Hidden Markov Models with Time-Continuous Output Behavior

Vom Fachbereich Mathematik und Informatik
der Freien Universität Berlin
zur Erlangung des akademischen Grades
eines Doktors der Naturwissenschaften
genehmigte Dissertation

vorgelegt von

Diplom–Mathematikerin Evelyn Dittmer

Berlin, 11. November 2008

Betreuer: Prof. Dr. Christof Schütte
Freie Universität Berlin
Fachbereich Mathematik und Informatik
Arnimallee 2-6,
14195 Berlin

Gutachter: Prof. Dr. Christof Schütte
Dr. Wilhelm Huisinga

Datum der Disputation: 11. Februar 2009

Contents

1 Introduction

Molecular dynamics (MD) analysis plays an important role in materials science, biophysics and biochemistry. It is applied to protein structure determination and also to real-life applications such as drug design.

For a better understanding of a biomolecule's dynamical behavior, time series obtained from MD simulations [31, 1] are investigated with the goal to find dynamical and structural patterns. In our application context such patterns are conformations.

Conformations can be seen as almost invariant or metastable sets in the configuration space [43], which are characterized either geometrically as possible shapes of a molecule, statistically in terms of distributions over the configuration space or kinetically by means of the local dynamics that describe the oscillation and relaxation behavior of a molecule. Biomolecular systems possess only few dominant conformations that can be modeled as metastable states.

Typical conformations provide information about the function of a biomolecule [66, 76]. Biomolecules do not exist in a unique structure. They fluctuate within a conformation, and occasionally perform a transition to another one [28]. After a conformational change the molecular system takes a certain time to equilibrate within the respective conformation (towards a "local equilibrium"). The changes occur rarely on a time scale on which the dynamical behavior of a molecule usually is simulated.

We aim at a characterization of the system with regard to the behavior on different time scales: on a macroscopic time scale, on which conformational changes take place, and on a microscopic time scale, on which the biomolecule is relaxing towards an local equilibrium and oscillates within a conformation in local equilibrium. That is, we have to select a model which represents both levels – long-term and short-term behavior – properly.

We will approach the task of *model selection* in an "a priori" manner: the model structure is specified based on previous knowledge about the kind of processes we have to handle. As a first step the most important features of the systems under consideration have to be identified. The model should provide a reduced description, which on one hand essentially characterizes the behavior of the system in terms of rare and instantaneous conformational changes and on the other hand reflects the local behavior within different conformations appropriately.

Once we have selected a model structure, we must specify model parameters. The problem of *model parameterization* is addressed with an "a posteriori" method. Biomolecular systems (or reduced test systems) are simulated by MD techniques. This way we obtain a time series and based on this we can determine model parameters. Eventually, conformational patterns are recognized by a time series analysis including parameter estimation and clustering of the time series into metastable sets.

However, a satisfying quality of the analysis results is achieved only, if the model is designed carefully and fits to the effective dynamics of the system. In the following we will characterize the systems under consideration in more detail.

System specification and objectives. The processes analyzed in the scope of this work exhibit a certain dynamical behavior: Namely, the state space possesses only few metastable sets and the macroscopic dynamics can be modeled by a Markov jump process. That is, on an appropriate time scale the jump process describing the transitions between metastable states is Markovian. This Markov jump process is a kind or flipping process with instantaneous transitions between metastable states. Typical correlation times in the system are sufficiently smaller than the waiting times between the hops amongst the metastable sets. On a microscopic time scale, the process may possess memory.

The crucial task in conformational analysis is the identification of metastable conformations. The metastable sets are not known a priori and therefore have to be extracted from the time series [53]. They can be recovered by certain geometrical, statistical or kinetical patterns. However, these patterns are unknown a priori as well.

Methods. For the identification of the unknown metastable states, which are understood as hidden in the data, we will apply hidden Markov models (HMMs). An HMM consists of two stochastic processes, of which only one is observable. These processes are referred to as "hidden" and "output" processes. The combination of different processes fits to our application context since we aim to model two different levels of the system: conformational changes and oscillatory behavior within a conformation.

The concept of hidden Markov models has been developed in the recent 50 years. Invented by Baum et al. [6, 7, 8] in the late sixties it was applied first in areas such as speech recognition [64] as well as communication and control theory [70]. Later, HMMs have also been applied in bioinformatics to identify genome or protein sequences [14, 4, 17]. In recent works it has also been applied to biomolecular dynamics [41] and climate data [37, 30].

As the history of HMMs shows, they have been invented in completely different application areas. Patterns used as output distributions so far provide a geometrical or statistical description of the data only.

The novel approach in this thesis is that we employ *kinetic models* ([72], Chap. 6) as HMM output processes. In this context a kinetic model means either a dynamical model specified by an SDE [48, 69], or a rate process realized by a random walk [44, 69, 71]. Thus, we can also express a transition behavior in the output process. We obtain a tool that allows for the identification of different metastable sets, which in general will not be dis-

tinguishable geometrically or statistically. The distinction of the metastable patterns is based on the detection of kinetic signals. By means of standard HMMs these patterns are not recognizable.

Furthermore, the model described above provides a deeper insight into the dynamical behavior of the process within the metastable states. The crucial distinction to standard HMM is illustrated in figure 1.1.

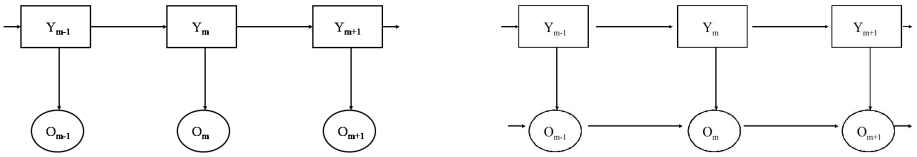

Figure 1.1: Standard HMM (left) and HMM with kinetic output process (right).

The kinetic processes considered in this thesis are *time-continuous Markov processes*: The *Ornstein-Uhlenbeck* process and the *Markov jump process* (MJP). They are used here in combination with HMMs. Depending on the time series under investigation we select an appropriate output process. We will call HMMs with Ornstein-Uhlenbeck output process *HMMSDE*, HMMs with Markov jump output process *HMM-MJP*.

The Ornstein-Uhlenbeck process is a good choice for modeling diffusive molecular dynamics in a harmonic well in the energy landscape (Smoluchowski dynamics). HMMSDE therefore fits well in the biomolecular application context. For this approach no discretization of the data is necessary. The MJP on the other hand requires discretized data. Even if the box-discretization is rather coarse-grained it is still able to recover metastable patterns. The MJP has the advantage that no assumption about the distribution of the observables is necessary. Only the transition behavior is taken into account.

As the state space of an MJP is discrete it can be employed not only as output, but also as hidden process. HMMs with hidden MJP are also discussed in this thesis. It can handle time series with non-equidistant time steps. The Ornstein-Uhlenbeck process is used as output process only since it has a continuous state space. However, in an HMM the hidden state space is discrete – in our case even finite.

Algorithmic proceeding. The algorithmic proceeding includes two steps: Recovering a model from data and inferring the metastable states of each data point. The first step is carried out by the estimation of model parameters. A central technique for this purpose is the well-known EM algorithm [22]. Whenever we derive an EM algorithm in Sections 3.1, 3.1.1, 5.2 - 5.3, 6.2 - 6.3, 7.1, and 7.2 we use the following paradigm:

1. First, we specify the joint likelihood of observable and hidden variables.
2. Then, we take partial observability into account.
    - Expectation-step: Compute the expectation value of the likelihood over the hidden variables. Write down the functional Q, which contains the expected likelihood.
    - Maximization-step: Establish the reestimation formulas by differentiating Q with respect to the model parameters.
3. Eventually, we give the algorithmic scheme to exemplify the iteration procedure.

As a second step we identify the unknown metastable states as hidden states of the respective HMM. The assignment of each data point to a metastable state implies a clustering of the time series.

The EM algorithm is relevant not only for the identification of HMM parameters but also for the estimation of the generator matrix of a Markov jump process. Generator estimation is a research area on its own. In this work we discuss and compare several approaches to this topic suggested in [58, 18, 11]. It is closely related to the *Imbedding Problem* [47]. The Imbedding Problem deals with the question when a discrete transition matrix is imbeddable into a time-continuous Markov process. A new result about imbedding of perturbed transition matrices is derived in Section 2.1.5.

Examples. The presented models have been designed for the analysis of biomolecular systems. In this work they will be applied to test examples, where complex systems are reduced to smaller ones that still reflect the main features of biomolecular time series: They are metastable and on a macroscopic time scale, on which the transitions between metastable states take place, they fulfill the Markov property.

Small test examples are expedient for illustrative examples. However, the algorithms are all applicable to higher dimensional molecule time series.

In several examples we will analyze time series that arise either from an HMM, from a discrete generator, or from Smoluchowski dynamics. The analysis of an HMM realization can be understood as reliability test. Estimated and original parameters are directly comparable. The realizations of Smoluchowski dynamics are of most physical relevance, as the resulting dynamics reflect a given potential. These time series are closely related to MD simulations, even though we take small toy potentials.

Alternative approaches. The combination of HMMs with other stochastic processes has been investigated in several other works. An overview is given in Section 7.3. The distinct feature of the techniques presented in this work is the combination of an HMM with kinetic models. Standard

HMMs combine a dynamical analysis by means of the hidden process with a geometrical analysis by means of the output process. The models presented here, namely HMMSDE and HMM-MJP, contain a dynamical aspect in the output process as well. This novel approaches allow for the recognition of kinetic patterns, even if the distributions are geometrically identical or largely overlapping as illustrated in the second part of the example from Section 7.2.4.

Outline. This thesis is organized as follows: Chapter 2 and 3 provide the theoretical part introducing the concepts that are used in the following chapters. In Chapter 2 the relevant stochastic processes are described. Chapter 3 presents the hidden Markov model framework as well as associated problems and algorithms. We place emphasis on the EM algorithm, used for the parameter estimation beyond the HMM context also for generator estimation.

In the remaining part of this thesis we will investigate how the concepts of HMMs and kinetic models (here, time-continuous Markov processes) can be combined. Chapter 4 introduces Ornstein-Uhlenbeck processes. The maximum likelihood estimators for a fully observable Ornstein-Uhlenbeck process are all given there.

Chapter 5 describes the combination of HMMs with Ornstein-Uhlenbeck processes, the *HMMSDE*. The corresponding EM algorithm is specified and finally elucidated by an example.

In Chapter 6 we discuss Markov jump processes in detail. We describe and compare different approaches for the parameter estimation: A method using the resolvent, a quadratic programming method and a maximum likelihood estimation method. The latter approach is preferable in combination with hidden Markov models and hence we describe it in more detail. In Chapter 7 we will combine HMMs with Markov jump processes. We call this approach *HMM-MJP*. Once a maximum likelihood estimator (MLE) is defined, it is compatible to the HMM framework. A time-continuous Markov process fits either as hidden process into an HMM or as output process. Both possibilities are discussed and exemplified by several examples.

Acknowledgement At this point I want to thank all those, who accompanied and supported me during the past four years and contributed to the success of this thesis. First of all I want to thank my advisor Christof Schütte, who guided me through this research field between numerics and stochastic. His great mathematical and personal support enabled me to make my scientific goals compatible with my family.

I wish to express special gratitude to Illia Horenko for his advise and for his constant presence. Working with him taught me a lot about time series analysis and innovative ways to approach numerical problems. I thank very much Tim Conrad for his patience and valuable suggestions. To Philipp Metzner and Tobias Jahnke I am indebted for the stimulating scientific exchange. I obtained many inspirations in the course of our common work. Special thanks go to the Biocomputing group for the inspiring atmosphere.

For mental and mathematical counsel as well as for proof reading I warmly want to thank my room-mate Andrea Weiße. Furthermore, my gratitude applies to Lars Dittmer for his language assistance.

Last but definitely not least I want to express my thanks to my family: My son Livian and his father Tibor for giving me energy and balance as well as my parents and my sister for just being there and accompanying me all the time.

The work was funded by a Konrad-Zuse scholarship and by the DFG Priority Program "Analysis, Modeling and Simulation of Multiscale Problems".

2 Markov Processes

Consider a stochastic process on $(S, \mathcal{B}, P)$, where S denotes the state space of the process $X(t),\ t \in \mathbb{R}$, (B) the σ-algebra and P the probability measure.

Definition 2.1 (Markov property [69], Chap. 1.5, p. 9)**.** $X(t)$ is called a Markov process with state space S, if for any $0 \leq t_1 < t_2 < ... < t_n$ and $B \in \mathcal{B}$

$$P(X(t_n) \in B | X(t_1), ..., X(t_{n-1})) = P(X(t_n) \in B | X(t_{n-1})).$$

A Markov process is continuous in time ([33], Chap.3.3.1, p. 46), if for any $\epsilon > 0$ and $z \in S$ we have

$$\mathbb{P}\left(\lim_{\Delta t \to 0} \frac{1}{\Delta t} \int_{|x-z|>\epsilon} P(x, t + \Delta t | z, t) dx = 0 \right) = 1.$$

Markov processes are kinetic processes. They describe the evolution of phase transitions of a system over time.

2.1 Markov Jump Processes

Let S be a finite set of states and $\{X(t)\}_{t \geq 0}$ an S-valued time-continuous Markov process.

Definition 2.2. A time-continuous Markov process with the following regularity conditions $\forall i, j \in S$

1.) $\mathbb{P}(0, i, j) = \delta_{ij}$

2.) $\lim_{t \to 0+} \mathbb{P}(t, i, j) = \delta_{ij}$

3.) its realization $X(t)$ is right-continuous and the limit from the left exists,

is called **Markov jump process**.

Regularity ensures right-continuity of the process. That is, there exists a $\tau > 0$ such that

$$X(t + s) = X(t)\ \forall s < \tau\ a.s.$$

For a detailed introduction of Markov jump processes see [44, 69, 71]. A Markov jump process can be considered as composition of two stochastic processes. The first is a Poisson process $N(t)$ with parameter λ, which controls the time between two state transitions (jump times) and the second is a time-independent Markov chain with transition matrix K, which controls the state transitions.

Jump times are exponentially distributed but in contrast to an ordinary Poisson process the state transitions are not deterministic from i to $i+1$ but rather stochastic from i to j with probability K_{ij}. We can write the probability density for a transition exactly at time t from state $i \in S$ to state $j \in S$ as

$$\mathbb{P}(i, j, \text{jump at } t) = K_{ij}\lambda \exp(-\lambda t), \tag{2.1}$$

since the jump times are independent from the state transitions.

The transition probability from i to j within a time interval $[0, t]$ is the probability for n jumps multiplied with the probability for a state transition from i to j in n steps, for $n = 0, ..., \infty$ [44].

$$\mathbb{P}(i, j, t) = \sum_{n=0}^{\infty} \mathbb{P}(N(t) = n) K_{ij}^n = \sum_{n=0}^{\infty} \frac{(t\lambda)^n}{n!} \exp(-\lambda t) K_{ij}^n \tag{2.2}$$

2.1.1 The Generator Matrix

With $\lambda < \infty$ the transition probability is differentiable. In the following we denote with $P(t) = (\mathbb{P}(i, j, t))_{ij}$ the time-dependent transition operator.

Definition 2.3. We call the differential of the transition probability in time point 0 the **infinitesimal generator**.

$$\lim_{t \to 0} \frac{P(t) - P(0)}{t} = L$$

With (2.2) in matrix notation we get

$$\begin{aligned} L &= \frac{\partial}{\partial t} \sum_{n=0}^{\infty} \frac{(t\lambda K)^n}{n!} \exp(-\lambda t)\big|_{t=0} \\ &= \frac{\partial}{\partial t} \text{expm}\left(t\lambda(K - Id)\right)\big|_{t=0} \\ &= \lambda(K - Id)\text{expm}\left(t\lambda(K - Id)\right)\big|_{t=0} \\ &= \lambda(K - Id). \end{aligned} \tag{2.3}$$

The notation expm denotes here the matrix exponential. In particular we obtain a concise expression for the transition probability:

$$\mathbb{P}(i, j, t) = \text{expm}(tL)_{ij}. \tag{2.4}$$

From (2.3) follows immediately that the generator matrix has the following structure

$$\begin{aligned} L_{ij} &\geq 0 && \forall i, j \text{ with } i \neq j \\ L_{ii} &= -\sum_{j \neq i} L_{ij} && \forall i. \end{aligned} \tag{2.5}$$

Definition 2.4 ([49])**.** A family of stochastic transition matrices $P(t)$ satisfying the properties

1.) $P(0) = Id$,

2.) $P(s)P(t) = P(s+t)$

is called a **transition semi-group**.

The transition probability defined in (2.4) is apparently a transition semi-group. From the semi-group property follows immediately

$$\frac{P(t+h) - P(t)}{h} = P(t)\frac{P(h) - Id}{h} = \frac{P(h) - Id}{h}P(t).$$

The first equality gives the **Kolmogorov forward equation**

$$P'(t) = P(t)L, \tag{2.6}$$

the second equality on the other hand the **Kolmogorov backward equation**

$$P'(t) = LP(t).$$

2.1.2 The Imbedding Problem

A Markov process $X(t)$ observed on equidistant time points $t_1 = \tau, ..., t_n = n\tau$ defines a discrete Markov chain $Y_1 = X(\tau), ..., Y_n = X(n\tau)$ with transition matrix $\tilde{P} = \text{expm}(\tau L)$. Yet conversely not every discrete Markov chain belongs to a continuous Markov process. In other words, not every discrete Markov chain can be imbedded into a time-continuous Markov process. The **imbedding problem** addresses just that question:

Does a time-continuous Markov process $P(t)$ for a given discrete Markov chain $\tilde{P}$ exist, such that for an appropriate τ $P(\tau) = \tilde{P}$ holds?

If such a process exists, it can be expressed in terms of the generator L, which is up to a scalar factor τ the logarithm of $\tilde{P}$, since

$$P(\tau) = \text{expm}(\tau L).$$

But $\tilde{L} = \tau^{-1}\text{logm}(\tilde{P})$ is not guaranteed to satisfy the generator constraints (2.5). Furthermore, the complex logarithm is not unique. Hence, if a matrix has complex eigenvalues, the matrix logarithm is also not unique. This means, there exist possibly no, one or a finite number of generators (that an infinite number of generators is impossible will be elucidated later on page 13). Firstly we will define the set of generators by

$$\mathcal{G} = \left\{ L \in \mathbb{R}^{d\times d} : L_{ij} \geq 0 \quad \text{for all } i \neq j, \quad L_{ii} = -\sum_{j\neq i} L_{ij} \right\} \tag{2.7}$$

and the set of imbeddable Markov chains by

$$\mathcal{P}_\tau = \left\{ \tilde{P} \in \mathbb{R}^{d\times d} : \exists L \in \mathcal{G} \text{ such that } \tilde{P} = \operatorname{expm}(\tau L) \right\}.$$

In the following some criteria for a stochastic matrix $\tilde{P}$ are listed, that guarantee that a generator possibly exists and if it is unique or not. Note that we have assumed S to be finite, such that the transition matrices are finite-dimensional.

But first we give some preliminary remarks about the computation of the logarithm. The logarithm of a matrix P, which is sufficiently close to the identity, such that $\|P - Id\| < 1$ holds, can be computed by

$$\operatorname{logm}(P) = -\sum_{r=1}^{\infty} \frac{(Id - P)^r}{r}. \tag{2.8}$$

If $P = VDV^{-1}$ is diagonizable, the matrix logarithm is simply

$$\operatorname{logm}(P) = V(\operatorname{logm}(D))V^{-1}.$$

Since $D = \operatorname{diag}\{\Lambda_i\}_{i\in S}$ is a diagonal matrix that contains the eigenvalues $\Lambda_1, ..., \Lambda_n$ of P, its logarithm is simply the scalar logarithm of the eigenvalues

$$\operatorname{logm}(D) = \operatorname{diag}\{\log(\Lambda_i)\}_{i\in S}.$$

The real scalar logarithm is unique, but the complex logarithm is not.

$$\log(\Lambda) = \log|\Lambda| + i\left(\arg(\Lambda) + 2\pi k\right),\ \forall k \in \mathbb{Z} \tag{2.9}$$

It is defined up to a multiple of $2\pi i$. That is, we can add to the eigenvalues of a matrix $L = \operatorname{logm}(P)$ any multiple of $2\pi i$ and obtain a matrix $\bar{L}$ with

$$\operatorname{expm}(L) = \operatorname{expm}(\bar{L}).$$

The logarithm (2.9) with $k = 0$ is called the **principal branch** of the logarithm or **principal logarithm**.

