National and Kapodistrian University of Athens

Master's Thesis

Hidden Semi-Markov and applications in Time series

Author:
Nikolaos Motis

Supervisor:
Samis Trevezas

A thesis submitted in partial fulfillment of the requirements for the degree of M.Sc. in Statistics and Operations Research

in the

Faculty of Sciences
Department of Mathematics

December 18, 2019

NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

Abstract

Faculty of Sciences
Department of Mathematics

M.Sc. in Statistics and Operations Research

Hidden Semi-Markov and applications in Time series

by Nikolaos Motis

In the context of this thesis, the discrete semi-Markov models will be presented, so as to connect them thereafter with the hidden semi Markov models. After this presentation, statistics estimations methods with basic tool the EM algorithm (for whom a brief description is given) will be pointed out. A detailed presentation of the latter, under specific assumptions whether on component distribution or on sojourn time distribution in hidden states, is followed. In particular, observation component distribution such as normal and student are under review, whereas, as far as for the sojourn time in hidden states are concerned, negative binomial is examined. Finally, hidden semi markov models are fitted in the prices of index SnP 500 exploring how effectively these models reproduce the stylized facts that Granger and Ding pointed out.

Περίληψη

Στα πλαίσια αυτής της διπλωματικής θα περιγραφούν αρχικά τα διακριτά ημι-μαρκοβιανά μοντέλα, ώστε στη συνέχεια να γίνει η σύνδεσή τους με τα κρυμμένα ημι-μαρκοβιανά μοντέλα. Μετά από αυτή την παρουσίαση θα αναδειχθούν μέθοδοι στατιστικής εκτίμησης με βασικό εργαλείο τον αλγόριθμο EM (Expectation-Maximization) για τον οποίο γίνεται μια σύντομη περιγραφή. Έπειτα θα γίνει λεπτομερής παρουσίαση του τελευταίου με συγκεκριμένες υποθέσεις είτε στις κατανομές των παρατηρούμενων καταστάσεων είτε στις κατανομές των χρόνων παραμονής στις κρυμμένες καταστάσεις. Ειδικότερα εξετάζονται κατανομές παρατήρησης όπως η κανονική και η student, ενώ για το χρόνο παραμονής στις κρυμμένες καταστάσεις εφαρμόζεται η αρνητική διωνυμική . Σε τελευταίο βήμα θα γίνει εφαρμογή σε πραγματικά δεδομένα από τον δείκτη SnP 500, η οποία θα πραγματοποιηθεί με τη χρήση του λογισμικού R, εξερευνώντας τα εμπειρικά ευρήματα (Stylized facts) που διατύπωσαν ο Granger και ο Ding και θα εξεταστεί κατά πόσο τα κρυμμένα ημι-μαρκοβιανά μοντέλα, αναπαράγουν αυτά τα ευρήματα.

Acknowledgements

Θα ήθελα να ευχαριστήσω όλους τους καθηγητές μου για τις γνώσεις που μου χαρίσανε, για όλες τις στιγμές που περάσαμε στις αίθουσες αλλά πάνω όλα για την μύηση στην μαθηματική σκέψη , κάτι που μένει, οι γνώσεις πάνε και έρχονται η μόρφωση όμως είναι κάτι που μένει και σε ακολουθεί. Θα ήθελα συγκεκριμένα να ευχαριστώ τον επιβλέποντα Δρ. Σάμη Τρέβεζα για όλη την βοήθεια και καθοδήγηση που μου παρήχε. Ακόμη θα ήθελα να ευχαριστήσω τον καθηγητή Απόστολο Μπουρνετα και την επίκουρη καθηγήτρια Λουκία Μελιγκοτσίδου που με τίμησαν με την συμμετοχή τους στην εξεταστική επιτροπή αλλά και γενικότερα για την καθοδήγηση που μου προσέφεραν καθ' όλη την διάρκεια των σπουδών μου σε προπτυχιακό και μεταπτυχιακό επίπεδο. Τέλος θα ήθελα να ευχαριστήσω τους γονείς μου για όλη την υποστήριξη που μου παρέχουν.

Contents

Abstract iii

Acknowledgements v

1 Semi-Markov and Hidden Semi-Markov models 1
- 1.1 Markov renewal chains and semi-markov processes . . . 1
- 1.2 Hidden Semi-Markov Models . . . 4
- 1.3 The Likelihood Function of a Hidden Semi-Markov Model . . . 5
- 1.4 The Partial Likelihood Estimator . . . 6
- 1.5 The Complete Likelihood Estimator . . . 7

2 Estimation in a Hidden Semi-Markov model 9
- 2.1 The EM Algorithm . . . 9
- 2.2 The Q function of an HSMM . . . 11
- 2.3 The Forward-Backward Algorithm . . . 13
  - 2.3.1 HMM-Model . . . 13
    - **The forward iteration for an HMM** . . . 13
    - The Backward Iteration for an HMM . . . 14
  - 2.3.2 HSMM-Model . . . 15
    - The forward iteration . . . 15
    - The Backward Iteration . . . 18
- 2.4 The Sojourn Time Distribution . . . 20
- 2.5 Parameter Re-estimation . . . 21
  - The initial Parameters . . . 22
  - The Transition Probabilities . . . 22
  - Non-Parametric State Occupancy Distribution with the Complete Likelihood Estimator . . . 22
  - Geometric State Occupancy distribution . . . 23
  - Negative Binomial State Occupancy Distribution . . . 23
  - The Observation Component . . . 24

3 Application to Real data of the SnP 500 31
- 3.1 Stylized facts . . . 31
- 3.2 Data . . . 32
- 3.3 Descriptive Statistics . . . 33
- 3.4 The models . . . 33
  - 3.4.1 HMM . . . 38
  - 3.4.2 HSMM . . . 38
- 3.5 Conclusion . . . 40

Στην μνήμη του πατέρα μου Αντώνη
18/1/2019

Chapter 1

Semi-Markov and Hidden Semi-Markov models

1.1 Markov renewal chains and semi-markov processes

Consider a random system with finite state space E={1,...,J}. We denote by M_E the set of real matrices on $E \times E$ and by $M_E(\mathbb{N})$ the set of matrix valued functions defined on $\mathbb{N}$, with values in M_E. For $A \in M_E(\mathbb{N})$, we write A = (A(u); $u \in \mathbb{N}$), where, for $u \in \mathbb{N}$ fixed, $A(u) = (A_{ij}(u); i, j \in E) \in M_E$. Set $I_E \in M_E$ for the identity matrix and $0_E \in M_E$ for the null matrix. When the space E is clear from the context, we will write I and 0 instead of I_E and 0_E.

We assume that the time evolution of the system is described by the following chains:

- The chain $J = (J_n)_{n\in\mathbb{N}}$ with state space E, where J_n is the system state at the nth jump time.
- The chain $R = (R_n)_{n\in\mathbb{N}}$ where R_n is the nth jump time.
- The chain $U = (U_n)_{n\in\mathbb{N}}$ where $U_n := R_n - R_{n-1}$, $n \geqslant 1$ is the sojourn time in state J_{n1}, before the nth jump.

Definition 1.1.1. *A matrix-valued function $\boldsymbol{q} = (q_{ij}(u)) \in M_E(\mathbb{N})$ is said to be a discrete-time semi-Markov kernel if it satisfies the following three properties:*

1. $0 \leqslant q_{ij}(u), i, j \in E, u \in \mathbb{N}$.
2. $q_{ij}(0) = 0, i, j \in E$.
3. $$\sum_{u=0}^{\infty} \sum_{j\in E} q_{ij}(u) = 1, i \in E.$$

Definition 1.1.2. *The chain $(J, R) = (J_n, R_n)_{n\in\mathbb{N}}$ is said to be a Markov renewal chain if $\forall n \in \mathbb{N}, \quad \forall i, j \in E, \quad \forall u \in \mathbb{N}$ it satisfies almost surely :*

$$P(J_{n+1} = j, R_{n+1} - R_n = u | J_0, ...J_n, R_0, ...R_n) = P(J_{n+1} = j, R_{n+1} - R_n = u | J_n). \tag{1.1.1}$$

Moreover, if Equation (1.1.1) is independent of n, then (J, R) is said to be homogeneous and the discrete-time semi-Markov kernel q is defined by

$$q_{ij}(u) := P(J_{n+1} = j, R_{n+1} - R_n = u | J_n = i).$$

Remark 1.1.1.

1. *Note that, if (J,R) is a (homogeneous) Markov renewal chain, we can easily see that $(J_n)_{n\in\mathbb{N}}$ is a (homogeneous) Markov chain, called the embedded Markov chain (EMC) associated with the MRC (J,R). We denote by*
   $p=(p_{ij})_{i,j\in E}\in M_E$ the transition matrix of (J_n), defined by

$$p_{ij}:=P(J_{n+1}=j|J_n=i), i,j\in E, n\in\mathbb{N}.$$

2. *The transition probabilities p_{ij} can be expressed in terms of the semi Markov kernel by*

$$p_{ij}=\sum_{u=0}^{\infty}q_{ij}(u). \tag{1.1.2}$$

3. *The second property of the definition of the semi- Markov kernel specifies that $q_{ij}(0)=0$, for all $i,j\in E$. The interpretation of this property is that the instantaneous transitions are not allowed. This is a direct consequence of the fact that the chain $(R_n)_{n\in\mathbb{N}}$ is supposed to be increasing $(0=R_0<R_1<R_2<...)$ or, equivalently, the random variables $U_n, n\in\mathbb{N}^*$, are strictly positive. Nor do we allow transitions to the same state, i.e., $p_{ii}=0, i\in E$.*

Let us also introduce some other interesting characteristics.

Definition 1.1.3. *Let $Q=(Q(u);u\in\mathbb{N})$ be the matrix-valued function defined by*

$$Q_{ij}(u):=P(J_{n+1}=j,U_{n+1}u|J_n=i)=\sum_{l=0}^{u}q_{ij}(l). \tag{1.1.3}$$

for all $i,j\in E$ and $u\in\mathbb{N}$.

It is called the cumulated semi-Markov kernel and it expresses the probability that the system starting from the state i will move to the state j in at most u time units.

When investigating the evolution of a Markov renewal chain we are interested in two types of holding time distributions: the sojourn time distributions in a given state and the conditional distributions depending on the next visited state .

Definition 1.1.4. *For all $i,j\in E$, let us define:*

1. *$f_{ij}()$, the conditional distribution of U_{n+1} given on the current state $J_n=i$ and the next visited state $J_{n+1}=j$. In particular, the probability function of this distribution is given by :*

$$f_{ij}(u):=P(U_{n+1}=u|J_n=i,J_{n+1}=j), u\in\mathbb{N}. \tag{1.1.4}$$

2. *$F_{ij}()$, the conditional cumulative distribution of $X_{n+1}, n\in\mathbb{N}$ given that $J_n=i$ and $J_{n+1}=j$. In particular,this function is given by :*

$$F_{ij}(u):=P(U_{n+1}\leqslant u|J_n=i,J_{n+1}=j)=\sum_{l=0}^{u}f_{ij}(l), u\in\mathbb{N}. \tag{1.1.5}$$

Obviously, for all $i,j\in E$ and for all $u\in\mathbb{N}$, we have

$$f_{ij}(u)=\begin{cases}\frac{q_{ij}(u)}{p_{ij}}, & \text{if} \quad p_{ij}\neq 0\\ 0, & \text{otherwise}\end{cases} \tag{1.1.6}$$

Definition 1.1.5. *For all $i \in E$, let us denote by:*

1. *$d_i()$ the sojourn time distribution in state i:*

$$d_i(u) := P(U_{n+1} = u | J_n = i) = \sum_{j \in E} q_{ij}(u), \quad u \in \mathbb{N}. \tag{1.1.7}$$

2. *$D_i()$ the sojourn time cumulative distribution in state i:*

$$D_i(u) := P(U_{n+1} \leqslant u | J_n = i) = \sum_{l=1}^{u} d_i(l), u \in \mathbb{N}. \tag{1.1.8}$$

As we saw in Equation (1.1.6), the semi-Markov kernel introduced in Definition 1.1.2 verifies the relation $q_{ij}(u) = p_{ij} f_{ij}(u)$.

We can also define two particular semi-Markov kernels for which $f_{ij}(u)$ does not depend on i or j, by setting the semi-Markov kernels of the form $q_{ij}(u) = p_{ij}\widetilde{f_j}(u)$ or $q_{ij}(u) = p_{ij} f_i(u)$. In particular, if

$$f_{ij}(u) = P(U_{n+1} = u | J_n = i, J_{n+1} = j) = P(U_{n+1} = u | J_n = i)$$

then there is no dependence on j. Note that this $f_i()$ is simply $d_i()$, the sojourn time distribution in state i, as defined above. These particular types of Markov renewal chains could be adapted for some applications, where practical arguments justify that the sojourn times in a state only depend on the current state or on the next visited state. Note also that these particular Markov renewal chains can be obtained by transforming the general Markov renewal chain (i.e., with the kernel $q_{ij}(u) = p_{ij} f_{ij}(u)$).

