}^{\tau-1} z_{jt} \Bigg\{ \frac{1}{2} p \log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_j| \\
&\qquad - \frac{1}{2} u_j (\boldsymbol{x}_t - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x}_t - \boldsymbol{\mu}_j) \Bigg\}. \qquad (5.54)
\end{aligned}$$

The calculation of $Q(\theta \mid \theta^{(k)})$ is also affected by the modified complete-data likelihood of the observation part. The conditional expectation of the complete-data log-likelihood is performed in parts. First, the expectation conditioned on the observations $x_0^{\tau-1}$ and $z_0, \ldots, z_{\tau-1}$ is taken. Then, the conditional expectation of z_t given $x_0^{\tau-1}$ is evaluated; hereby $P(Z_t = 1 \mid x_0^{\tau-1}) = L_j(t)$ holds true. From Equation (5.53) and (5.54), it is clear that

$$E(U_t \mid \boldsymbol{x_t}, z_t, \theta^{(k)})$$

and

$$E(\log U_t \mid \boldsymbol{x_t}, z_t, \theta^{(k)})$$

have to be calculated.

The calculation of $E(U_t \mid \boldsymbol{x_t}, z_t, \theta^{(k)})$ is based on the fact that the conjugate prior distribution of U_t is the gamma distribution. It can be shown that the distribution of U_t given $X_t = \boldsymbol{x}_t$ and $Z_{jt} = 1$ is

$$U_t \mid \boldsymbol{x}_t, z_{jt} = 1 \sim \text{ gamma}(m_{1j}, m_{2j}),$$

where

$$m_{1j} := \frac{1}{2}(\nu_j + p)$$

and

$$m_{2j} := \frac{1}{2}\{\nu_j + \delta(\boldsymbol{x}_t, \boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)\}.$$

From the definition of the gamma distribution, it follows that

$$E(U_t \,|\, \boldsymbol{x}_t, z_{jt} = 1) = \frac{{\nu_j}^{(k)} + p}{\nu_j + \delta(\boldsymbol{x}_t, \boldsymbol{\mu}, \boldsymbol{\Sigma})},$$

which yields the desired result:

$$E(U_t \,|\, \boldsymbol{x}_t, z_{jt} = 1, \theta^{(k)}) = \frac{{\nu_j}^{(k)} + p}{{\nu_j}^{(k)} + \delta(\boldsymbol{x}_t^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)})}.$$

To calculate the term $E(\log U_t \,|\, \boldsymbol{x_t}, z_t, \theta^{(k)})$ we have to use the result that if a random variable R is distributed by gamma(α, β), then

$$E(\log R) = \psi(\alpha) - \log \beta,$$

where $\psi(s)$ is again the digamma function given by

$$\psi(s) = \frac{\frac{\partial \Gamma(s)}{\partial s}}{\Gamma(s)}.$$

As shown for the derivation of $E(U_t \,|\, \boldsymbol{x_t}, z_t, \theta^{(k)})$, the conditional density of U_t given $\boldsymbol{x}_t$ and $z_{jt} = 1$ is given by gamma(m_{1j}, m_{2j}). Applying the above result to the conditional density of U_t yields

$$
\begin{aligned}
&E(\log U_t \,|\, \boldsymbol{x_t}, z_t, \theta^{(k)}) \\
&\qquad = \psi\left(\frac{\nu_j^{(k)}+p}{2}\right) - \log\left(\frac{1}{2}\left\{\nu_j^{(k)} + \delta(\boldsymbol{x}_t, \boldsymbol{\mu}_j^{(k)}, \boldsymbol{\Sigma}_j^{(k)})\right\}\right) \\
&\qquad = \log u_{jt}^{(k)} + \left\{\psi\left(\frac{\nu_j^{(k)}+p}{2}\right) - \log\left(\frac{\nu_j^{(k)}+p}{2}\right)\right\}
\end{aligned}
$$

with

$$
u_{jt}^{(k)} \;:=\; \frac{{\nu_j}^{(k)}+p}{{\nu_j}^{(k)} + \delta(\boldsymbol{x}_t^{(k)}, \boldsymbol{\mu}^{(k)}, \boldsymbol{\Sigma}^{(k)})}.
$$

Applying the results for $E(\log U_t \,|\, \boldsymbol{x_t}, z_t, \theta^{(k)})$ and $E(U_t \,|\, \boldsymbol{x_t}, z_t, \theta^{(k)})$, the observation part of the Q-function in Equation (5.42) can be split up into two parts:

$$
Q_b(\{b_j(x_0^{\tau-1})\} \,|\, \theta^{(k)}) = \sum_{t=0}^{\tau-1} L_j(t) Q_{1t}(\nu_j \,|\, \theta^{(k)}) + \sum_{t=0}^{\tau-1} L_j(t) Q_{2t}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j \,|\, \theta^{(k)}), \tag{5.55}
$$

where, ignoring all terms not involving ν_j, yields

$$
\begin{aligned}
Q_{1t}(\nu_j \,|\, \theta^{(k)}) \;=\; & -\log\Gamma\left(\frac{1}{2}\nu_j\right) + \frac{1}{2}\nu_j \log\left(\frac{1}{2}\nu_j\right) \\
& + \frac{1}{2}\nu_j\left[\sum_{t=0}^{\tau-1}\left(\log u_{jt}^{(k)} - u_{jt}^{(k)}\right)\right] \\
& + \psi\left(\frac{\nu_j^{(k)}+p}{2}\right) - \log\left(\frac{\nu_j^{(k)}+p}{2}\right)
\end{aligned}
$$

and

$$
\begin{aligned}
Q_{2t}(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j \,|\, \theta^{(k)}) \;=\; & -\frac{1}{2}p\log(2\pi) - \frac{1}{2}\log\boldsymbol{\Sigma}_j + \frac{1}{2}p\log u_{jt}^{(k)} \\
& -\frac{1}{2}u_{jt}^{(k)}(\boldsymbol{x}_t - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma_j^{-1}} (\boldsymbol{x}_t - \boldsymbol{\mu}_j).
\end{aligned}
$$

The re-estimation procedure consists of a maximization of the two terms of equation (5.55) w.r.t. the parameters $\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j$ and ν_j.

The re-estimation formulae for $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ can be derived explicitly, yielding

$$\boldsymbol{\mu}_j^{(k+1)} = \frac{\sum_{t=0}^{\tau-1} L_j(t) u_{jt}^{(k)} \boldsymbol{x}_t}{\sum_{t=0}^{\tau-1} L_j(t) u_{jt}^{(k)}} \tag{5.56}$$

and

$$\boldsymbol{\Sigma}_j^{(k+1)} = \frac{\sum_{t=0}^{\tau-1} L_j(t) u_{jt}^{(k)} (\boldsymbol{x}_t - \boldsymbol{\mu}_j^{(k+1)})(\boldsymbol{x}_t - \boldsymbol{\mu}_j^{(k+1)})^T}{\sum_{t=0}^{\tau-1} L_j(t)}. \tag{5.57}$$

Note that, according to Kent et al. (1994), for the case of a single component t distribution, the denominator of (5.57) can also be replaced by $\sum_{t=0}^{\tau-1} L_j(t) u_{jt}^{(k)}$ to increase the speed of convergence.

The re-estimation of the degrees of freedom ν_j is a bit more complicated. The estimator $\nu_j^{(k+1)}$ is the (unique) solution of the equation

$$\begin{aligned} &-\psi\left(\frac{1}{2}\nu_j^{(k)}\right) + \log\left(\frac{1}{2}\nu_j^{(k)}\right) + 1 \\ &+ \frac{1}{\sum_{t=0}^{\tau-1} L_j(t)} \left[\sum_{t=0}^{\tau-1} L_j(t) \left(\log u_{jt}^{(k)} - u_{jt}^{(k)}\right)\right] \\ &+ \psi\left(\frac{\nu_j^{(k)} + p}{2}\right) - \log\left(\frac{\nu_j^{(k)} + p}{2}\right) = 0, \end{aligned} \tag{5.58}$$

which can be found, e.g., by a bisection algorithm or by quasi-Newton methods as the left hand side expression is monotonically increasing.

Remark: Most of the quantities involved in the re-estimation formulae for the initial values in Equation (5.31), for the transition probabilities in Equation (5.32), for the dwell time distributions in the Equations (5.33), (5.34), (5.35), (5.39), (5.40), (5.41) and for the state-dependent distributions in the Equations (5.43), (5.44), (5.46), (5.47), (5.48), (5.49), (5.50), (5.56), (5.57), (5.58) can be calculated more or less directly from the quantities involved in the backward recursion with only a few additional computations. One exception is the t distribution which requires a numerical root finding procedure. However, in this

case the EM algorithm is not significantly slowed down because the function to be approximated is monotonically increasing.
Moreover, minor computations concern the contributions at time $t = 0$ and the contributions of the time spent in the last visited state to the re-estimation quantities of the state occupancy distributions (see Sections 5.3.3.1 and 5.3.3.2).

5.4 Asymptotic properties of the maximum likelihood estimators

The asymptotic properties of the maximum likelihood estimators for HMMs are well established. Asymptotic normality of the (consistent) maximum likelihood estimator of a HMM was proved for the first time by Baum & Petrie (1966). They treated a HMM with finite state space under certain assumptions made on some technical conditions, e.g., stationarity and ergodicity. More than thirty years later, Bickel et al. (1998) succeeded in relaxing the restriction to a finite state space; in 1998 they proved asymptotic normality for a general HMM. Stationarity and ergodicity were also assumed and some regularity conditions were imposed on the conditional distributions. However, most of the common distributions fulfill these conditions.

On the other hand, the literature on the asymptotic properties of the maximum likelihood estimators for extensions of the HMM is relatively thin. Bickel et al. (1998), Douc & Matias (2001) and Douc et al. (2004) treat some generalized HMMs. Currently, the only reference which treats the asymptotic properties of HSMMs, namely consistency and asymptotic normality of non-parametric estimators, is Barbu & Limnios (2005).

5.5 Stationary Hidden Semi-Markov Models

The fitting of stationary HSMMs can be achieved by different approaches. If it can be assumed that the beginning of the observations coincides with the entry in a new state, it is possible to arrive at an extension of the M-step similar to the fitting procedure for stationary HMMs, presented in Section 3.1.2.

However, the start of the state sequence can also be independent of the generation of the data series. In that case it is appropriate to introduce a left-censoring to the observed data. The EM algorithm for this process, namely the Hidden Equilibrium Semi-Markov Model described by Guédon (2005), relies

on a complete-data likelihood of the form

$$L_C\big(s_{-v}^{\tau-1+u}, x_0^{\tau-1} \,|\, \theta\big) = P\big(S_{-v} \neq s_0, S_{-w} = s_0, w = 1 \ldots, v-1, S_0^{\tau-1} = s_0^{\tau-1}, \\ S_{\tau-1+w} = s_{\tau-1}, w = 1, \ldots, u-1, S_{\tau-1+u} \neq s_{\tau-1}, X_0^{\tau-1} = x_0^{\tau-1} \,|\, \theta\big).$$

Compared to estimation procedures introduced in the preceding sections, the M-step is more demanding in both its theoretical and the computational aspects.

The easiest and only asymptotically correct method to implement a stationary model is the following. For many long-running processes it is reasonable to simply assume that the estimate for the initial distribution π is given by the steady state distribution of the semi-Markov chain:

$$\hat{\pi} = \frac{\boldsymbol{\pi}^s \times_c \big(E(d_0(u)), \ldots, E(d_{J-1}(u))\big)}{\sum_{j=0}^{J-1} \left[\pi_j^s \cdot E(d_j(u))\right]},$$

where $\boldsymbol{\pi}^s$ denotes the stationary distribution of the TPM of the embedded Markov chain satisfying $\boldsymbol{\pi}^s \boldsymbol{T} = \boldsymbol{T}$ and $\times_c$ indicates the component-wise multiplication.

Chapter 6

Stylized Facts of Daily Return Series and Hidden Semi-Markov Models

Applications related to Financial Econometrics like risk measurement, pricing of derivatives, margin setting and many other financial indicators rely on a suitable modeling of the distributional and temporal properties of the daily return series of stocks, indices or other assets.