Remark 2.1. Note that the principal logarithm is always computed by (2.8).

Multiple branches of the logarithm imply the problem: If we compute the principal logarithm $\tilde{L}$ of a given stochastic matrix $\tilde{P}$ and

$$\tilde{L} \notin \mathcal{G}$$

we can not infer, that

$$\tilde{P} \notin \mathcal{P}.$$

Example 1. The transition matrix

$$\tilde{P} = \begin{pmatrix} 0.2742 & 0.5360 & 0.1858 & 0.0018 & 0.0005 & 0.0006 & 0.0011 \\ 0.2755 & 0.5349 & 0.1860 & 0.0017 & 0.0003 & 0.0004 & 0.0012 \\ 0.2755 & 0.5360 & 0.1849 & 0.0014 & 0.0004 & 0.0007 & 0.0011 \\ 0.0003 & 0.0005 & 0.0002 & 0.9977 & 0.0002 & 0.0004 & 0.0006 \\ 0.0007 & 0.0010 & 0.0004 & 0.0004 & 0.9957 & 0.0009 & 0.0008 \\ 0.0004 & 0.0007 & 0.0004 & 0.0008 & 0.0007 & 0.9963 & 0.0007 \\ 0.0005 & 0.0011 & 0.0004 & 0.0005 & 0.0003 & 0.0006 & 0.9967 \end{pmatrix}$$

has the principal logarithm

$$\tilde{L} = \begin{pmatrix} -5.2440 & 7.0526 & \boxed{\text{-1.8141}} & 0.0020 & 0.0015 & 0.0022 & \boxed{\text{-0.0001}} \\ 1.4349 & -3.8862 & 2.4481 & 0.0021 & 0.0000 & \boxed{\text{-0.0006}} & 0.0019 \\ 3.6256 & 0.7442 & -4.3722 & 0.0000 & 0.0001 & 0.0012 & 0.0011 \\ 0.0005 & 0.0002 & 0.0004 & -0.0023 & 0.0002 & 0.0004 & 0.0006 \\ 0.0015 & \boxed{\text{-0.0006}} & 0.0013 & 0.0004 & -0.0043 & 0.0009 & 0.0008 \\ 0.0003 & 0.0003 & 0.0009 & 0.0008 & 0.0007 & -0.0037 & 0.0007 \\ 0.0000 & 0.0019 & 0.0000 & 0.0005 & 0.0003 & 0.0006 & -0.0033 \end{pmatrix},$$

which is certainly not in $\mathcal{G}$, since there are several negative off-diagonal entries. The eigenvalues of $\tilde{L}$ are

$$(0.0000, -0.0032, -0.0053, -0.0047, -0.0041, -6.7493{+}3.0545i, -6.7493{-}3.0545i).$$

If we add to the 6-th eigenvalue $2\pi i$ and on the 7-th eigenvalue $-2\pi i$, we obtain another matrix logarithm of $\tilde{P}$

$$\tilde{\tilde{L}} = \begin{pmatrix} -4.5096 & 0 & 4.5042 & 0.0035 & 0.0008 & 0 & 0.0011 \\ 2.3155 & -2.3184 & 0 & 0.0013 & 0 & 0 & 0.0016 \\ 0 & 6.6707 & -6.6743 & 0 & 0.0009 & 0.0027 & 0 \\ 0.0005 & 0.0004 & 0.0002 & -0.0023 & 0.0002 & 0.0004 & 0.0006 \\ 0.0017 & 0.0005 & 0.0000 & 0.0004 & -0.0043 & 0.0009 & 0.0008 \\ 0.0007 & 0 & 0.0008 & 0.0008 & 0.0007 & -0.0037 & 0.0007 \\ 0 & 0.0012 & 0.0007 & 0.0005 & 0.0003 & 0.0006 & -0.0033 \end{pmatrix},$$

for which $\tilde{\tilde{L}} \in \mathcal{G}$ and $\text{expm}(\tilde{\tilde{L}}) = \tilde{P}$ holds.

2.1.3 Necessary Conditions for the Existence of a Generator

Due to the periodicity of the complex matrix logarithm, there are an infinite number of matrices $L_i \neq L$ with $\text{expm}(L_i) = \text{expm}(L)$. However, we will see in the following, that only a finite number of these matrices satisfy the generator constraints. For this purpose we study the connections between the spectral and algebraic properties of matrices from $\mathcal{G}$ and $\mathcal{P}$.

Theorem 2.1 (of Geršgorin [54])**.** Let A be an $n \times n$-matrix with entries $A_{ij} \in \mathbb{C}$ and define the Geršgorin discs by

$$\mathcal{D}_i = \Big\{ z \in \mathbb{C} : |z - A_{ii}| \leq \sum_{\substack{j=1 \\ j \neq i}}^{n} |A_{ij}| \Big\}, \text{ for } 1 \leq i \leq n.$$

Then all eigenvalues of A are contained in $\bigcup_{i=1}^{n} \mathcal{D}_i$, the union of the Geršgorin discs.

Applied to stochastic matrices P, the Geršgorin discs become

$$\mathcal{D}_i = \Big\{ z \in \mathbb{C} : |z - P_{ii}| \leq 1 - P_{ii} \Big\}, \text{ for } 1 \leq i \leq n.$$

Let now $g(t) = \min_i P_{ii}(t)$. Then all eigenvalues lie in the disc around $g(t)$ with radius $1 - g(t)$:

$$|\Lambda_i - g(t)| \leq 1 - g(t).$$

Analogously, for generator matrices the Geršgorin discs have the shape

$$\mathcal{D}_i = \Big\{ z \in \mathbb{C} : |z - L_{ii}| \leq |L_{ii}| \Big\}, \text{ for } 1 \leq i \leq n.$$

With $l_{min} = \min_i L_{ii}$ all eigenvalues lie in the disc around l_{min} with radius $-l_{min}$:

$$|\lambda_i - l_{min}| \leq -l_{min}. \tag{2.10}$$

Theorem 2.2. Let L be an $n \times n$ matrix in $\mathcal{G}$, with eigenvalues $|\lambda_1| \leq ... \leq |\lambda_n|, \lambda_k \in \mathbb{C}$ and $P = P(1) = \text{expm}(L)$ the associated transition matrix. We denote the eigenvalues of the transition matrix P by $\Lambda_1, ..., \Lambda_n \in \mathbb{C}$. Then the following properties are satisfied

1. Complex eigenvalues λ_i (resp. Λ_i) can occur only in complex conjugated pairs.
2. $\text{trace}(L) = \log[\det(P)]$.
3. $|\Lambda_i| \leq 1$, $\forall i$
4. $|\lambda_i - \text{trace}(L)| \leq -\text{trace}(L)$, $\forall i$.

The first property follows from the fact that L is a real matrix and hence the characteristic polynomial has real coefficients [63]. The second property can be found in [47]:

$$\text{trace}(L) = \sum_i \lambda_i = \sum_i \log[\Lambda_i] = \log[\prod_i \Lambda_i] = \log[\det(P)].$$

Note, that complex eigenvalues can only occur in complex conjugated pairs. The sum of the eigenvalues therefore is real. The last two properties are

given in [46]. Property 3 follows from the well known theorem of Frobenius ([68], Chap. 1.1, p. 3-4) and property 4 can be checked by

$$\begin{aligned} & |\lambda - l_{ii}| && \leq -l_{ii} \\ \Leftrightarrow \quad & |\lambda - l_{ii}| + |\textstyle\sum_{j\neq i} l_{jj}| && \leq -l_{ii} + |\displaystyle\sum_{j\neq i} l_{jj}|. \end{aligned}$$

Since

$$|\lambda - l_{ii}| + |\sum_{j\neq i} l_{jj}| \geq |\lambda - \sum_{j} l_{jj}|$$

and all diagonal entries are negative, thus

$$-l_{ii} + |\sum_{j\neq i} l_{jj}| = |\sum_{j} l_{jj}|$$

holds. We finally obtain

$$\begin{aligned} |\lambda - \sum_{j} l_{jj}| \quad &\leq \quad |\sum_{j} l_{jj}| \\ \Leftrightarrow |\lambda - \mathrm{trace}(L)| \quad &\leq \quad |\mathrm{trace}(L)|. \end{aligned}$$

The disc defined by (2.10) is smaller than $|\lambda - \mathrm{trace}(L)| \leq |\mathrm{trace}(L)|$. However, since by property 2 a relation between the eigenvalues of the generator and the transition matrix is given, the rougher estimates from Theorem (2.2) provide some nice necessary conditions. We will exploit that in the following. Putting together properties 2 and 4 we get

$$|\lambda_i - \log[\det(P)]| \leq |\log[\det(P)]|. \tag{2.11}$$

Since $\mathrm{trace}(L)$ is real, $\log[\det(P)]$ also is real. And as the real logarithm of $\det(P)$ is defined, follows $\det(P) > 0$. And hence we obtain criteria for the eigenvalues of P: $\Lambda_i \neq 0$ and the number of $\Lambda_i < 0$ has to be even.

Further, if we write the eigenvalues λ_i of L as (2.9) and bear in mind that $\log[\det(P)]$ is purely real, we see easily that only a finite number of multiples of $2\pi i$ can be added without leaving the disk defined by (2.11) and that $|\Lambda_i| = 1$ is true if and only if $\Lambda_i = 1$. These observations have already been discussed in [26].

Other criteria have been proven by Goodman [34]:

$$\prod_{i=1}^{n} P_{ii} \geq \det(P),$$

and by Chung ([16], Chap. II.1, p.126)

$$\exists\ t_0, \text{ such that } P_{ij}(t_0) = 0 \Leftrightarrow P_{ij}(t) = 0\ \forall t > t_0$$

and

$$\exists\ t_0,\ \text{such that } P_{ij}(t_0) \neq 0 \Leftrightarrow P_{ij}(t) \neq 0\ \forall t > t_0.$$

Altogether we obtain the following corollary:

Corollary 2.1. Let P be a stochastic matrix with state space S and eigenvalues $(\Lambda_i)_{i\in S}$. If one of the following **necessary criteria** is *not* satisfied,

1. $\prod_{i=1}^{n} P_{ii} \geq \det(P) > 0$,
2. $\exists\ t_0,\ \text{such that } P_{ij}(t_0) = 0 \Leftrightarrow P_{ij}(t) = 0 \forall t > t_0$,
3. $\exists\ t_0,\ \text{such that } P_{ij}(t_0) \neq 0 \Leftrightarrow P_{ij}(t) \neq 0 \forall t > t_0$,
4. $\Lambda_i \neq 0,\ \forall i \in S$,
5. the number of $i \in S$ with $\Lambda_i < 0$ has to be even.
6. $|\Lambda_i| = 1$, if and only if $\Lambda_i = 1,\ \forall i \in S$,
7. if Λ_i are real distinct eigenvalues $\Rightarrow \Lambda_i > 0,\ \forall i \in S$

P is not imbeddable.

2.1.4 Uniqueness of the Generator

Eventually, we will address the question, when a unique generator exists. Cuthbert [19] gave some results about the uniqueness of the matrix logarithm for finite-state Markov processes and formulated the following two theorems:

Theorem 2.3. [19] Let P be a stochastic matrix with $\inf\{P_{ii}\} > \frac{1}{2}$, then P is imbeddable in a uniformly continuous way. For the principal logarithm

$$-\sum_{r=1}^{\infty} \frac{(Id - P)^r}{r} \in \mathcal{G}$$

holds.

Convergence of the series (2.8) is guaranteed by

$$\begin{aligned} 1 - \frac{1}{2}\|P(t) - Id\| &= 1 - \frac{1}{2} \max_i \left(\sum_{j\neq i} |P(t_{ij})| + |P(t_{ii}) - 1| \right) \\ &= 1 - \frac{1}{2} \max_i \left(1 - P(t_{ii}) + 1 - P(t_{ii})\right)) \\ &= 1 - \frac{1}{2}\left(2 - 2g(t)\right) = g(t), \end{aligned}$$

with $g(t) = \min_i P(t)_{ii}$. From $g(t) > \frac{1}{2}$, $\|P(t) - Id\| < 1$ follows and hence the principal logarithm converges.

Theorem 2.4. [20] Let $P(t)$ be a finite-state Markov semi-group. For those t for which

$$\det(P(t)) > \exp(-\pi)$$

holds, a unique generator exists.

Here the relation

$$|\lambda_i - \log[\det(P(t))]| \leq -\log[\det(P(t))] \Rightarrow |Im(\lambda_i)| \leq -\log[\det(P(t))]$$

was exploited. If now $-\log[\det(P(t))] \leq \pi$, we obtain $-\pi < Im(\lambda_i) < \pi$. Thus $\lambda_i = t^{-1}\log[\Lambda_i(t)]$ is unique. Both theorems provide conditions for the principal logarithm to be the unique generator. Another uniqueness criterion is given by the following theorem.

Theorem 2.5 ([45])**.** Let P be a stochastic matrix. If $\det(P) > \frac{1}{2}$, then P has at most one generator. If further $\|P - Id\| < \frac{1}{2}$ holds, the only possible generator is the principal logarithm of P.

According to Israel, Rosenthal and Wei the eigenvalues in Theorem 2.4 have to be distinct, which is not claimed in theorem 2.5.

Remark 2.2. Unfortunately, the existence and uniqueness theorems can not be generalized to the case $\exp(\tau F)$, where F is no generator matrix. This is due to all these statements being based on

$$|\lambda_i - \text{trace}(F)| \leq |\text{trace}(F)|,$$

which does not hold in general for $F \notin \mathcal{G}$.

2.1.5 Imbedding of Perturbed Generator Matrices

We aim at investigating how the error of a perturbed propagator

$$\epsilon = P(\tau) - \hat{P}(\tau)$$

is transferred to the generator matrix. $P(\tau)$ denotes an imbeddable transition matrix generated by

$$P(\tau) = \exp(\tau\ L),$$

where L is a generator matrix with $L \in \mathcal{G}$. Further, let the matrix $\hat{P}(\tau)$ be stochastic but not necessarily imbeddable. The rows of the error matrix must therefore all sum to zero.

Since the logarithm series converges for $\|P(\tau) - Id\| < 1$ to the principal logarithm, and the generator is unique under this condition, we restrict the error analysis to matrices that are sufficiently near to the identity.

Remark 2.3. For each L we can find a transition matrix $P(\tau) = \exp(\tau L)$, such that $\|P(\tau) - Id\| < 1$ is fulfilled. We just have to choose τ small enough.

Furthermore, suppose P and ϵ commutate. We compute the distance between the logarithm of the perturbed and the unperturbed matrix:

$$\begin{aligned}
& \log[P(\tau)+\epsilon] - \log[P(\tau)] \\
= & \sum_{n=1}^{\infty} \frac{(Id - P(\tau) - \epsilon)^n}{n} - \sum_{n=1}^{\infty} \frac{(Id - P(\tau))^n}{n} \\
= & \sum_{n=1}^{\infty} \frac{1}{n} \left(\sum_{k=0}^{n} \binom{n}{k} (Id - P(\tau))^k (-\epsilon)^{n-k} \right) - \frac{(Id - P(\tau))^n}{n} \\
= & \sum_{n=1}^{\infty} \frac{1}{n} \left(\sum_{k=0}^{n-1} \binom{n}{k} (Id - P(\tau))^k (-\epsilon)^{n-k} \right) \\
= & \sum_{n=1}^{\infty} \frac{1}{n} \Big(n(Id - P(\tau))^{n-1}(-\epsilon) + \underbrace{\sum_{k=0}^{n-2} \binom{n}{k} (Id - P(\tau))^k (-\epsilon)^{n-k}}_{o(\epsilon)} \Big) \\
= & -\sum_{n=1}^{\infty} (Id - P(\tau))^{n-1} \epsilon + o(\epsilon).
\end{aligned}$$

For the second equation was exploited that P and ϵ commutate, such that the binomial expansion is possible. The resulting term $-\sum_{n=1}^{\infty}(Id-P(\tau))^{n-1}$ is exactly the Neumann series, which converges to $P^{-1}(\tau)$ for $\|P(\tau)-Id\| < 1$. We obtain the error estimate

$$\begin{aligned}
\left\| \tau(\hat{L} - L) \right\| = & \left\| \log[P(\tau)+\epsilon] - \log[P(\tau)] \right\| \\
= & \left\| P^{-1}(\tau)\epsilon + o(\epsilon) \right\| \leq \left\| P^{-1}(\tau)\epsilon \right\| + o(\|\epsilon\|) \\
\leq & \|\epsilon\| \ \left\| P^{-1}(\tau) \right\|.
\end{aligned} \tag{2.12}$$

Remark 2.4. The commutativity of P and ϵ is ensured if for example ϵ has the same structure as L:

$$\epsilon = cL,$$

where the constant c is a scaling factor. Since P is generated by L, P and L commutate. Thus, P and ϵ also commutate.

We end this section with the following proposition

Proposition 2.1. Let $P(\tau)$ be an imbeddable propagator matrix $P(\tau) = \mathrm{expm}(\tau L)$, such that $\|P(\tau) - Id\| < 1$. Then for each ϵ commutable with P and satisfying

$$\tau^{-1} \|P^{-1}(\tau)\| \ \|\epsilon\| \leq \min_{i,j \in S} |L_{ij}|$$

all perturbed transition matrices in

$$\mathcal{P}_\epsilon = \{ \tilde{P}(\tau), \text{ with } \|P(\tau) - \tilde{P}(\tau)\| \leq \|\epsilon\| \}$$

are imbeddable.

Proof. Unless otherwise specified we use here as everywhere else in this thesis the 2-norm as matrix norm. First note that

$$\|A\| \geq \max_{j,j \in S} |A_{ij}|$$

for any matrix A holds. Now denote the error matrix of the generator by

$$\delta = \hat{L} - L.$$

We get

$$\max_{i,j \in S} |\delta| \leq \|\delta\| \leq \tau^{-1} \left\| P^{-1}(\tau) \right\| \ \|\epsilon\|.$$

Per definition this term is bounded by

$$\tau^{-1} \left\| P^{-1}(\tau) \right\| \ \|\epsilon\| \leq \min_{i,j \in S} |L_{ij}|,$$

finally we obtain

$$\max_{i,j \in S} |\delta| \leq \min_{i,j \in S} |L_{ij}|.$$

But this means nothing else than that the generator properties are preserved. □

2.2 The Ornstein-Uhlenbeck Process

The focus in this section is on another class of Markov processes with continuous time and continuous state space. In particular we will concentrate on diffusion processes, which are described in more detail e.g. in [48, 69]. The following summary is based on [49]. This class of kinetic processes is essentially based on the so-called **Wiener process**.