Example 1.1.1. *From a Markov chain we can have a particular case of a MRC with semi-Markov kernel*

$$q_{ij}(u) = \begin{cases} p_{ij}(p_{ii})^{u-1} & , if \quad i \neq j \quad and \quad u \in \mathbb{N}^*, \\ 0 & , elsewhere. \end{cases} \tag{1.1.9}$$

Proof:
Let Z_t be a Markov chain with transition matrix

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1S} \\ \vdots & \ddots & & \vdots \\ p_{S1} & p_{S2} & \cdots & p_{SS} \end{bmatrix}.$$

Let J_n be the system state at the nth jump time, R_n be the nth jump time, then

$$\begin{aligned} P(J_{n+1} = j, R_{n+1} - R_n = u | J_0, ..., J_n = i; R_0, ...R_n) = \\ = P(J_{n+1} = j, R_{n+1} - R_n = u | J_n = i) = \\ = P(Z_{R_n+1} = i, ..., Z_{R_n+u-1} = i, Z_{R_n+u} = j | Z_{R_n} = i) = \\ = P(Z_{R_n+1} = i | Z_{R_n} = i)...P(Z_{R_n+u} = j | Z_{R_n+u-1} = i) = \\ = p_{ii}^{u-1} p_{ij}. \end{aligned} \tag{1.1.10}$$

Definition 1.1.6. *Let (J, R) be a Markov renewal chain. The chain $S = (S_t)_{t \in \mathbb{N}}$ is said to be a semi-Markov chain associated with the MRC (J, R) if*

$$S_t := J_{N(t)}, t \in \mathbb{N},$$

where

$$N(t) := \max\{t \in \mathbb{N} | R_n \leqslant t\}, \tag{1.1.11}$$

corresponds to the discrete-time counting process of the number of jumps in [1, t] $\subset \mathbb{N}$. *Thus* S_t *gives the system state at time t. We have also* $J_n = S_{R_n}$ *and* $R_n = \min\{k > R_{n-1} | S_k \neq S_{k-1}\} n \in \mathbb{N}$.

It can be easily shown that $(J_n)_{n\in\mathbb{N}}$ is a Markov chain, also $(S_t)_{t\in\mathbb{N}}$ preserves the Markov property at least at jump times $(R_n)_{n\in\mathbb{N}}$. For this reason, J_n is known as an embedded Markov chain and that's why $(S_n)_{n\in\mathbb{N}}$ is referred to as a semi-Markov chain. In what follows the semi-Markov chain is defined by the semi-Markov kernel

$$q_{ij}(u) = p_{ij} f_i(u), \quad \forall \quad i, j \in E, \tag{1.1.12}$$

so the sojourn time in a state depends only on the current state.

1.2 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.

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 necessarily follows a geometric distribution as a consequence of the Markov property of the underlying Markov chain. In the sequel, 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.

An HSMM consists of a pair of discrete-time stochastic processes $\{S_t\}$ and $\{X_t\}$, observed at times $t = 0, \ldots, T-1$, so T corresponds to the observations length of the processes. The observed process $\{X_t\}$ is linked to the hidden, unobserved state process $\{S_t\}$ by the conditional distribution depending on the state process, where S_t is a finite 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.

In the discrete case the output process $\{X_t\}$ is related to the semi-Markov chain $\{S_t\}$ by the observation (or emission) probabilities

$$b_j(x_t) = P(X_t = x_t | S_t = j),$$

where $\sum_{x_t} b_j(x_t) = 1$.

The observation process is characterized by the conditional independence property,

$$P(X_0^{T-1} = x_0^{T-1} | S_0^{T1} = s_0^{T1}) = \prod_{t=0}^{T-1} P(X_t = x_t | S_t = s_t),$$

where

$$\{X_{t_0}^{t_1} = x_{t_0}^{t_1}\} := \{X_{t_0} = x_{t_0}, ..., X_{t_1} = x_{t_1}\},$$
$$\{S_{t_0}^{t_1} = s_{t_0}^{t_1}\} := \{S_{t_0} = s_{t_0}, ..., S_{t_1} = s_{t_1}\},$$

which implies the fact that the output process at time t only depends on the state of the underlying semi-Markov chain at time t.

1.3 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 the following, we consider the case of a single observed sequence which is relevant for later applications. As a first step we consider the classical form of the complete-data likelihood $\tilde{L_c}$, introduced by Ferguson (1980), which only allows 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^{T-1} and the states s_0^{T-1} of the underlying semi-Markov chain are known, and thus

$$\tilde{L_c}(s_0^{T-1}, x_0^{T-1} | \theta) = P(S_0^{T-1} = s_0^{T-1}, X_0^{T-1} = x_0^{T-1} | \theta)$$

$$= P(S_0^{T-1} = s_0^{T-1} | \theta) P(X_0^{T-1} = x_0^{T-1} | S_0^{T-1} = s_0^{T-1}, \theta)$$

$$= \pi_{\tilde{s_0}} d_{\tilde{s_0}}(u_0) \prod_{r=1}^{R} p_{\tilde{s}_{r-1}\tilde{s_r}} d_{\tilde{s_r}}(u_r) \prod_{t=0}^{T-1} b_{s_t}(x_t), \tag{1.3.1}$$

where the number of visited states, R+1, is a fixed, but unknown number, $\tilde{s_r}$ is the $(r+1)^{th}$ visited state , u_r denotes the time spent in state $\tilde{s_r}$, θ denotes the vector of all parameters and $d_j(u)$ corresponds to the probability of staying at state j for exactly u time units, that is,

$$d_j(u) := P(S_{t+u+1} \neq j, S_{t+u-v} = j, v = 0, \ldots, u-1 | S_{t+1} = j, S_t \neq j). \tag{1.3.2}$$

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 T, that is, we consider state sequences of the form:

$$\pi_{\tilde{s}_0} d_{\tilde{s}_0}(u_0) \prod_{r=1}^{R} p_{\tilde{s}_{r-1}\tilde{s_r}} d_{\tilde{s}_r}(u_r) \mathbf{1}_{\{\sum_r u_r = T\}}(u_r)_{r\geq 1} \tag{1.3.3}$$

Actually,for all possible R,states $\tilde{s}_0, \tilde{s}_1, \ldots, \tilde{s_R}$ and durations $u_0, u_1, \ldots, u_R$,the indicator function $\mathbf{1}_{\{\sum_r u_r = T\}}(u_r)_{r\geq 1}$ 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 $T-1$ is obtained by enumeration of the complete-data likelihood over all possible state sequences, which yields

$$\tilde{L}(\theta) = \sum_{s_0, \ldots, s_{T-1}} \tilde{L_c}(s_0^{T-1}, x_0^{T-1} | \theta). \tag{1.3.4}$$

In this representation, the difficulty of solving equation (1.3.4) explicitly is obvious: the sum includes all admissible paths of the form given by Equation (1.3.3), which effectively eliminates any chance of obtaining an analytic solution.

1.4 The Partial Likelihood Estimator

The standard formulation (1.3.3) 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 to T. 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 a first step to generalize the classical approach, Guédon (2003) proposed to write the contribution of the state sequence to the complete-data likelihood as

$$\pi_{\tilde{s_0}} d_{\tilde{s_0}}(u_0) \{ \prod_{r=1}^{R-1} p_{\tilde{s}_{r-1}\tilde{s_r}} d_{\tilde{s_r}}(u_r) \} p_{\tilde{s}_{R-1}\tilde{s}_R} \bar{D}_{\tilde{s}_R}(u_R) \mathbf{1}_{\{\sum_r u_r = T\}}(u_r)_{r\geq 0}, \tag{1.4.1}$$

where

$$\bar{D}_j(u) := \sum_{v \geqslant u} d_j(u).$$

The main difference with the classical approach presented in Equation (1.3.3) lies in the substitution of the ordinary sojourn time probability by the survival function $\bar{D}$ for the last visited state. The survival function performs a right-censoring of the sojourn time in the last visited state. In the partial likelihood estimator, the contribution from the survival function is ignored.

1.5 The Complete Likelihood Estimator

In this setting the complete-data likelihood incorporates both the outputs x_0^{T-1} and the state sequence s_0^{T1} . The difference with 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_{T-1} from time $T-1$ to $T-1+u$, $u = 0, 1, \ldots$. The exit from the last visited state takes place at time $T-1+u$, which yields the complete-data likelihood of the underlying semi-Markov chain

$$L_c(s_0^{T-1+u}, x_0^{T-1}|\theta) = P(S_0^{T-1} = s_0^{T-1}, S_{T-1} + v = s_{T-1}, v = 1, ..., u-1 \\ , S_{T-1+u} \neq s_{T-1}, X_0^{T-1} = x_0^{T-1}|\theta).$$

The estimator based on this specification of the complete-data problem is called complete likelihood estimator. Compared to formula (1.4.1), the contribution of the state sequence to the complete-data likelihood has to be modified to

$$\pi_{\tilde{s_0}} d_{\tilde{s_0}}(u_0) \prod_{r=0}^{R-1} p_{\tilde{s}_{r-1}\tilde{s}_r} d_{\tilde{s}_r}(u_r) \mathbf{1}_{\{\sum_{r=0}^{R-1} u_r < T \leq \sum_{r=0}^{R} u_r\}}(u_0, ..., u_r). \tag{1.5.1}$$

Compared to the original likelihood given by Equation (1.5.1), the completed state sequence complicates the likelihood function by an additional sum over all possible prolongations of the state sequence, that is,

$$L(\theta) = \sum_{s_0, \ldots s_{T-1}} \sum_{u_{T+}} L_c(s_0^{T-1+u}, x_0^{T-1}|\theta). \tag{1.5.2}$$

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.

Chapter 2

Estimation in a Hidden Semi-Markov model

2.1 The EM Algorithm

The estimation problem in HSMMs corresponds to an incomplete-data problem, since the underlying path of the hidden semi-Markov chain remains inaccessible and only a part of the data related to another observable process are accessible. Therefore, estimation via the EM algorithm is a suitable way to perform maximum-likelihood estimation in HSMMs. For this reason, we first introduce the basic principles of the EM algorithm.

Let Y be the random vector corresponding to the observed data y, with p.d.f. denoted by $g(y;\theta)$, where $\theta = (\theta_1, ..., \theta_d)$ is a vector of unknown parameters with parameter space Ω . The EM algorithm is a broadly applicable algorithm that provides an iterative procedure for computing MLE's in situations where the presence of some additional data would render the problem of ML estimation straightforward. Hence, in this context, the observed data vector y can be viewed as being incomplete and thus be regarded as an observable function of the so-called complete data. The notion of incomplete data includes the conventional sense of missing data, but it also applies to situations where the complete data represent what would be available from some hypothetical experiment. In the latter case, the complete data may contain some variables that are never observable in a data sense, but are added only artificially in order to facilitate the estimation procedure. When a problem does not at first appear to be an incomplete-data one, computation of the MLE is often greatly facilitated by artificially formulating it to be as such. This is because the EM algorithm exploits the reduced complexity of ML estimation given the complete data. For many statistical problems the complete-data likelihood has a nice form.

Within this framework, let x denote the vector containing the augmented or the so-called complete data, and let z denote the vector containing the additional data, referred to as the unobservable, or missing, or latent data. We let $g_c(x;\theta)$ denote the p.d.f. of the random vector X corresponding to the complete data vector x. The complete-data log-likelihood function is given by

$$\log L_c(\theta) = \log g_c(x;\theta).$$

Formally, we have two sample spaces $\mathscr{X}$ and $\mathscr{Y}$ and a many-to-one mapping from $\mathscr{X}$ to $\mathscr{Y}$. Instead of observing the complete-data vector x in $\mathscr{X}$, we observe the incomplete-data vector $y = y(z)$ in $\mathscr{Y}$. It follows that

$$g(y;\theta) = \int_{\mathscr{X}(y)} g_c(x;\theta)dx,$$

where $\mathscr{X}(y)$ is the subset of $\mathscr{X}$ determined by the equation $y = y(x)$.

The EM algorithm intends to solve the problem of maximising the incomplete-data likelihood indirectly by proceeding iteratively in terms of the complete-data log–likelihood function, $\log L_c(\theta)$. Since the complete data are unobservable, and hence the complete-data log–likelihood function is a random variable (for a given value of θ), it is replaced by its conditional expectation given the observed data y, where the expectation is computed under the current fit for θ.

More specifically, let θ^0 be some initial value for θ. Then, at the first iteration, the E-step requires the computation of

$$Q(\theta;\theta^0) = E_{\theta^{(0)}}\{\log L_c(\theta)|y\}. \tag{2.1.1}$$

The M-step requires the maximization of $Q(\theta;\theta^0)$ with respect to θ over the parameter space Ω, that is, we choose θ^1 such that

$$Q(\theta^1;\theta^0) \geq Q(\theta;\theta^0),$$

for all $\theta \in \Omega$. The E- and M-steps are then carried out again, but this time with $\theta^{(0)}$ replaced by the current fit $\theta^{(1)}$.