The normal distribution with stationary parameters has often been chosen to model daily return series in financial theory. After the classical paper of Fama (1965), which observed more kurtosis and higher peaks contradicting the assumption of normality, many authors proposed solutions to overcome this drawback. For example, Praetz (1972) and Blattberg & Gonedes (1974) preferred the t distribution while Mittnik & Rachev (1993) examined various stable distributions. The recent articles of Gettinby et al. (2004) and Harris & Küçüközmen (2001) provide a good overview of the several approaches available in the literature.

As with the distributional properties, the temporal properties of daily return series have also been examined by many authors. The most popular models of the last decade fall into the class of ARCH-type models; for a review see Bollerslev et al. (1992) or Franses & van Dijk (2000). Other suitable alternative approaches include the Stochastic Volatility models introduced by Taylor (1986) and have been applied in many contexts, e.g., by Koopman et al. (2005), and Yu (2002). State-space models based on the Kalman filter were investigated inter alia by Faff et al. (2000) or Yao & Gao (2004). The large class of Markov-switching models was introduced by Hamilton (1989, 1990). Turner et al. (1989) first considered a Markov mixture of normal distributions, and

many other studies followed, e.g., Cecchetti et al. (1990), Linne (2002), and Bialkowski (2003).

In this chapter, we focus on modeling the distributional and temporal properties of daily return series by HSMMs. Rydén et al. (1998) show that a HMM – mixing normal variables according to the states of an unobserved Markov chain – reproduces most of the stylized facts for daily return series established by Granger & Ding (1995*a*,*b*). However, the analysis of Rydén et al. (1998), henceforth abbreviated RY, also illustrates that the stylized fact of the very slowly decaying autocorrelation for absolute (or squared) returns cannot be described by a HMM. The lack of flexibility of a HMM to model the temporal higher order dependence can be explained by the implicit geometric distributed sojourn time in the hidden states.
The two hidden HSMMs explored in this chapter are generalization of the model presented by RY. After fitting them to daily return series from 18 sectors of the EUROSTOXX, we show a significantly improved fit of the autocorrelation function. For more detailed results we refer to Bulla & Bulla (2006).

The remainder of this chapter is organized as follows. In Section 6.1 we present the estimation procedures for our specific models for financial time series. Section 6.2 contains a short description of the data. Section 6.3 outlines the results of our analysis while Section 6.4 summarizes the findings. Section 6.5 presents some tables with detailed results.

6.1 Modeling Daily Return Series

RY fitted HMMs with normal component distributions to subseries of the well-known S&P 500 index. They noticed that the autocorrelation function of the estimated model does not satisfactorily capture the behavior of the empirical autocorrelation function, mainly due to the much slower decay of the latter. The temporal dependence properties of a HMM rely on the values of the TPM (MacDonald & Zucchini 1997, Chapter 2.4). However, the geometric sojourn time distribution is a fixed feature of these models. In contrast to many other applications like speech recognition, there exists no test data for financial time series where the "real" sojourn time distribution is known. The approach of Sansom & Thomson (2001), who fitted HSMMs with non-parametric state occupation in a first step to deduce the shape of a parametric distribution in the context of rainfall data, does not yield satisfactory results for daily return series. The nonparametric estimates are too unstable to select a suitable parametric alternative. We therefore generalize the model of RY by fitting a

HSMM with negative binomial sojourn time distributions of the form

$$d_j(u) = \binom{u-2+r_j}{u-1} \pi_j^{r_j}(1-\pi_j)^{u-1},\ u \in \{1,2,\dots\},$$

where j denotes the state. The number of parameters only increases by one per state and, for $r = 1$, our model reduces to a HMM.

While Granger & Ding (1995*a*,*b*) suggested a double exponential distribution to characterize daily returns, RY proposed mixtures of normal variables. We fit HSMMs with variables following normal and t distributions. For the remainder of this chapter, the HMM of RY will be denoted by M_{RY}, the HSMM with conditional normal distributions by SM_N and the HSMM with conditional t distributions by SM_t. RY investigated HMMs with two and three states. They noticed that the three-state models were less similar to each other because the estimation results show a strong dependence on outliers. For this reason, and for better comparability, all models estimated in this chapter have two states.

6.2 The Data Series

The data used here are the daily returns calculated for 18 Pan-European industry portfolios, covering the period from 1$^{\text{st}}$ January 1987 to 5$^{\text{th}}$ September 2005. All sector indices are from STO (2004), and the common currency used is the Euro. The daily returns of the period from $t-1$ to t are computed continuously by

$$R_t = \ln(P_t) - \ln(P_{t-1}),$$

where P_t represents the index closing price on day t and ln is the natural logarithm. All data are obtained from Thomson Financial Datastream.[6]
Descriptive statistics for the data are provided in Table 6.1. It was found that all sector indices are leptokurtic and negatively skewed. The Jarque-Bera statistic confirms the departure from normality for all return series at the 1% level of significance.

[6]For further information see *http://www.thomson.com/*

Table 6.1: Descriptive statistics of daily sector returns

Summary of the daily returns data of the 18 DJ STOXX sector indices, covering the period from 1st January 1987 to 5th September 2005.

Sector	N[a]	Mean·10^4	$S.D. \cdot 10^2$	Skew.	Ex. Kurt.	JB[b]
Automobiles	4824	0.534	1.48	−0.36	5.99	7315
Banks	4824	2.284	1.20	−0.29	6.98	9863
Basics	4824	2.087	1.25	−0.40	6.82	9474
Chemicals	4824	2.536	1.29	−0.12	5.60	6314
Construction	4824	2.881	1.09	−0.56	6.39	8478
Financials	4824	1.993	1.16	−0.51	8.04	13217
Food	4824	2.838	1.10	−0.25	14.25	40877
Healthcare	4824	3.537	1.31	−0.39	6.30	8096
Industrials	4824	2.740	1.16	−0.42	6.42	8423
Insurance	4824	1.247	1.45	−0.26	7.39	11041
Media	4824	2.419	1.49	−0.44	7.34	10990
Oil & Gas	4824	3.895	1.24	−0.25	4.48	4098
Personal	4824	3.317	1.05	−0.16	4.78	4627
Retail	4824	1.676	1.87	−0.26	5.26	5614
Technology	4824	3.330	1.63	−0.13	4.27	3683
Telecom	3521	2.246	1.48	−0.36	5.38	4339
Travel	3521	2.380	1.48	−0.19	2.47	916
Utilities	3521	2.070	1.19	−0.43	7.21	7740

[a] In September 2004, STOXX Ltd. replaced the sectors Cyclical Goods & Services, Non-Cyclical Goods & Services, and Retail (old) by the new sectors Travel & Leisure, Personal & Household Goods, and Retail (new), respectively. The history of the newly formed sectors (with 3521 observations) dates back to 31st December 1991.

[b] JB is the Jarque-Bera statistic for testing normality.

6.3 Empirical Results

All estimation results for the HMM are reported in Table 6.5 of Section 6.5. The corresponding results for the HSMMs with conditional normal and t distributions are reported in Tables 6.6 and 6.7, respectively.

All three models considered fit the marginal distributions of the returns reasonably well. As expected for daily return series, the empirical mean and the mean of the fitted models lie very close to zero for all 18 sectors. As shown in Table 6.2, the empirical standard deviation is also reproduced very well by the three models.

Table 6.2: Standard deviation of the data and the fitted models

The empirical standard deviation of the 18 DJ STOOXX sectors in comparison to the standard deviation of the fitted HMM and the standard deviation of the two fitted HSMMs. All results are multiplied by 100.

Sector	*Obs.*	M_{RY}	SM_N	SM_t
Automobiles	1.48	1.47	1.47	1.46
Banks	1.20	1.19	1.19	1.18
Basics	1.25	1.24	1.24	1.24
Chemicals	1.29	1.29	1.28	1.27
Construction	1.09	1.09	1.09	1.08
Financials	1.16	1.16	1.15	1.15
Food	1.10	1.10	1.09	1.08
Healthcare	1.31	1.31	1.31	1.30
Industrials	1.16	1.16	1.16	1.15
Insurance	1.45	1.45	1.44	1.44
Media	1.49	1.49	1.48	1.48
Oil & Gas	1.24	1.24	1.24	1.23
Personal	1.05	1.05	1.04	1.04
Retail	1.87	1.87	1.86	1.85
Technology	1.63	1.63	1.63	1.60
Telecom	1.48	1.48	1.48	1.47
Travel	1.48	1.48	1.47	1.44
Utilities	1.19	1.19	1.17	1.16

The three models exhibit a clear tendency towards the kurtosis. Even though all models are subject to excess kurtosis, SM_t provides the best results. The average empirical excess kurtosis of the 18 sectors is 6.41 and the average excess kurtosis of SM_t is 7.00. However, the two models M_{RY} and SM_N based on normal conditional distributions, only achieve an average excess kurtosis of 2.95 and 3.45, respectively. The results for all sectors are shown in Table 6.3. The ability of SM_t to capture excess kurtosis is displayed for the three sectors Food, Industrials, and Travel & Leisure (henceforth mentioned only as 'Travel') in the Figures 6.1, 6.2, and 6.3, respectively. These three sectors were selected because they are subject to high, medium and low excess kurtosis (Food: 14.25, Industrials: 6.42, Travel & Leisure: 2.47). The results are similar for the remaining return series. The density of M_{RY} has the lowest peak of the three models, followed by SM_N and SM_t, whereby the latter reproduces the concentration of returns close to zero far better than the competitors.

Figure 6.1: Observed and fitted distributions for the Food sector

Histogram of the returns for the Food sector and the marginal distribution of the three models M_{RY}, SM_N, and SM_t.

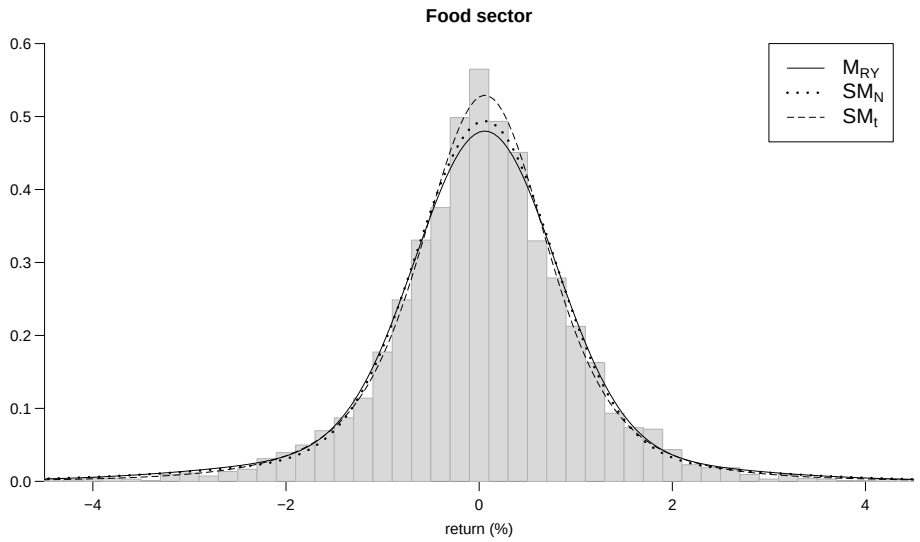


Figure 6.2: Observed and fitted distributions for the Industrials sector

Histogram of the returns for the Industrials sector and the marginal distribution of the three models M_{RY}, SM_N, and SM_t.