Definition 2.5 (Wiener process – Brownian Motion)**.** A standard d-dimensional *Wiener process* $W(t)$ is a stochastic process with the following properties:

- $\mathbb{P}(W(0) = 0) = 1$,
- for all s, t with $0 \leq s \leq t$ the increments $W(t) - W(s)$ are independent and normally distributed

$$\mathbb{P}(W(t) - W(s) = x) = \frac{1}{(\sqrt{2\pi(t-s)})^d} \exp\left(-\frac{x^2}{2(t-s)}\right). \qquad (2.13)$$

Due to the independence of the increments the Wiener process is Markovian. Further, for the increments

$$\mathbb{E}\left[W(t) - W(s)\right] = 0 \qquad (2.14)$$

$$\mathrm{Var}\left[W(t) - W(s)\right] = t - s \qquad (2.15)$$

holds. These properties follow immediately from the Gaussian distribution of the increments (2.13). The Wiener process is almost nowhere differentiable. To see this we consider the absolute value of the differential quotient $\mathbb{E}\left[\left|\frac{W(t)-W(t+\Delta t)}{\Delta t}\right|\right]$. The expectation of the absolute value of the increments amounts to

$$\mathbb{E}\left[|W(t)-W(s)|\right] \;=\; \sqrt{\frac{2(t-s)}{\pi}}.$$

This implies that the differential quotient is unbounded

$$\lim_{\Delta t\to 0}\mathbb{E}\left[\left|\frac{W(t)-W(t+\Delta t)}{\Delta t}\right|\right] \;=\; \lim_{\Delta t\to 0}\sqrt{\frac{2}{\pi(\Delta t)}}=\infty.$$

Hence the SDE

$$\dot{X}(t)=f(X(t))+\sigma(X(t))\dot{W}(t),\; X(0)=x_0 \tag{2.16}$$

does not exist formally. Notwithstanding an integral equation exists.

$$X(t)=X_0+\int_0^t f(X(s))\;ds+\int_0^t \sigma(X(s))\;dW(s).$$

The differential of the Wiener process $\dot{W}(t)$ is also referred to as **white noise**. The stochastic integral is defined similarly to the Riemann integral as the limit of the sums

$$\int_0^T g(t)\;dW(t)=\lim_{n\to\infty}\sum_{i=0}^{n-1} g(\hat{t})(W(t_{i+1})-W(t_i)) \tag{2.17}$$

over an infinitesimal fine partition of the interval $[0,t]$ into intervals $[t_i,t_{i+1}]$ with $t_0=0$ and $t_n=T$. However, in contrast to the Riemann integral the value of the stochastic integral depends on the choice of the evaluation point $\hat{t}$ between t_i and t_{i+1}. This is due to the fact that the Wiener process is not of bounded variation. The Itô integral is defined by (2.17) with evaluation point $\hat{t}=t_i$ and the Stratonovich integral with $\hat{t}=(t_{i+1}-t_i)/2$. The sums in (2.17) are effectively computable for finite n since the increments $W(t_{i+1})-W(t_i)$ are normally distributed random variables. Unless otherwise specified, the Itô integral will be used in the following. Before we focus on SDEs in more detail, we keep hold of some preliminary properties of the Itô integral.

Proposition 2.2. [49] Let $f:[\alpha,\beta]\to\mathbb{R}$ be a continuous function and $\xi=\int_\alpha^\beta f(t)\;dW(t)$, then for the expectation value holds

$$\mathbb{E}\left[\xi\right]=0$$

and for the variance

$$\mathrm{Var}\left[\xi\right]=\int_\alpha^\beta f^2(t)\;dt.$$

Proof. The proof is a simple calculation:

$$\begin{aligned}
\mathbb{E}[\xi] &= \mathbb{E}\left[\int_\alpha^\beta f(t)\ dW(t)\right] \\
&= \mathbb{E}\left[\lim_{n\to\infty}\sum_{i=0}^{n} f(t_i)(W(t_{i+1}) - W(t_i))\right] \\
&= \lim_{n\to\infty}\sum_{i=0}^{n} f(t_i)\mathbb{E}\left[(W(t_{i+1}) - W(t_i))\right],
\end{aligned}$$

By (2.14) follows $\mathbb{E}\left[(W(t_{i+1}) - W(t_i))\right] = 0$, so that the whole expression vanishes and we get

$$\mathbb{E}[\xi] = 0.$$

For the variance the independence of the increments is exploited:

$$\begin{aligned}
\mathrm{Var}[\xi] &= \mathbb{E}\left[\left(\int_\alpha^\beta f(t)\ dW(t)\right)^2\right] \\
&= \mathbb{E}\left[\left(\lim_{n\to\infty}\sum_{i=0}^{n} f(t_i)(W(t_{i+1}) - W(t_i))\right)^2\right] \\
&= \lim_{n\to\infty}\sum_{i,j=0}^{n} f(t_i)f(t_j)\mathbb{E}\left[(W(t_{i+1}) - W(t_i))(W(t_{j+1}) - W(t_j))\right].
\end{aligned}$$

With the independence assumption from above and (2.15) we obtain

$$\mathbb{E}\left[(W(t_{i+1}) - W(t_i))(W(t_{j+1}) - W(t_j))\right] = \delta_{ij}(t_{i+1} - t_i)$$

and thus

$$\mathrm{Var}[\xi] = \lim_{n\to\infty}\sum_{i,j=0}^{n} f^2(t_i)(t_{i+1} - t_i) = \int_\alpha^\beta f^2(t)\ dt.$$

□

The SDE (2.16) specifies a **diffusion process**. If f is linear and σ constant in X we have a special kind of diffusion process, the **Ornstein-Uhlenbeck process**

$$\dot{X}(t) = -D(X(t) - \mu) + \sigma\dot{W}(t),\ X(0) = x_0. \tag{2.18}$$

Remark 2.5. The common representation of the Ornstein-Uhlenbeck process is the SDE

$$\dot{X}(t) = DX(t) + \sigma\dot{W}(t),\ X(0) = x_0,$$

but in the scope of this thesis the constant μ will be required.

The following calculations are carried out for the one-dimensional case. The explicit solution of the SDE can be calculated simply by variation of constants:

$$X(t) = \exp(-Dt)X_0 + \mu(Id - \exp(-Dt)) + \int_0^t \exp(-D(t-s))\sigma \; dW(s).$$

By means of proposition 2.2 we can now compute the expectation value $\mathbb{E}[X(t)]$ and the variance $\mathrm{Var}[X(t)]$ of the Ornstein-Uhlenbeck process. The solution $X(t)$ depends on the random variables X_0 and $W(s)$, which are independent.

$$\mathbb{E}[X(t)] = \mathbb{E}[\exp(-Dt)X_0] + \mathbb{E}[\mu(Id - \exp(-Dt))] + \mathbb{E}\left[\int_0^t \exp(-D(t-s))\sigma \; dW(s)\right].$$

The last term on the right hand side vanishes according to proposition 2.2, hence the expectation value reduces to

$$\mathbb{E}[X(t)] = \exp(-Dt)\ \mathbb{E}[X_0] + \mu(Id - \exp(-Dt)).$$

The variance is calculated as follows

$$\mathrm{Var}[X(t)] = \mathbb{E}\left[\left(\exp(-Dt)(X_0 - \mathbb{E}[X_0]) + \int_0^t \exp(-D(t-s))\sigma \; dW(s)\right)^2\right].$$

By the independence of X_0 and $W(s)$ and proposition 2.2 we obtain

$$\begin{aligned} \mathrm{Var}[X(t)] &= \exp(-2Dt)\ \mathrm{Var}[X_0] + \int_0^t \exp(-2D(t-s))\sigma^2 \; ds \\ &= \exp(-2Dt)\ \mathrm{Var}[X_0] + \frac{\sigma^2}{2D}(Id - \exp(-2Dt)). \end{aligned}$$

To obtain the moments for a process in equilibrium we consider the limits:

$$\begin{aligned} \lim_{t\to\infty} \mathbb{E}[X(t)] &= \mu \\ \lim_{t\to\infty} \mathrm{Var}[X(t)] &= \frac{\sigma^2}{2D}. \end{aligned}$$

Corollary 2.2. The Ornstein-Uhlenbeck process is a Gaussian process with mean value

$$\mathbb{E}[X(t)] = \exp(-Dt)\ \mathbb{E}[X_0] + \mu(Id - \exp(-Dt)).$$

and variance

$$\mathrm{Var}[X(t)] = \exp(-2Dt)\ \mathrm{Var}[X_0] + \frac{\sigma^2}{2D}(Id - \exp(-2Dt)).$$

An alternative approach to determine the probability density for a canonical ensemble of SDE-systems (so called diffusion processes) as specified in (2.16) is given by the solution of the Fokker-Planck equation. Let $\rho(t, x)$ denote the probability density function of the process above. It depends on the process itself and on the time t. The Fokker-Planck equation (FPE) specifies how this density evolves in time:

$$\frac{\partial \rho(x,t)}{\partial t} = -\sum_{i=1}^{d} \frac{\partial}{\partial x_i} \left(f_i(x)\rho(x,t) \right) + \frac{1}{2} \sum_{i,j=1}^{d} \frac{\partial^2}{\partial x_i x_j} \left(\sigma_{ij}(x)\rho(x,t) \right). \quad (2.19)$$

The coefficient of the deterministic part of the diffusion process $f(x)$ is called **drift** vector, the coefficient of the stochastic part $\sigma(x)$ **diffusion** tensor. The Fokker-Planck equation is also referred to as Kolmogorov forward equation, which we have already introduced for a discrete state space in (2.6). Under certain conditions it possesses analytic solutions ([65], p.7). If the drift vector is linear and the diffusion tensor is constant as in (2.18), the solutions are Gaussian densities. We will use this technique in Section 4.1 ff.

3 Hidden Markov Models

A hidden Markov model consists of two stochastic processes. A Markov chain $Y_0, ..., Y_M$ and a second process $O_0, ..., O_M$, of which random variables are distributed according to some parametric distribution. In what follows we will use a Gaussian normal distribution. The random variables $(O_m)_{m=0,...,M}$ are independent from each other, but the parameters of their distribution depend on the state of the Markov chain at that time:

$$\mathbb{P}(O_m|Y_m = k) = \mathcal{N}(O_m, \mu_k, \Sigma_k).$$

Remark 3.1. In case of continuous output the output probabilities are actually probability density functions (pdf). We use here the notation of Rabiner [64].

Only the random variables of the second process $(O_m)_{m=0,...,M}$ are observable, the states of the Markov chain $(Y_m)_{m=0,...,M}$ are hidden. A schematic representation of HMM is shown in Figure 3.1. For an introduction to HMMs see [64]. The parameters of the hidden Markov process are specified by a transition matrix P and an initial distribution π. To each state of the Markov chain an output distribution is assigned. This output distribution is specified by the parameters of the particular distribution for each state the Markov chain can assume. In case of a Gaussian output distribution these are the mean value μ_k and the covariance matrix Σ_k for each state $k \in S$, where S is the state space of the Markov chain. An HMM with Gaussian output is thus fully specified by the parameter tuple $\lambda = (\pi, P, (\mu_k)_{k\in S}, (\Sigma_k)_{k\in S})$. The likelihood of a realization from an HMM is given by

$$\begin{aligned} \mathbb{P}(O, Y|\lambda) &= \pi_{Y_0}\, \mathcal{N}(O_0, \mu_{Y_0}, \Sigma_{Y_0}) \\ &\quad \prod_{m=0}^{M-1} P_{Y_m Y_{m+1}} \mathcal{N}(O_{m+1}, \mu_{Y_{m+1}}, \Sigma_{Y_{m+1}}). \end{aligned} \tag{3.1}$$

In the context of HMMs there are essentially three problems:

1. the calculation of the likelihood for a given observation data O and fix model parameters λ
$$\mathbb{P}(O|\lambda),$$
2. the inference of the most likely hidden state path, given the observation data O and model parameters λ
$$\text{argmax}_Y \mathbb{P}(O, Y|\lambda),$$
3. the estimation of the model parameters λ, that maximize the likelihood of the observed data O
$$\text{argmax}_\lambda \mathbb{P}(O|\lambda).$$

The first question is addressed by a dynamic programming technique, the **forward-backward algorithm**, the second problem can be solved by a similar technique – the **Viterbi algorithm** – and the third task is carried out by an **EM algorithm**. The next section concentrates on the latter mentioned algorithm.

Computation of the likelihood for a given observation sequence. At this point the forward-backward algorithm will be elucidated. It is used to solve the first problem and is also employed in the EM algorithm.

The forward variables α denote the likelihood

$$\alpha_i(t_m) = \mathbb{P}(O_0, \ldots, O_m, Y_m = i|\lambda),$$

the backward variables β denote

$$\beta_i(t_m) = \mathbb{P}(O_{m+1}, \ldots, O_M|Y_m = i, \lambda),$$

for $i \in S$ and $m = 1, .., M$. Both auxiliary variables are computable recursively

$$\begin{aligned} \alpha_i(t_m) &= \sum_{Y_{m-1} \in S} \alpha_{Y_{m-1}}(t_{m-1})\ P_{Y_{m-1}i}\ \mathcal{N}(O_m, \mu_i, \Sigma_i) \\ \beta_i(t_m) &= \sum_{Y_{m+1} \in S} P_{iY_{m+1}}\ \mathcal{N}(O_{m+1}, \mu_{Y_{m+1}}, \Sigma_{Y_{m+1}})\ \beta_{Y_{t+1}}(t_{m+1}) \end{aligned} \tag{3.2}$$

with initial values

$$\begin{aligned} \alpha_i(t_0) &= \pi_i \mathcal{N}(O_0, \mu_{i_0}, \Sigma_{i_0}), \\ \beta_i(t_M) &= 1, \qquad \forall i \in S. \end{aligned}$$

Both together provide an occupancy probability of being in state i at time t_m, having observed the time series $O_0, \ldots, O_M$.

$$\begin{aligned} \alpha_i(t_m)\beta_i(t_m) &= \mathbb{P}(O_0, \ldots, O_m, Y_m = i|\lambda)\mathbb{P}(O_{m+1}, \ldots, O_M|Y_m = i, \lambda) \\ &= \mathbb{P}(O_0, \ldots, O_M, Y_m = i|\lambda). \end{aligned} \tag{3.3}$$

This way we obtain the likelihood $\mathbb{P}(O_0, \ldots, O_M|\lambda)$ by means of the forward backward variables:

$$\mathbb{P}(O_0, \ldots, O_M|\lambda) = \sum_{i \in S} \alpha_i(t_m)\beta_i(t_m). \tag{3.4}$$

Remark 3.2. The likelihood (3.4) is independent of the time point t_m, at which the forward-backward variables are evaluated since

$$\sum_{i \in S} \alpha_i(t_m)\beta_i(t_m) = \sum_{i \in S} \alpha_i(t_{m+1})\beta_i(t_{m+1})$$

holds. In particular

$$\mathbb{P}(O_0, \ldots, O_M|\lambda) = \sum_{i \in S} \alpha_i(t_M)$$

holds.

Hence the likelihood $\mathbb{P}(O_0, \ldots, O_M|\lambda)$ is computable in the complexity of the forward-backward recursions, which is $\mathcal{O}(N^2M)$, where N is the number of states contained in S. This is indeed feasible in contrast to the straightforward evaluation of the likelihood (3.1)

$$\mathbb{P}(O_0, \ldots, O_M|\lambda) = \sum_{Y \in S^M} \mathbb{P}(O_0, \ldots, O_M, Y_0, ..., Y_M|\lambda),$$

with complexity $\mathcal{O}(N^M)$. As we will see in the next subsection these auxiliary variables containing occupancy probabilities act as weights in the parameter reestimation as stated in Algorithm 3.2.

Optimal sequence of hidden states. Problem (2) can be solved by applying the Viterbi algorithm [70]. For given λ and O this algorithm computes the most probable hidden path $Y^* = (Y_0^*, \ldots, Y_M^*)$. This path is called the *Viterbi path*. For an efficient computation we define the highest probability along a single path, for the first m observations, ending in the hidden state i at the time t_m,

$$\delta_i(t_m) = \max_{Y_0, Y_1, \ldots Y_{m-1}} P(Y_0, Y_1 \ldots Y_m = i, O_0, O_1 \ldots O_m|\lambda).$$

This quantity is given by induction as

$$\delta_j(t_m) = \max_{i \in S}[\delta_i(t_{m-1})P(i,j)]\ \mathcal{N}(O_m, \mu_j, \Sigma_j). \tag{3.5}$$

In addition, the argument i that maximizes (3.5) is stored in ψ in order to actually retrieve the hidden state sequence. These quantities are calculated for each t and j, and then the Viterbi path will be given by the sequence of the arguments in ψ, obtained from backtracking. For more details see [64]. The dynamic programming technique is the same as in the forward-backward algorithm. The only difference is the substitution of the sum with a maximization. The complexity remains $\mathcal{O}(N^2M)$.

3.1 The EM Algorithm

The expectation maximization (EM) algorithm is designed to determine the maximum likelihood estimator for partially observable data. In this section we will introduce the algorithm generically and afterwards apply it to HMMs. For the abstract introduction we denote the observables with X_1,

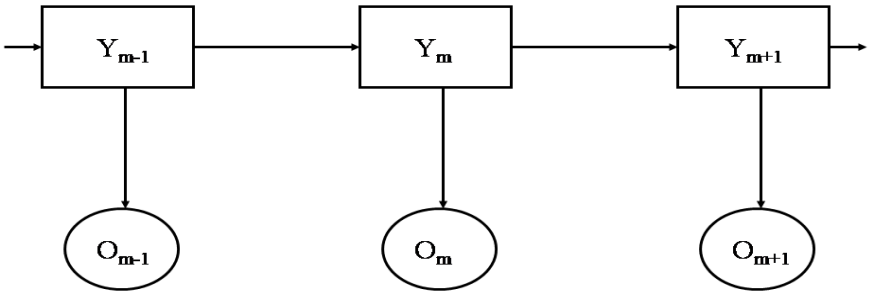

Figure 3.1: Hidden Markov model with output process $O_0, ..., O_{m-1}, O_m, O_{m-1}, ..., O_M$ and hidden process $Y_0, ..., Y_{m-1}, Y_m, Y_{m-1}, ..., Y_M$.

the hidden variables with X_2 and the parameters with Θ. The parameterized likelihood to be maximized is the likelihood of the observed data

$$\mathcal{L}(X_1|\Theta) = \int \mathcal{L}(X_1, X_2|\Theta)\ dX_2, \tag{3.6}$$

specified as marginal distribution of the model data, integrated over the hidden part. This likelihood is too complex to maximize it directly. The whole likelihood is not available since it is partially hidden. Thus the expectation value $\mathbb{E}\left[\mathcal{L}(X_1, X_2|\Theta)\right]$ over the hidden part of the data has to be computed. However, to compute the expectation a probability measure $\mathbb{P}(X_2)$ for the hidden part is required a priori. To formulate this prior probability measure it is necessary to make an initial assumption on the parameters Θ_0. Furthermore, for computational reasons we build the expectation value of the log-likelihood and eventually obtain the functional

$$Q(\Theta, \Theta_0) = \mathbb{E}\big[\log[\mathcal{L}(X_1, X_2|\Theta)]|\Theta_0\big] = \int \log[\mathcal{L}(X_1, X_2|\Theta)]\ \mathcal{L}(X_2|\Theta_0)\ dX_2. \tag{3.7}$$

Finding the maximum of Q is much easier than finding the maximum of $\mathcal{L}(X_1|\Theta)$ directly, because the functional Q has a unique global maximum as shown by Baum [9]. Accordingly we obtain an iterative procedure

$$\hat{\Theta}_{k+1} = \text{argmax}_\Theta Q(\Theta, \Theta_k).$$

Moreover it is shown in the same article that this iteration causes also an increase in the original likelihood $\mathcal{L}(X_1|\Theta)$ and converges finally in a (local) maximum of $\mathcal{L}(X_1|\Theta)$. We will come back to this point later in Chapter 7.1. Overall, the EM algorithm is a two step iteration: In the **expectation step** the expected log-likelihood is computed given the current parameter guess

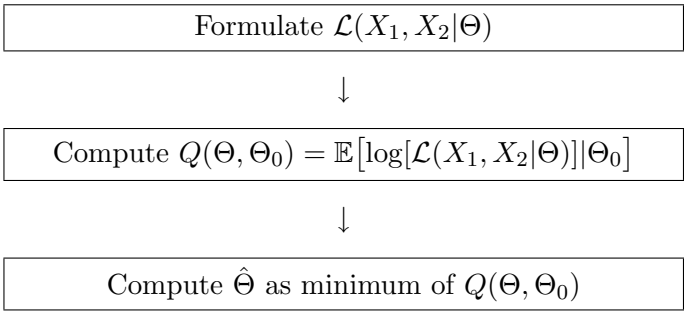

Figure 3.2: Proceeding to determine the reestimation formulas in the EM algorithm.