The iterative scheme of the EM-algorithim can be summarized as follows:

- **E-step**: Calculate $Q(\theta;\theta^{(k)})$ where

$$Q(\theta;\theta^{(k)}) = E_{\theta^{(k)}}\{\log L_c(\theta)|y\}.$$

- **M-Step**: Choose $\theta^{(k+1)}$ to be any value of $\theta \in \Omega$ that maximizes $Q(\theta;\theta^{(k)})$ that is,

$$Q(\theta^{(k+1)};\theta^{(k)}) \geq Q(\theta;\theta^{(k)}),$$

for all $\theta \in \Omega$.

- Repeat until $L(\theta^{(k+1)}) - L(\theta^{(k)}) < \epsilon$

The E- and M-steps are alternated repeatedly until convergence, which is assumed to be achieved if the difference $L(\theta^{(k+1)}) - L(\theta^{(k)})$ is lower than an arbitrarily small value ϵ. Dempster, Laird and Rubin (1977) have shown that the (incomplete-data) likelihood function $L(\theta)$ does not decrease after an EM iteration; that is,

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

for $k = 0, 1, 2 \ldots$

Remark 2.1.1. *If bounded from above, the sequence $L(\theta^{(k)})$ converges to some L^*. Furthermore, under certain regularity 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.*

Remark 2.1.2. *Another issue is the dependence on the initial value. Very often the log-likelihood function has multiple maxima or even other type of stationary points. Hence the convergence of the EM algorithm depends strongly on the initial value. To increase the probability of obtaining good estimates, different initial values should be tried.*

2.2 The Q function of an HSMM

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

$$\begin{aligned} Q(\theta;\theta^{(k)}) &= E\{\log L_c(S_0^{T-1+u}, X_0^{T-1};\theta)|X_0^{T-1} = x_0^{T-1};\theta^{(k)}\} \\ &= \sum_{s_0,\dots s_{T-1}} \sum_{u_{T+}} \log L_c(s_0^{T-1+u}, x_0^{T-1};\theta) P(S_0^{T-1+u} = s_0^{T-1+u}|X_0^{T-1} = x_0^{T-1};\theta^{(k)}). \end{aligned}$$

For the remainder of this work we change the notation and we replace $\{X_0^t = x_0^t\}$ by $\{x_0^t\}$ for notational convenience.

To obtain a mathematically tractable formulation of the Q-function, the conditional expectation has to be rewritten path-wise. In fact,

$$\begin{aligned} L_c(s_0^{T-1+u}, x_0^{T-1};\theta) &= P(S_0^{T-1+u} = s_0^{T-1+u}, x_0^{T-1};\theta) \\ &= \pi_{\tilde{s_0}} d_{\tilde{s}_0}(u_0) \left(\prod_{t=0}^{T-1} b_{s_t}(x_t)\right) \left(\prod_{r=1}^{R} p_{\tilde{s}_{r-1}\tilde{s}_r} d_{\tilde{s}_r}(u_r)\right), \end{aligned} \tag{2.2.1}$$

and consequently

$$\begin{aligned} Q(\theta;\theta^{(k)}) = & \\ & \sum_{s_0,\dots s_{T-1}} \sum_{u_{T+}} P(S_0^{T-1+u} = s_0^{T-1+u}|x_0^{T-1};\theta^{(k)}) \log \pi_{\tilde{s_0}} \\ + & \sum_{s_0,\dots s_{T-1}} \sum_{u_{T+}} \sum_{r=1}^{R} P(S_0^{T-1+u} = s_0^{T-1+u}|x_0^{T-1};\theta^{(k)}) \log p_{\tilde{s}_{r-1}\tilde{s}_r} \\ + & \sum_{s_0,\dots s_{T-1}} \sum_{u_{T+}} \sum_{r=0}^{R} P(S_0^{T-1+u} = s_0^{T-1+u}|x_0^{T-1};\theta^{(k)}) \log d_{\tilde{s}_r}(u_r) \\ + & \sum_{s_0,\dots s_{T-1}} \sum_{u_{T+}} \sum_{t=0}^{T-1} P(S_0^{T-1+u} = s_0^{T-1+u}|x_0^{T-1};\theta^{(k)}) \log b_{s_t}(x_t). \end{aligned} \tag{2.2.2}$$

We have decomposed the Q-function into four terms that we can maximize individually. So the first term of (2.2.2) becomes

$$\sum_{j=0}^{J-1} P(S_0 = j|x_0^{T-1};\theta^{(k)}) \log \pi_j,$$

because summing over all possible paths is equivalent to repeatedly selecting the different π_j $(j = 0,\dots,J-1)$ and can therefore be marginalized to $t = 0$. The second term in Equation (2.2.2) is transformed similarly by marginalizing the full paths to the transitions

from i to j at time t for all $t \in \{0, \dots, T-2\}$:

$$\sum_{i=0}^{J-1} \sum_{j \neq i} \sum_{t=0}^{T-2} P(S_{t+1} = j, S_t = i | x_0^{T-1}, \theta^{(k)}) \log p_{ij}. \tag{2.2.3}$$

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:

$$\sum_{j=0}^{J-1} \sum_{u} \left\{ \sum_{t=0}^{T-2} P(S_{t+u+1} \neq j, S_{t+u-v} = j, v = 0, \dots u-1, S_t \neq j | x_0^{T-1}, \theta^{(k)}) + \right.$$
$$\left. P(S_u \neq j, S_{u-v} = j, v = 1, \dots, u | x_0^{T-1}, \theta^{(k)}) \right\} \log d_j(u). \tag{2.2.4}$$

The last term of equation (2.2.2) including the conditional distributions is also transformed to the sum of the marginal distributions of the observations at time t in state j by

$$\sum_{j=0}^{J-1} \sum_{t=0}^{T-1} P(S_t = j | x_0^{T-1}, \theta^{(k)}) \log b_j(x_t). \tag{2.2.5}$$

At this point, we introduce the re-estimation quantities for the initial probabilities, the transition probabilities, the sojourn times and observation components respectively:

$$Q_1(\pi; \theta^{(k)}) := \sum_{j=0}^{J-1} P(S_0 = j | X_0^{T-1} = x_0^{T-1}; \theta^{(k)}) \log \pi_j, \tag{2.2.6}$$

$$Q_2(p; \theta^{(k)}) := \sum_{i=0}^{J-1} \sum_{j \neq i} \sum_{t=0}^{T-2} P(S_{t+1} = j, S_t = i | x_0^{T-1}, \theta^{(k)}) \log p_{ij}, \tag{2.2.7}$$

$$Q_3(d; \theta^{(k)}) :=$$

$$\sum_{j=0}^{J-1} \sum_{u} \left\{ \sum_{t=0}^{T-2} P(S_{t+u+1} \neq j, S_{t+u-v} = j, v = 0, \dots u-1, S_t \neq j | x_0^{T-1}, \theta^{(k)}) \right.$$
$$\left. + P(S_u \neq j, S_{u-v} = j, v = 1, \dots, u | x_0^{T-1}, \theta^{(k)}) \right\} \log d_j(u), \tag{2.2.8}$$

$$Q_4(b; \theta^{(k)}) := \sum_{j=0}^{J-1} \sum_{t=0}^{T-1} P(S_t = j | x_0^{T-1}, \theta^{(k)}) \log b_j(x_t). \tag{2.2.9}$$

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 couples of states i and j.

Subsequently, the M-step maximizes each of the terms with respect to the corresponding parameters to obtain the next set of initial values for the E-step of the following iteration. The difficulty of the maximization varies with the choice of the component distributions and may

also involve numerical maximization methods when an explicit solution is not available.

2.3 The Forward-Backward Algorithm

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 couple of states. First, we present the forward-backward algorithm in the case of a HMM to make it smoother for the reader. The comptutations are performed under the same parameter value $\theta^{(k)}$, so it is supressed for simplicity of notation.

2.3.1 HMM-Model

In the case of an HMM-model the basic idea is the decomposition of the probabilities $L_j(t) := P(S_t = j|x_0^{T-1})$ and $L_{ij}(t) := P(S_{t-1} = i, S_t = j|x_0^{T-1})$.

$$\begin{aligned} L_j(t) &= P(S_t = j|x_0^{T-1}) \\ &= \frac{P(S_t = j, x_0^t, x_{t+1}^{T-1})}{P(x_0^{T-1})} \\ &= \frac{P(S_t = j, x_0^t)P(x_{t+1}^{T-1}|S_t = j, \cancel{x_0^t})}{P(x_0^{T-1})} \\ &= \frac{A_j(t)B_j(t)}{L_n}, \end{aligned} \tag{2.3.1}$$

and

$$\begin{aligned} L_{ij}(t) &= P(S_{t-1} = i, S_t = j|x_0^{T-1}) \\ &= \frac{P(S_{t-1} = i, S_t = j, x_0^{t-1}, x_t, x_{t+1}^{T-1})}{P(x_0^{T-1})} \\ &= \frac{P(S_{t-1} = i, x_0^{t-1})P(S_t = j|S_{t-1} = i, \cancel{x_0^{t-1}})}{L_n} \\ &\times P(x_t|\cancel{S_{t-1} = i}, \cancel{x_0^{t-1}}, S_t = j)P(x_{t+1}^{T-1}|\cancel{S_{t-1} = i}, \cancel{x_t}, , \cancel{x_0^{t-1}}, S_t = j) \\ &= \frac{A_{t-1}(i)p_{ij}b_j(x_t)B_t(j)}{L_n} \end{aligned}$$

where

$$\begin{aligned} A_j(t) &= P(S_t = j, x_0^t) \\ B_j(t) &= P(x_{t+1}^{T-1}|S_t = j) \\ b_j(x_t) &= P(x_t|S_t = j) \\ L_n &= P(x_0^{T-1}) \end{aligned}$$

The forward iteration for an HMM

The forward iteration involves the computation of the forward-probabilities $A_j(t) = P(S_t = j, x_0^t)$ for each state j forward from time 0 to time $T-1$ and can be given as follows.

- Start:

$$A_j(0) = P(S_0 = j, x_0) = P(S_0 = j)P(x_0|S_0 = j) = \pi_j b_j(x_0).$$

- Iteration:

$$A_j(t) = \sum_{i=0}^{J-1} A_i(t-1)p_{ij}b_j(x_t), \tag{2.3.2}$$

for all $t \in \{1, \ldots, T-1\}$ and all $j \in \{0, \ldots, J-1\}$. This holds due to the following decomposition:

$$\begin{aligned} A_j(t) = P(S_t = j, x_0^t) &= \sum_{i=0}^{J-1} P(S_{t-1} = i, S_t = j, x_0^{t-1}, x_t) \\ &= \sum_{i=0}^{J-1} P(S_{t-1} = i, x_0^{t-1})P(S_t = j|S_{t-1} = i, \cancel{x_0^{t-1}})P(x_k|\cancel{S_{t-1} = i}, S_t = j, \cancel{x_0^{t-1}}) \\ &= \sum_{i=0}^{J-1} A_i(t-1)p_{ij}b_j(x_t). \end{aligned}$$

The Backward Iteration for an HMM

The backward iteration performs the computation of the conditional probabilities $B_j(t) = P(x_{t+1}^{T-1}|S_t = j)$ for each state j , backwards from time $T-1$ to time 0.