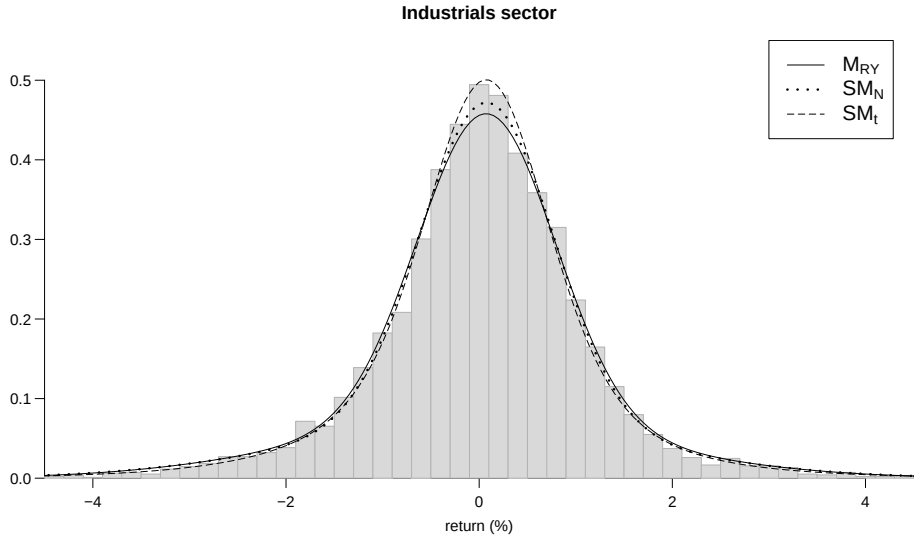
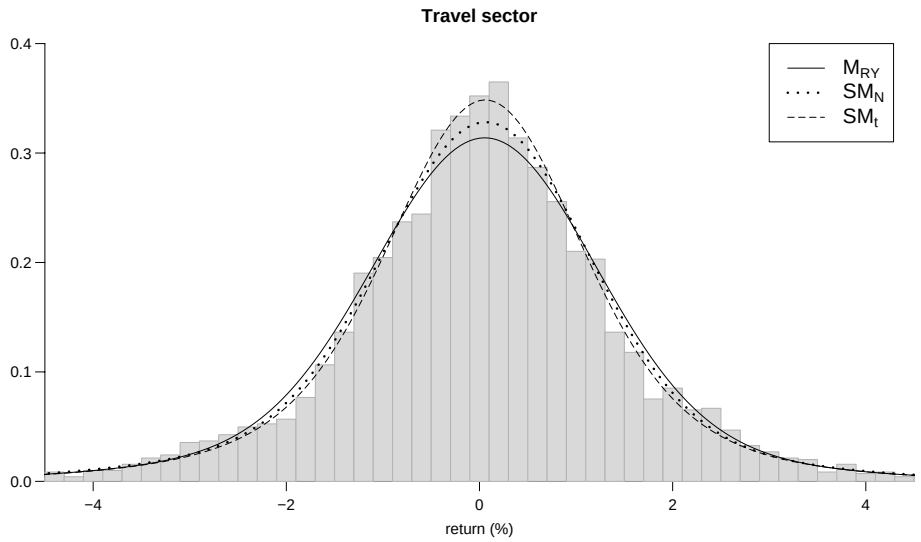


Figure 6.3: Observed and fitted distributions for the Travel & Leisure sector

Histogram of the returns for the Travel & Leisure sector and the marginal distribution of the three models M_{RY}, SM_N, and SM_t.



The average log-likelihood of M_{RY} is 14200 and it increases to 14236 and 14271 for SM_N and SM_t, respectively. Comparing M_{RY} and SM_N, the Technology (Personal) sector is subject to the maximum (minimum) increase of the log-likelihood of 56.5 (13.8). For the Food (Personal) sector the extension from SM_N to SM_t leads to an increase of 89.5 (12.5).

A standard procedure for comparing two nested models is the Likelihood Ratio Test (LRT). It compares a relatively more complex model to a simpler model to assess whether it fits the data significantly better. As the three models are hierarchically nested, the LRT may be applied with the null hypothesis of $r_1, r_2 = 1$ for the comparison M_{RY}/SM_N, and $\nu_1, \nu_2 = \infty$ for SM_N/SM_t. For each of the 18 sectors, SM_N is significantly better than M_{RY} at 0.1% level of significance. The same statement holds true for the comparison SM_N/SM_t, indicating that SM_t provides the best fit to the data. In addition, this statement is supported by the Akaike information criterion which, on an average, decreases from -28388 for the HMM to -28456 and -28522 for the two HSMMs.

Table 6.3: Kurtosis of the data and the fitted models

The empirical excess kurtosis of the 18 DJ STOOXX sectors in comparison to the excess kurtosis of the fitted HMM and of the two fitted HSMMs.

Sector	*Obs.*	M_{RY}	SM_N	SM_t
Automobiles	5.99	3.15	3.70	6.53
Banks	6.98	4.10	4.66	8.42
Basics	6.82	3.00	3.39	8.78
Chemicals	5.60	2.50	2.95	5.70
Construction	6.39	2.60	3.19	7.25
Financials	8.04	4.16	5.00	9.87
Food	14.25	3.48	4.26	17.56
Healthcare	6.30	2.26	2.58	5.17
Industrials	6.42	2.70	2.96	7.07
Insurance	7.39	4.31	5.29	9.02
Media	7.34	3.79	4.11	9.30
Oil & Gas	4.48	2.13	2.49	4.41
Personal	4.78	2.86	3.44	4.96
Retail	5.26	3.07	3.40	5.51
Technology	4.27	1.99	2.32	3.67
Telecom	5.38	2.66	3.17	5.26
Travel	2.47	1.80	2.04	2.53
Utilities	7.21	2.57	3.20	5.04

For the HMM, the mean and variance of the sojourn times (in each state) are controlled by the parameters of the geometric distributions. The additional parameter of the HSMM allows more flexibility of mean and variance. The results are displayed in Figure 6.4.

For every model, the expected sojourn time is higher in the low-risk state, where risk is measured in terms of variance of the respective conditional distribution. This seems reasonable because periods of high volatility reflect a nervous market and are historically less persistent than periods of low volatility.

It is remarkable that the average sojourn times for the HSMMs are significantly lower than for M_{RY}, i.e. the persistence of both the high- and the low-risk state is much lower. The higher sojourn times of SM_t w.r.t. SM_N are a consequence of the heavier tails of the component distributions – in most cases the degrees of freedom take values between 5 and 10. On the other hand, the standard deviations of the sojourn time distributions show a smaller difference between the models.

Figure 6.4: Mean and standard deviation of the sojourn time distributions of the sectors, grouped by model and high-risk (HR)/low-risk (LR) state

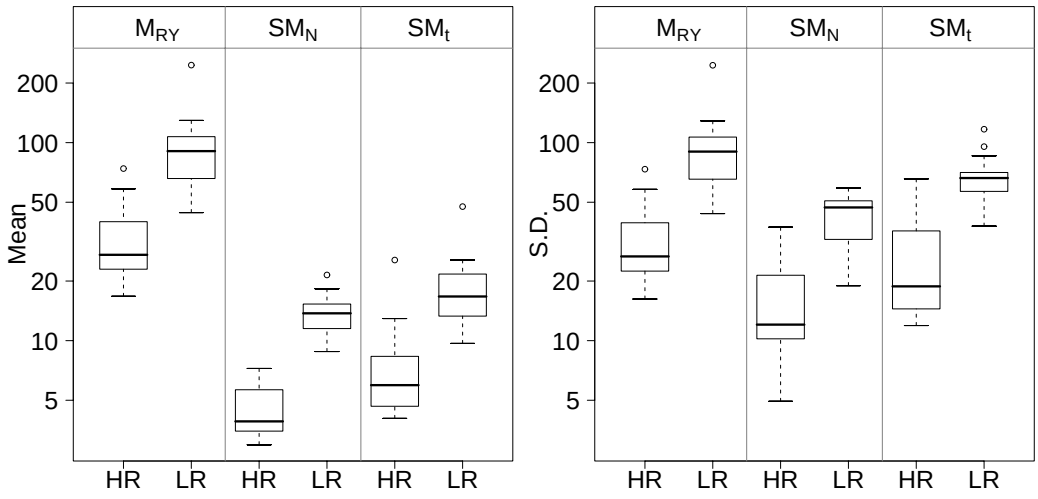
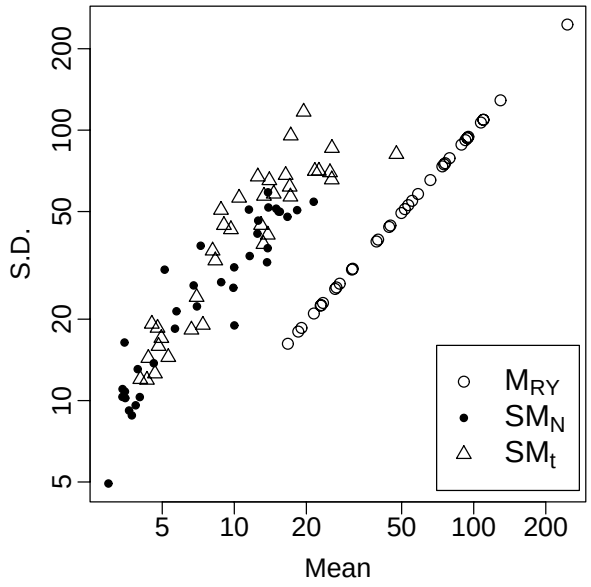


Figure 6.5: Mean and standard deviation of the sojourn times



Taking both mean and standard deviation into account, the coefficient of variation equals $1 - \pi_1$ and $1 - \pi_2$ for the two states of M_{RY}. Due to the high persistence of the states, this yields values close to one. However, the aver-

age coefficient of variation of the $SM_N(SM_t)$ is 3.23 (3.72). Plotting mean and standard deviation of the sojourn time against each other shows a clear separation of the HMM and the two HSMMs, as shown in Figure 6.5.

To analyze the temporal properties of the three models we compare the empirical autocorrelation function (ACF) of the 18 sectors to the model ACF. The slow decay of the ACF for series of absolute/squared daily return is difficult to model. RY stated that this stylized fact cannot be reproduced by the HMM because the decay of the autocorrelations is (much) faster than that observed in reality. He considered this stylized fact to be "the most difficult (...) to reproduce with a HMM". Figures 6.6, 6.7 and 6.8 show the empirical ACF of squared returns as well as the ACF of the three models for the 18 sectors. The solid line represents the ACF of M_{RY}, while the dotted and dashed lines represent SM_N and SM_t, respectively.

The HMM shows the typical strong decay of the autocorrelations and is far from the gray empirical ACF, which confirms the results of RY. Both SM_N and SM_t show a good fit in the tail of the ACF and reproduce this stylized fact much better than the HMM. However, the SM_t looses some of its credibility due to the bad fit for the lags of lower order. Here, SM_N performs clearly better. To measure the fit of the ACF, we calculate the mean squared error (MSE) of the models and a weighted mean squared error ($wMSE$). The $wMSE$ reweights the error at lag i by $0.95^{(100-i)}$ to increase the influence of higher order lags. The results reported in Table 6.4 confirm the visual impression that SM_N provides the best fit with respect to both criteria. Compared to M_{RY}, the MSE of the ACF for SM_N is reduced by approximately 40%.

Table 6.4: Average mean squared error and weighed mean squared error for the ACF of the 18 sectors

Criterion	M_{RY}	SM_N	SM_t
$\overline{MSE} \cdot 10^2$	5.37	3.22	5.47
$\overline{wMSE} \cdot 10^3$	10.36	5.06	5.46

A remarkable observation is the bad fit of all models for the sectors Banks, Insurance, and Financial Services. We can offer no plausible explanation as to why HMMs and HSMMs should not be an appropriate way to model the daily returns for the sectors from the financial industry. In our observation period between the years 1987 to 2005, this part of the economy was severely affected by the boom and subsequent crash of the Internet bubble.