$Q(\Theta, \Theta_k)$ and in the **maximization step** the parameter reestimation is carried out. The maximum is found by differentiating Q with respect to each parameter component of Θ and find the root.

$$\frac{\partial Q(\Theta, \Theta_k)}{\partial \Theta} = 0 \tag{3.8}$$

Since we know that Q has an unique global maximum (3.8) is sufficient for its identification. The implementation is listed in Algorithm 3.1.

Algorithm 3.1 Generic EM algorithm

Require: Time series X_1 ,initial guess of parameters Θ_0, accuracy ε.
(1) Set $\tilde{\Theta} := \Theta_0$.
(2) Compute $Q(\Theta, \tilde{\Theta})$
(3) Reestimation: determine Θ, such that (3.8) holds.
(4) $\Delta\mathcal{L} = \log[\mathcal{L}(X_1|\Theta)] - \log[\mathcal{L}(X_1|\tilde{\Theta})]$;
if $\Delta\mathcal{L} > \varepsilon$ **then**
Set $\tilde{\Theta} := \Theta$.
Go to step (2)
end if
return Θ

Summarizing, we will give the scheme, which will be used in the Sections 3.1.1, 5.2 - 5.3, 6.2 - 6.3, 7.1, and 7.2, whenever an EM algorithm is derived (cf. Figure 3.2).

- First, specify the **joint likelihood** of the entire (observable and hidden) data given an initial guess for the model parameters Θ:

$$\mathcal{L}(X_1, X_2|\Theta).$$

- Then compute the **expectation value** for the joint likelihood over the hidden data with respect to a prior probability $\mathcal{L}(X_2|\Theta_0)$ (also

$\mathcal{L}(X_1, X_2|\Theta_0)$ or $\mathcal{L}(X_2|X_1, \Theta_0)$ works here). This step is also referred to as the **expectation step**. Θ_0 stands for the initial parameters, which are required to state the prior probability

$$Q(\Theta, \Theta_0) = \mathbb{E}\left[\mathcal{L}(X_1, X_2|\Theta)|\Theta_0\right].$$

- Finally compute the partial derivatives of Q with respect to each component of Θ and find the root. The estimator $\hat{\Theta}$ that solves

$$\frac{\partial Q(\Theta, \Theta_0)}{\partial \Theta} = 0$$

is the unique maximum of Q. This way a **reestimation formula** for Θ is obtained. Parameter reestimation is the actual **maximization step** of the EM algorithm.

3.1.1 The Baum-Welch Algorithm

In this thesis the EM algorithm is applied to two different model structures. The first is an HMM, the corresponding EM algorithm will be described in more detail now. The EM algorithm for the special case of HMMs is also known as the **Baum-Welch algorithm**. The second is a time-continuous Markov chain, the corresponding EM algorithm will be pointed out in Chapter 6.

Likelihood. In the context of HMMs we denote the observables by $O_0, ..., O_M$, the hidden variables by $Y_0, ..., Y_M$ and the parameters by λ. The likelihood that has to be maximized is $\mathcal{L}(O|\lambda)$. It is defined as a marginal distribution of the likelihood for the entire HMM as specified in (3.6)

$$\mathcal{L}(O|\lambda) = \sum_{Y \in S^M} \mathcal{L}(O, Y|\lambda).$$

The entire likelihood is explicitly computable by (3.1) and also its marginal distribution is effectively computable by means of forward-backward variables (3.4).

Expectation-Step. To estimate parameters that maximize $\mathcal{L}(O|\lambda)$, we apply the EM iteration, where in each iteration a functional (3.7) with a unique global maximum as mentioned above is maximized. This functional denotes the expected log-likelihood. To build the expectation value the prior probability $\mathcal{L}(O, Y|\lambda_0)$ instead of $\mathcal{L}(Y|\lambda_0)$ was used. The expected

log-likelihood has the concrete shape

$$\begin{aligned} Q(\lambda,\lambda_0) &= \sum_{Y\in S^M} \Big[\log[\pi_{Y_0}] + \log[\mathcal{N}(O_0,\mu_{Y_0},\Sigma_{Y_0})] + \\ &\qquad \sum_{m=0}^{M-1} \log[P_{Y_m Y_{m+1}}] + \log[\mathcal{N}(O_{m+1},\mu_{Y_{m+1}},\Sigma_{Y_{m+1}})]\Big]\ \mathcal{L}(O,Y|\lambda_0) \\ &= \sum_{i\in S} \left(\log[\pi_i]\alpha_i(t_0)\beta_i(t_0) + \sum_{m=0}^{M} \log[\mathcal{N}(O_m,\mu_i,\Sigma_i)]\alpha_i(t_m)\beta_i(t_m)\right) + \\ &\sum_{i,j\in S} \sum_{m=0}^{M-1} \log[P_{ij}]\ \alpha_i(t_m)\ (P_0)_{ij}\ \mathcal{N}(O_{m+1},(\mu_0)_j,(\Sigma_0)_j)\ \beta_i(t_m). \quad (3.9) \end{aligned}$$

In the last equation was exploited that $\mathcal{L}(O,Y_m=i|\lambda_0)$ is expressible by forward-backward variables (3.3) and alike

$$\mathcal{L}(O,Y_m=i,Y_{m+1}=j|\lambda_0) = \alpha_i(t_m)\ (P_0)_{ij}\ \mathcal{N}(O_{m+1},(\mu_0)_j,(\Sigma_0)_j)\ \beta_i(t_m).$$

This follows immediately from the recursion (3.2).

Maximization-Step. Eventually we obtain the reestimation formulas straightforward as the MLEs:

$$\begin{aligned} \frac{\partial Q(\lambda,\lambda_k)}{\partial \pi_i} = 0 \Leftrightarrow (\hat{\pi}_{k+1})_i &= \frac{\alpha_i(t_0)\beta_i(t_1)}{\sum_i \alpha_i(t_0)\beta_i(t_0)} \\ \frac{\partial Q(\lambda,\lambda_k)}{\partial P_{ij}} = 0 \Leftrightarrow (\hat{P}_{k+1})_{ij} &= \frac{\alpha_i(t_m)\ (P_k)_{ij}\ \mathcal{N}(O_{m+1},(\mu_k)_j,(\Sigma_k)_j)\ \beta_i(t_m)}{\sum_{m=0}^{M-1} \alpha_i(t_m)\beta_i(t_m)} \\ \frac{\partial Q(\lambda,\lambda_k)}{\partial \mu_i} = 0 \Leftrightarrow (\hat{\mu}_{k+1})_i &= \frac{\sum_{m=0}^{M} O_m\ \alpha_i(t_m)\beta_i(t_m)}{\sum_{m=1}^{M} \alpha_i(t_m)\beta_i(t_m)} \\ \frac{\partial Q(\lambda,\lambda_k)}{\partial \Sigma_i} = 0 \Leftrightarrow (\hat{\Sigma}_{k+1})_i &= \frac{\sum_{m=0}^{M} (O_m-(\mu_{k+1})_i)(O_m-(\mu_{k+1})_i)'\ \alpha_i(t_m)\beta_i(t_m)}{\sum_{m=1}^{M} \alpha_i(t_m)\beta_i(t_m)}, \end{aligned} \quad (3.10)$$

where for π and P Lagrange multipliers have been imposed to assure that $\sum_i \pi_i = 1$ and $\sum_j P_{ij} = 1$. The calculations are to be found in [50]. For the convenience of the reader is given the implementation of the EM algorithm in Algorithm 3.2 below. The explicit computation of $Q(\lambda,\lambda_k)$ is needless since only the forward backward variables enter the reestimation formulas and at the time facilitate the computation of $\mathcal{L}(O|\lambda)$. So we restrain the expectation step to the computation of α and β.

Complexity and Convergence. How does the numerical effort of the algorithmic realization scale with the size of the problem, i.e., with the length of the observation sequence M, its dimension d, and the number of hidden states N? The literature on the application of EM, Viterbi, and forward-backward algorithms to the parameterization of HMMs demonstrates that

Algorithm 3.2 Baum-Welch algorithm

Require: Time series O, initial guess of parameters λ_0, accuracy ε.

(1) Set $\tilde{\lambda} := \lambda_0$.
(2) Compute forward-backward variables via (3.2).
(3) Reestimation: determine λ according to (3.10).
(4) Compute $\Delta\mathcal{L} = \log[\mathcal{L}(O|\lambda)] - \log[\mathcal{L}(O|\tilde{\lambda})]$ via (3.4).
if $\Delta\mathcal{L} > \varepsilon$ **then**
  Set $\tilde{\lambda} := \lambda$.
  Go to step (2)
end if
return λ

one step of EM and the entire forward-backward algorithm scale linearly in M and quadratically in N. The scaling with respect to d is not so obvious; it depends mainly on the distribution of the observables or on its maximum likelihood estimators respectively. Whenever the distribution is Gaussian, the scaling will be $\mathcal{O}(d^2M + d^3)$. Carefully putting all terms together one finds an asymptotic estimate of the form [25]

$$\mathcal{O}\Big((d^2M + N^2)M \; + \; d^3\Big) \; \times \text{ number of EM iterations.}$$

Here, the necessary number of iterations of the EM procedure should be determined by a certain accuracy requirement on the error of the underlying optimization problem, i.e., the maximum likelihood problem. There is a variety of results on the convergence of the EM algorithm [73, 74]. One of the basic pitfalls of EM algorithms is the following: given the initial values, they often get trapped by a local maximum; this phenomenon is quite typical for HMMs, or more generally, mixture models. There is copious literature on the subject and how to address it, see for instance [51]. In this thesis convergence is controlled by the following termination criterion: When the increase in likelihood in the last EM iteration does not exceed a certain preset threshold level, the iteration is stopped.

The accuracy of the results will also critically depend on the length of the observation sequence. In the context of the problems considered herein this means that we will have to have "enough" time steps in each of the metastable states.

3.1.2 Metastability Analysis with HMMs

A central issue in many application areas of the following modified HMMs is the identification of metastable states. That is, states in which the time series stays for a long time. If the number of metastable states is known a priori, we can assume each state of the hidden Markov chain as metastable

and the identification is carried out by the Viterbi algorithm (problem (2) stated in the beginning of this chapter).

The problem of identifying the number of dominant metastable states can be formulated as the problem of *aggregating* states O_m from the time series into metastable states (i.e., clustering states that belong to the same metastable state). The identification of an optimal aggregation based on the observation of the dynamical behavior is an important algorithmic problem. There are no general solutions to this problem and the best way to handle it often is a mixture of insight and preliminary analysis. However, this task could in general be handled by using algorithmic concepts from finite mixture models or comparable approaches, cf. [52]. For the problems considered herein, we will see that optimal aggregates can be identified via the dominant eigenmodes of a so-called *transfer* or *transition matrix*, which describes the overall transition probabilities between all states of the system under consideration. The identification is possible by considering the largest eigenvalues of the transition matrix and by exploiting an intriguing property of dominant eigenmodes: they exhibit significant jumps between different metastable aggregates, while varying only slowly within them [21, 67]. This has led to the construction of an aggregation technique called "Perron Cluster Cluster Analysis" (PCCA) [23, 24].

We will use PCCA within the HMM framework as follows: In the setup of an HMM for a given observation sequence, one is confronted with the task of selecting the number N of hidden states *in advance*. Since our goal is to identify metastable states we can proceed as suggested in [29]: Start the EM algorithm with an appropriate number of hidden states, say N, that should be greater than the expected number of metastable states. After termination of the EM algorithm, take the resulting transition matrix P and aggregate the N hidden states into $N_{\text{meta}} \leq N$ metastable states by means of PCCA. The resulting conformation states will then allow for an interpretation of the results in terms of metastable states.

4 Parameter Estimation for Ornstein-Uhlenbeck Processes

Before we consider the parameter estimation of the modified HMM in Chapter 5, we firstly will derive the maximum likelihood estimators for the output process stand-alone assuming full observability. In this chapter we will focus on the process given by the SDE:

$$\dot{X}(t) \quad = \quad -\nabla_X V(X(t)) + \sigma \dot{W}(t).$$

The left term on the right hand side denotes the derivative with respect to X of a potential. If the potential has the shape

$$V(X) = (X-\mu)' D (X-\mu), \tag{4.1}$$

with a positive definite symmetric form D, it is called **harmonic**; its derivative with respect to X is linear and thus we obtain an Ornstein-Uhlenbeck process (2.18). The Ornstein-Uhlenbeck process is Markovian, hence the likelihood $\mathcal{L}(O) = \mathbb{P}(O_0, ..., O_M)$ factorizes as follows:

$$\mathbb{P}(O_0, ..., O_M) = \prod_{i=1}^{M} \mathbb{P}(O_i | O_{i-1}) \mathbb{P}(O_0). \tag{4.2}$$

To determine the maximum likelihood estimator, we finally have to specify the particular factors of the likelihood $\mathbb{P}(O_i = x | O_{i-1} = y) = \rho(x, t_i | y)$.

4.1 Propagation of the Probability Density

Considering a statistical density function $\rho(x, t|x_0)$ of an ensemble of SDE solutions (2.18) for different realizations of the stochastic process W with initial values $X(0) = y$, we get an equivalent representation of the dynamics in terms of the Fokker-Planck operator:

$$\partial_t \rho = \triangle_x V(x) \rho + \nabla_x V(x) \cdot \nabla_x \rho + \frac{1}{2} \nabla_x \cdot B \nabla_x \rho, \tag{4.3}$$

where $B = (\sigma^2 \in \mathbf{R}^1$ denotes the variance of the white noise (for $\mathbf{R}^d$ it is a positive definite self-adjoint matrix $B = \sigma' \sigma$). In the following considerations the matrices B and D are supposed to commutate. In the one-dimensional case they do definitely, the multi-dimensional case is considered in the end of this section. The equation (4.3) is the Fokker-Planck equation from (2.19) with $f(x) = \nabla_x V(x) = D(x - \mu)$. In the case of harmonic potentials the drift vector is linear and the partial differential equation can be solved analytically whenever the initial density function can be represented as a superposition of Gaussian distributions: the solution of the Fokker-Planck equation (4.3) remains to be a sum of Gaussians whenever the initial

probability function $\rho(\cdot, t=0)$ is one, see also ([65], p.7). Therefore, let us apply the variational principle (Dirac-Frenkel-MacLachlan principle [27]) to (4.3) restricted to functions ρ of the form

$$\rho(x,t) = A(t)\exp\left(-(x-y(t))'\Sigma(t)(x-y(t))\right).$$

The particular terms of (4.3) are

$$\begin{aligned}
\partial_t\rho &= \Big(\dot{A}(t) + 2A(t)\dot{y}(t)'\Sigma(t)(x-y(t)) - A(t)(x-y(t))'\dot{\Sigma}(t)(x-y(t))\Big) \\
&\quad \exp\Big(-(x-y(t))'\Sigma(t)(x-y(t))\Big), \\
\triangle_x V(x) &= \mathrm{trace}(D), \\
\nabla_x V(x) &= D(x-\mu) = D(x-y(t)) + D(y(t)-\mu), \\
\nabla_x\rho &= -2\Sigma(t)(x-y(t))A(t)\exp\Big(-(x-y(t))'\Sigma(t)(x-y(t))\Big), \\
\nabla_x \cdot B\nabla_x\rho &= \Big(-2(x-y(t))'B\Sigma^2(t)(x-y(t)) + \mathrm{trace}(B\Sigma(t))\Big) \\
&\quad A(t)\exp\Big(-(x-y(t))'\Sigma(t)(x-y(t))\Big).
\end{aligned}$$

Putting these terms together into equation (4.3), it leads to the solution of the system of ordinary differential equations. We get:

$$\begin{aligned}
\dot{y} &= -D(y-\mu), \\
\dot{\Sigma} &= -2B\,\Sigma^2 + 2D\Sigma, \\
\dot{A} &= (\mathrm{trace}(D - B\Sigma))\,A,
\end{aligned}$$

for the time-dependent parameters $\{y, \Sigma, A\}$. The explicit solution of this system of equations on the time-interval $[t, t+\tau]$ is:

$$\begin{aligned}
y(t+\tau) &= \mu + \exp(-D\tau)\,(y(t)-\mu), \\
\Sigma(t+\tau) &= \left(D^{-1}B - \exp(-2D\tau)\left(D^{-1}B - \Sigma(t)^{-1}\right)\right)^{-1}, \\
A(t+\tau) &= \frac{1}{\sqrt{\pi}}\,\det(\Sigma(t+\tau))^{1/2},
\end{aligned} \tag{4.4}$$

For y the solution is straightforward, for Σ the ordinary differential equation (ODE) was transformed to

$$\dot{\Sigma}\Sigma^{-2} = -2B + 2D\Sigma^{-1},$$

with $\tilde{\Sigma} = \Sigma^{-1}$ we get the ODE

$$\dot{\tilde{\Sigma}} = 2B - 2D\tilde{\Sigma},$$

and the solution follows immediately. The ODE for A has the solution

$$A(t+\tau) = A(t)\exp\left(\mathrm{trace}\left(\int_t^{t+\tau} D - B\Sigma(s)\,ds\right)\right).$$

Another transformation of the ODE system for Σ yields

$$\frac{1}{2}\dot{\Sigma}\Sigma^{-1} \quad = \quad D - B\Sigma.$$

Now, the left side can be easily integrated

$$\frac{1}{2}\mathrm{logm}(\Sigma) \quad = \quad \int D - B\Sigma(s)\ ds.$$

Thus we get for A

$$\begin{aligned} A(t+\tau) &= A(t)\exp\Big(\frac{1}{2}\mathrm{trace}\big[\mathrm{logm}\big(\Sigma(t+\tau)\big) - \mathrm{logm}(\Sigma(t))\big]\Big) \\ &= \frac{A(t)}{\det\Big[\mathrm{expm}\Big(\mathrm{logm}\big(\Sigma(t)\big)\Big)\Big]^{\frac{1}{2}}}\left(\det\Big[\mathrm{expm}\Big(\mathrm{logm}\big(\Sigma(t+\tau)\big)\Big)\Big]\right)^{\frac{1}{2}} \\ &= \frac{1}{\sqrt{\pi}}\det(\Sigma(t+\tau))^{\frac{1}{2}}. \end{aligned}$$

The coefficient has to be $\frac{1}{\sqrt{\pi}}$ to obey the constraint that ρ should be a normal distribution. In case of initial states that are sums of Gaussians, each Gaussian would move independently according to (4.4) and we would get the solution of (4.3) by superposition.

However, in the case considered herein, we are interested in the probability of output O_{j+1} under the condition that the system has been in state O_j at the time t_j. For this, we can now use (4.4) with $y(t_j) = O_j$ and $\Sigma(t_j)^{-1} = 0$. Therefore, the output probability distribution results to be

$$\rho(O_{j+1}|O_j) = A(t_{j+1})\exp\Big(-(O_{j+1} - y(t_{j+1}))'\Sigma(t_{j+1})(O_{j+1} - y(t_{j+1}))\Big),$$

with

$$\begin{aligned} y(t_{j+1}) &= \mu + \exp(-D\tau)\,(O_j - \mu), \\ \Sigma(t_{j+1}) &= \big(D^{-1}B - \exp(-2D\tau)\,D^{-1}B\big)^{-1} \\ &= (Id - \exp(-2D\tau))^{-1}\,DB^{-1}, \\ A(t_{j+1}) &= \frac{1}{\sqrt{\pi}}\,\Sigma(t_{j+1})^{1/2}, \end{aligned} \tag{4.5}$$

with $\tau = t_{j+1} - t_j$.