- **Start:** The backward iteration starts at $t = T-1$ with

$$B_j(T-1) = 1 \quad \forall \quad j.$$

- **Iteration:**

$$\begin{aligned} B_i(t) = \sum_{j=0}^{J-1} p_{ij}b_j(x_{t+1})B_j(t+1) \\ \forall \quad t = T-2, \ldots, 0. \end{aligned}$$

the above recursion holds due to the following decompositions.

$$\begin{aligned} B_i(t) = P(x_{t+1}^T|S_t = i) &= \sum_{j=0}^{J-1} P(S_{t+1} = j, x_{t+1}^{T-1}|S_t = i) \\ &= \sum_{j=0}^{J-1} P(S_{t+1} = j|S_t = i)P(x_{t+1}, x_{t+2}^{T-1}|S_{t+1} = j, S_t = i) \\ &= \sum_{j=0}^{J-1} p_{ij}P(x_{t+1}|S_{t+1} = j)P(x_{t+2}^{T-1}|S_{t+1} = j, S_t = i, x_{t+1}) \\ &= \sum_{j=0}^{J-1} p_{ij}b_j(x_{t+1})B_j(t+1). \end{aligned}$$

Finally, the term L_n is computed with the help of the intermediate quantities $A_i(t)$ and $B_i(t)$. In particular,

$$L_n = P(x_0^{T-1}) = \sum_{i=0}^{J-1} A_i(n) = \sum_{i=0}^{J-1} A_i(t)B_i(t). \tag{2.3.3}$$

2.3.2 HSMM-Model

In addition to the quantities $L_j(t)$ and their associated decompositions, the forward-backward algorithm for HSMMs requires the computation of the quantities.

$$L1_j(t) := P(S_{t+1} \neq j, S_t = j | X_0^{T-1} = x_0^{T-1}).$$

The subsequent decompositions are possible:

$$\begin{aligned} L1_j(t) &= P(S_{t+1} \neq j, S_t = j | X_0^{T-1} = x_0^{T-1}) \\ &= \frac{P(X_{t+1}^{T-1} = x_{t+1}^{T-1} | S_{t+1} \neq j, S_t = j)}{P(X_{t+1}^{T-1} = x_{t+1}^{T-1} | x_0^t)} P(S_{t+1} \neq j, S_t = j | x_0^t) \\ &= \bar{B}_j(t) F_j(t), \end{aligned}$$

where

$$\begin{aligned} F_j(t) &= P(S_{t+1} \neq j, S_t = j | x_0^t) \\ \bar{B}_j(t) &= \frac{P(X_{t+1}^{T-1} = x_{t+1}^{T-1} | S_{t+1} \neq j, S_t = j)}{P(X_{t+1}^{T-1} = x_{t+1}^{T-1} | x_0^t)}. \end{aligned}$$

The forward iteration

To utilize the forward iteration we have to decompose $F_j(t)$ in the following way:

$$\begin{aligned} F_j(t) &= P(S_{t+1} \neq j, S_t = j | x_0^t) \\ &= \sum_{u=1}^{t} \sum_{i \neq j} P(S_{t+1} \neq j, S_{t-v} = j, v = 0, \ldots, u-1, S_{t-u} = i | x_0^t) \\ &+ P(S_{t+1} \neq j, S_{t-v} = j, v = 0, \ldots, t | x_0^t) \\ &= F_{1j}(t) + F_{2j}(t). \end{aligned}$$

where $F_{2j}(t)$ corresponds to the possibility that the system started at state j, and no transition was made until time t, followed by a state change at time $t+1$. For the term $F_{1j}(t)$ we have:

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

and for $F_{2j}(t)$ we have

$$\begin{aligned}
F_{2j}(t) &= P(S_{t+1}\neq j, S_{t-v}=j, v=0,\ldots,t|x_0^t) \\
&= \frac{P(x_o^t|S_{t-v}=j, v=0\ldots.,t)}{P(x_0^t)} \\
&\times P(S_{t+1}\neq j, S_{t-v}=j, v=0,\ldots,t) \\
&= \frac{b_j(x_t)}{N_t}\left(\prod_{v=1}^{t}\frac{b_j(x_{t-v})}{N_{t-v}}\right)d_j(t+1)\pi_j,
\end{aligned}$$

where

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

The quantity N_t is the so-called normalizing factor.
The forward iteration involves the computation of the forward-probabilities for each state j forward from time 0 to time $T-1$ and can be given as follows.

- The start of the loop at t = 0

$$F_j(0) = P(S_1\neq j, S_0=j|X_0=x_0) = \pi_j d_j(1).$$

- The iteration procedure is:

$$F_j(t) = P(S_{t+1} \neq j, S_t = j | x_0^t) =$$

$$\frac{b_j(x_t)}{N_t}\left[\sum_{u=1}^{t}\left\{\prod_{v=1}^{u-1}\frac{b_j(x_{t-v})}{N_{t-v}}\right\}d_j(u)\sum_{i\neq j}p_{ij}F_i(t-u) + \left\{\prod_{v=1}^{t}\frac{b_j(x_{t-v})}{N_{t-v}}\right\}d_j(t+1)\pi_j\right], \tag{2.3.4}$$

for all $t \in \{0, ...T-2\}$ and $j \in \{0, ..., J-1\}$. Using arguments similar to those for the derivation of (2.3.4), the last step of the iteration can be written as

$$F_j(T-1) = P(S_{T-1} = j | X_0^t = x_0^t) =$$

$$\frac{b_j(x_{T-1})}{N_{T-1}}\left[\sum_{u=1}^{T-1}\left\{\prod_{v=1}^{u-1}\frac{b_j(x_{T-1-v})}{N_{T-1-v}}\right\}\bar{D}_j(u)\sum_{i\neq j}p_{ij}F_i(T-1-u)\right.$$
$$\left. + \left\{\prod_{v=1}^{T-1}\frac{b_j(x_{T-1-v})}{N_{T-1-v}}\right\}\bar{D}_j(T)\pi_j,\right.$$

for $j \in \{0, ..., J-1\}$. 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 (2.3.4) is replaced by the corresponding survival functions $\bar{D}_j(u)$.

Note that N_t are directly obtained during the forward recursion.

$$\begin{aligned}
N_t &= P(x_t|x_0^{t-1}) = \sum_j P(S_t = j, x_t|x_0^{t-1}) = \\
&= \sum_j P(x_t|S_t = j, \cancel{x_0^{t-1}})P(S_t = j|x_0^{t-1}) = \\
&= \sum_j b_j(x_t)\left[\sum_{u=1}^{t}\sum_{i\neq j}P(S_{t-v} = j, v = 0, \ldots, u-1, S_{t-u} = i|x_0^{t-1})\right. \\
&\left. + P(S_{t-v} = j, v = 0, \ldots, t|x_0^{t-1})\right] \\
&= \sum_j b_j(x_t)\left[\sum_{u=1}^{t}\left[\frac{P(x_{t-u+1}^{t-1}|S_{t-v} = j, v = 0, \ldots, u-1)}{P(x_{t-u+1}^{t-1}|x_0^{t-u})}\right.\right. \\
&\times P(S_{t-v} = j, v = 0, \ldots, u-2|S_{t-u+1} = j, S_{t-u} \neq j) \\
&\times \sum_{i\neq j}\{P(S_{t-u+1} = j|S_{t-u+1} \neq i, S_{t-u} = i) \\
&\left.\left.\times P(S_{t-u} = i)|x_0^{t-u})\ \}\right] \right. \\
&+ \frac{P(x_o^{t-1}|S_{t-v} = j, v = 0. \ldots, t)}{P(x_0^{t-1})} \\
&\left.\times P(S_{t-v} = j, v = 0, \ldots, t)\right]
\end{aligned}$$

$$= \sum_j b_j(x_t) \left[\sum_{u=1}^{t} \left\{ \prod_{v=1}^{u-1} \frac{b_j(x_{t-v})}{N_{t-v}} \right\} \bar{D}_j(u) \sum_{i \neq j} p_{ij} F_i(t-u) \right.$$
$$\left. + \left\{ \prod_{v=1}^{t} \frac{b_j(x_{t-v})}{N_{t-v}} \right\} \bar{D}_j(t+1)\pi_j \right] \tag{2.3.5}$$

and can be used for forecasting procedures by setting $t = T$.

The Backward Iteration

The backward iteration performs the computation of the smoothing probabilities $L_j(t) = P(S_{t-1} = j|x_0^{T-1})$ for each state j, backward from time $T-1$ to time 0.

- **Start:** The backward iteration starts at $t = T-1$ with

$$L_j(T-1) = P(S_{T-1} = j|x_0^{T-1}) = F_j(T-1).$$

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

$$\begin{aligned} L_j(t) &= P(S_t = j|x_0^{T-1}) \\ &= P(S_{t+1} \neq j, S_t = j|x_0^{T-1}) + P(S_{t+1} = j|x_0^{T-1}) \\ &- P(S_{t+1} = j, S_t \neq j|x_0^{T-1}) \\ &= L1_j(t) + L_j(t+1) - P(S_{t+1} = j, S_t \neq j|x_0^{T-1}). \end{aligned} \tag{2.3.6}$$

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 $P(S_{t+1} = j, S_t \neq j|x_0^{T-1})$ which represents the entrance into state j, require a bit more attention.
$L1_j(t)$ can be decomposed into two terms:

$$\begin{aligned} L1_j(t) &= P(S_{t+1} \neq j, S_t = j|x_0^{T-1}) \\ &= \sum_{k \neq j} \left[\sum_{u=1}^{T-2-t} P(S_{t+u+1} \neq k, S_{t+u-v} = k, v = 0 \dots, u-1, S_t = j|x_0^{T-1}) \right. \\ &\left. + P(S_{T-1-v} = k, v = 0, \dots T-2-t, S_t = j|x_o^{T-1}) \right]. \end{aligned} \tag{2.3.7}$$

The first term in Equation (2.3.7) can be decomposed as follows :

$$
\begin{aligned}
&P(S_{t+u+1} \neq k, S_{t+u-v} = k, v = 0 \ldots, u-1, S_t = j | x_0^{T-1}) \\
&= \frac{P(S_{t+u+1} \neq k, S_{t+u-v} = k, v = 0 \ldots, u-1, S_t = j, x_0^{T-1})}{P(S_{t+u+1} \neq k, S_{t+u} = k, x_0^{T-1})} \\
&\times P(S_{t+u+1} \neq k, S_{t+u} = k | x_0^{T-1}) \\
&= \frac{P(x_{t+u+1}^{T-1} | S_{t+u+1} \neq k, S_{t+u} = k)}{P(x_{t+u+1}^{T-1} | S_{t+u+1} \neq k, S_{t+u} = k)} \\
&\times \frac{P(S_{t+u+1} \neq k, S_{t+u} = k | x_0^{T-1})}{P(S_{t+u+1} \neq k, S_{t+u} = k | x_0^{t+u})} \\
&\times \frac{P(x_{t+1}^{t+u} | S_{t+u-v} = k, v = 0, \ldots, u-1)}{P(x_{t+1}^{t+u} | x_0^T)} \\
&\times P(S_{t+u+1} \neq k, S_{t+u-v} = k, v = 0, \ldots, u-2 | S_{t+1} = k, S_t \neq k) \\
&\times P(S_{t+1} = k | S_{t+1} \neq j, S_t = j) P(S_{t+1} \neq j | x_o^t) \\
&= \frac{L1_k(t+u)}{F_k(t+u)} \Big\{ \prod_{v=0}^{u-1} \frac{b_k(x_{t+u-v})}{N_{t+u-v}} \Big\} d_k(u) p_{jk} F_j(t).
\end{aligned}
$$

The second term of Equation (2.3.7), corresponding to the last visited state, can be decomposed using a similar argument.This yields

$$
\begin{aligned}
&P(S_{T-1-v} = k, v = 0, \ldots T-2-t, S_t = j | x_o^{T-1}) \\
&= \Big\{ \prod_{v=0}^{T-2-t} \frac{b_k(x_{T-1-v})}{N_{T-1-v}} \Big\} \bar{D}_k(T-1-t) p_{jk} F_j(t).
\end{aligned}
$$

Combining the two decompositions, $L1_j(t)$ becomes

$$
\begin{aligned}
&\Big[\sum_{k \neq j} \Big[\sum_{u=1}^{T-2-t} \frac{L1_k(t+u)}{F_k(t+u)} \Big\{ \prod_{v=0}^{u-1} \frac{b_k(x_{t+u-v})}{N_{t+u-v}} \Big\} d_k(u) \\
&+ \Big\{ \prod_{v=0}^{T-2-t} \frac{b_k(x_{T-1-v})}{N_{T-1-v}} \Big\} \bar{D}_k(T-1-t) \Big] p_{jk} \Big] F_j(t). \qquad (2.3.8)
\end{aligned}
$$

The third term of Equation (2.3.6) can also be transformed as follows:

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

To perform the calculation we introduce the auxiliary quantities

$$G_j(t+1,u) := \frac{L1_j(t+u)}{F_j(t+u)}\Big\{\prod_{v=0}^{u-1}\frac{b_j(x_{t+u-v)}}{N_{t+u-v}}\Big\}d_j(u), u \in \{1,...,T-2-t\}$$

$$G_j(t+1,T-1-t) := \Big\{\prod_{v=0}^{T-2-t}\frac{b_j(x_{T-1-v})}{N_{T-1-v}}\Big\}\bar{D}_j(T-1-t)$$

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

Then, the backward iteration can be performed as follows :

- At each time t, $G_j(t+1,u)$, $G_j(t+1,T-1-t)$ and $G_j(t+1)$ are precomputed.
- $L1_j(t)$ and $P(S_{t+1}=j, S_t \neq j|x_0^{T-1})$ can be transformed to