Figure 6.6: Empirical (gray bars) and model ACF for the first six sectors at lag 1 to 100

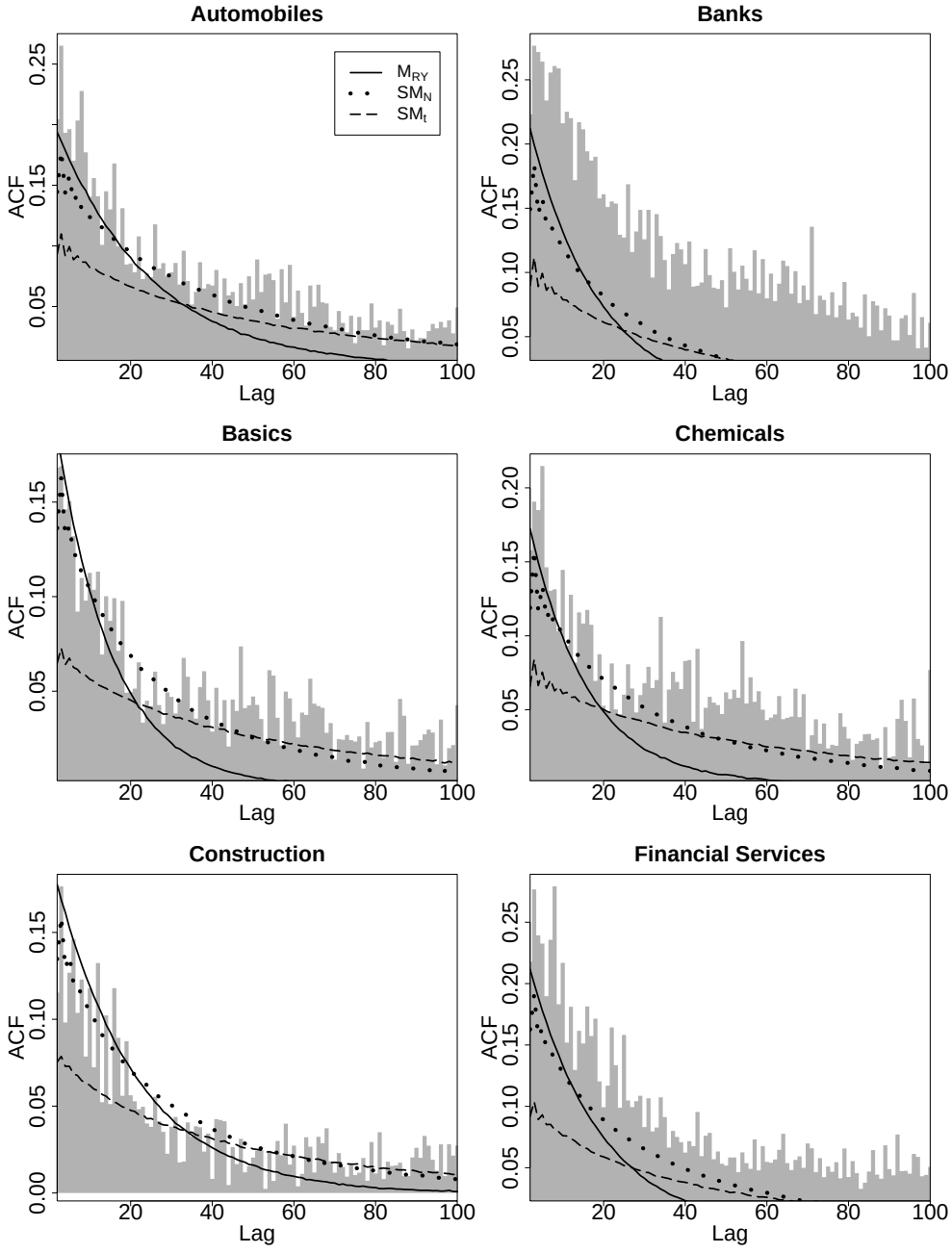


Figure 6.7: Empirical (gray bars) and model ACF for the sectors seven to twelve at lag 1 to 100

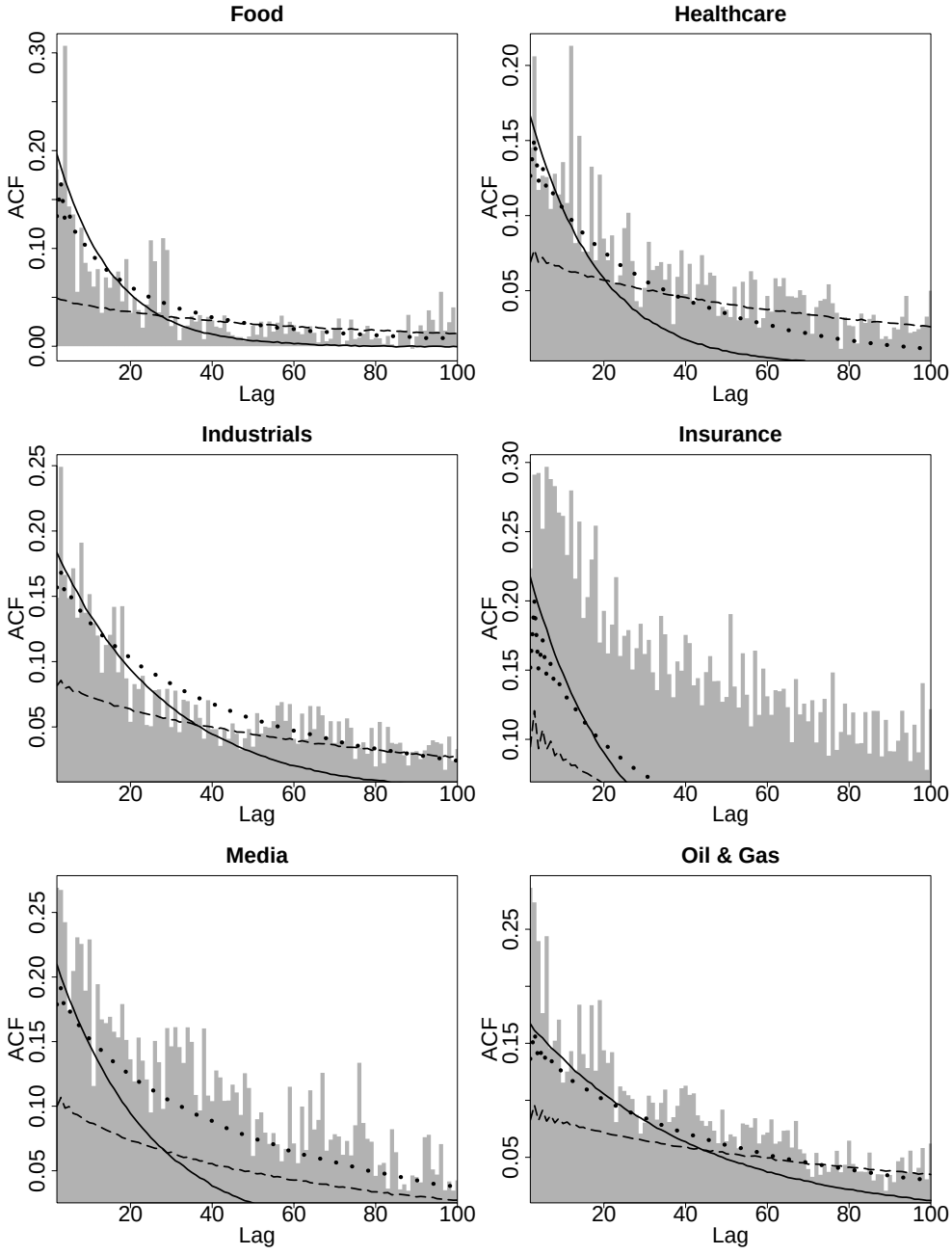
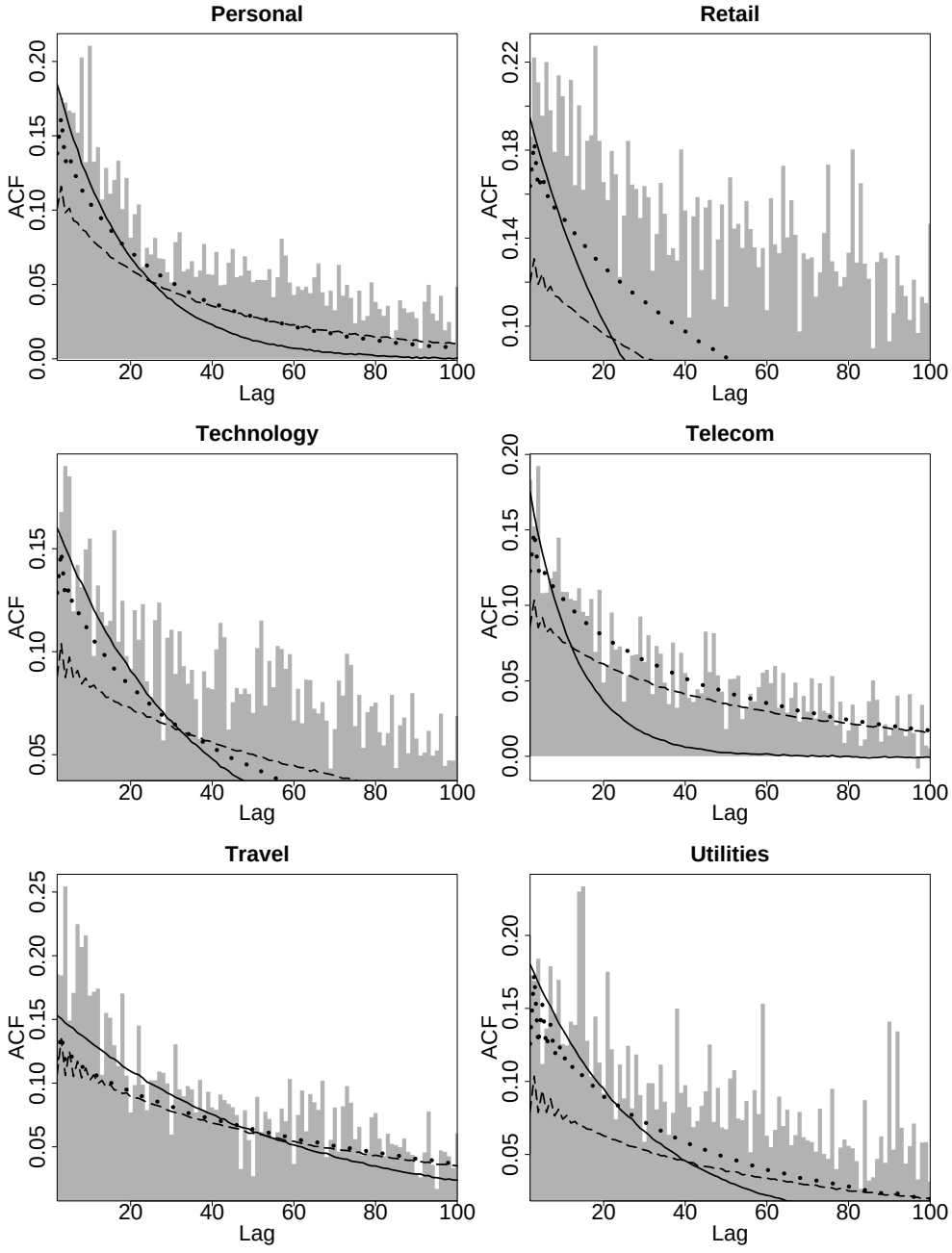


Figure 6.8: Empirical (gray bars) and model ACF for sectors thirteen to eighteen at lag 1 to 100



On the other hand, the TMT (Technology, Media and Telecommunications) stocks also experienced extreme fluctuations during and after the Internet bubble and, for these sectors, the autocorrelation function is modeled reasonably well. However, alternative approaches should be considered for the sectors from the financial industry. Retail is the only example for a poorly fitting ACF from the remaining sectors. We are unable to offer a fundamental explanation for this anomaly.

To summarize, it may be stated that there is a trade off between the distributional and the temporal properties of the two HSMMs. While t conditional distributions are the first choice while considering the model selection criteria AIC and LRT, the HSMM with normal conditional distributions reproduces the shape of the empirical autocorrelation function much better than SM_t and, naturally, than M_{RY}.