Remark 4.1. The obtained solutions have been already stated in Corollary 2.2 with $y(t) = \mathbb{E}[O(t)]$ and $\frac{1}{2}\Sigma^{-1}(t) = \mathrm{Var}[O(t)]$. The corollary was stated for the one-dimensional case, where B and D are commutable.

Now, that we have $\rho(x, t|x_0)$, we can construct the maximum likelihood estimator. However, in order to obtain an analytical solution, we have to make another assumption.

4.1.1 Further Simplification

The formula (4.5) for the parameters of the output distribution can be further simplified by the assumption that we do only want to know about the evolution of the system within a short time interval $[t, t+\tau)$. We then can apply an Euler discretization resulting in

$$\begin{aligned} y(t+\tau) = &\quad O_t - D(O_t - \mu)\tau \\ \Sigma(t+\tau) = &\quad \frac{1}{2\tau} B^{-1} \\ A(t+\tau) = &\quad \frac{1}{\sqrt{\pi}} \Sigma^{1/2}(t+\tau), \end{aligned} \tag{4.6}$$

which is not necessary but simplifies the following steps significantly.

Therefore we have the following probability distribution for the observation sequence $O_0, ..., O_M$ (cf. 4.2)

$$\begin{aligned} \mathcal{L}(O_0, ..., O_M) &= A(t_0) \exp\Big(- (O_0 - y(t_0))' \Sigma(t_0)(O_0 - y(t_0))\Big) \\ &\quad \prod_{i=1}^{M} A(t_i) \exp\Big(- (O_i - y(t_i))' \Sigma(t_i)(O_i - y(t_i))\Big). \end{aligned}$$

Due to (4.6) the Gaussian observation likelihood reduces to

$$\begin{aligned} \rho(O_i | O_{i-1}) &= \frac{1}{(2\pi\tau)^{d/2}} \det(B)^{-1/2} \\ &\quad \exp\left(-(O_i - y)' \frac{1}{2\tau} B^{-1}(O_i - y))\right), \end{aligned}$$

with

$$y = (O_{i-1} - D(O_{i-1} - \mu)\tau).$$

The initial density function

$$\rho_0(O_0) = A(t_0) \exp\Big(- (O_0 - y(t_0))' \Sigma(t_0)(O_0 - y(t_0))\Big)$$

has mean $y(t_0) = O_0$ and variance $\Sigma(t_0)^{-1} = 0$. It is a Gaussian distribution, which has density 1 at O_0 and 0 elsewhere. The density function results in a Dirac delta function.

4.2 Optimal Parameters via the Maximum Likelihood Principle

The likelihood now is expressible by the time-independent parameters $\{\mu, D, B\}$. To determine the parameters that maximize the likelihood $\mathcal{L}(O_0, ..., O_M)$

for a given observation sequence, we take for the sake of simplicity the log-likelihood into consideration.

$$\begin{aligned}\log[\mathcal{L}(O_0,...,O_M)] = \quad & \log[\rho_0(O_0)] + \sum_{i=1}^{M} \log\left[\frac{1}{(2\pi\tau)^{d/2}}\ \det(B)^{-1/2}\right] \\ & - (O_i - O_{i-1} + D\tau(O_{i-1}-\mu))'\,(2\tau B)^{-1} \\ & (O_i - O_{i-1} + D\tau(O_{i-1}-\mu))\end{aligned} \tag{4.7}$$

ρ_0 does not depend on $\{\mu, D, B\}$ and is therefore negligible. We derive partial derivatives of the likelihood $\log[\mathcal{L}(O)]$ defined above, which then are used to compute the maximum of $\log[\mathcal{L}(O)]$.

$$\begin{aligned}\frac{\partial \log[\mathcal{L}(O)]}{\partial \mu} = & \sum_{i=1}^{M} D\frac{1}{2}B^{-1}(O_i - O_{i-1} + D\tau(O_{i-1}-\mu)) \\ \frac{\partial \log[\mathcal{L}(O)]}{\partial D} = & \sum_{i=1}^{M}(-O_i + O_{i-1} - D(O_{i-1}-\mu)\tau)(O_{i-1}-\mu)'\frac{1}{2}B^{-1} \\ \frac{\partial \log[\mathcal{L}(O)]}{\partial B^{-1}} = & \frac{1}{2}\Big[\sum_{i=1}^{M} -B^{-1} + (-O_i + O_{i-1} - D\tau(O_{i-1}-\mu)) \\ & (-O_i + O_{i-1} - D\tau(O_{i-1}-\mu))'\frac{1}{\tau}\Big]\end{aligned}$$

The maximum is obtained by finding the root of the above system of derivatives as follows:

$$\frac{\partial \log[\mathcal{L}(O)]}{\partial \mu} = 0 \quad \Leftrightarrow \quad \mu = \frac{\sum_{i=1}^{M}(O_i - O_{i-1} + DO_{i-1}\tau)}{D\tau\sum_{i=1}^{M}}, \tag{4.8}$$

$$\frac{\partial \log[\mathcal{L}(O)]}{\partial D} = 0 \quad \Leftrightarrow \quad D = \frac{\sum_{i=1}^{M}(O_i - O_{i-1})(O_{i-1}-\mu)}{-\tau\sum_{i=1}^{M}(O_{i-1}-\mu)^2}, \tag{4.9}$$

$$\frac{\partial \log[\mathcal{L}(O)]}{\partial B^{-1}} = 0 \quad \Leftrightarrow \quad B = \frac{1}{\tau M}\sum_{i=1}^{M}(-O_i + O_{i-1} - D(O_{i-1}-\mu)\tau)^2.$$

The parameters μ and D are independent of B. Hence it suffices to solve the smaller system of equation (4.8) and (4.9):

$$\begin{aligned}\Rightarrow \quad & {\scriptstyle (4.8),(4.9)} \quad \frac{\sum_{i=1}^{M}(O_i - O_{i-1})(O_{i-1}-\mu)}{-\tau\sum_{i=1}^{M}(O_{i-1}-\mu)^2} = \frac{\sum_{i=1}^{M}(O_i - O_{i-1})}{\tau\sum_{i=1}^{M}(O_{i-1}-\mu)} \\ \Leftrightarrow \quad & \mu = \frac{X_1X_2 - X_3X_4}{X_1X_4 - X_3X_5}\end{aligned}$$

with

$$X_1 = \sum_{i=1}^{M}(O_i - O_{i-1}), \qquad X_2 = \sum_{i=1}^{M}(O_{i-1}^2),$$

$$X_3 = \sum_{i=1}^{M}(O_i - O_{i-1})O_{i-1}, \quad X_4 = \sum_{i=1}^{M} O_{i-1},$$

$$X_5 = M.$$

Summarizing we obtain the estimators:

$$\begin{aligned}
\hat{\mu} &= \frac{X_1X_2 - X_3X_4}{X_1X_4 - X_3X_5}, \\
\hat{D} &= \frac{\sum_{i=1}^{M}(O_i - O_{i-1})(O_{i-1} - \mu)}{-\tau \sum_{i=1}^{M}(O_{i-1} - \mu)^2} \\
\hat{B} &= \frac{1}{\tau M}\sum_{i=1}^{M}(-O_i + O_{i-1} - D(O_{i-1} - \mu)\tau)^2
\end{aligned}$$

Remark 4.2. The further simplification (4.6) by means of Euler discretization with constant time step is *not* necessary. It leads to the previous explicit formula for the maximizing parameters. When omitting it, we would have to solve some low-dimensional algebraic equations. This is possible without significant numerical effort but includes an additional Newton iteration. Details of this *discretization-free approach* to HMMSDE and its derivation are published elsewhere [40]. A more general approach to a discretization-free multi-dimensional parameter estimation was derived in [41]. We will explicate it in the following section.

4.3 Multi-Dimensional Parameter Estimation without Euler Discretization

We end this chapter with a generalization of the parameter estimation for Ornstein-Uhlenbeck processes. This approach actually was designed in [41] for systems governed by Langevin dynamics

$$\begin{aligned}
\dot{q}(t) &= M^{-1}p(t) \\
\dot{p}(t) &= -\nabla V(q(t)) - \gamma M^{-1}p(t) + \sigma\dot{W}(t).
\end{aligned}$$

The variables q and p denote the positions and momenta of the system, M stands for the mass matrix, V for the potential and γ for the friction matrix. For high friction the system is transferred to the overdamped Langevin (or Smoluchowski) dynamics. With an harmonic potential V as given in (4.1) we obtain an Ornstein-Uhlenbeck process

$$\dot{X}(t) = -D(X(t) - \mu) + \sigma\dot{W}(t).$$

The variable X is defined on the position space only. In contrast to the previous sections, we define here a multi-dimensional process. As pointed out in section 2.2 the solution of the SDE above is

$$X(t+\tau) \quad = \quad \mu + \mathrm{expm}(-D\tau)(X(t)-\mu) + \int_0^\tau \mathrm{expm}(-D(\tau-s))\sigma\dot{W}(s).$$

In Corollary 2.2 was stated that X is normal distributed with mean

$$y_m(\tau) \quad = \quad \mu + \mathrm{expm}(-D\tau)(X(t_m)-\mu), \tag{4.10}$$

and variance

$$R(\tau) \quad = \quad \int_0^\tau \mathrm{expm}(-Ds)\ B\ \mathrm{expm}(-Ds)\dot{W}(s).$$

B stands for the squared noise intensity $B = \sigma\sigma'$. In Corollary 2.2 the integral (4.11) was solved under the assumption that D and σ are commutable. Without this restriction we obtain by partial integration the linear matrix equation

$$-\left(R(\tau)\ D + D\ R(\tau)\right) = \mathrm{expm}(-D\tau)\ B\ \mathrm{expm}(-D\tau) + B.$$

Suppose the time lag between the particular observation points t_m is constant (i.e. $t_{m+1} - t_m = \tau$). We can specify the likelihood of the process in terms of the parameter triple $(\mu, \mathrm{expm}(-D\tau), R(\tau))$:

$$\log \mathcal{L}(O_0, ..., O_M) = \log[A(\tau)] + \Big(-\frac{1}{2}(O_0 - y_0(\tau))' R(\tau)^{-1}(O_0 - y_0(\tau))\Big)$$
$$+ \sum_{i=1}^{M} \log[A(\tau)] - \frac{1}{2}\left(O_m - y_m(\tau)\right)' R(\tau)^{-1}\left(O_m - y_m(\tau)\right), \tag{4.11}$$

with $y_m(\tau)$ from (4.10) and a normalization constant

$$A(\tau) = \frac{1}{\sqrt{(2\pi)^d \det(R(\tau))}}.$$

The dimension of the observables is here denoted by d. Differentiating (4.11) with respect to $(\mu, \mathrm{expm}(-D\tau), R(\tau))$ and determining the roots produces

the linear matrix equation system:

$$\begin{aligned}
\mu &= \frac{1}{M-1}(Id - \mathrm{expm}(-D\tau))^{-1} \sum_{m=1}^{M-1} (O_{m+1} - \mathrm{expm}(-D\tau)O_m) \\
\mathrm{expm}(-D\tau) &= C_1(\mu)C_2^{-1}(\mu) \\
C_1(\mu) &= \Big(\sum_{m=1}^{M-1}(O_{m+1}-\mu)\Big)\Big(\sum_{m=1}^{M-1}(O_m-\mu)\Big)' \\
C_2(\mu) &= \Big(\sum_{m=1}^{M-1}(O_m-\mu)\Big)\Big(\sum_{m=1}^{M-1}(O_m-\mu)\Big)' \\
R(\tau) &= \frac{1}{M-1}\sum_{m=1}^{M-1}\Big(O_{m+1}-\mu-\mathrm{expm}(-D\tau)(O_m-\mu)\Big) \\
&\quad \Big(O_{m+1}-\mu-\mathrm{expm}(-D\tau)(O_m-\mu)\Big)'.
\end{aligned}$$

The solution of the system above determines the optimal parameters $(\hat{\mu}, \mathrm{expm}(-\hat{D}\tau), \hat{B})$.

$$\begin{aligned}
\hat{\mu} &= \bar{O} - (Id - \mathrm{Cor}(O))^{-1}\delta \\
\mathrm{expm}(-\hat{D}\tau) &= \mathrm{Cor}(O) \\
\hat{B} &= (\mathrm{Cov}(O) + E)D + D(\mathrm{Cov}(O) + E),
\end{aligned}$$

$$\text{with } \delta = \frac{1}{M-1}(O_M - O_1).$$

The moving average, the positive definite covariance matrix and the normalized autocorrelation are defined as

$$\begin{aligned}
\bar{O} &= \sum_{m=1}^{M-1} O_m \\
\mathrm{Cov}(O) &= \sum_{m=1}^{M-1}(O_m - \bar{O})(O_m - \bar{O})' \\
\mathrm{Cor}(O) &= \sum_{m=1}^{M-1}(O_{m+1} - \bar{O})(O_m - \bar{O})'\mathrm{Cov}(O)^{-1}.
\end{aligned}$$

The symmetric matrix E is obtained by solving the Sylvester equation

$$\mathrm{Cor}(O)E\mathrm{Cor}(O) - E = \delta\delta' + \frac{1}{M-1}\left((O_M - \bar{O})(O_M - \bar{O})' - (O_1 - \bar{O})(O_1 - \bar{O})'\right)$$

Whenever the Eigenvalues of $\hat{D}$ are on the positive complex half-plane $\mathbb{C}^+$, the Sylvester equation has an unique solution. This condition is fulfilled as D is symmetric and positive definite the Eigenvalues are even positive and real.

In this approach it is no longer necessary to use an Euler discretization to compute the explicit estimators. It applies also to multi-dimensional processes. What makes this improvement possible is the parameter estimation of the operator $\mathrm{expm}(-D\tau)$ rather than a direct estimation of D. The crucial drawback of this proceeding is that it implies the requirement of a constant time lag τ.

Furthermore, D is in general not unique since the matrix logarithm of $\mathrm{expm}(-D\tau)$ is not. In case of Ornstein-Uhlenbeck dynamics as stated above this problem does not occur since D is symmetric and has therefore solely real eigenvalues. Thus the matrix logarithm is unique but in the more general context of Langevin dynamics this constraint does not apply.

5 HMMSDE

Diffusion processes are kinetic models that can be combined with an HMM framework (c.f. [38]). In the foregoing chapter the parameter estimators for an observable diffusion process have been derived. We will see now that the embedding of this process into an HMM is possible in a natural way and provides a powerful tool for the time series analysis. In contrast to the standard HMM, here the output process consists instead of independent identically distributed random variables of a SDE governed dynamic process, such as an Ornstein-Uhlenbeck process.

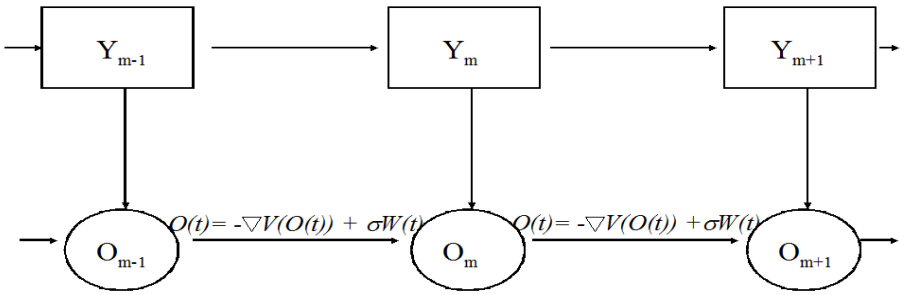

Figure 5.1: Hidden Markov model with Ornstein-Uhlenbeck output process. The hidden states are $Y_0, ..., Y_{m-1}, Y_m, Y_{m+1}, ..., Y_M$, the observable data points are $O_0 ..., O_{m-1}, O_m, O_{m+1}, ..., O_M$. The continuous process between the observables is given by the SDE (5.1).

5.1 Model Design

Now we explain in more detail how the HMM concept has been combined with a special class of kinetic models, namely diffusion processes (2.2). The effective dynamics here are approximated by stochastic differential equations (SDEs) of the following type for the state $X \in \mathbb{R}^n$ of the system:

$$\begin{aligned} \dot{X}(t) &= -\nabla_X V^{(Y(t))}(X(t)) + \sigma^{(Y(t))}\ dW(t) \\ Y(t) &= \text{Markov jump process with states } 1, \ldots, N. \end{aligned} \tag{5.1}$$

The standard Brownian motion is denoted by $W(t)$, ∇_X stands for the gradient with respect to X, $\sigma = (\sigma^{(1)}, \ldots, \sigma^{(N)})$ contains noise intensities, and $\mathcal{V} = (V^{(1)}, \ldots, V^{(N)})$ interaction potentials. The jump process $Y(t)$ is intended to mimic the hopping of the effective dynamics from one metastable set to another metastable set such that its hopping rates have to be related

to the transition rates between the sets. It thus can be represented by an $N \times N$ rate matrix L. The SDEs (5.1) then have to approximate the (more rapidly mixing) dynamics within the metastable states, and thus must have correlation times that are significantly smaller than the typical waiting times between hops of the jump process. Altogether, the model is completely characterized by the tuple $(\pi, L, \mathcal{V}, \sigma)$, where π denotes the initial distribution of the hidden Markov process Y. In the following we will assume that $\mathcal{V}$ only contains potentials $V^{(Y)}$ from a certain family of potentials that is given by a (not too large) tuple of parameters $\theta^{(Y)}$ (e.g., polynomial potentials) such that $\mathcal{V}$ is completely determined by the parameters $\Theta = (\theta^{(1)}, \ldots, \theta^{(N)})$.

Consequently, we have to find a procedure that can determine the optimal model $\lambda = (\pi, L, \Theta, \sigma)$ for the complex system under consideration. The goal of the algorithmic approach to be presented herein is to identify the optimal model (π, L, Θ, σ) from time series resulting from long-term simulation of the complex system under consideration. Thereby, the information about which and how many metastable sets being present in the time series is understood as being *hidden* within the data. Then, metastability is identified in the following way: we try to assign to any state from the given time series the hidden metastable state to which it belongs. The metastable sets then are represented by aggregates containing those states that are assigned to the same hidden state. We will present a procedure that solves the assignment problem *and* the estimation problem for the parameters (π, L, Θ, σ) simultaneously and iteratively via the well-known EM algorithm. This procedure will result from some HMM approach as introduced in Section 3.1.1.

The approach proposed herein (called HMMSDE in the following) can be thought of as an extension of the HMM approach in the sense that the "output" is assumed to result from stochastic differential equations. Equation (5.1) describes the process in continuous time; the HMM, however, already reflects the situation that the observation sequence is given in discrete time.

5.2 Concept

Our goal is to identify optimal parameters for our model (5.1) for given observation data $(O_m)_{m=0,\ldots,M}$. That is, the states O_m, the system is in at times t_m, are known already, and we have to define the functional with respect to which we then will have to determine the optimal parameters λ. This will be done by means of the *maximum likelihood principle*, i.e., the functional will be given by a likelihood function $\mathcal{L}$ that will be constructed in the following way: For given parameters λ, the likelihood $\mathcal{L}(O, Y|\lambda)$ has to be the probability of output $X(t_m) = O_m$, $m = 0, \ldots, M$, and the associated sequence of metastable states (Y_m) (the state sequence of the Markov jump process at times t_m, $m = 0, \ldots, M$). The model structure is illustrated in Figure 5.1. In contrast to the standard HMM the HMMSDE introduced an

additional dependency between the observables. Thus, in order to construct $\mathcal{L}$ appropriately, we have to know the probability of output of state $X(t_m)$ under the condition of being in metastable state Y_m for given parameters λ. We will see that we can determine this probability by considering the propagation of probability densities by the SDE associated with metastable state Y_m analogously to Chapter 4.