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

and

$$L1_j(t) = \Big\{\sum_{k\neq j} G_k(t+1)p_{jk}\Big\}F_j(t).$$

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

2.4 The Sojourn Time Distribution

The aim of this section is to show how $Q_3(d;\theta^{(k)})$ – the part of the Q-function dealing with the sojourn time can be computed using the quantities derived in Section 2.2. As long as we deal with non-stationary HSMMs, this is the part of the estimation procedure which is affected by the use of either the partial likelihood estimator or the complete likelihood estimator from Sections 1.4 and 1.5 . Recall from Equation (2.2.8) that the the state occupancy distribution for each state j is given by :

$$Q_3(d;\theta^{(k)}) :=$$

$$\sum_u\{\sum_{t=0}^{T-2} P(S_{t+u+1}\neq j, S_{t+u-v}=j, v=0,...u-1, S_t\neq j|x_0^{T-1},\theta^{(k)})$$

$$+P(S_u \neq j, S_{u-v} = j, v = 1, ..., u | x_0^{T-1}, \theta^{(k)})\} \log d_j(u). \tag{2.4.1}$$

$$= \sum_u n_{ju}^{(k)} \log d_j(u). \tag{2.4.2}$$

The computation of the two terms involved in (2.4.1) 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

$$\begin{aligned} &u \leq T-2-t: \\ &P(S_{t+u+1} \neq j, S_{t+u-v} = j, v = 0, ...u-1, S_t \neq j | x_0^{T-1}, \theta^{(k)}) \\ &= G_j(t+1, u) \sum_{i \neq j} F_i(t), \end{aligned}$$

which is directly available from the computation of $L_j(t)$.

$$\begin{aligned} &u > T-2-t: \\ &P(S_{t+u+1} \neq j, S_{t+u-v} = j, v = 0, ...u-1, S_t \neq j | x_0^{T-1}, \theta^{(k)}) \\ &= \Big\{ \prod_{v=0}^{T-2-t} \frac{b_j(x_{T-1-v})}{N_{T-1-v}} \Big\} d_j(u) \sum_{i \neq j} p_{ij} F_i(t). \end{aligned}$$

The second term in (2.4.1) corresponds to the time spent in the initial state from $t = 0$, and it can be computed by the already known quantities from the forward-backward algorithm. Again, we consider two separate cases.

$$\begin{aligned} &u \leq T-1: \\ &P(S_u \neq j, S_{u-v} = j, v = 1, ..., u | x_0^{T-1}, \theta^{(k)}) \\ &= \frac{L1_j(u-1)}{F_j(u-1)} \Big\{ \prod_{v=1}^{u} \frac{b_j(x_{u-v})}{N_{u-v}} \Big\} d_j(u) \pi_j, \end{aligned}$$

$$\begin{aligned} &u > T-1: \\ &P(S_u \neq j, S_{u-v} = j, v = 1, ..., u | x_0^{T-1}, \theta^{(k)}) \\ &= \Big\{ \prod_{v=1}^{T} \frac{b_j(x_{T-v})}{N_{T-v}} \Big\} d_j(u) \pi_j. \end{aligned}$$

2.5 Parameter Re-estimation

Now that we defined the Q-function it is time for the re-estimation procedure of the EM algorithm, the M-step. This step determines the likelihood-increasing next set of parameters $\theta^{(k+1)}$ by

$$\theta^{(k+1)} = \underset{\theta}{\text{argmax}} Q(\theta | \theta^{(k)})$$

As we show the Q-function of an HSMM can be decomposed in four terms each depending on a given subset of θ. Hence, the re-estimation formulae for the parameters can be derived by maximizing each one of the different terms separately.

The initial Parameters

We start with the parameters involved in the underlying hidden semi-Markov chain. The term of the Q-function corresponding to the initial parameters is given by

$$Q_1(\pi;\theta^{(k)}) := \sum_{j=0}^{J-1} P(S_0 = j|x_0^{T-1},\theta^{(k)}) \log \pi_j. \tag{2.5.1}$$

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

$$\pi_j = P(S_0 = j|X_0^{T-1} = x_0^{T-1},\theta^{(k)}),$$

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

$$\pi_j^{(k+1)} = P(S_0 = j|x_0^{T-1},\theta^{(k)}) = L_j(0). \tag{2.5.2}$$

The Transition Probabilities

The term of $Q(\theta;\theta^{(k)})$ corresponding to the transition probabilities parameters is given by

$$Q_2(p;\theta^{(k)}) := \sum_{i\neq j} \sum_{t=2}^{T-2} P(S_{t+1} = j, S_t = i|x_0^{T-1},\theta^{(k)}) \log p_{ij}. \tag{2.5.3}$$

Adding Lagrange multipliers for all i, with the constraint $\sum_{j=0}^{J-1} p_{ij} = 1$ differentiating w.r.t. p_{ij} and solving the resulting equations we get

$$p_{ij}^{(k)} = \frac{\sum_{t=0}^{T-2} P(S_{t+1} = j, S_t = i|X_0^{T-1} = x_0^{T-1},\theta^{(k)})}{\sum_{t=0}^{T-2} P(S_{t+1} \neq i, S_t = i|X_0^{T-1} = x_0^{T-1},\theta^{(k)})}. \tag{2.5.4}$$

The re-estimation formula can be written as

$$p_{ij}^{(k+1)} = \frac{\sum_{t=0}^{T-2} G_j(t+1) p_{ij} F_i(t)}{\sum_{t=0}^{T-2} L1_i(t)}. \tag{2.5.5}$$

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

Non-Parametric State Occupancy Distribution with the Complete Likelihood Estimator

The term of the Q-function treating the non-parametric state occupancy distribution is given by :

$$Q_3(d;\theta^{(k)}) := \sum_{j=0}^{J-1} \sum_{u} n_{ju}^{(k)} \log d_j(u).$$

It has to be maximized under the constraint $\sum_u d_j(u) = 1$, for all $j = 0, 1, \ldots J-1$, which leads to the following re-estimation formula for the state occupancy probabilities :

$$d_j^{(k+1)}(u) = \frac{n_{ju}^{(k)}}{\sum_{u=1}^{M_j} n_{ju}^{(k)}} = \frac{n_{ju}^{(k)}}{\sum_{t=0}^{T-2} L1_j(t) + L_j(T-1)}. \tag{2.5.6}$$

Geometric State Occupancy distribution

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

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

where $j \in \{0, 1, \ldots, J-1\}$ and $u \in \{1, \ldots, M_j = T-1\}$. We rewrite $\sum_{j=0}^{T-1} Q_3(d_j; \theta^{(k)})$, where for each $j = 0, 1, \ldots, J-1$, the corresponding part of the Q-function becomes:

$$Q_3(d; \theta) = \sum_{u=1}^{T-1} n_{ju}^{(k)} [(u-1)\log(1-p_j) + \log p_j].$$

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

$$p_j = \frac{\sum_{u=1}^{T-1} n_{ju}^{(k)}}{\sum_{u=1}^{T-1} u n_{ju}^{(k)}}.$$

Negative Binomial State Occupancy Distribution

The negative binomial distribution is an extension of the geometric distribution. The resulting probabity function of the sojourn times 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 $\Gamma(\cdot)$ 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 so as to increase numerical stability. Then, for each j the corresponding part of the Q-function becomes

$$\begin{aligned} Q_3(d_j; \theta^{(k)}) = \sum_{u=1}^{T-1} n_{ju}^{(k)} [&\log\Gamma(u-1+r) - \log\Gamma(r) - \log\Gamma(u) \\ &+ r\log\pi + (u-1)\log(1-\pi)]. \end{aligned}$$

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

$$\pi = \frac{r \sum_{u=1}^{T-1} n_{ju}^{(k)}}{\sum_{u=1}^{T-1} n_{ju}^{(k)} (r+u-1)}. \tag{2.5.7}$$

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

$$\frac{\partial}{\partial r} Q_3(d_j; \theta^{(k)}) = 0 \Rightarrow \sum_{u=1}^{T-1} n_{ju}^{(k)} (\psi(u-1+r) - \psi(r) + \log \pi) = 0. \tag{2.5.8}$$

Substituting π from Equation (2.5.7) in Equation (2.5.8) yields the expression

$$\sum_{u=1}^{T-1} n_{ju}^{(k)} \Big(\psi(u-1+r) - \psi(r) + \log \Big[\frac{r \sum_{u=1}^{T-1} n_{ju}^{(k)}}{\sum_{u=1}^{T-1} n_{ju}^{(k)} (r+u-1)} \Big] \Big) = 0, \tag{2.5.9}$$

which has to be solved numerically, e.g. by a bisectioning algorithm. The estimation of π follows directly from Equation (2.5.7).

The Observation Component

The conditional distributions of the observed states given the hidden ones 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 (2.2.7) is given by

$$\begin{aligned} Q_4(b_j; \theta^{(k)}) &= \sum_{t=0}^{T-1} P(S_t = j | x_0^{T-1}; \theta^{(k)}) \log b_j(x_t) = \\ &= \sum_{t=0}^{T-1} L_j(t) \log b_j(x_t). \end{aligned} \tag{2.5.10}$$

Depending on the distributional assumptions imposed on $b_j(x_t)$, the maximization of the corresponding term of $Q(\theta|\theta^{(k)})$ may also involve numerical methods. In the following we deal with some common distributions, e.g., the Poisson, Bernoulli, normal, and t distribution. If the conditional distributions are a Bernoulli distributions then

$$b_j(x_t) = p_j^{x_t} (1-p_j)^{1-x_t}, \quad x_t = 0, 1, \ldots \tag{2.5.11}$$

In this section, p_j denotes the parameter of the Bernoulli conditional distribution.

$$Q_4(b; \theta^{(k)}) = \sum_{t=0}^{T-1} L_j(t) [x_t \log p_j + (1-x_t) \log(1-p_j)],$$

which has to be maximized w.r.t. p_j to perform the M-step and we get

$$p_j^{(k+1)} = \frac{\sum_{t=0}^{T-1} L_j(t) x_t}{\sum_{t=0}^{T-1} L_j(t)} \tag{2.5.12}$$

If the component distributions are assumed to be univariate Poisson distributions with parameters λ_j then :

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

$$Q_4(b_j; \theta^{(k)}) = \sum_{t=0}^{T-1} L_j(t)[x_t \log \lambda_j - \lambda_j - \log x_t!].$$

The maximization w.r.t. λ_j yields

$$\lambda_j = \frac{\sum_{t=0}^{T-1} L_j(t) x_t}{\sum_{t=0}^{T-1} L_j(t)}, \tag{2.5.13}$$

and the re-estimation quantity is

$$\lambda_j^{(k+1)} = \frac{\sum_{t=0}^{T-1} L_j(t) x_t}{\sum_{t=0}^{T-1} L_j(t)}. \tag{2.5.14}$$

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

$$b_j(t) = \frac{1}{(2\pi)^{p/2}} \exp\Big(-1/2(x_t - \mu_j)^T \Sigma_j^{-1} (x_t - \mu_j)\Big)$$

with mean μ_j and positive definite covariance matrix Σ_j. The dimension of the observations is denoted by p and all vectors are column vectors. Representing the constant terms by C, Equation (2.5.10) becomes

$$Q_4(b_j; \theta^{(k)}) = \sum_{t=0}^{T-1} L_j(t)[C - \frac{1}{2}\log(|\Sigma_j|) - \frac{1}{2}(x_t - \mu_j)^T \Sigma_j^{-1}(x_t - \mu_j)]. \tag{2.5.15}$$

We first treat the maximization w.r.t. μ_j, by taking the derivative of Equation (2.5.15) w.r.t. μ_j and setting it equal to zero, this yields

$$\mu_j = \frac{\sum_{t=0}^{T-1} L_j(t) x_t}{\sum_{t=0}^{T-1} L_j(t)}.$$

The first step of the maximization w.r.t. Σ consists in transforming Equation (2.5.15) to

$$\begin{aligned} &\frac{1}{2}\log(|\Sigma_j^{-1}|) \sum_{t=0}^{T-1} L_j(t) - \frac{1}{2} \sum_{t=0}^{T-1} L_j(t) tr(\Sigma_j^{-1}(x_t - \mu_j)(x_t - \mu_j)^T) \\ &\qquad = \frac{1}{2}\log(|\Sigma_j^{-1}|) \sum_{t=0}^{T-1} L_j(t) - \frac{1}{2} \sum_{t=0}^{T-1} L_j(t) tr(\Sigma_j^{-1} N_{jt}), \end{aligned}$$

where $N_{jt} = (x_t - \mu_j)(x_t - \mu_j)^T$. Differentiating w.r.t. Σ yields

$$\frac{1}{2}\sum_{t=0}^{T-1} L_j(t)\big(2\Sigma_j - diag(\Sigma_j)\big) - \frac{1}{2}\sum_{t=0}^{T-1} L_j(t)(2N_{jt} - diag(N_{jt}))$$
$$= \frac{1}{2}\sum_{t=0}^{T-1} L_j(t)(2M_{jt} - diagM_{jt}) = 2S - diag(S),$$

where $M_{jt} := \Sigma_j N_{jt}$ and $S := \sum_{t=0}^{T-1} L_j(t)(\Sigma_j - N_{jt})$. Setting the derivative equal to zero w.r.t *Sigma* yields

$$2S - diag(S) = 0 \Rightarrow S = 0.$$

This is equivalent to $\sum_{t=0}^{T-1} L_j(t)(\Sigma_j - N_{jt}) = 0$ and it follows that

$$\Sigma_j = \frac{\sum_{t=0}^{T-1} L_j(t) N_{jt}}{\sum_{t=0}^{T-1} L_j(t)}.$$

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

$$\mu_j^{(k+1)} = \frac{\sum_{t=0}^{T-1} L_j(t) x_t}{\sum_{t=0}^{T-1} L_j(t)}, \tag{2.5.16}$$

$$\Sigma_j^{(k+1)} = \frac{\sum_{t=0}^{T-1} L_j(t)(x_t - \mu_j^{(k+1)})(x_t - \mu_j^{(k+1)})^T}{\sum_{t=0}^{T-1} L_j(t)}. \tag{2.5.17}$$