6.4 Conclusion

We present a generalized approach of RY to model daily return series and, in particular, improve the temporal dependence properties. We show that the one stylized fact, the slowly decaying autocorrelation function which could not be reproduced by a HMM, is well described by a HSMM with negative binomial sojourn time and normal conditional distributions.
The negative binomial sojourn time distribution is only one of the many possibilities to extend a HMM. It certainly seems to be a more appropriate choice than the geometric sojourn time distribution of a HMM. Future research may show that other, parametric or non-parametric alternatives are better.

6.5 Estimation Results

Table 6.5: Parameter estimates for the HMM

Estimated parameters of the HMM with normal component distributions for the eighteen DJ STOXX sector indices. For state i, $i = 1, 2$, p_i is the parameter of the sojourn time distribution, μ_i and σ_i^2 model the component distribution.

Sector	$1-p_1$	$1-p_2$	$\mu_1 \cdot 10^3$	$\mu_2 \cdot 10^3$	$\sigma_1^2 \cdot 10^4$	$\sigma_2^2 \cdot 10^4$	$S.D. \cdot 10^2$	E.K.	AIC
Automobiles	0.968	0.991	−1.777	0.585	6.288	0.978	1.47	3.15	−28473
Banks	0.957	0.987	−1.204	0.667	4.427	0.502	1.19	4.10	−31114
Basics	0.948	0.982	−1.490	0.803	4.160	0.635	1.24	3.00	−30061
Chemicals	0.953	0.980	−0.925	0.761	3.947	0.657	1.29	2.50	−29645
Construction	0.962	0.989	−1.416	0.766	3.259	0.601	1.09	2.60	−31058
Financials	0.957	0.987	−1.229	0.640	4.172	0.460	1.16	4.16	−31483
Food	0.946	0.985	−0.870	0.608	3.646	0.514	1.10	3.48	−31398
Healthcare	0.963	0.981	−0.622	0.842	3.835	0.663	1.31	2.26	−29407
Industrials	0.975	0.989	−0.845	0.757	3.288	0.506	1.16	2.70	−30931
Insurance	0.964	0.989	−1.141	0.496	6.785	0.739	1.45	4.31	−29360
Media	0.968	0.989	−1.273	0.740	6.520	0.778	1.49	3.79	−28989
Oil & Gas	0.983	0.992	−0.195	0.651	3.482	0.666	1.24	2.13	−29955
Personal	0.956	0.991	−1.303	0.681	3.375	0.599	1.05	2.86	−31418
Retail	0.978	0.987	−1.166	0.923	8.173	0.822	1.87	3.07	−27222
Technology	0.981	0.989	−0.546	0.843	5.511	1.010	1.63	1.99	−27351
Telecom	0.940	0.977	−0.493	0.495	5.559	0.922	1.48	2.66	−20559
Travel	0.986	0.996	−0.962	0.598	5.274	1.255	1.48	1.80	−20337
Utilities	0.975	0.991	−0.743	0.547	3.594	0.627	1.19	2.57	−22230

Table 6.6: Parameter estimates for the HSMM with normal conditional distributions

Estimated parameters of the HSMM with normal component distributions for the eighteen DJ STOXX sector indices. For state i, $i = 1, 2$, π_i and r_i are the parameters of the sojourn time distribution, μ_i and σ_i^2 model the component distribution.

Sector	$1-\pi_1$	$1-\pi_2$	$r_1 \cdot 10$	$r_2 \cdot 10$	$\mu_1 \cdot 10^3$	$\mu_2 \cdot 10^3$	$\sigma_1^2 \cdot 10^4$	$\sigma_2^2 \cdot 10^4$	$S.D. \cdot 10^2$	E.K.	AIC
Automobiles	0.983	0.995	0.518	0.746	−1.866	0.576	6.763	0.924	1.47	3.70	−28537
Banks	0.969	0.993	0.847	0.775	−1.102	0.625	4.656	0.459	1.19	4.66	−31203
Basics	0.969	0.991	0.922	0.966	−1.571	0.812	4.366	0.592	1.24	3.39	−30132
Chemicals	0.979	0.990	0.544	0.821	−1.111	0.804	4.213	0.611	1.28	2.95	−29736
Construction	0.965	0.994	0.995	0.824	−1.553	0.748	3.614	0.575	1.09	3.19	−31104
Financials	0.971	0.994	0.890	0.848	−1.413	0.631	4.652	0.448	1.15	5.00	−31569
Food	0.919	0.995	1.752	0.629	−0.950	0.581	4.110	0.500	1.09	4.26	−31464
Healthcare	0.981	0.991	0.703	0.843	−0.801	0.897	4.035	0.631	1.31	2.58	−29468
Industrials	0.988	0.993	0.728	1.081	−0.940	0.793	3.375	0.475	1.16	2.96	−30979
Insurance	0.977	0.995	0.558	0.622	−1.233	0.469	7.630	0.707	1.44	5.29	−29464
Media	0.992	0.993	0.471	1.427	−1.329	0.749	6.713	0.742	1.48	4.11	−29055
Oil & Gas	0.986	0.996	0.644	0.476	−0.367	0.705	3.701	0.633	1.24	2.49	−30006
Personal	0.976	0.993	0.615	1.176	−1.616	0.711	3.712	0.575	1.04	3.44	−31454
Retail	0.996	0.990	0.280	1.232	−1.316	0.960	8.534	0.777	1.86	3.40	−27331
Technology	0.990	0.987	0.495	1.178	−0.503	0.825	5.704	0.873	1.63	2.32	−27423
Telecom	0.991	0.975	0.231	2.313	−0.568	0.502	5.993	0.859	1.48	3.17	−20640
Travel	0.996	0.988	0.183	1.549	−0.996	0.710	5.074	1.074	1.47	2.04	−20361
Utilities	0.980	0.996	0.488	0.429	−0.793	0.522	3.975	0.600	1.17	3.20	−22279

Table 6.7: Parameter estimates for the HSMM with Student t conditional distributions

Estimated parameters of the two-state HSMM with t component distributions for the eighteen DJ STOXX sector indices. For state i, $i = 1, 2$, π_i and r_i are the parameters of the sojourn time distribution, μ_i, σ_i^2 and ν_i model the component distribution.

Sector	$1-\pi_1$	$1-\pi_2$	$r_1 \cdot 10$	$r_2 \cdot 10$	$\mu_1 \cdot 10^3$	$\mu_2 \cdot 10^3$	$\sigma_1^2 \cdot 10^4$	$\sigma_2^2 \cdot 10^4$	ν_1	ν_2	$S.D. \cdot 10^2$	E.K.	AIC
Automobiles	0.985	0.997	0.588	0.511	−1.562	0.622	4.525	0.769	7.55	10.94	1.46	6.53	−28594
Banks	0.977	0.996	0.792	0.466	−0.684	0.628	3.039	0.383	7.04	10.38	1.18	8.42	−31258
Basics	0.983	0.996	0.962	0.550	−1.045	0.816	2.283	0.482	5.53	9.94	1.24	8.78	−30224
Chemicals	0.986	0.995	0.549	0.410	−0.592	0.744	2.490	0.484	6.71	8.81	1.27	5.70	−29814
Construction	0.982	0.996	1.146	0.935	−1.027	0.805	1.966	0.490	5.90	13.49	1.08	7.25	−31170
Financials	0.979	0.996	0.902	0.678	−1.011	0.671	2.810	0.366	6.33	11.03	1.15	9.87	−31625
Food	0.994	0.993	1.405	3.281	−0.263	0.658	1.351	0.370	4.37	10.37	1.08	17.56	−31639
Healthcare	0.997	0.991	0.238	1.052	−0.320	0.932	2.299	0.516	6.61	11.89	1.30	5.17	−29564
Industrials	0.994	0.997	0.722	0.823	−0.441	0.787	1.999	0.401	6.21	12.13	1.15	7.07	−31066
Insurance	0.979	0.997	0.659	0.399	−0.892	0.516	4.898	0.575	7.20	11.10	1.44	9.02	−29512
Media	0.993	0.995	0.499	1.199	−1.084	0.767	4.380	0.642	6.32	12.34	1.48	9.30	−29121
Oil & Gas	0.994	0.998	0.397	0.289	−0.049	0.640	2.523	0.526	8.15	10.56	1.23	4.41	−30066
Personal	0.977	0.996	0.868	0.863	−1.457	0.710	2.624	0.495	9.26	12.80	1.04	4.96	−31483
Retail	0.996	0.995	0.328	0.815	−1.077	0.983	6.548	0.631	9.35	7.99	1.85	5.51	−27386
Technology	0.990	0.997	0.616	0.294	−0.538	0.834	4.358	0.709	10.46	6.83	1.60	3.67	−27497
Telecom	0.990	0.992	0.340	0.981	−0.070	0.465	4.349	0.742	8.29	7.90	1.47	5.26	−20679
Travel	0.989	0.999	0.419	0.250	−1.036	0.619	5.311	0.987	47.02	10.20	1.44	2.53	−20382
Utilities	0.984	0.997	0.562	0.285	−0.288	0.500	2.427	0.503	7.98	20.67	1.16	5.04	−22313

Chapter 7

Conclusion and Future Work

In this final chapter we review our main results and discuss possible extensions to our work. The common theme of all chapters of this thesis is the application of HMMs and HSMMs to financial time series. We examine different classes of models, e.g., regime switching models in the CAPM framework or semi-Markovian mixture distribution for daily return series. Moreover, we address various computational issues in the parameter estimation of HMMs.

In Chapter 3 we present the most familiar methods of parameter estimation for HMMs. We investigate the coverage probability of bootstrap-based confidence bands and the effect of different parameterizations on the performance of estimation algorithms. We show that a hybrid algorithm provides an excellent compromise to overcome the trade-off between stability of the EM algorithm and the speed of DNM. If a choice has to be made between the EM algorithm and DNM, the latter is preferable if one can provide accurate initial values, or if the estimation is time-critical. Clearly, if the formulae required for the EM algorithm are too difficult to derive, or if one wishes to avoid deriving these, then one has to use DNM. In all other cases the EM algorithm is the preferred method due to its greater stability.
An interesting aspect for further analysis would be an extension to HSMMs. For these models, neither the estimation results of DNM, nor the coverage probabilities of confidence bands, are available in the literature.

The analysis of different approaches to model time-varying beta risk in Chapter 4 yielded various interesting results. It is seen that the Markov switching approach provides an efficient method to reduce the number of parameters when a large number of time series have to be modeled jointly. However, in comparison to other modeling techniques, the Markov regime switching models showed an unsatisfactory out-of-sample forecast performance.

The results could be improved significantly by increasing the number of states of the models substantially. Special attention has to be paid to the number of parameters of such a model, in particular the entries of the TPM. A possibility that could be considered in future research is to reduce the number of parameters by imposing appropriate restrictions on the TPM.

In Chapter 5 and 6 we introduce the theoretical basis of right-censored HSMMs and present a generalization of the approach of Rydén et al. (1998) to model daily return series. We succeed to improve the temporal properties and show that the one stylized fact that a HMM cannot reproduce, namely the slowly decaying autocorrelation function, is significantly better described by a HSMM with negative binomial sojourn time and normal conditional distributions.
The semi-Markov approach is, in our opinion, the field with the largest number of possible extensions. While HMMs allow, basically, only the conditional distributions to be changed (apart from certain generalizations such as feedback or duration dependent models), the HSMM additionally allows one to modify distribution of the state-dependent sojourn time distribution. Finally, the asymptotic properties of maximum likelihood estimators are far from completely explored and offer a challenging problem requiring further research.

Appendix A

The EM Algorithm

A popular method for estimating the parameters of HMMs is the Baum-Welch algorithm, a technique similar to what became known as the EM algorithm later (Baum et al. 1970). Originally, the EM algorithm was developed for estimating the parameters of a model in the case of missing data. However, it could also be adopted in the estimation of HMMs and HSMMs.
In this section, we briefly describe the EM algorithm in a general form which is applicable to a variety of other problems dealing with hidden/missing/incomplete data. Mathematical strictness is relaxed for the benefit of a better, intuitive comprehension of the subject. For further reading, a good introduction is given by Bilmes (1998) and the references therein.