For this purpose let us assume again that the potentials V^i, $i \in S$ are of harmonic form:

$$V^i(X) \quad = \quad \frac{1}{2} D^i (X - \mu^i)^2 + V_0^i.$$

This assumption simplifies the derivation of the parameterization algorithms significantly. Furthermore, and also for the sake of simplicity, we will present the derivation for a one-dimensional state space. As we will point out in Section 5.4, both assumptions are not necessary.

5.2.1 Likelihood Function

Whenever we assume the potential to be harmonic the model is characterized by the parameter tuple $\lambda = (\pi, L, y, \Sigma, A)$, where π denotes the initial distribution of the Markov chain, P its transition matrix, and y, Σ, A the parameters of the output distributions according to (4.5). Suppose that the observed data $(O_m)_{m=0,\ldots,M}$ is given with constant time stepping τ, i.e., $t_m = t_{m-1} + \tau$ for all $m = 0, \ldots, M$. Setting $t_0 = 0$ we have $t_m = m\tau$ and especially $T = t_M = M\tau$. In addition to the observation sequence $O = O_0, ..., O_M$ we also have the sequence of hidden metastable states $Y = Y_0, ..., Y_M$ which herein are given by the N possible states of the Markov jump process, i.e., we have $Y_m \in S$.

The restriction on equidistant time steps τ will become dispensable in the next chapter. For the time being we assume a constant time step τ and confine ourselves to estimating the transition matrix P instead of the generator matrix L.

Let L be the generator matrix of jumps between the hidden states. Then the transition probability between hidden states within two consecutive steps of the observations, i.e., the transition probability from hidden state i to hidden state j after time τ under the condition to be in i at time $t = 0$, is given by the ij-th entry of the transition matrix

$$P = \mathrm{expm}(\tau L).$$

Therefore for given model $\lambda = (\pi, P, y, \Sigma, A)$ we have the following joint probability distribution for the observation and hidden state sequences:

$$\mathbb{P}(O, Y|\lambda) \quad = \quad \pi(Y_0)\ \rho_0(O_0|Y_0) \prod_{m=1}^{M} P(Y_{m-1}, Y_m)\ \rho(O_m|Y_m, O_{m-1}),$$

wherein the probability distributions ρ have the form

$$\rho(O_m|Y_m, O_{m-1}) = A^{(Y_m)}(t_m) \exp\Big(-\Big(O_m - y^{(Y_m)}(t_m)\Big)' \Sigma^{(Y_m)}(t_m) \Big(O_m - y^{(Y_m)}(t_m)\Big)\Big),$$

where the superindex refers to the hidden state Y_m of the system at time t_m, and y, Σ and A have to be computed from (4.5). As y, Σ and A actually depend on the time-independent parameters μ, D and B, we can also formulate the model parameters as

$$\lambda = (\pi, P, \mu, D, B). \tag{5.2}$$

The joint Likelihood function for the model given the complete data reads

$$\mathcal{L}(O, Y|\lambda) = \mathbb{P}(O, Y|\lambda).$$

5.3 Partial Observability

Since only the diffusion process O is observable but the Markov chain Y is hidden, the parameter estimation will be carried out by the EM algorithm as introduced in Chapter 3.1.

5.3.1 Expectation Step

We aim to estimate the parameters that maximize the expectation Q of the log-likelihood $\log[\mathcal{L}(O, Y|\lambda)]$ of the complete data with respect to the hidden sequence Y. According to [10] (Chap. 4.2) the expectation value Q can be rewritten as

$$Q(\lambda, \lambda_0) = \sum_{Y \in S^{M+1}} \mathbb{P}(O, Y|\lambda_0) \log[\mathbb{P}(O, Y|\lambda)],$$

where S denotes the state space of the hidden states. As described in Chapter 3.1 this form will allow us to find very efficient maximizers. To simplify notation we will use $\lambda = (\pi, P, \mu, D, B)$ for the a posteriori parameters and λ_0=$(\pi^0, P^0, \mu^0, D^0, B^0)$ for the a priori parameters. The expected likelihood has a similar shape as (3.9)

$$\begin{aligned} Q(\lambda, \lambda_0) &= \sum_{i \in S} \Big(\big[\log[\pi_i] + \log[\rho_0(O_0|i)]\big] \alpha_i(t_0)\beta_i(t_0) \\ &+ \sum_{m=1}^{M} \log[\rho(O_m, \mu_i, D_i, B_i)] \alpha_i(t_m)\beta_i(t_m) \Big) \\ &+ \sum_{i,j \in S} \sum_{m=0}^{M-1} \log[P_{ij}]\ \alpha_i(t_m)\ P^0_{ij}\ \rho(O_{m+1}, \mu^0_j, D^0_j, B^0_j)\ \beta_i(t_m). \end{aligned}$$

The modification only affects the output probability in ρ and ρ_0. The parameter estimation of ρ_0 is omitted since only one data point O_0 is concerned. An estimator for ρ_0 would be a Dirac-delta distribution.

$$\delta(x) = \begin{cases} 1 & x = O_0 \\ 0 & \text{else} \end{cases}$$

5.3.2 Maximization Step

To maximize the functional Q we have to determine the partial derivatives with respect to the particular parameters and find the roots. We obtain the maximum likelihood estimators for the hidden process π and P from the standard Baum-Welch formulas (3.10), the parameters of the output process are determined by the partial derivatives of Q with respect to μ_i, D_i and B_i. Since only the term $\sum_{m=1}^{M} \log[\rho(O_m, \mu_i, D_i, B_i)]\alpha_i(t_m)\beta_i(t_m)$ depends on these parameters, we can neglect the remaining terms. The term $\sum_{m=0}^{M} \log[\rho(O_m, \mu_i, D_i, B_i)]$ is exactly the log-likelihood from (4.7). Under consideration of the forward backward variables we obtain the estimators:

$$\begin{aligned} \hat{\mu}_i &= \frac{X_1 X_2 - X_3 X_4}{X_1 X_4 - X_3 X_5}, \\ \hat{D}_i &= \frac{\sum_{m=1}^{M} (O_m - O_{m-1})(O_{m-1} - \mu)\ \alpha_i(t_m)\beta_i(t_m)}{-\tau \sum_{m=1}^{M} (O_{m-1} - \mu)^2\ \alpha_i(t_m)\beta_i(t_m)} \\ \hat{B}_i &= \frac{\sum_{m=1}^{M} (-O_m + O_{m-1} - D(O_{m-1} - \mu)\tau)^2\ \alpha_i(t_m)\beta_i(t_m)}{\tau \sum_{m=1}^{M} \alpha_i(t_m)\beta_i(t_m)} \end{aligned} \tag{5.3}$$

with

$$X_1 = \sum_{m=1}^{M} (O_m - O_{m-1})\ \alpha_i(t_m)\beta_i(t_m), \qquad X_2 = \sum_{m=1}^{M} (O_{m-1}^2)\ \alpha_i(t_m)\beta_i(t_m),$$

$$X_3 = \sum_{m=1}^{M} (O_m - O_{m-1})O_{m-1}\ \alpha_i(t_m)\beta_i(t_m), \quad X_4 = \sum_{m=1}^{M} O_{m-1}\ \alpha_i(t_m)\beta_i(t_m),$$

$$X_5 = \sum_{m=1}^{M} \alpha_i(t_m)\beta_i(t_m).$$

The resulting EM algorithm is a modification of the Baum-Welch algorithm for HMMs with Gaussian output. The only difference lies in the maximization step due to the reestimation formulas (5.3). The EM-iteration is summarized in Algorithm 5.1 below.

5.4 Enhancements and Application

The HMMSDE as defined here is restricted to one-dimensional time series. This approach has been enhanced in a recent publication [41], where the HMMSDE turned out to be a special case of an HMM with output governed by a second order Langevin dynamics. The process given by (2.16) is also

Algorithm 5.1 EM algorithm for HMMSDE

Require: Time series O, initial guess of parameters λ_0, accuracy ε.

(1) Set $\tilde{\lambda} := \lambda_0$.

(2) Compute forward-backward variables via (3.2).

(3) Reestimation: determine λ according to (5.3).

(4) Compute $\Delta\mathcal{L} = \log[\mathcal{L}(O|\lambda)] - \log[\mathcal{L}(O|\tilde{\lambda})]$ via (3.4).

if $\Delta\mathcal{L} > \varepsilon$ **then**

Set $\tilde{\lambda} := \lambda$.

Go to step (2)

end if

return λ

referred to as high-friction Langevin process. For this more general model have been derived multi-dimensional parameter estimates and furthermore in this approach no Euler discretization as in Section 4.1.1 was required any more.

In case of full observability the optimal estimators have been derived in section 4.3. For the partially observable case the HMMSDE results as a special case of the HMM-Langevin. The optimal estimators are specified in Theorem 3.1 of [41].

The HMMSDE as presented here allows for varying time steps if the hidden process of the HMM is modified as described in Section 7.1. For multivariate time series we can use one-dimensional projections and combine them as proposed in [25]. Assume that we already applied HMMSDE to several one-dimensional observation time series of the system under consideration, but to each one independently. Suppose that the different time series simply are resulting from different projections of the full time series in state space. In this situation one may be interested in combining the hidden states from each of the single projections into "higher dimensional" metastable states of the system. This can be done by analyzing the Viterbi paths derived from the single one-dimensional observation time series: Suppose we are concerned with J one-dimensional time series and therefore J Viterbi paths. The J Viterbi paths can be understood as a J-dimensional discrete time series. Every state of this time series lies in the discrete state space consisting of all possible combinations of the metastable states of the single one-dimensional time series. We obviously can take this time series, compute its transfer matrix by counting transitions between its discrete states, determine the dominant eigenmodes of this transfer matrix, and again apply PCCA to identify metastable decompositions of the discrete state space. The sets in such a metastable decomposition have to be interpreted as aggregates of the metastable states from the low-dimensional time series where the aggregation is done based on additional insight coming from the combination of all of the low-dimensional information. This

concept leads to the algorithm 5.2.

Algorithm 5.2 Metastability analysis of multi-dimensional time series with HMMSDE

Require: Multi-dimensional time series $O_0, ..., O_M$.

(1) Determine model parameters and Viterbi paths for each one-dimensional observation time series.

(2) Combine the Viterbi paths and compute the transfer matrix in the discrete state space of combined metastable states.

(3) Determine metastable decompositions via PCCA.

return Metastable sets.

5.5 Illustrative Example: Three-Hole Potential

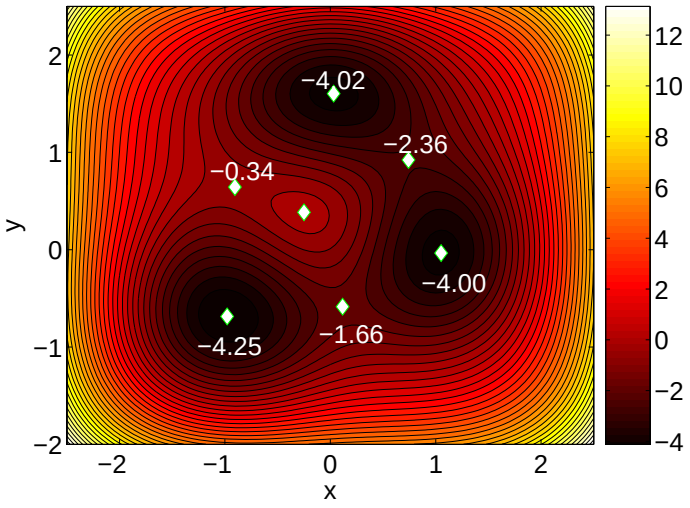

Figure 5.2: Potential V used for illustrative example. We observe three wells in the potential landscape (see colorbar). The tags indicate the minima and saddle points of the potential; the numbers give the value of the potential at these points. We observe that the leftmost minimum is the deepest well separated by the most pronounced energy barrier from the two other ones.

In the following example we will test how HMMSDE performs in the analysis of time series exhibiting similar features to molecular dynamics time series. However, the potentials describing molecular force field are in general more complex. Here we use for illustrative means the three-hole potential V illustrated in Figure 5.2 (thus setting $d = 2$). Figure 5.3 shows typical realizations of the Smoluchowski process associated with this potential (setting $\sigma = 0.131$). We observe that the vicinity of the wells in the potential energy landscape can approximately be identified with the metastable sets

of the process; it is well-known from large deviation theory that in fact, for small enough noise intensity, the vicinity of the wells of the potential energy landscape is formed by the metastable sets of Smoluchowski processes (at least such wells that are separated from each other by significant energy barriers) [42].

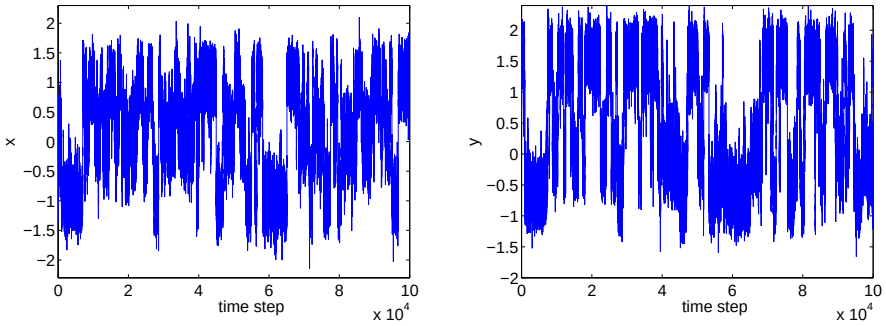

Figure 5.3: Typical realization of the Smoluchowski dynamics (both components (left/right) of the state versus time) in the potential energy landscape V shown in Figure 5.2 for $\sigma = 0.131$.

Next, we discretized the transfer operator of the process (fine grid with 100×100 discretization boxes in discretization domain $[-3, 3] \times [-3, 3]$) for different values of τ which results in the dominant eigenvalues listed in Table 5.1.

$\sigma(P(\tau))$	λ_1	λ_2	λ_3	λ_4
$\tau = 0.01$	1.000	0.999	0.997	0.959
$\tau = 0.10$	1.000	0.994	0.975	0.656
$\tau = 1.00$	1.000	0.937	0.776	0.015

Table 5.1: Leading four eigenvalues of transfer operator $P(\tau)$ for different values of τ for Smoluchowski motion with potential and parameters as described in the text.

While the eigenvector of the largest eigenvalue is constant, the corresponding second and third eigenvectors of $P(\tau)$ are shown in Figures 5.4 (they are identical for all values of τ because of the semi-group property).

Having computed the dominant eigenvectors we can determine the optimal metastable decomposition by means of PCCA as introduced above. The results on the spectrum (see $\tau = 0.1$ for example) exhibit a hierarchy of metastability that is in perfect agreement with the general insight on metastability of Smoluchowski motion: We can apply PCCA to the first *two* eigenvectors of the transfer operator; this results in the metastable decomposition that distinguishes between the vicinity of the deepest well and the remaining state space (see Figure 5.5, left). When applying PCCA to the first *three* eigenvectors, however, the resulting metastable decomposition identifies the vicinities of all three wells as the metastable regions of

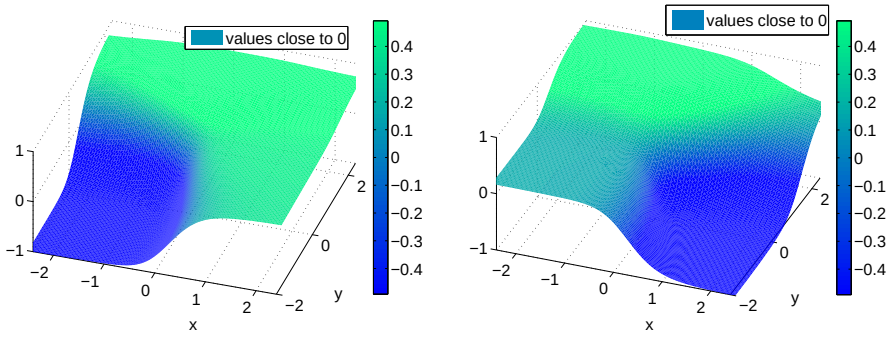

Figure 5.4: Second and third eigenvectors of the transfer operator for Smoluchowski process in potential of Figure 5.2 (details see text).

the system (cf. Figure 5.5, right). This outcome is desirable and typical: metastable decomposition via spectral properties of the transfer operator are hierarchical in the sense that the process of including more and more leading eigenvalues uncovers finer and finer details of metastability within the system, see [66, 42].

So, what happens if we take the first four eigenvectors? This we can immediately understand by comparing the values of the functional meta

$$\mathrm{meta}(\mathbf{D}_m) = \sum_{j=1}^{m} \mathbb{P}(\tau, A_j, A_j)/m.$$

for the optimal metastable decompositions $\mathbf{D}_m$ into $m = 1, 2, 3, 4$ sets $A_1, ..., A_m$ and $\tau = 1$ as given in Table 5.2: Between $m = 3$ and $m = 4$ there is a significant drop in metastability indicating that it makes no real sense to speak of four metastable sets for the system under consideration.

m	1	2	3	4
$\mathrm{meta}(\mathbf{D}_m)$	1.000	0.967	0.899	0.613
$\frac{1}{k}\sum_{k=1}^{m} \lambda_k$	1.000	0.969	0.904	0.682

Table 5.2: Metastabilities of the optimal metastable decomposition $\mathbf{D}_m$ into $m = 1, 2, 3, 4$ sets (as computed by PCCA from the dominant eigenvectors) and its theoretical upper bound [39].

We now assume a time series $(X(t))_{t=t_0,...,t_M}$ with $t_k - t_{k-1} = \tau = 0.01$ and $M = 10^5$ being given in the test system introduced above. For given $t = t_0, ..., t_M$ let $O_m = X(t_m) \in \mathbb{R}^2$ be the full state of the system.

For this choice of τ the transfer operator $P(\tau) = \exp(\tau L)$ of the Smoluchowski motion considered above has the following dominant eigenvalues

$$\sigma(P^t) = \{1.000, 0.999, 0.997, 0.959, \ldots\}.$$

Let us consider the two observation time series $(O_m^{(j)})_{m=0,...,M}$, $j = 1, 2$, with $O_m^{(j)} = X_j(t_m)$ (the first and second components of the state of system).

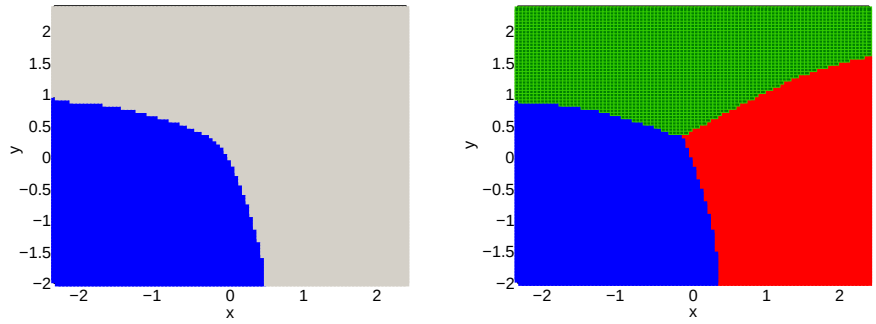

Figure 5.5: Optimal metastable decomposition resulting from PCCA based on the first two (left) and first three (right) eigenvectors of the transfer operator.

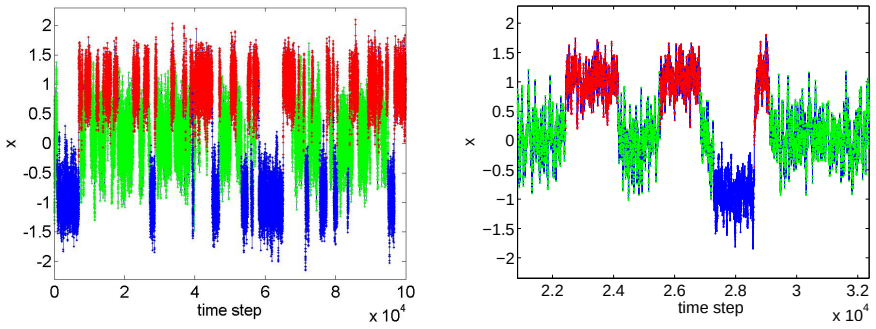

Figure 5.6: Observation time series ($O_m^{(1)}$). Left: Entire time axis. Right: Magnification clearly exhibiting metastability and overlapping. Color/grey scale due to Viterbi path (see text below).