For the case of mixtures of normal distributions as component distributions, we do not provide all the calculations here but instead we give a short overview and report the resulting re-estimation formulae. We refer to Sansom and Thomson (2000) for the details. The density of the normal distribution with mean μ and positive definite covariance matrix Σ be given by

$$f(x_t; \mu, \Sigma) = \frac{1}{(2\pi)^{p/2}} \exp\Big\{ -1/2(x_t - \mu)^T \Sigma^{-1}(x_t - \mu)\Big\},$$

where p is the dimension of the observations. The component distribution associated with each state j is assumed to be a finite mixture of M normal densities. Let $m \in \{0, ..., M-1\}$ denote the mixture components, thus for each state j, there are M means, denoted by μ_{jm}, and M covariance matrices, denoted by Σ_{jm}. Then, Equation (2.5.10) becomes

$$b_j(x_t) = \sum_{m=0}^{M-1} \phi_{jm} f(x_t; \mu_{jm}, \Sigma_{jm}),$$

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

$$L_{jm} := \frac{L_j(t)}{\sum_{m=0}^{M-1} \phi_{jm} f(x_t; \mu_{jm}, \Sigma_{jm})} \phi_{jm} f(x_t; \mu_{jm}, \Sigma_{jm}),$$

which can be interpreted as weighted probability of observing x_t in the mixing component m of state j. Then the re-estimation formulae for μ, Σ and ψ can be written as

$$\phi_{jm}^{(k+1)} = \frac{\sum_{t=0}^{T-1} L_{jm}(t)}{\sum_{m=0}^{M-1} \sum_{t=0}^{T-1} L_{jm}(t)} = \frac{\sum_{t=0}^{T-1} L_{jm}(t)}{\sum_{t=0}^{T-1} L_j(t)}, \tag{2.5.18}$$

$$\mu_{jm}^{(k+1)} = \frac{\sum_{t=0}^{T-1} L_{jm}(t) x_t}{\sum_{t=0}^{T-1} L_{jm}(t)}, \tag{2.5.19}$$

$$\Sigma_{jm}^{(k+1)} = \frac{\sum_{t=0}^{T-1} L_{jm}(t)(x_t - \mu_{jm}^{(k+1)})(x_t - \mu_{jm}^{(k+1)})^T}{\sum_{t=0}^{T-1} L_{jm}(t)}. \tag{2.5.20}$$

The derivations for (2.5.18), (2.5.19) and (2.5.20) are similar to that of the normal component distributions. 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 and McLachlan (2000) for the estimation of mixtures of t distributions can be adopted to the case of an HSMM and we follow their approach. Recall that the t distribution is derived from a Normal mixture model of the form

$$\int g(x; \mu, \Sigma/u) dU(u) \tag{2.5.21}$$

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

$$U \sim gamma\Big(\frac{1}{2}\nu, \frac{1}{2}\nu\Big),$$

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

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

where $\Gamma(\cdot)$ denotes the gamma function. Evaluating the integral given in (2.5.21) yields the density of the t distribution with location parameter μ, ν degrees of freedom and positive definite inner product matrix $Sigma$. The density is given by

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

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

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

and p corresponds to the dimension of the observations. Note that f converges to the density function of a Normal distribution with mean μ and covariance Σ as ν tends to infinity. The mean μ of the t distribution exists for all ν > 1, and the covariance matrix is given by $\nu/(\nu-2)\Sigma$ for $\nu > 2$. In the case of component t distribution, the observation distribution from Equation (2.5.10) is

$$b_j(x_t) = \frac{\Gamma(\frac{\nu_j+p}{2})|\Sigma_j|^{-1/2}}{(\pi\nu)^{1/2}\Gamma(\frac{\nu_j}{2})\{1+\delta(x_j,\mu_j,\Sigma_j)/\nu_j\}^{1/2(\nu_j+p)}}. \tag{2.5.22}$$

Unfortunately, the re-estimation formulae of the parameters involved in (2.5.22) 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 and 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 x_t belongs to component j and zero otherwise. Secondly, the missing data from the gamma distributed random variable U, denoted by $u_0, ..., u_{T-1}$, has to be added to the complete-data with

$$[X_t|u_t, z_{jt} = 1] \sim N(\mu_j, \Sigma_j/u_t),$$

for $t \in \{0, .., T-1\}$, and

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

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

$$\log L_c(\mu_j, \Sigma_j, \nu_j) = \log L_{c1}(\nu_j) + \log L_{c2}(\mu_j, \Sigma_j),$$

where

$$\log L_{c1}(\nu_j) = \sum_{t=0}^{T-1} z_{jt}\Big\{ -\log\Gamma(\frac{1}{2}\nu_j) + \frac{1}{2}\nu_j \log(\frac{1}{2}\nu_j) + \frac{1}{2}\nu_j(\log u_t - u_t) - \log u_t\Big\}, \tag{2.5.23}$$

with

$$\log L_{c2}(\mu_j, \Sigma_j) = \sum_{t=0}^{T-1} z_{jt}\Big\{\frac{1}{2}p\log(2\pi) - \frac{1}{2}\log|\Sigma_j| - \frac{1}{2}u_j(x_t-\mu_j)^T\Sigma_j^{-1}(x_t-\mu_j)\Big\}. \tag{2.5.24}$$

The calculation of $Q(\theta|\theta^{(k)})$ 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^{T-1} and $z_0, ..., z_{T-1}$ is taken. Then, the conditional expectation of z_t given x_0^{T-1} is evaluated; hereby $P(Z_t =$

$1|x_0^{T-1}) = L_j(t)$ holds true. From Equations (2.5.23) and (2.5.24), it is clear that

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

and

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

have to be calculated. The calculation of $E(U_t|x_t, z_t, \theta^{(k)})$ 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 $X_t = x_t$ and $Z_{jt} = 1$ is

$$[U_t|x_t, z_{jt} = 1] \sim gamma(m_{1j}, m_{2j}),$$

where $m_{1j} := \frac{1}{2}(\nu_j + p)$ and $m_{2j} := \frac{1}{2}\{\nu_j + \delta(x_t, \mu_t, \Sigma_t)\}$. From the definition of the gamma distribution, it follows that

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

which yields the desired result:

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

To compute the term $E(\log U_t|x_t, z_t, \theta^{(k)}$ we have to use the result that if a random variable R is distributed as $gamma(,)$, then $E(\log R) = \psi(a) - log\beta$ where $\psi(s)$ is the digamma function. As shown in the derivation of $E(U_t|x_t, z_t, {}^{(k)})$, the conditional density of U_t given x_t and $z_{jt} = 1$ is in $gamma(m_{1j}, m_{2j})$. Applying the above result to the conditional density of U_t yields

$$\begin{aligned} E(\log U_t|x_t, z_t, \theta^{(k)}) &= \psi\Big(\frac{\nu_j^{(k)} + p}{2}\Big) - \log\Big(\frac{1}{2}\{\nu_j^{(k)} + \delta(x_t, \mu_j^{(k)}, \Sigma_j^{(k)})\}\Big) \\ &= \log u_{jt}^{(k)} + \left\{\psi(\frac{\nu_j^{(k)} + p}{2}) - \log(\frac{\nu_j^{(k)} + p}{2})\right\}, \end{aligned}$$

where

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

Applying the results for $E(\log U_t|x_t, z_t, \theta^{(k)})$ and $E(U_t|x_t, z_t, \theta^{(k)})$ the observation part of the Q-function in Equation (2.5.10) can be split up into two parts:

$$Q_4(b_j; \theta^{(k)}) = \sum_{t=0}^{T-1} L_j(t) Q_{1t}(\nu_j; \theta^{(k)}) + \sum_{t=0}^{T-1} L_j(t) Q_{2t}(\mu_j, \Sigma_j; \theta^{(k)}), \tag{2.5.25}$$

where, ignoring all terms not involving ν_j, yields

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

and

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

The re-estimation procedure consists of the maximization of the two terms of Equation (2.5.25) w.r.t. the parameters μ_j , Σ_j and ν_j. The re-estimation formulae for μ_j and Σ_j can be derived explicitly, yielding

$$\mu_j^{(k+1)} = \frac{\sum_{t=0}^{T-1} L_j(t) u_{jt}^{(k)} x_t}{\sum_{t=0}^{T-1} L_j(t) u_{jt}^{(k)}} \tag{2.5.26}$$

and

$$\Sigma_j^{(k+1)} = \frac{\sum_{t=0}^{T-1} L_j(t) u_{jt}^{(k)} (x_t-\mu_j^{(k+1)})(x_t-\mu_j^{(k+1)})^T}{\sum_{t=0}^{T-1} L_j(t)}. \tag{2.5.27}$$

Note that, according to Kent et al. (1994), for the case of a single component t distribution, the denominator of (2.5.27) can also be replaced by $\sum_{t=0}^{T-1} L_j(t) u_{jt}^{(k)}$ to increase the speed of convergence. The re-estimation of the degrees of freedom ν_j is a bit more complicated. The estimator is the (unique) solution of the equation

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

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

Chapter 3

Application to Real data of the SnP 500

In this chapter, we employ two-state and three-state hidden semi-Markov models and hidden Markov models to explain the time-varying distribution of the stock market returns during the period 1987 until 2006. Our results indicate that the time-varying distribution depends on the hidden states, which are represented by three market conditions, namely the bear, sidewalk, and bull markets. We use the R package "hsmm" created by Bulla, J., and Bulla, I.

3.1 Stylized facts

Most of the empirical studies on modelling stock returns focus on certain properties of absolute and squared daily returns. Dating back to Granger and Ding (1995a,b), the temporal properties can be summarized as follow:

- TP1: returns are not autocorrelated (except for, possibly, the first lag).
- TP2: the autocorrelation function of the absolute and quadratic returns are slow decaying, and $cor(|X_t|, |X_{tl}|) > cor(X_t^2, X_{tl}^2)$, with l being a positive integer denoting the lag.
- TP3: autocorrelations of powers of absolute returns are highest at power one. (Taylor effect)
- TP4: the autocorrelations of sign (X_t) are negligibly small.

Further distributional properties are:

- DP1: $|X_t|$ and sign (X_t) are independent.
- DP2: $E(X_t) = Var(X_t)$.
- DP3: the marginal distribution of $|X_t|$ is exponential.

Note that an exponentially distributed variable (DP3) X_t has the following properties

- $E(X_t) = \sqrt{Var(X_t)}$
- $E(X_t - E(X_t))^3 = 2$
- $E(X_t - E(X_t))^4 = 9$

Distributional properties relate to the non- Gaussianity of the distribution of asset returns, whilst temporal properties refer to the time dependence of asset returns and of the squared/absolute asset returns. Rydén et al (1998) and Bulla (2011) showed that an HMM with normal and t

conditional distributions, respectively, satisfies TP1 and that TP4 is not violated in practice. The absence of correlation between returns must not be mistaken for a property of independence and identical distribution: return fluctuations are not identically distributed and the properties of the distribution change with time. In particular, absolute returns or squared returns exhibit a long-range slowly decaying autocorrelation function. This phenomenon is widely known as "volatility clustering", as "large changes tend to be followed by large changes, of either sign, and small changes tend to be followed by small changes" (Mandelbrot, 1963)

3.2 Data

The data used here are the daily returns covering the period from 1st January 1987 to 5th September 2005 . All sector indices are from STOXX Ltd. and the common currency used is Euro. The daily returns of the period from t to $t-1$ 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. 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.

The following plots shows the prices of the index banks and the daily returns.