A.1 Prerequisites

Consider a sample of τ i.i.d. observations from a random variable X:

$$x_0, \ldots, x_{\tau-1}.$$

It is assumed that the observations $x_0^{\tau-1} = x_0, \ldots, x_{\tau-1}$ are incomplete; i.e., we deal with a data set one part of which is observable and the other part is missing. The complete data are denoted by C and the missing data by S:

$$C = (X, S).$$

Let the density functions of C, X and S be denoted by

$$p_C(c \,|\, \theta),\ p_X(x \,|\, \theta) \text{ and } p_S(s \,|\, \theta),$$

respectively. For better readability, we omit the subscripts of the above functions in the following.

The aim of the EM-algorithm is the maximization of the log-likelihood of the observed data

$$\begin{aligned} \log L(\theta) &:= \log p(x_0^{\tau-1} \,|\, \theta) \\ &= \sum_{t=0}^{\tau-1} \log p(x_t \,|\, \theta). \end{aligned} \tag{A.1}$$

Analytic maximization of Equation (A.1) can be difficult. In many cases, an analytic solution is unavailable or the calculation is extremely difficult. The EM algorithm, a special iterative procedure, offers one solution to handle this problem. It makes use of an auxiliary function, the so called "Q-function", to deal with the missing observations. The Q-function plays an important role in the calculation of the complete-data log-likelihood $\log L_c(\theta)$, which is defined as

$$\log\, L_c(\theta) := \log(\, p(X, S) \,|\, \theta).$$

As a part of the data is unknown, $\log L_c(\theta)$ cannot be calculated directly. However, the complete-data likelihood can be handled in a reasonable way by estimating the distribution of the missing data S and calculating the expectation of the complete-data likelihood as follows.
Let $\theta^{(k)}$ be a predetermined parameter for the distribution $p(c|\theta)$ of the complete data. The Q-function is defined by

$$Q(\theta \,|\, \theta^{(k)}) := E_S \left[\log \left(L_c(\theta) \,|\, \theta^{(k)}, x_0^{\tau-1} \right) \right].$$

This function gives the expectation of the complete-data likelihood w.r.t. the density of the missing data, conditioned on the observations $x_{\tau-1}^0$ and the predetermined (set of) parameter(s) $\theta^{(k)}$. The Q-function depends only on $\theta \in \Theta$ and maps from Θ to $\mathbb{R}$.

In many cases, the Q-function cannot be directly used for explicit calculations if it is given in the form given above. However, it can be transformed in the

following way:

$$\begin{aligned} Q(\theta|\theta^{(k)}) &= E_S\left[\log\left(L_c(\theta)\,|\,\theta^{(k)}, x_{\tau-1}^0\right)\right] \\ &= E_S\left[\sum_{t=0}^{\tau-1}\log\left(p(x_t, s\,|\,\theta)\right)|\theta^{(k)}, x_0^{\tau-1})\right] \\ &= \int_s\left[\sum_{t=0}^{\tau-1}\log\left(p(x_t, s\,|\,\theta)\right)\right]\, p(s\,|\,\theta^{(k)}, x_0^{\tau-1})\, ds. \end{aligned} \tag{A.2}$$

The density $p(s\,|\,\theta^{(k)}, x_0^{\tau-1})$ represents another difficulty because an explicit expression is often difficult to obtain. However, it may be substituted by the density $p(s\,|\,\theta^{(k)}, x_0^{\tau-1})\cdot p(x_0^{\tau-1}\,|\,\theta^{(k)})$ which is obtained as a product of Equation (A.2) with $p(x_0^{\tau-1}\,|\,\theta^{(k)})$. The last mentioned expression is independent of θ and therefore does not affect the subsequent steps of the algorithm. The Q-function becomes

$$Q(\theta\,|\,\theta^{(k)}) = \int_s\left[\sum_{t=0}^{\tau-1}\log\left(p(x_t, s\,|\,\theta)\right)\right]\, p(s, x_0^{\tau-1}\,|\,\theta^{(k)})\, ds. \tag{A.3}$$

A.2 Implementation of the EM Algorithm

The EM algorithm is an iterative algorithm which splits up into the so-called E- and M-step.

1. Enter the observations $x_0^{\tau-1}$, the appropriate density functions of the corresponding Q-function, the initial value for θ, $\theta^{(0)}$, and a stopping criterion. The stopping criterion may be determined by the number of repetitions that have to be executed, or by a minimum increase of the log-likelihood.

2. This is the main iteration step of the estimation procedure. The E-step and the M-step are carried out in a loop.

   - **E-Step**: Compute the conditional expectation of the missing observations given the observed data and $\theta^{(k)}$. Then evaluate the complete-data log-likelihood by substituting the functions depending on S by the corresponding functions depending on the conditional expectation; this yields the expected log-likelihood. That is,

the Q-function from Equation (A.2) has to be calculated, considering $\theta^{(k)}$ as fixed parameters:

$$Q(\theta \mid \theta^{(k)}) := E_S \left[\log(L_c(\theta) \mid \theta^{(k)}, x_0^{\tau-1}) \right].$$

$Q(\theta, \theta^{(k)})$ depends on two arguments: The parameter θ that maximizes the likelihood in the next step, and $\theta^{(k)}$, which is used to evaluate the expectation.

- **M-Step**: To determine $\theta^{(k+1)}$s the Q-function has to be maximized w.r.t. θ:

$$\theta^{(k+1)} = \arg\max_{\theta} Q(\theta \mid \theta^{(k)}).$$

3. The two steps are repeated until the prescribed stopping criterion is fulfilled. This may be when the increase of the log-likelihood falls below a given limit or when a certain number of repetitions has been carried out.

Note that the M-step strongly depends on the distributions chosen for the data. It may happen that an explicit solution for the M-step cannot be obtained by straightforward calculus. In that case the algorithm may be slowed down by numerical approximations that are necessary in every iteration.

A.3 Convergence properties of the EM Algorithm

In this section we briefly state some general results for the convergence of the EM algorithm in short form. For a detailed outline see, e.g., Dempster et al. (1977), Little & Rubin (1987), Liporace (1982), and Wu (1983).

Let $\theta^{(k)}$ and $\theta^{(k+1)}$ denote two elements of the sequence of estimates obtained by the EM algorithm. The log-likelihood increases at each step until it reaches a stationary point, i.e.,

$$\log L(\theta^{(k+1)}) \geq \log L(\theta^{(k)}).$$

If bounded from above, the sequence $L(\theta^{(k)}))$ converges to some L^*. Furthermore, under certain technical conditions, L^* is a stationary value of the likelihood. To ensure that L^* is a stationary value, the Q-function must be continuous in both arguments. This holds true, e.g. in the case of the curved exponential family (Wu 1983).

In general the speed of convergence of the EM algorithm depends on the quantity of unknown information. For $\theta^{(k)}$ sufficiently close to the true parameter $\tilde{\theta}$, the convergence behavior can be described by

$$|\theta^{(k+1)} - \tilde{\theta}| = r_c |\theta^{(k)} - \tilde{\theta}|,$$

where r_c is given by the ratio of the missing to the complete information (cf. Dempster et al. 1977). Thus the rate of convergence can become very low if many observations are missing (see Little & Rubin 1987).

Another issue is the dependence on the initial value. Very often the log-likelihood function has multiple maxima. Hence the convergence of the EM algorithm depends strongly on the initial value (see e.g. Hasselblad 1966, Laird 1978). To increase the probability of obtaining good estimates, different initial values should be tried.

In the context of HMMs/HSMMs, the unobserved state sequence of the underlying (semi-)Markov chain are regarded as the missing data. The two steps of the EM algorithm can be implemented by means of the forward and backward probabilities. The effort required to estimate the parameters of the state-dependent distributions depends on the particular model used for the observation process. For example, models with t observation distributions, or negative binomial sojourn time distributions, require numerical maximization procedures (which may be avoided by the One-Step-Late algorithm of Green (1990)).

Appendix B

The Forward-Backward Algorithm

The forward-backward algorithm is a very elegant solution for the computation of the quantities required for the calculation of the E-step. The form presented here was introduced by Guédon (2003) and has a number of appealing properties: immunity to numerical underflow and, compared to most other approaches, a low computational complexity. Furthermore this implementation of the forward-backward algorithm calculates the smoothing probability

$$L_j(t) = P\left(S_t = j \,|\, X_0^{\tau-1} = x_0^{\tau-1}\right),\ t \in \{0, 1, \ldots, \tau - 1\}$$

directly during the backward pass through the observations. The sequence of observations is denoted by $x_0^{\tau-1} := x_0, \ldots, x_{\tau-1}$ and the hidden state sequence by $s_0^{\tau-1} := s_0, \ldots, s_{\tau-1}$.

The main components involved in the derivation are the quantities

- $F_j(t) = P\left(S_{t+1} \neq j, S_t = j \,|\, X_0^t = x_0^t\right),$
- $L1_j(t) = P\left(S_{t+1} \neq j, S_t = j \,|\, X_0^{\tau-1} = x_0^{\tau-1}\right),$ and
- $P\left(S_{t+1} = j, S_t \neq j \,|\, X_0^{\tau-1} = x_0^{\tau-1}\right).$

The first expression is calculated by the forward iteration, the latter two result from the backward iteration.

Forward Iteration

For $t = 0, \ldots, \tau - 2$ and $j = 0, \ldots, J - 1$, the forward iteration is given by

$$
\begin{aligned}
F_j(t) &= P\left(S_{t+1} \neq j, S_t = j \,|\, X_0^t = x_0^t\right) \\
&= \sum_{u=1}^{t} \sum_{i \neq j} P\left(S_{t+1} \neq j, S_{t-v} = j, v = 0, \ldots, u-1, S_{t-u} = i \,|\, X_0^t = x_0^t\right) \\
&\quad + P\left(S_{t+1} \neq j, S_{t-v} = j, v = 0, \ldots, t \,|\, X_0^t = x_0^t\right) \\
&= \sum_{u=1}^{t} \Bigg[\frac{P\left(X_{t-u+1}^t = x_{t-u+1}^t \,|\, S_{t-v} = j, v = 0, \ldots, u-1\right)}{P\left(X_{t-u+1}^t = x_{t-u+1}^t \,|\, X_0^{t-u} = x_0^{t-u}\right)} \\
&\qquad \times P\left(S_{t+1} \neq j, S_{t-v} = j, v = 0, \ldots, u-2 \,|\, S_{t-u+1} = j, S_{t-u} \neq j\right) \\
&\qquad \times \sum_{i \neq j} \Bigg\{ P\left(S_{t-u+1} = j \,|\, S_{t-u+1} \neq i, S_{t-u} = i\right) \\
&\qquad\qquad \times P\left(S_{t-u+1} \neq i, S_{t-u} = i \,|\, X_0^{t-u} = x_0^{t-u}\right) \Bigg\} \Bigg] \\
&\quad + \frac{P\left(X_0^t = x_0^t \,|\, S_{t-v} = j, v = 0, \ldots, t\right)}{P\left(X_0^t = x_0^t\right)} \\
&\qquad \times P\left(S_{t+1} \neq j, S_{t-v} = j, v = 0, \ldots, t\right) \\
&= \frac{b_j(x_t)}{N_t} \Bigg[\sum_{u=1}^{t} \left\{ \prod_{v=1}^{u-1} \frac{b_j(x_{t-v})}{N_{t-v}} \right\} d_j(u) \sum_{i \neq j} p_{ij} F_i(t-u) \\
&\qquad\qquad + \left\{ \prod_{v=1}^{t} \frac{b_j(x_{t-v})}{N_{t-v}} \right\} d_j(t+1)\, \pi_j \Bigg],
\end{aligned}
$$

where

$$
\begin{aligned}
b_j(x_t) &= P\left(X_t = x_t \,|\, S_t = j\right), \\
d_j(u) &= P\left(S_{t+u+1} \neq j, S_{t+u-v} = j, v = 0, \ldots u-2 \,|\, S_{t+1} = j, S_t \neq j\right), \text{ and} \\
N_t &= P\left(X_t = x_t \,|\, X_0^{t-1} = x_0^{t-1}\right).
\end{aligned}
$$