We first apply HMMSDE to observation time series $(O_m^{(1)})$ (see Figure 5.6 for illustration) and set $N = 3$. Eleven iterations of the EM algorithm result in the following transition matrix

$$P = \begin{pmatrix} 0.9983 & 0.0013 & 0.0004 \\ 0.0017 & 0.9983 & 0.0000 \\ 0.0008 & 0.0000 & 0.9992 \end{pmatrix}.$$

that has the spectrum

$$\sigma(P) = \{1.000, 0.999, 0.997\},$$

which perfectly agrees with the results of the transfer operator approach (that is based on the full two-dimensional information instead of on the reduced observation time series). The HMMSDE results for the parameters of the potential and the noise intensities are given in the table below and are in very good agreement with the results to be expected.

parameter	$j = 1$	$j = 2$	$j = 3$
$\mu^{(j)}$	0.0552	1.0169	-0.9584
$\sigma^{(j)^2}$	0.1325	0.1321	0.1302
$D^{(j)}$	0.5589	1.0507	0.9324

Table 5.3: Parameters of HMMSDE for training with $(O_m^{(1)})$.

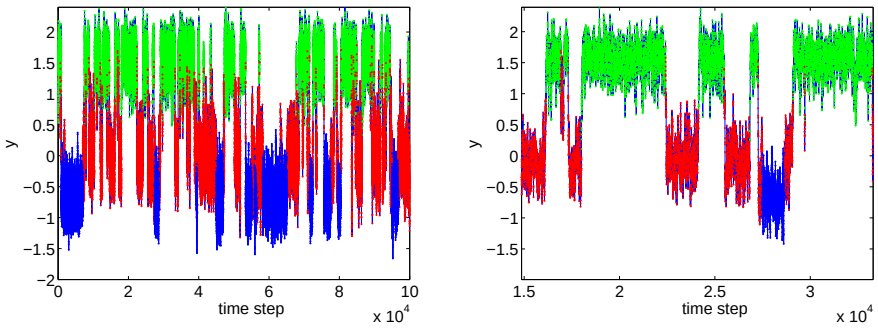

Figure 5.7: Observation time series $(O_m^{(2)})$. Left: Entire time axis. Right: Magnification clearly exhibiting metastability and overlapping. Color/grey scale due to Viterbi path (see text below).

Next we apply HMMSDE to observation time series $(O_m^{(2)})$ (see Figure 5.6) and set $N = 3$. Nine iterations of the EM algorithm result in the following transition matrix

$$P = \begin{pmatrix} 0.9987 & 0.0013 & 0.0000 \\ 0.0014 & 0.9981 & 0.0005 \\ 0.0000 & 0.0007 & 0.9993 \end{pmatrix}.$$

with spectrum

$$\sigma(P) = \{1.000, 0.999, 0.997\}.$$

The HMMSDE results now are again in good agreement with the results to

parameter	$j = 1$	$j = 2$	$j = 3$
$\mu^{(j)}$	1.5526	-0.0084	-0.6693
$\sigma^{(j)^2}$	0.1318	0.1347	0.1343
$D^{(j)}$	1.0607	0.5018	1.1037

Table 5.4: Parameters of HMMSDE for training with $(O_m^{(2)})$.

be expected.

Next we compute the Viterbi paths for the two HMMSDE results based on $(O_m^{(1)})$ or $(O_m^{(2)})$ respectively. This renders the assignment to metastable states as illustrated in Figs. 5.6 and 5.7, and in a two-dimensional representation in Figure 5.8. We observe that the agreement of the assignment with the metastable states resulting from the transfer operator approach (see Figure 5.4) is good. However, as the picture shows, the assignment of the points in the transition regions gets ambiguous. The algorithm for combining the results of our two different projections (as of page 20) yields the results shown in Figure 5.9, where all points which are not clearly assigned to any of the metastable states are identified as belonging to some "transition state".

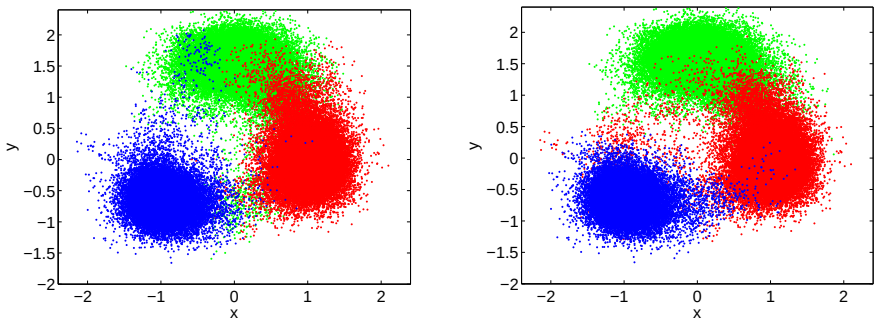

Figure 5.8: Visualization of the assignment of states to the three metastable states as resulting from the Viterbi paths computed via HMMSDE based on $(O_m^{(1)})$ (left) and $(O_m^{(2)})$ (right).

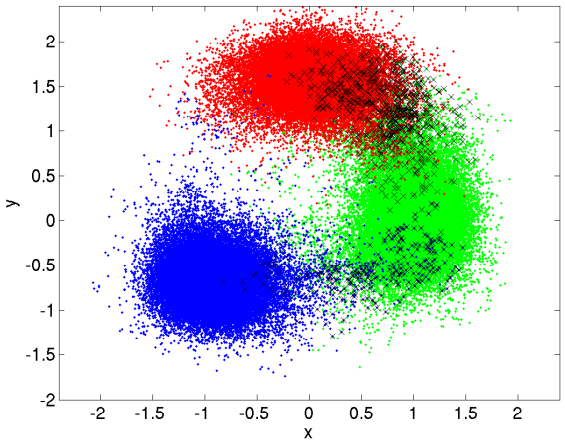

Figure 5.9: Visualization of the assignment of states to the three metastable states (points of three different grey tones) and transition states (black crosses) as resulting from the clustering of both one-dimensional Viterbi paths computed according to the transfer operator approach.

6 Parameter Estimation for Markov Jump Processes

In this section we will define a maximum likelihood estimator for a time-continuous Markov chain. For the sake of illustration we firstly assume that the discrete Markov process $X(t_0) = Y_0, ..., X(t_M) = Y_M$ is directly observable. In the next Chapter the Markov jump process will be combined with HMMs either as output process in section 7.2 or as hidden process in section 7.1. In the latter case even the discrete data points are hidden. However, for the time being we assume the data points $Y_0, ..., Y_M$ to be known and are interested in the underlying time-continuous Markov process. To determine a maximum likelihood estimator we will first set up the likelihood function.

We will begin with the likelihood of a discrete Markov chain $Y_0, ..., Y_M$ and then develop the likelihood for a continuous Markov process, which is of course only partially (namely at discrete time points) observable. Next we will state an EM algorithm for the parameter estimation of a Markov jump process and finally compare it with two other estimation approaches.

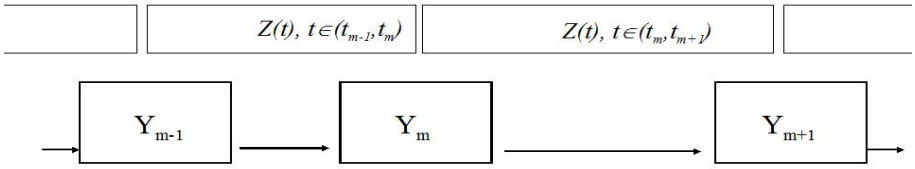

Figure 6.1: Time-continuous Markov process Z with discrete observed data points Y_{m-1}, Y_m and Y_{m+1}.

6.1 Discrete Likelihood

Let $t_0, ..., t_M$ be a series of time points and $Y_m, m = 0, ..., M$ be a discrete Markov chain. The log-likelihood function in this case is simply

$$\mathcal{L}(Y) = \sum_{m=0}^{M-1} \log[P_{Y_m Y_{m+1}}] = \sum_{i,j \in S} \log[P_{ij}] C_{ij}, \tag{6.1}$$

where C denotes the frequency matrix

$$C_{ij} = \sum_{m=0}^{M-1} \mathbb{1}_{\{Y_m=i, Y_{m+1}=j\}}. \tag{6.2}$$

Taking into account that $\sum_j P_{ij} = 1$ holds, the MLE for the discrete Markov chain can be easily identified by counting the relative frequencies:

$$\hat{P}_{ij} = \frac{C_{ij}}{\sum_j C_{ij}}. \tag{6.3}$$

But for a discrete Markov $\hat{P}$ chain it is not clear whether or not an underlying continuous process $P(t)$ exists, such that for certain time increments Δt $\hat{P} = P(\Delta t)$ holds (see 2.1.2). In particular we are interested in the underlying continuous Markov process $X(t)$ with $X(t_m) = Y_m$ for a discrete Markov chain Y. Therefore a reformulation of the likelihood in terms of $P(t) = \text{expm}(tL)$ is required

$$\log[\mathcal{L}(Y)] = \sum_{m=0}^{M-1} \log[\text{expm}((t_{m+1} - t_m)L)_{Y_m Y_{m+1}}]. \tag{6.4}$$

For the sake of simplicity we assume equidistant time points $t_{m+1} - t_m = \tau, \forall m$. This way we can write (6.4) as

$$\log[\mathcal{L}(Y)] = \sum_{i,j \in S} \log[\text{expm}(\tau L)_{ij}] C_{ij}. \tag{6.5}$$

In case of different time increments τ_k we have to compute a frequency matrix for each τ_k and sum up (6.5) over k.

Even for this simplified case, the derivative of (6.5) with respect to the entries of L

$$\frac{\partial}{\partial L} \log[\mathcal{L}] = \sum_{n=1}^{\infty} \sum_{q=0}^{n} \frac{\tau^n}{n!} (L)^{q-1} Z (L')^{n-q},$$

has such a complicated form that the root can not be found analytically. Hence no analytical expression for the MLE with respect to L is available. The notation L' denotes here the transpose of the matrix L and Z is a matrix with entries $Z_{ij} = C_{ij} / \exp(\tau L)_{ij}$.

6.2 Continuous Likelihood

The authors of [3, 11] took as a remedy the likelihood of the whole continuous process $X(t)_{\{t \in [0,T]\}}$ into account. Let $[0, T]$ be the entire time interval, $t_0 = 0 < t_1 < ... < t_M = T$ the time points at which the process is observable

and $t_0 \leq \tau_1 < ... < \tau_{M'} \leq T$ the number of jump times at which the Markov jump process performs the state transitions. In (2.1) the probability density for a single jump was stated. If the whole process contains M' jumps the likelihood results as

$$\mathcal{L}(X) = \prod_{m=0}^{M'} \left(K_{X(\tau_m)X(\tau_{m+1})} \lambda \exp\left(-(\tau_{m+1} - \tau_m)\lambda\right) \right) \exp\left(-(T - \tau_{M'})\lambda\right). \tag{6.6}$$

The last term $\exp\left(-(T - \tau_{M'})\lambda\right)$ comprises the probability density that no jump between the last jump time $\tau_{M'}$ and T occurs. Now we set

$$N_{ij}(\tau_{m+1} - \tau_m) = \{\text{Number of jumps from i to j within } [\tau_m, \tau_{m+1}]\} \tag{6.7}$$

and

$$R_i(\tau_{m+1} - \tau_m) = \int_{\tau_m}^{\tau_{m+1}} \mathbb{1}_{X(s)=i} \, ds. \tag{6.8}$$

Since τ_m and τ_{m+1} are consecutive jump times, (6.7) is either 0 or 1 and (6.8) is either 0 or $\tau_{m+1} - \tau_m$. For arbitrary time points (6.7) denotes the number of transitions $i \to j$ in that time interval and (6.8) the sojourn time in i. The likelihood becomes

$$\begin{aligned} \mathcal{L}(X) \quad = \quad & \prod_{m=0}^{M'} \Big(\prod_{i,j \in S} (K_{ij}\lambda)^{N_{ij}(\tau_{m+1} - \tau_m)} \prod_{i \in S} \exp\left(-R_i(\tau_{m+1} - \tau_m)\lambda\right) \Big) \\ & \prod_{i \in S} \exp\left(-R_i(T - \tau_{M'})\lambda\right). \end{aligned}$$

Since $[\tau_m, \tau_{m+1}], m = 0, ..., M' - 1$ and $[\tau_{M'}, T]$ is a disjoint decomposition of $[0, T]$ and further the $N_{ij}(\tau_{m+1} - \tau_m)$ as well as the $R_i(\tau_{m+1} - \tau_m)$ are independent for each m, (6.7) and (6.8) are additive. We get

$$\mathcal{L}(X) = \prod_{i,j \in S} (K_{ij}\lambda)^{N_{ij}(T)} \prod_{i \in S} \exp\left(-R_i(T)\lambda\right).$$

Finally unify both products,

$$\mathcal{L}(X) = \prod_{i,j \in S} (K_{ij}\lambda)^{N_{ij}(T)} \exp\left(-R_i(T)\lambda K_{ij}\right)$$

exploit (2.3) and obtain the likelihood function in terms of the generator matrix L:

$$\mathcal{L}(X) = \prod_{i,j \in S} (L_{ij} + \delta_{ij}\lambda)^{N_{ij}(T)} \exp\left(-(L_{ij} + \delta_{ij}\lambda)R_i(T)\right). \tag{6.9}$$

If we make furthermore the assumption that no self transitions occur in the Markov jump process, that is no jumps from i to i take place, we can write instead of (2.3)

$$L = \Lambda(\tilde{K} - Id).$$

$\tilde{K}$ is a stochastic matrix with $\tilde{K}_{ii} = 0, \forall i \in S$ and jump rate λ turned into a diagonal matrix $\Lambda = \text{diag}(\lambda_i)_{i \in S}$. In this case $N_{ii}(T) = 0$ and $L_{ii} = -\lambda_i$ holds, thus we obtain exactly the likelihood from [11]

$$\mathcal{L}(X) = \prod_{\substack{i,\, j \,\in\, S \\ j \,\neq\, i}} (L_{ij})^{N_{ij}(T)} \exp\left(-(L_{ij}) R_i(T)\right). \tag{6.10}$$

Remark 6.1. Since for the maximum likelihood estimation the derivatives of the likelihood with respect to $L_{ij}, i \neq j$ suffice, we obtain the same estimators regardless whether (6.9) or (6.10) was taken.

To obtain the maximum likelihood estimator we derive the log-likelihood

$$\log[\mathcal{L}(X)] = \sum_{\substack{i,\, j \,\in\, S \\ j \,\neq\, i}} N_{ij}(T) \log[L_{ij}] - (L_{ij}) R_i(T)$$

with respect to L_{ij} and find the root

$$\frac{N_{ij}(T)}{L_{ij}} - R_i(T) = 0.$$

The MLE is simply

$$\hat{L}_{ij} = \frac{N_{ij}(T)}{R_i(T)}. \tag{6.11}$$

The estimator for the diagonal entries follows simply from the generator constraint

$$\hat{L}_{ii} = -\sum_{j \neq i} \hat{L}_{ij}.$$

6.3 Finding an Optimal Generator under Partial (Discrete) Observation

The likelihood (6.10) takes the whole continuous process $X(t)$ into account. If we want to estimate the Maximum likelihood estimator with respect to discrete data $Y_0 = X(t_0), ..., Y_{M'} = X(t_{M'})$, we can not compute (6.11) since neither N nor R are available. On the other hand the discrete likelihood function (6.1) does not allow for an analytical maximum likelihood estimator.

In short, we want to maximize the likelihood $\mathcal{L}(Y|L) = \mathbb{P}(Y|L)$, which is the marginal distribution of the continuous likelihood $\mathcal{L}(X|L) = \mathbb{P}(X|L)$, but we can only observe the discrete data points Y.

A standard technique to determine the MLE under partial observability is the **EM algorithm** (see Chap. 3.1). As a first step we will show that $\mathcal{L}(Y|L)$ is indeed the marginal distribution of $\mathcal{L}(X|L)$.

In the following let Y denote the discrete Markov chain, X the continuous process and $Z = X \backslash Y$ the continuous process between the observable data points Y (see also Figure 6.1).

$Y_m = X(t_m)$,	$m = 0, ..., M';\ t_0 = 0, t_{M'} \leq T$	the discrete observed random variables
$X(t)$,	$t \in [0, T]$	the entire continuous process
$Z =$	$X(t), t \in [0, T] \backslash \{t_0, ..., t_{M'}\}$	"the hidden continuous random variables"
$\mathcal{Z}$	$\mathcal{Z}_t$ $\sigma-$ algebra over S	Filtration generated by Z

Theorem 6.1. With the above notation as well as (6.1) and (6.10) holds

$$\mathcal{L}(Y|L) = \int_{z \in \mathcal{Z}} \mathcal{L}(z, Y|L) dz.$$

Proof. To show that (6.1) is the marginal distribution of (6.10), we integrate over all realizations Z of the continuous process between the observables Y. Due to the Markov property we have

$$\mathbb{P}(Y|L) = \prod_{m=0}^{M'-1} \mathbb{P}(Y_{m+1}|Y_m, L).$$

On the right hand side even inter-independence between the intervals

$$Z_{(t_m, t_{m+1})} = Z(t) : t \in (t_m, t_{m+1}).$$

is fulfilled. Hence it is factorizable analogously:

$$\int_{z \in \mathcal{Z}} \mathbb{P}(z, Y|L) dz = \prod_{m=0}^{M'-1} \int_{z \in \mathcal{Z}_{(t_m, t_{m+1})}} \mathbb{P}(z, Y_{m+1}|Y_m, L) dz.$$

Thus it is sufficient to show for each interval that

$$\mathcal{L}(Y_{m+1}|Y_m, L) = \int_{z \in \mathcal{Z}_{(t_m, t_{m+1})}} \mathcal{L}(z, Y_{m+1}|Y_m, L) dz,$$

wherein $\mathcal{Z}_{(t_m, t_{m+1})}$ denotes the filtration $\mathcal{Z}$ restricted on the interval (t_m, t_{m+1}).