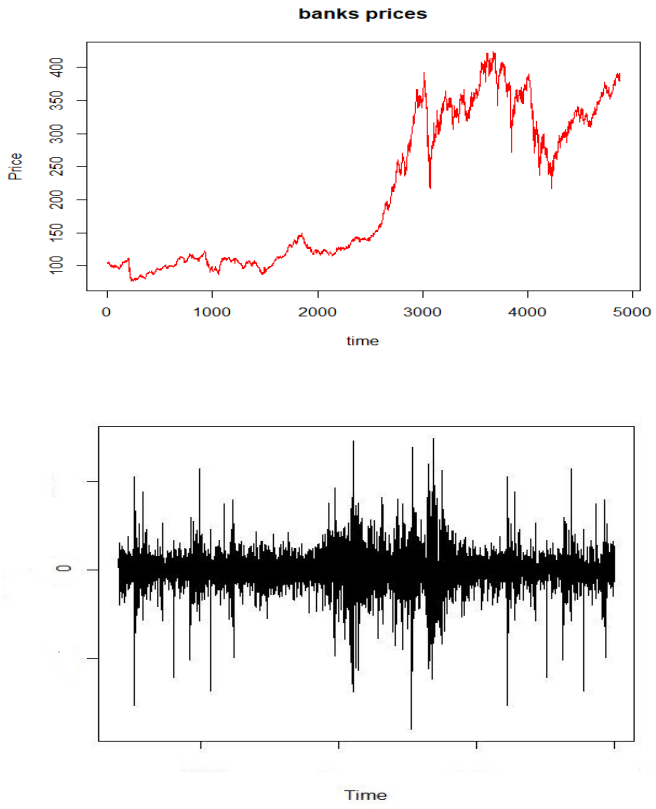

Figure 3.1: Prices and daily Returns

3.3 Descriptive Statistics

The following table presents the four moments of the daily returns of the sectors banks and chemicals. The third moment, skewness, shows that the daily returns are negatively skewed. The fourth moments, kurtosis, are larger than 3, this implies that the daily returns have the leptokurtosis and the fat tails. The third and fourth moments indicate that the distribution of the daily returns deviates from the normal distribution.

	Banks	Chemicals
Mean	2725×10^{-3}	2537×10^{-3}
Standard deviation	0.0115	0.01279
Skewness	-0.3537	-0.1249
Excess Kurtosis	6.7	5.677949

3.4 The models

In this section, we fit Hidden markov and Hidden-semi Markov models with negative binomial sojourn time distributions because the Hidden Markov models do not capture the behavior of the empirical autocorrelation function satisfactorily, mainly due to the much slower decay of the latter. The temporal dependence properties of a HMM depend on the values of the transition probability matrix but the sojourn times are always geometrically distributed.

Excess kurtosis of the data and the fitted models				
Sector	obs	HSMMt-3s	HSMMn-3s	HMMn-3s
Banks	6.7	6.33	4.97	5.02
Chemicals	5.67	9.42	3.14	4.20

Chemicals

model	AIC	BIC	likelihood
HMM-2s	-29991.58	-29946.14	15002.79
HSMMn-2s	-30080.15	-30021.73	15049.08
HSMMt-2s	-30160.05	-30088.64	15091.02
HMM-3s	-30199.56	-30108.68	15113.78
HSMMn-3s	-30234.43	-30124.08	15134.22
HSMMt-3s	**-30273.49**	**-30143.66**	15156.74

Banks

model	AIC	BIC	likelihood
HMM-2s	-31596.23	-31550.79	15805.11
HSMMn-2s	-31690.53	-31632.11	15854.27
HSMMt-2s	-31749.02	-31677.61	15885.51
HMM-3s	-31936.92	**-31846.04**	15982.46
HSMMn-3s	-31953.513	-31843.15	15993.75
HSMMt-3s	**-31962.71**	-31832.88	16001.35

In the above tables, HMM corresponds to the Hidden Markov Model, HSMMn to the Hidden Semi-Markov Model ending in n for the normal component distribution and in t for the student distribution. In the case of the HMM the component distribution is normally distributed. Figure (3.2) depicts the autocorrelation plot of the absolute values in the case of the 2-states models. Figure (3.3) depicts of the autocorrelation plot of squared values in the case of the 2states-models. Finally figures (3.4), (3.5) depict the autocorelation plot of absolute and

squared values in the case of 2-states and 3-states models, respectively. The HSSMt is represented by the blue line, the HMM with the green line and the HSMMn with the red line. The grey bars represent the empirical autocorelations.
All the 2-states models can't reproduce the long-memory observed in the ACF plots, because the autocorrelation tends to zero after the 25th lag in contrast to the 3 state models that are more efficient.