Backward Iteration

The backward recursion is based on the probabilities $L1_j(t)$, $t \in \{0, \ldots, \tau-2\}$, which are given by

$$\begin{aligned} L1_j(t) &= P\left(S_{t+1} \neq j, S_t = j \mid X_0^{\tau-1} = x_0^{\tau-1}\right) \\ &= \sum_{k \neq j} \Bigg[\sum_{u=1}^{\tau-2-t} P\left(S_{t+u+1} \neq k, S_{t+u-v} = k, v = 0, \ldots, u-1, \right. \\ &\qquad \left. S_t = j \mid X_0^{\tau-1} = x_0^{\tau-1}\right) \\ &\qquad + P\left(S_{\tau-1-v} = k, v = 0, \ldots, \tau-2-t, S_t = j \mid X_0^{\tau-1} = x_0^{\tau-1}\right) \Bigg]. \end{aligned} \tag{B.1}$$

According to Guédon (2003), the first term in equation (B.1) can be decomposed to

$$\begin{aligned} & P\left(S_{t+u+1} \neq k, S_{t+u-v} = k, v = 0, \ldots, u-1, S_t = j \mid X_0^{\tau-1} = x_0^{\tau-1}\right) \\ &= \frac{P\left(S_{t+u+1} \neq k, S_{t+u-v} = k, v = 0, \ldots, u-1, S_t = j, X_0^{\tau-1} = x_0^{\tau-1}\right)}{P\left(S_{t+u+1} \neq k, S_{t+u} = k, X_0^{\tau-1} = x_0^{\tau-1}\right)} \\ &\quad \times P\left(S_{t+u+1} \neq k, S_{t+u} = k \mid X_0^{\tau-1} = x_0^{\tau-1}\right) \\ &= \frac{P\left(X_{t+u+1}^{\tau-1} = x_{t+u+1}^{\tau-1} \mid S_{t+u+1} \neq k, S_{t+u} = k\right)}{P\left(X_{t+u+1}^{\tau-1} = x_{t+u+1}^{\tau-1} \mid S_{t+u+1} \neq k, S_{t+u} = k\right)} \\ &\quad \times \frac{P\left(S_{t+u+1} \neq k, S_{t+u} = k \mid X_0^{\tau-1} = x_0^{\tau-1}\right)}{P\left(S_{t+u+1} \neq k, S_{t+u} = k \mid X_0^{t+u} = x_0^{t+u}\right)} \\ &\quad \times \frac{P\left(X_{t+1}^{t+u} = x_{t+1}^{t+u} \mid S_{t+u-v} = k, v = 0, \ldots, u-1\right)}{P\left(X_{t+1}^{t+u} = x_{t+1}^{t+u} \mid X_0^t = x_0^t\right)} \\ &\quad \times P\left(S_{t+u+1} \neq k, S_{t+u-v} = k, v = 0, \ldots, u-2 \mid S_{t+1} = k, S_t \neq k\right) \\ &\quad \times P\left(S_{t+1} = k \mid S_{t+1} \neq j, S_t = j\right) P\left(S_{t+1} \neq j, S_t = j \mid X_0^t = x_0^t\right) \qquad \text{(B.2)} \\ &= \frac{L1_k(t+u)}{F_k(t+u)} \left\{ \prod_{v=0}^{u-1} \frac{b_k\left(x_{t+u-v}\right)}{N_{t+u-v}} \right\} d_k(u)\, p_{jk} F_j(t). \end{aligned}$$

The second term in (B.1), corresponding to the last visited state, can be decomposed using a similar argument. This yields

$$P\left(S_{\tau-1-v}=k, v=0,\ldots,\tau-2-t, S_t=j \,|\, X_0^{\tau-1}=x_0^{\tau-1}\right) = \left\{\prod_{v=0}^{\tau-2-t}\frac{b_k\left(x_{\tau-1-v}\right)}{N_{\tau-1-v}}\right\} D_k\left(\tau-1-t\right) p_{jk} F_j\left(t\right)$$

where $D_j\left(u\right)=\sum_{v\geq u} d_j\left(v\right)$ denotes the survivor function. Combining the two decompositions, $L1_j\left(t\right)$ becomes

$$L1_j\left(t\right) = \left[\sum_{k\neq j}\left[\sum_{u=1}^{\tau-2-t}\frac{L1_k\left(t+u\right)}{F_k\left(t+u\right)}\left\{\prod_{v=0}^{u-1}\frac{b_k\left(x_{t+u-v}\right)}{N_{t+u-v}}\right\} d_k\left(u\right) + \left\{\prod_{v=0}^{\tau-2-t}\frac{b_k\left(x_{\tau-1-v}\right)}{N_{\tau-1-v}}\right\} D_k\left(\tau-1-t\right)\right] p_{jk}\right] F_j\left(t\right).$$

The third term of equation (5.24) can also be transformed using a similar decomposition as in (B.2). We obtain

$$\begin{aligned} & P\left(S_{t+1}=j, S_t\neq j \,|\, X_0^{\tau-1}=x_0^{\tau-1}\right) \\ = & \sum_{u=1}^{\tau-2-t}\sum_{i\neq j} P\left(S_{t+u+1}\neq j, S_{t+u-v}=j, v=0,\ldots,u-1, S_t=i \,|\, X_0^{\tau-1}=x_0^{\tau-1}\right) \\ & + \sum_{i\neq j} P\left(S_{\tau-1-v}=j, v=0,\ldots,\tau-2-t, S_t=i \,|\, X_0^{\tau-1}=x_0^{\tau-1}\right) \\ = & \left[\sum_{u=1}^{\tau-2-t}\frac{L1_j\left(t+u\right)}{F_j\left(t+u\right)}\left\{\prod_{v=0}^{u-1}\frac{b_j\left(x_{t+u-v}\right)}{N_{t+u-v}}\right\} d_j\left(u\right) \right. \\ & \left. + \left\{\prod_{v=0}^{\tau-2-t}\frac{b_j\left(x_{\tau-1-v}\right)}{N_{\tau-1-v}}\right\} D_j\left(\tau-1-t\right)\right] \sum_{i\neq j} p_{ij} F_i\left(t\right). \end{aligned}$$

Appendix C

Source Code for the Estimation Procedures

The software for the estimation procedures of HSMMs is available on the homepage of the Institute for Statistics and Econometrics at *http://www.statoek.wiso.uni-goettingen.de*. It includes:

- The source code for the algorithms (C++).
- A dynamically linked library (dll) which can be loaded by the software package `R`.
- An interface written in `R`-code to access the dll.

Please take into account that the estimation algorithms are still under construction.

Appendix D

Notational Conventions and Abbreviations

DNM	:	Direct numerical maximization
EM	:	Expectation maximization
HMM(s)	:	Hidden Markov Model(s)
HSMM(s)	:	Hidden Semi-Markov Model(s)
MAE	:	Mean average error
MR	:	Mean reversion
MS	:	Markov switching
MSM	:	Markov switching Market
MSE	:	Mean squared error
TPM	:	Transition probability matrix
OLS	:	Ordinary least squares
KF	:	Kalman filter
RW	:	Random walk
$\mathbf{1}_{\{\dots\}}(\cdot)$	:	Indicator function
J	:	Number of hidden states
L	:	Likelihood
L_c	:	Complete-data Likelihood
$Q(\cdot)$	:	Q-function
$\{S_t\}$	:	Hidden/state process
$S_{t_1}^{t_2}$	:	Sequence $S_{t_1}, \dots, S_{t_2}$
$\boldsymbol{T}$	:	Transition probability matrix
$\{X_t\}$	:	Observed process
$X_{t_1}^{t_2}$	:	Sequence $X_{t_1}, \dots, X_{t_2}$
$(\cdot)^T$	:	Transposition
$\Gamma(\cdot)$	:	Gamma function
$\Psi(\cdot)$	:	Digamma function
τ	:	Number of observations

Bibliography

Aalen, O. O. & Husebye, E. (1991), 'Statistical analysis of repeated events forming renewal processes', *Statistics in Medicine* **10**, 1227–1240.

Archer, G. E. B. & Titterington, D. M. (2002), 'Parameter estimation for hidden Markov chains', *Journal of Statistical Planning and Inference* **108**(1-2), 365–390. C. R. Rao 80th birthday felicitation volume, Part II.

Barbu, V. & Limnios, N. (2005), 'Maximum likelihood estimation for hidden semi-markov models', *Comptes rendues de l'Académie des sciences* **Ser. I 342**, 201–205.

Baum, L. E. & Petrie, T. (1966), 'Statistical inference for probabilistic functions of finite state Markov chains', *Annals of Mathematical Statistics* **37**, 1554–1563.

Baum, L. E., Petrie, T., Soules, G. & Weiss, N. (1970), 'A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains', *Annals of Mathematical Statistics* **41**, 164–171.

Bialkowski, J. (2003), 'Modelling returns on stock indices for western and central european stock exchanges - markov switching approach', *South-Eastern Europe Journal of Economics* **2**(2), 81–100.

Bickel, P. J., Ritov, Y. & Rydén, T. (1998), 'Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models', *The Annals of Statistics* **26**(4), 1614–1635.

Bilmes, J. A. (1998), 'A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models'. International Computer Science Institute, Berkeley, California.

Blattberg, R. C. & Gonedes, N. J. (1974), 'A comparison of the stable and student distributions as statistical models for stock prices', *Journal of Business* **52**(1–2), 244–280.

Böhning, D. (1999), *Computer-assisted analysis of mixtures and applications*, Vol. 81 of *Monographs on Statistics and Applied Probability*, Chapman & Hall/CRC, Boca Raton, FL. Meta-analysis, disease mapping and others.

Böhning, D., Schlattmann, P. & Lindsay, B. (1992), 'Computer assisted analysis of mixtures (c.a.man): Statistical algorithms', *Biometrics* **48**, 283–303.

Bollerslev, T., Chou, R. Y. & Kroner, K. F. (1992), 'Arch modeling in finance: A review of the theory and empirical evidence', *Journal of Econometrics* **47**(2), 5–59.

Bos, T. & Newbold, P. (1984), 'An empirical investigation of the possibility of stochastic systematic risk in the market model', *Journal of Business* **57**(1), 35–41.

Bulla, J. & Berzel, A. (2006), Computational issues in parameter estimation for stationary hidden markov models. Submitted.

Bulla, J. & Bulla, I. (2006), Stylized facts of financial time series and hidden semi-markov models. To appear.

Campbell, J. Y., Lo, W. & Craig, M. A. (1997), *The Econometrics of Financial Markets*, Princeton University Press, Princeton, NJ, USA.

Campillo, F. & Le Gland, F. (1989), 'MLE for partially observed diffusions: direct maximization vs. the EM algorithm', *Stochastic Processes and their Applications* **33**(2), 245–274.

Cecchetti, S. G., Lam, P.-S. & Mark, N. C. (1990), 'Mean reversion in equilibrium asset prices', *The American Economic Review* **80**(3), 398–418.

Collins, D. W., Ledolter, J. & Rayburn, J. D. (1987), 'Some further evidence on the stochastic properties of systematic risk', *J. Bus.* **60**(3), 425–48.

Cox, D. R. (1975), 'Partial likelihood', *Biometrika* **62**(2), 269–276.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977), 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society. Series B. Methodological* **39**(1), 1–38. With discussion.

Dennis, Jr., J. E. & Moré, J. J. (1977), 'Quasi-Newton methods, motivation and theory', *SIAM Review* **19**(1), 46–89.

Devijver, P. A. (1985), 'Baum's forward-backward algorithm revisited', *Pattern Recognition Letters* **3**, 369–373.

Douc, R. & Matias, C. (2001), 'Asymptotics of the maximum likelihood estimator for general hidden Markov models', *Bernoulli. Official Journal of the Bernoulli Society for Mathematical Statistics and Probability* **7**(3), 381–420.

Douc, R., Moulines, É. & Rydén, T. (2004), 'Asymptotic properties of the maximum likelihood estimator in autoregressive models with Markov regime', *The Annals of Statistics* **32**(5), 2254–2304.

Dunmur, A. P. & Titterington, D. M. (1998), 'The influence of initial conditions on maximum likelihood estimation of the parameters of a binary hidden Markov model', *Statistics & Probability Letters* **40**(1), 67–73.

Durbin, R., Eddy, S. R., Krogh, A. & Mitchison, G. (1998), *Biological sequence analysis. Probabilistic models of proteins and nucleic acids*, Cambridge University Press, Cambridge, UK.

Efron, B. & Tibshirani, R. J. (1993), *An introduction to the bootstrap*, Vol. 57 of *Monographs on Statistics and Applied Probability*, Chapman and Hall, New York.

Elliott, R. J., Aggoun, L. & Moore, J. B. (1995), *Hidden Markov models*, Vol. 29 of *Applications of Mathematics (New York)*, Springer-Verlag, New York. Estimation and control.

Engle, R. F. (1982), 'Autoregressive conditional heteroskedasticity with estimates of the variance of UK inflation', *Econometrica* **50**, 987–1008.

Ephraim, Y. & Merhav, N. (2002), 'Hidden Markov processes', *Institute of Electrical and Electronics Engineers. Transactions on Information Theory* **48**(6), 1518–1569. Special issue on Shannon theory: perspective, trends, and applications.

Fabozzi, F. J. & Francis, J. C. (1978), 'Beta as a random coefficient', *Journal of Financial and Quantitative Analysis* **13**(1), 101–116.

Faff, R. W., Hillier, D. & Hillier, J. (2000), 'Time varying beta risk: An analysis of alternative modelling techniques', *Journal of Business Finance & Accounting* **27**(5), 523–554.

Fama, E. F. (1965), 'The behavior of stock-market prices', *Journal of Business* **38**(1), 34–105.

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

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

Fridman, M. (1994), A two state capital asset pricing model, Ima preprint series, Institute of Mathematics and its Applications, University of Minnesota.

Gettinby, G., Sinclair, C., Power, D. & Brown, R. (2004), 'An analysis of the distribution of extreme share returns in the UK from 1975 to 2000', *Journal of Business Finance & Accounting* **31**(5), 607–646.

Granger, C. W. & Ding, Z. (1995*a*), 'Some properties of absolute return: An alternative measure of risk', *Annales d'Économie et de Statistique* **40**, 67–91.

Granger, C. W. & Ding, Z. (1995*b*), Stylized facts on the temporal and distributional properties of daily data from speculative markets. Department of Economics, University of California, San Diego, unpublished paper.

Green, P. J. (1990), 'On use of the EM algorithm for penalized likelihood estimation', *Journal of the Royal Statistical Society. Series B. Methodological* **52**(3), 443–452.

Grimmett, G. R. & Stirzaker, D. R. (2001), *Probability and random processes*, third edn, Oxford University Press, New York.

Guédon, Y. (1992), 'Review of several stochastic speech unit models', *Computer Speech and Language* **6**, 377–402.

Guédon, Y. (1999), 'Computational methods for discrete hidden semi-Markov chains', *Applied Stochastic Models in Business and Industry* **15**(3), 195–224.

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

Guédon, Y. (2005), Hidden equilibrium semi-Markov chains. Preprint.

Guédon, Y. & Cocozza-Thivent, C. (1990), 'Explicit state occupancy modelling by hidden semi-markov models: Application of derin's scheme', *Computer Speech and Language* **4**, 167–192.

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

Hamilton, J. D. (1990), 'Analysis of time series subject to changes in regime', *Journal of Econometrics* **45**(1-2), 39–70.

Harris, R. D. & Küçüközmen, C. C. (2001), 'The empirical distribution of UK and US stock returns', *Journal of Business Finance & Accounting* **28**(5–6), 715–740.

Hasselblad, V. (1966), 'Estimation of parameters for a mixture of normal distributions', *Technometrics. A Journal of Statistics for the Physical, Chemical and Engineering Sciences* **8**, 431–446.

Hathaway, R. J. (1986), 'A constrained EM algorithm for univariate normal mixtures', *Journal of Statistical Computation and Simulation* **23**, 211–230.

Huang, H.-C. (2000), 'Tests of regimes-switching CAPM', *Applied Financial Economics* **10**, 573–578.

Jamshidian, M. & Jennrich, R. I. (1997), 'Acceleration of the EM algorithm by using quasi-Newton methods', *Journal of the Royal Statistical Society. Series B. Methodological* **59**(3), 569–587.

Kent, J. T., Tyler, D. E. & Vardi, Y. (1994), 'A curious likelihood identity for the multivariate t-distribution', *Communications in Statistics. Simulation and Computation* **23**(2), 441–453.

Koopman, S. J., Jungbacker, B. & Hol, E. (2005), 'Forecasting daily variability of the S&P 100 stock index using historical, realised and implied volatility measurements', **12**(3), 445–475.

Koski, T. (2001), *Hidden Markov models for bioinformatics*, Vol. 2 of *Computational Biology Series*, Kluwer Academic Publishers, Dordrecht.

Kulkarni, V. G. (1995), *Modeling and analysis of stochastic systems*, Texts in Statistical Science Series, Chapman and Hall Ltd., London.

Laird, N. (1978), 'Nonparametric maximum likelihood estimation of a mixed distribution', *Journal of the American Statistical Association* **73**(364), 805–811.

Lange, K. (1995), 'A quasi-Newton acceleration of the EM algorithm', *Statistica Sinica* **5**(1), 1–18.

Lange, K. & Weeks, D. E. (1989), 'Efficient computation of lod scores: genotype elimination, genotype redefinition, and hybrid maximum likelihood algorithms', *Annals of Human Genetics* **53**(1), 67–83.

Levinson, S. E. (1986), 'Continuously variable duration hidden markov models for automatic speech recognition', *Computer Speech and Language* **1**, 29–45.

Linne, T. (2002), A markov switching model of stock returns: an application to the emerging markets in Central and Eastern Europe, *in* 'East European Transition and EU Enlargement', Physica-Verlag, pp. 371–384.

Lintner, J. (1965), 'The valuation of risky assets and the selection of risky investments in stock portfolios and capital budgets', *Review of Economics and Statistics* **47**, 13–37.

Liporace, L. A. (1982), 'Maximum likelihood estimation for multivariate observations of Markov sources', *Institute of Electrical and Electronics Engineers. Transactions on Information Theory* **28**(5), 729–734.

Little, R. J. A. & Rubin, D. B. (1987), *Statistical analysis with missing data*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons Inc., New York.

MacDonald, I. L. & Zucchini, W. (1997), *Hidden Markov and other models for discrete-valued time series*, Vol. 70 of *Monographs on Statistics and Applied Probability*, Chapman & Hall, London.

Matsumoto, M. & Nishimura, T. (1998), 'Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator', *ACM Transactions on Modeling and Computer Simulation* **8**(1), 3–30.

Mengersen, K. L. & Robert, C. P. (1996), 'Testing for mixtures: a Bayesian entropic approach', pp. 255–276.

Mengersen, K. L. & Robert, C. P. (1999), 'Reparameterisation issues in mixture modelling and their bearing on MCMC algorithms', *Computational Statistics & Data Analysis* **29**(3), 325–343.

Mergner, S. & Bulla, J. (2005), Time-varying beta risk of pan–european industry portfolios: A comparison of alternative modeling techniques. Submitted.

Mittnik, S. & Rachev, S. T. (1993), 'Modeling asset returns with alternative stable distributions', *Econometric Reviews* **12**(3), 261–389.

Nelder, J. A. & Mead, R. (1965), 'A simplex method for function minimization', *Computer Journal* **7**, 308–313.

Nityasuddhi, D. & Böhning, D. (2003), 'Asymptotic properties of the EM algorithm estimate for normal mixture models with component specific variances', *Computational Statistics & Data Analysis* **41**(3-4), 591–601.

Parzen, E. (1962), *Stochastic processes*, Holden-Day Series in Probability and Statistics, Holden-Day Inc., San Francisco, Calif.

Peel, D. & McLachlan, G. J. (2000), 'Robust mixture modelling using the t-distribution', *Stat. Comput.* **10**, 339–348.

Praetz, P. D. (1972), 'The distribution of share price changes', *Journal of Business* **45**(1), 49–55.

R Development Core Team (2005), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. **URL:** *http://www.R-project.org*

Rabiner, L. (1989), 'A tutorial on hidden markov models and selected applications in speech recognition', *Institute of Electrical and Electronics Engineers. Transactions on Information Theory* **77**(2), 257–284.

Redner, R. A. & Walker, H. F. (1984), 'Mixture densities, maximum likelihood and the EM algorithm', *SIAM Review. A Publication of the Society for Industrial and Applied Mathematics* **26**(2), 195–239.

Robert, C. P. & Titterington, D. M. (1998), 'Reparameterization strategies for hidden markov models and bayesian approaches to maximum likelihood estimation', *Statistics and Computing* **8**, 145–158.

Rydén, T. (1995*a*), 'Consistent and asymptotically normal parameter estimates for Markov modulated Poisson processes', *Scandinavian Journal of Statistics. Theory and Applications* **22**(3), 295–303.

Rydén, T. (1995*b*), 'Estimating the order of hidden Markov models', *Statistics. A Journal of Theoretical and Applied Statistics* **26**(4), 345–354.

Rydén, T., Terasvirta, T. & Asbrink, S. (1998), 'Stylized facts of daily return series and the hidden markov model', *Journal of Applied Econometrics* **13**(3), 217–244.

Sansom, J. & Thomson, P. (2000), 'Fitting hidden semi-Markov models', *NIWA Technical Report* **NTR77**. National Institute of Water and Atmospheric Research, New Zealand.

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

Schnabel, R. B., Koontz, J. E. & Weiss, B. E. (1985), 'A modular system of algorithms for unconstrained minimization', *Association for Computing Machinery. Transactions on Mathematical Software* **11**(4), 419–440.

Seneta, E. (1981), *Nonnegative matrices and Markov chains*, Springer Series in Statistics, second edn, Springer-Verlag, New York.

Sharpe, W. F. (1964), 'Capital asset prices: A theory of market equilibrium under conditions of risk', *Journal of Finance* **19**, 425–442.

STO (2004), *STOXX Ltd.'s shift to Industry Classification Benchmark (ICB)*. Available at http://www.stoxx.com/indexes/stoxx_shifto_icb.pdf.

Sunder, S. (1980), 'Stationarity of market risk: Random coefficients tests for individual stocks', *Journal of Finance* **35**(4), 883–896.

Taylor, S. J. (1986), *Modelling Financial Time Series*, John Wiley & Sons, Chichester, UK.

Turner, C. M., Startz, R. & Nelson, C. R. (1989), 'A markov model of heteroskedasticity, risk, and learning in the stock market', *Journal of Financial Economics* **25**(1), 3–22.

Visser, I., Raijmakers, M. E. J. & Molenaar, P. C. M. (2000), 'Confidence intervals for hidden markov model parameters', *British Journal of Mathematical & Statistical Psychology* **53**(2), 317–327.

Wang, P. & Puterman, M. L. (2001), 'Analysis of longitudinal data of epileptic seizure counts – a two-state hidden Markov regression approach', *Biometrical Journal. Methods. Applications. Case Studies. Regulatory Aspects. Reviews* **43**(8), 941–962.

Wu, C.-F. J. (1983), 'On the convergence properties of the EM algorithm', *The Annals of Statistics* **11**(1), 95–103.

Yao, J. & Gao, J. (2004), 'Computer-intensive time-varying model approach to the systematic risk of Australian industrial stock returns', *Australian Journal of Management* **29**(1), 121–146.

Yu, J. (2002), 'Forecasting volatility in the New Zealand stock market', *Applied Financial Economics* **12**, 193–202.

Yu, S.-Z. & Kobayashi, H. (2003), 'An efficient forward–backward algorithm for an explicit-duration hidden markov model', *IEEE Signal Processing Letters* **10**(1), 11–14.

Zucchini, W. & MacDonald, I. L. (1998), 'Hidden markov time series models: Some computational issues', *Computing Science and Statistics* **30**, 157–163. Interface Foundation of North America.