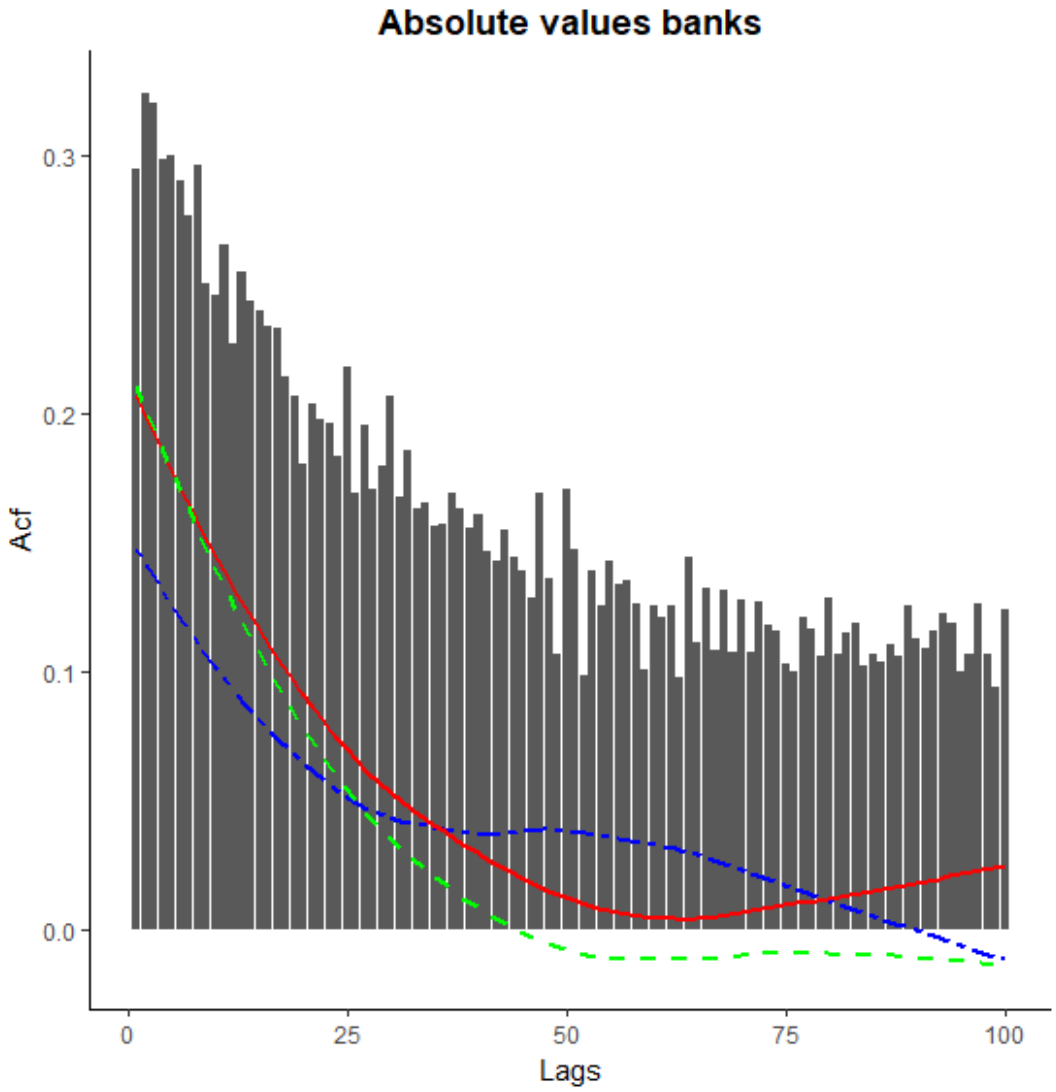

Figure 3.2: absolute values 2 states

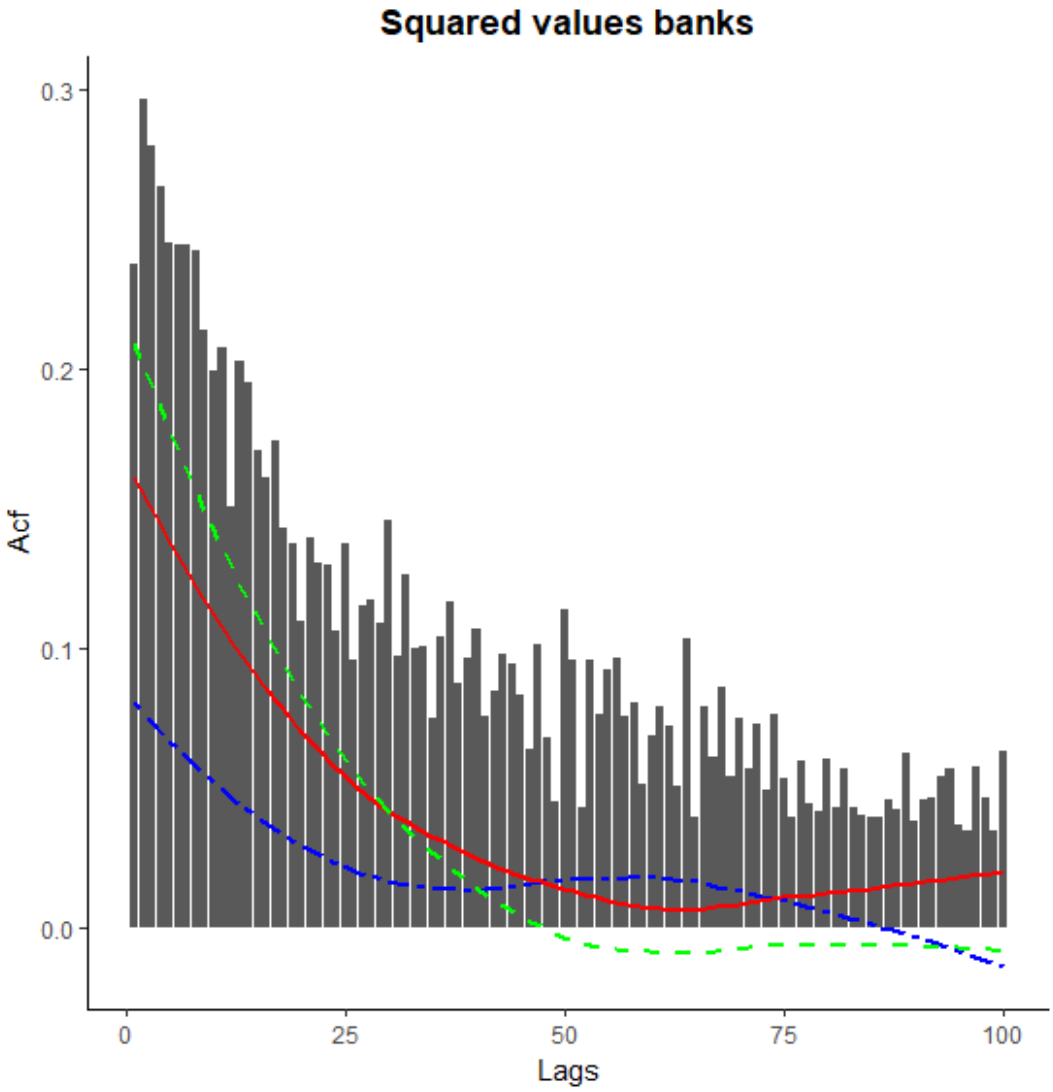

Figure 3.3: Squared values 2 states

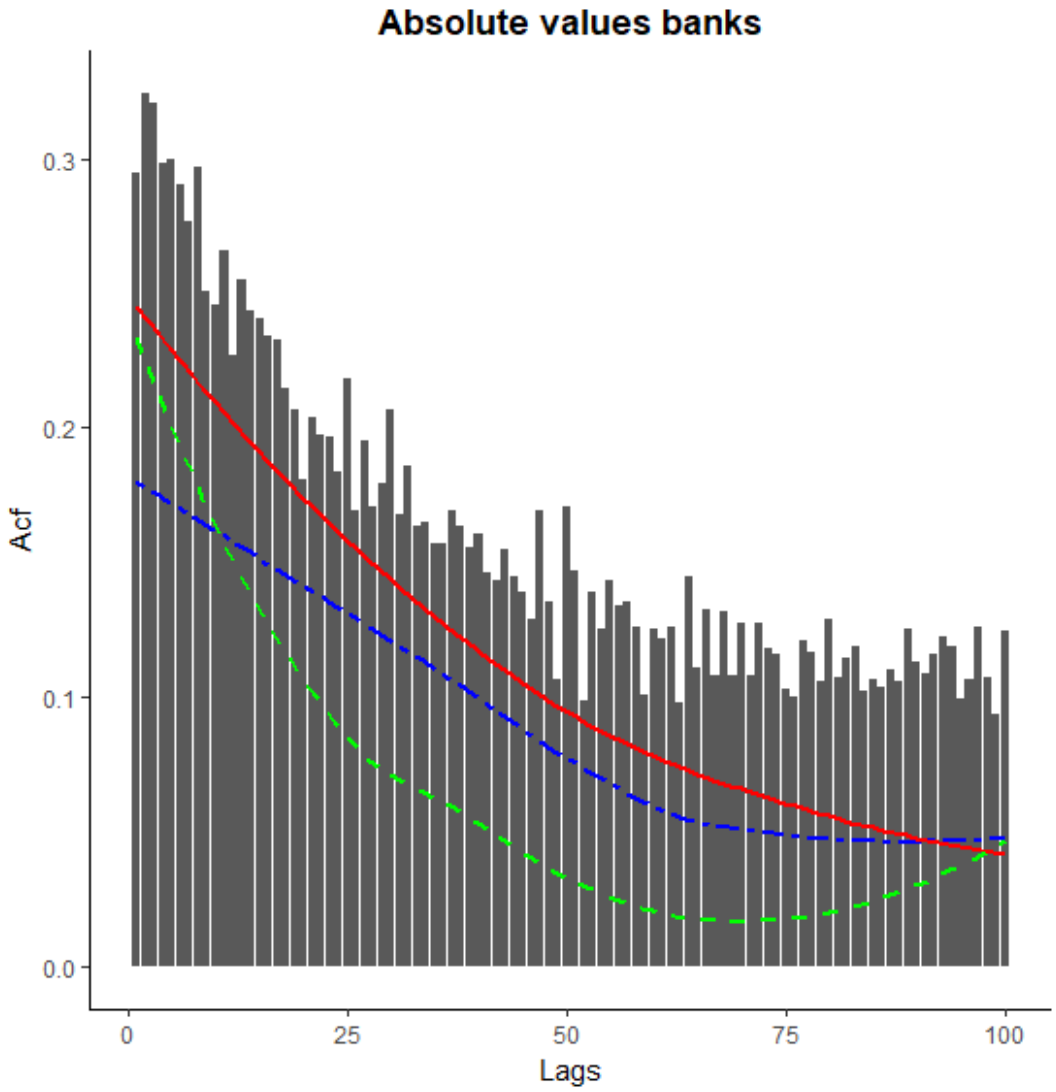

Figure 3.4: Absolute values 3 states

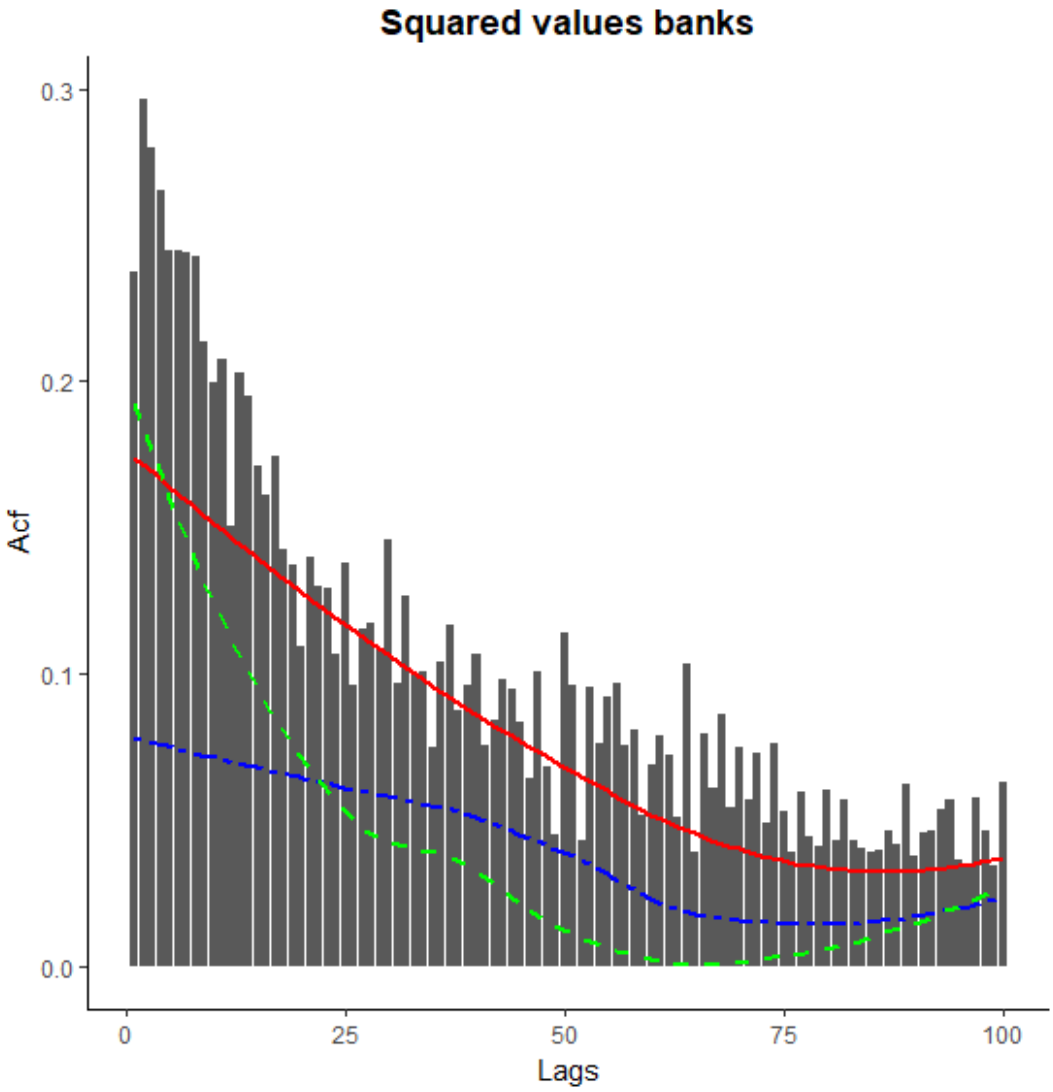

Figure 3.5: Squared values 3 states

Better fit in autocorrelation plots is perfomed by the 3 states models, with HSSM-n providing the best fit in contrast to the AIC criterion which displays HSMM-t as the best model. We used the three state models to intrepret the market conditions. Specifically, state 1 corresponds to the bear market, state 2 corresponds to the bull market and state 3 corresponds to the sidewalk market . We define the bear, sidewalk, and bull markets from the perspective of the distributional features.

Definition 3.4.1. *A bear market*

- *The mean of the distribution of the daily returns conditional on a bear market is significantly less than 0.*
- *The frequency of the positive returns is expected to be smaller than that of the negative returns.*
- *Because of the above statistical properties, the price in a bear market is generally decreasing.*

Definition 3.4.2. *A sidewalk market*

- *The mean of the distribution of the daily returns conditional on a sidewalk market should be insignificantly different from 0.*
- *It is expected to observe a roughly equal number of positive and negative returns.*
- *Because of the above statistical properties, the price in a sidewalk market stays in a band and shows a mean-reversion pattern.*

Definition 3.4.3. *A Bull Market*

- *The mean of the distribution of the daily returns conditional on a bull market should be signifi-cantly larger than 0.*
- *The frequency of the positive returns is expected to be larger than that of the negative returns.*
- *Because of the above statistical properties, the price in a bull market is generally increasing.*

3.4.1 HMM

Below we give in all cases of interest, the estimated values from the EM-algorithm as described in the previous chapter. The transition matrix :

$$P = \begin{bmatrix} 0 & 0.1 & 0.9 \\ 0.42 & 0 & 0.58 \\ 0.46 & 0.54 & 0 \end{bmatrix},$$

the observation component distribution:

$$\begin{aligned} [X_t|S_t = 1] &\sim N_1(-28 \times 10^{-4}, 74 \times 10^{-5}), \\ [X_t|S_t = 2] &\sim N_2(58 \times 10^{-5}, 384 \times 10^{-7}), \\ [X_t|S_t = 3] &\sim N_3(52 \times 10^{-5}, 135 \times 10^{-6}), \end{aligned}$$

and the sojourn time distributions :

$$\begin{aligned} d_1(u) &= (1 - 0.054)^{u-1}0.054, \\ d_2(u) &= (1 - 0.007)^{u-1}0.007, \\ d_3(u) &= (1 - 0.019)^{u-1}0.019. \end{aligned}$$

State 1 corresponds to bear market with positive return frequency 44%, state 2 corresponds to bull market with positive return frequency 54% and state 3 corresponds to sidewalk market with positive return frequency 50%.

3.4.2 HSMM

The transition matrix for the model with normal component was estimated as:

$$P = \begin{bmatrix} 0 & 0 & 1 \\ 0.98 & 0 & 0.02 \\ 0.75 & 0.25 & 0 \end{bmatrix},$$

the observation component distribution:

$$[X_t|S_t = 1] \sim N_1(-30 \times 10^{-4}, 72 \times 10^{-5}),$$
$$[X_t|S_t = 2] \sim N_2(63 \times 10^{-5}, 383 \times 10^{-7}),$$
$$[X_t|S_t = 3] \sim N_3(57 \times 10^{-5}, 127 \times 10^{-6}),$$

and the sojourn time distribution

$$d_1(u) = \binom{u-2+0.062}{u-1} 0.013^{0.062}(1-0.013)^{u-1},$$
$$d_2(u) = \binom{u-2+1.18}{u-1} 0.083^{1.18}(1-0.083)^{u-1},$$
$$d_3(u) = \binom{u-2+0.294}{u-1} 0.014^{0.294}(1-0.014)^{u-1}.$$

State 1 corresponds to bear market with positive return frequency 45%, state 2 corresponds to bull market with positive return frequency 53% and state 3 corresponds to sidewalk market with positive return frequency 49%. The two models had some differences in Hidden states but the results are pretty similar.

The transition matrix for the model with student component:

$$P = \begin{bmatrix} 0 & 0 & 1 \\ 0.95 & 0 & 0.05 \\ 0.72 & 0.28 & 0 \end{bmatrix}.$$

The first state follows a Student distribution with mean,variance, and degrees of freedom

$$(M_1, Var_1, Df_1) = (-30 \times 10^{-3}, 723 \times 10^{-7}, 100),$$

the second state :

$$(M_2, Var_2, Df_2) = (62 \times 10^{-5}, 34 \times 10^{-6}, 16.63),$$

and the third state:

$$(M_3, Var_3, Df_3) = (61 \times 10^{-5}, 117 \times 10^{-6}, 17.5).$$

The sojourn time distributions :

$$d_1(u) = \binom{u-2+0.092}{u-1} 0.014^{0.092}(1-0.014)^{u-1},$$
$$d_2(u) = \binom{u-2+1.28}{u-1} 0.008^{1.28}(1-0.008)^{u-1},$$
$$d_3(u) = \binom{u-2+0.32}{u-1} 0.011^{0.32}(1-0.011)^{u-1}.$$

Fitting a t-distribution for banks had many computational issues (good initial values were needed and we had to relax the stopping criterion). When fitting t-distribution with 3-states the 2 states looks too similar, so to capture the volatility clustering 3-states with normal or 2-states with Student will indicate good results.

3.5 Conclusion

In this work, we appplied hidden semi-Markov models (HSMM) and hidden Markov models (HMMs) to explain the time-varying distribution of stock market returns. Our results indicate that the time-varying distribution of the stock market returns depends on the market conditions, namely the bear, sidewalk, and bull market. Stylized facts, AIC and BIC indicated different models for best fit but all showed that the 3-states models are performing better than the 2-states. Through Monte Carlo simulations, we have found that our three-state HSMM models can reproduce the stylized facts of the long-memory and the Taylor effect, whereas the 2 state-models can't.

Bibliography

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

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

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

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

[5] 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.*

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

[7] *Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). Journal of the Royal Statistical Society B 39, 1-38.*

[8] 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.*

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

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

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

[12] Guedon, Y. (2003). *Estimating hidden semi-markov chains from discrete sequences. Journal of Computational and Graphical Statistics, 12 , 604-639.*

[13] Guedon, Y. (2005), *Hidden equilibrium semi-Markov chains. Preprint.*

[14] Guedon, 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.*

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

[16] 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.*

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

[18] Mandelbrot, B. B. (1997). *The variation of certain speculative prices. Springer.*

[19] Peel, D. & McLachlan, G. J. (2000),*'Robust mixture modelling using the t-distribution', Stat. Comput. 10, 339-348.*

[20] Ryden, T. (1995a), *'Consistent and asymptotically normal parameter estimates for Markov modulated Poisson processes', Scandinavian Journal of Statistics. Theory and Applications 22(3), 295–303.*

[21] Ryden, T. (1995b), *'Estimating the order of hidden Markov models', Statistics. A Journal of Theoretical and Applied Statistics 26(4), 345–354.*

[22] Ryden, 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.*

[23] Sansom, J. & Thomson, P. (2000), *'Fitting hidden semi-Markov models', NIWA Technical Report NTR77. National Institute of Water and Atmospheric Research, New Zealand.*

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

[25] Vlad Stefan Barbu, Nikolaos Limnios *Semi-Markov Chains and Hidden Semi-Markov Models toward Applications*

[26] Zhenya Liu,Shixuan Wang *Decoding Chinese Stock Market Returns: Three-State Hidden Semi-Markov Model*