For this purpose we assume that during the time interval $[t_m, t_{m+1}]$ occur M' jumps. By construction of a Markov jump process a realization Z is fully determined by the jump times $\tau_1, ..., \tau_{M'}$ together with the states $Z(\tau_m)$ visited at the times τ_m, for $m = 1, ..., M'$. We set $t_m = \tau_0 < \tau_1 < ... < \tau_{M'} \leq t_{m+1}$ and are now able to formulate $\mathcal{L}(z, Y_{m+1}|Y_m, L)$ as

$$\begin{aligned}
& \mathbb{P}(Y_{m+1}, Z((t_m, t_{m+1}))|Y_m) \\
= & \prod_{n=0}^{M'-1} K_{Z(\tau_n)Z(\tau_{n+1})}\ \lambda\ \exp(-(\tau_{n+1} - \tau_n)\lambda) \exp(-(t_{m+1} - \tau_{M'})\lambda) \\
= & \left(\prod_{n=0}^{M'-1} K_{Z(\tau_n)Z(\tau_{n+1})} \right) \lambda^{M'} \exp(-(t_{m+1} - t_m)\lambda),
\end{aligned}$$

(cf. 6.6). All possible realizations from $\mathcal{Z}_{(t_m,t_{m+1})}$ with M' jumps are obtained by integrating over all possible jump times $\tau_1, ..., \tau_{M'}$ and further by summing up over all possible states $Z(\tau_1), ..., Z(\tau_{M'})$.

$$\int_{t_m}^{\tau_{M'+1}} ... \int_{t_m}^{\tau_2} \underbrace{\sum_{Z(\tau_1)\in S} ... \sum_{Z(\tau_{M'})\in S} \left(\prod_{n=0}^{M'-1} K_{Z(\tau_n)Z(\tau_{n+1})} \right)}_{K^{M'}_{Y_m Y_{m+1}}} \lambda^{M'}$$
$$\exp(-(t_{m+1}-t_m)\lambda) d\tau_1 ... d\tau_{M'}$$

$$= K^{M'}_{Y_m Y_{m+1}} \lambda^{M'} \exp(-(t_{m+1}-t_m)\lambda) \int_{t_m}^{\tau_{M'+1}} ... \int_{t_m}^{\tau_2} d\tau_1 \ ... \ d\tau_{M'}$$

$$= K^{M'}_{Y_m Y_{m+1}} \frac{(\lambda(t_{m+1}-t_m))^{M'}}{M'!} \exp(-(t_{m+1}-t_m)\lambda)$$

In the underbraces the discrete version of the theorem of Chapman-Kolmogorov ([15], p.330) was used and in the second equation the independence of the jump times τ_n from the states $Z(\tau_n)$, resp. Y_m was exploited. This way we have determined the likelihood for all possible realizations with M' jumps. Eventually summing up this likelihood over $M' = 0, ..., \infty$ leads to the marginal distribution:

$$\sum_{M'=0}^{\infty} \frac{(K\lambda(t_{m+1}-t_m))^{M'}_{Y_m Y_{m+1}}}{M'!} \exp(-(t_{m+1}-t_m)\lambda)$$
$$= \operatorname{expm}\left((t_{m+1}-t_m)\lambda(K-Id)\right)_{Y_m Y_{m+1}}$$
$$= \operatorname{expm}\left((t_{m+1}-t_m)L\right)_{Y_m, Y_{m+1}} = \mathcal{L}(Y_{m+1}|Y_m, L).$$

□

To maximize the discrete likelihood we introduce again the expected log-likelihood

$$Q(L, L_0) = \mathbb{E}\big[\log[\mathcal{L}(X|L)]|Y, L_0\big]$$
$$= \int \log[\mathcal{L}(X|L)] \ \mathcal{L}(Z|Y, L_0).$$

To build the expectation was chosen $\mathcal{L}(Z|Y, L_0)$ as prior probability measure. This is the main difference to the Baum-Welch algorithm, where the prior probability measure was $\mathcal{L}(O, Y|\lambda_0)$ instead of $\mathcal{L}(Y|O, \lambda_0)$. The observables are the argument not the condition of the likelihood. However, both terms differ only by the factor $\mathcal{L}(O|\lambda_0)$, which is independent on the hidden variables. The analog prior probability measure in the time-continuous Markov process context would be $\mathcal{L}(Z, Y|L_0) = \mathcal{L}(X|L_0)$. That the EM algorithm

still works if $\mathcal{L}(Z|Y, L_0)$ is taken will be proven later in Chapter 7.1. If we insert (6.10) now into the Q-functional we obtain

$$\begin{aligned} Q(L, L_0) &= \int \sum_{i=1}^{d} \sum_{j \neq i} \log[L_{ij}] \; N_{ij}(T) \; \mathcal{L}(Z|Y, L_0) dZ \\ &\quad - \int \sum_{i=1}^{d} \sum_{j \neq i} L_{ij} \; R_i(T) \; \mathcal{L}(Z|Y, L_0) dZ \\ &= \sum_{i=1}^{d} \sum_{j \neq i} \log[L_{ij}] \; \mathbb{E}\left[N_{ij}(T)|Y, L_0\right] \\ &\quad - \sum_{i=1}^{d} \sum_{j \neq i} L_{ij} \; \mathbb{E}\left[R_i(T)|Y, L_0\right]. \end{aligned} \tag{6.12}$$

6.3.1 The Expectation Step

The non-trivial task which remains will be to evaluate the conditional expectations $\mathbb{E}\left[N_{ij}(T)|Y, L_0\right]$ and $\mathbb{E}\left[R_i(T)|Y, L_0\right]$ respectively. In the following two different approaches for that will be discussed. The first one is due to [11] and [3], the second one was presented in [56] and [36]. The first step towards their computation is the observation that by the Markov property, the homogeneity of the Markov jump process and a constant time lag τ the conditional expectations in (6.12) can be expressed as sums

$$\begin{aligned} \mathbb{E}\left[R_i(T)|Y, L_0\right] &= \sum_{k=1}^{d} \sum_{l=1}^{d} C_{kl} \; \mathbb{E}\left[R_i(\tau)|X(\tau) = l, X(0) = k, L_0\right], \\ \mathbb{E}\left[N_{ij}(T)|Y, L_0\right] &= \sum_{k=1}^{d} \sum_{l=1}^{d} C_{kl} \; \mathbb{E}\left[N_{ij}(\tau)|X(\tau) = l, X(0) = k, L_0\right], \end{aligned} \tag{6.13}$$

where C denotes the frequency matrix (6.2).

Remark 6.2. In (6.13) we have assumed a constant time lag τ. If the time lags are different, say $\tau_1, ..., \tau_n$, we have to sum up the expressions above (6.13) over $\tau_1, ..., \tau_n$. Note that also the frequency matrix depends on the time lag $C_{kl}(\tau)$, for $\tau = \tau_1, ..., \tau_n$.

Before we will elucidate the two proposed methods for the computation of the expectation values, we will describe in more detail the meaning of these expectations.

Recall the definition of R (6.8). The expected sojourn time is

$$\mathbb{E}\left[R_i(t)|X(t) = l, X(0) = k, L_0\right] = \int_0^t \mathbb{P}(X(s) = i|X(t) = l, X(0) = k, L_0) \; ds.$$

Due to the Markov property and the Bayesian theorem this can be transformed to

$$\begin{aligned}&\int_0^t \frac{\mathbb{P}(X(s)=i|X(0)=k,L_0)\mathbb{P}(X(t)=l|X(s)=i,L_0)}{\mathbb{P}(X(t)=l|X(0)=k,L_0)}\ ds\\ =&\int_0^t \frac{P_{ki}(s)P_{il}(t-s)}{P_{kl}(t)}\ ds,\end{aligned}$$

with $P(s)=\mathrm{expm}(sL_0)$. To compute the expectation value for N we specify at first the probability for a direct jump from i to j at the time s, given that the process is in the state k at 0 and in the state l at t. For a regular Markov jump process we can express this probability as limit from the left

$$\lim_{\nu\to s}\mathbb{P}(X(\nu)=i,X(s)=j|X(t)=l,X(0)=k,L_0).$$

Note that at the time s a direct jump from i to j is claimed. The regularity conditions in Definition 2.2, in particular the right continuity, prevent instantaneous state transitions, therefore the probability is expressible by the limit above. In this way we can state

$$\begin{aligned}&\mathbb{E}\left[N_{ij}(t)|X(t)=l,X(0)=k,L_0\right]\\ =&\lim_{\nu\to s}\int_0^t \mathbb{P}(X(\nu)=i,X(s)=j|X(t)=l,X(0)=k,L_0)\ dt\\ =&\lim_{\nu\to s}\int_0^t \frac{\mathbb{P}(X(\nu)=i,X(s)=j,X(t)=l|X(0)=k,L_0)}{\mathbb{P}(X(t)=l|X(0)=k,L_0)}\ ds\\ =&\lim_{\nu\to s}\int_0^t \frac{\mathbb{P}(X(\nu)=i|X(0)=k,L_0)\ \mathbb{P}(X(s)=j|X(\nu)=i,L_0)\ \mathbb{P}(X(t)=l|X(s)=j,L_0)}{\mathbb{P}(X(t)=l|X(0)=k,L_0)}\ ds.\end{aligned}$$

As mentioned above, $\lim_{\nu\to s}\mathbb{P}(X(s)=j|X(\nu)=i,L_0)$ denotes the probability for a direct jump, thus it is given by (2.1) instead of (2.4). Altogether we get

$$\begin{aligned}&\lim_{\nu\to s}\int_0^t \frac{P_{ki}(\nu)\ \lambda\ \exp(-\lambda(s-\nu))\ K_{ij}\ P_{jl}(t-s)}{P_{kl}(t)}\ ds\\ =&\int_0^t \frac{P_{ki}(s)\ \lambda\ K_{ij}\ P_{jl}(t-s)}{P_{kl}(t)}\ ds\\ =&\int_0^t \frac{P_{ki}(s)\ L_{0_{ij}}\ P_{jl}(t-s)}{P_{kl}(t)}\ ds.\end{aligned}$$

In the last equation the relation (2.3) was exploited.

These integrals can be solved analytically under certain conditions, but first we will have a closer look at the first approach, in which the expectations are propagated as system of ODEs. The conditional expectations on the

right hand side in (6.13) can be decomposed further by using the identities

$$\begin{aligned}\mathbb{E}\left[R_i(t)|X(t)=l,X(0)=k,L\right] = &\quad \frac{\mathbb{E}\left[R_i(t)\,\mathbb{1}_{X(t)=l}|X(0)=k,L\right]}{P_{kl}(t)},\\ \mathbb{E}\left[N_{ij}(t)|X(t)=l,X(0)=k,L\right] = &\quad \frac{\mathbb{E}\left[N_{ij}(t)\,\mathbb{1}_{X(t)=l}|X(0)=k,L\right]}{P_{kl}(t)}.\end{aligned} \tag{6.14}$$

Finally the auxiliary functions defined by

$$\begin{aligned}M^i_{kl}(t) &:= \mathbb{E}\left[R_i(t)\,\mathbb{1}_{X(t)=l}|X(0)=k,L\right],\\ F^{ij}_{kl}(t) &:= \mathbb{E}\left[N_{ij}(t)\,\mathbb{1}_{X(t)=l}|X(0)=k,L\right]\end{aligned}$$

satisfy systems of ordinary differential equations. The vectors $M^i_k(t) = (M^i_{k1}(t),\ldots,M^i_{km}(t))$ and $F^{ij}_k(t) = (F^{ij}_{k1}(t),\ldots,F^{ij}_{km}(t))$ respectively satisfy the two systems of ODEs

$$\begin{aligned}&\frac{d}{dt}M^i_k(t) = M^i_k(t)L + A^i_k(t), \quad M^i_k(0)=0\\ &\qquad \text{with } A^i_k(t) = P_{ki}(t)e_i,\\ &\frac{d}{dt}F^{ij}_k(t) = F^{ij}_k(t)L + A^{ij}_k(t), \quad F^{ij}_k(0)=0\\ &\qquad \text{with } A^{ij}_k(t) = L_{ij}P_{ki}(t)e_j,\end{aligned} \tag{6.15}$$

where e_i and e_j are the i-th and j-th unit vectors. To summarize, the computation of the function $Q(L,L_0)$ in the E-step reduces to solving the systems of ODEs given in (6.15). Solving these ODEs numerically, however, causes prohibitive computational costs when the number of states of the system is large. Another option is to approximate the matrix-exponentials which are involved in the analytic solutions of (6.15)

$$\begin{aligned}M^i_k(t) &= \int_0^t A^i_k(s)\exp((t-s)L)ds,\\ F^{ij}_k(t) &= \int_0^t A^{ij}_k(s)\exp((t-s)L)ds\end{aligned}$$

via the so-called uniformization method [61]. Choose $\alpha = \max_{i=1,\ldots,d}\{-L_{ii}\}$, and define $B = Id + \alpha^{-1}L$. Then $M^i(t) = (M^i_{kl}(t))_{k,l\in S}$ is given by

$$M^i(t) = \exp(-\alpha t)\alpha^{-1}\sum_{n=0}^{\infty}\frac{(\alpha t)^{n+1}}{(n+1)!}\sum_{j=0}^{n}B^j(e_ie_i')B^{n-j},$$

and $F^{ij}(t) = (F^{ij}_{kl}(t))_{k,l\in S}$ by

$$F^{ij}(t) = L_{ij}\exp(-\alpha t)\alpha^{-1}\sum_{n=0}^{\infty}\frac{(\alpha t)^{n+1}}{(n+1)!}\sum_{j=0}^{n}B^j(e_ie_j')B^{n-j},$$

with e_i' denoting the transpose of the unit vector e_i. However, this expansion is fairly time consuming and for high dimensional matrices intractable. Moreover the infinite sum has to be cut off at a finite n which entails inaccuracies.

In [56] and [36] an alternative way to compute the left hand sides in (6.14) was chosen, which avoids the use of ODEs. It will be explained in detail now.

It has been shown above that the conditional expectations $\mathbb{E}[N_{ij}(t)|X(t)=l,X(0)=k,L]$ and $\mathbb{E}[R_i(t)|X(t)=l,X(0)=k,L]$ can be expressed in terms of the generator L. Recalling the notation of the transition matrix $P(s)=\exp(sL)$, we have the identities

$$\begin{aligned}\mathbb{E}[R_i(t)|X(t)=l,X(0)=k,L] &= \frac{1}{P_{kl}(t)}\int_0^t P_{ki}(s)P_{il}(t-s)ds,\\ \mathbb{E}[N_{ij}(t)|X(t)=l,X(0)=k,L] &= \frac{L_{ij}}{P_{kl}(t)}\int_0^t P_{ki}(s)P_{jl}(t-s)ds.\end{aligned} \tag{6.16}$$

The crucial observation is now that a spectral decomposition of the generator L leads to closed form expressions of the integrals in (6.16). To be more precise, consider the spectral decomposition of a generator L, that is

$$L = UD_\lambda U^{-1}$$

where the columns of the matrix U consist of all eigenvectors to the corresponding eigenvalues of L in the diagonal matrix $D_\lambda=\operatorname{diag}(\lambda_1,\dots,\lambda_m)$. Consequently, the expression of the transition matrix $P(t)$ simplifies to

$$P(t)=\exp(tL)=U\exp(tD_\lambda)U^{-1}$$

and we finally end up with a closed form expression of the integrals in (6.16), that is

$$\int_0^t P_{ab}(s)P_{cd}(t-s)ds=\sum_{p=1}^N U_{ap}U_{pb}^{-1}\sum_{q=1}^N U_{cq}U_{qd}^{-1}\Psi_{pq}(t), \tag{6.17}$$

where the symmetric matrix $\Psi(t)=(\Psi_{pq}(t))_{p,q\in S}$ is defined as

$$\Psi_{pq}(t)=\begin{cases} te^{t\lambda_p} & \text{if } \lambda_p=\lambda_q\\ \frac{e^{t\lambda_p}-e^{t\lambda_q}}{\lambda_p-\lambda_q} & \text{if } \lambda_p\neq\lambda_q.\end{cases}$$

Putting (6.16) and (6.17) together we can compute the expectation values for N and R by

$$\begin{aligned}\mathbb{E}[R_i(t)|X(t)=l,X(0)=k,L] &= \frac{1}{P_{kl}(t)}\sum_{p=1}^N U_{kp}U_{pi}^{-1}\sum_{q=1}^N U_{iq}U_{ql}^{-1}\Psi_{pq}(t),\\ \mathbb{E}[N_{ij}(t)|X(t)=l,X(0)=k,L] &= \frac{L_{ij}}{P_{kl}(t)}\sum_{p=1}^N U_{kp}U_{pi}^{-1}\sum_{q=1}^N U_{jq}U_{ql}^{-1}\Psi_{pq}(t).\end{aligned} \tag{6.18}$$

6.3.2 The Maximization Step

At last the maximization step of the EM algorithm has to be focussed on. We have to determine the parameter L such that

$$\frac{\partial Q(L, L_0)}{\partial L} = 0.$$

Since L has to fulfill the generator constraints it suffices to derive (6.12) with respect to L_{ij} for all $i, j \neq i$ and set the diagonal entries L_{ii} to $-\sum_{i \neq j} L_{ij}$ afterwards. We obtain

$$\frac{\partial Q(L, L_0)}{\partial L_{ij}} = \frac{1}{L_{ij}} \mathbb{E}\left[N_{ij}(T)|Y, L_0\right] - \mathbb{E}\left[R_i(T)|Y, L_0\right].$$

The root is consequently in analogy to the fully observable case (6.11)

$$\hat{L}_{ij} = \frac{\mathbb{E}\left[N_{ij}(T)|Y, L_0\right]}{\mathbb{E}\left[R_i(T)|Y, L_0\right]}. \tag{6.19}$$

This reestimation formula is effectively computable by (6.15) or (6.18) respectively. Finally we will give the EM algorithm for a generator estimation in the pseudocode below (Algorithm 6.1).

Algorithm 6.1 EM algorithm for a generator estimation

Require: Time series Y, initial guess of parameters L_0, accuracy ε.
(1) Set $\tilde{L} := L_0$.
(2) Compute the expectation values for transitions and sojourn times via (6.18) or alternatively by solving the ODEs (6.15), if L_0 is not diagonizable.
(3) Reestimation: determine L according to (6.19).
(4) Compute $\Delta\mathcal{L} = \log[\mathcal{L}(Y|L)] - \log[\mathcal{L}(Y|\tilde{L})]$ via (6.1).
if $\Delta\mathcal{L} > \varepsilon$ **then**
  Set $\tilde{L} := L$.
  Go to step (2)
end if
return L

6.4 Comparison and Discussion of the Maximum Likelihood Estimator

Besides the MLE approach literature provides other possibilities to identify the generator of a Markov jump process from a given time series Y. Two of them shall be described in the next two sections.

The first one is the resolvent approach, which interpolates the Laplace transform. It is possible to compute the generator matrix directly from the resolvent estimator obtained by numerical quadrature.

The second one is a quadratic optimization method to determine a generator whose eigenvalue decomposition corresponds best to the logarithmized eigenvalue decomposition of a empirical transition matrix. The best correspondence is measured by the lowest Euclidian distance.

6.4.1 Resolvent Method

For any generator $L \in \mathcal{G}$ the parameter-dependent matrix

$$R(\alpha) = (\alpha Id - L)^{-1}, \qquad \alpha > 0 \tag{6.20}$$

is called the resolvent of L. The inverse exists for all $\alpha > 0$ since the real parts of the eigenvalues of a generator $L \in \mathcal{G}$ are non-positive. An alternative formula representing the resolvent in terms of the propagator $P(t) = \exp(tL)$ is given by the Laplace transform

$$R(\alpha) = \int_0^\infty \exp(-\alpha t) P(t)\ dt, \qquad \alpha > 0. \tag{6.21}$$

The integral exists since $\|P(t)\| = 1$ holds for all $t \geq 0$, and the equivalence of (6.21) and (6.20) follows from

$$(\alpha Id - L)R(\alpha) = -\int_0^\infty \frac{d}{dt}\Big(\exp(-\alpha t)P(t)\Big)\ dt = Id.$$

The main idea of the resolvent method is to approximate the resolvent using its integral representation (6.21) and then to estimate the underlying generator via the identity

$$L = \alpha Id - R^{-1}(\alpha). \tag{6.22}$$

Computing the integral in (6.21), however, requires an approximation of the propagator $P(t)$. Suppose that the process $X(t)$ has been observed at equidistant time points $t_k = k\tau$ with some fixed time lag $\tau > 0$ and $k = 0, \ldots, K$.

A simple estimate $\tilde{P}^{(k)} \approx P(t_k)$ is provided by

$$\tilde{P}^{(k)} = \Big(\tilde{P}_{ij}^{(k)}\Big)_{i,j} \quad \text{with entries} \quad \tilde{P}_{ij}^{(k)} = \frac{C_{ij}^{(k)}}{\sum_{j=1}^d C_{ij}^{(k)}}, \tag{6.23}$$

with frequency matrix C as defined in (6.2). In the approach of [59] and [58] these estimates are used to approximate $P(t)$ in the interval $[t_n, t_{n+1}]$ by linear interpolation:

$$P(t) \ \approx\ P(t_n) + (t - t_n)\frac{P(t_{n+1}) - P(t_n)}{\tau} \ \approx\ \tilde{P}^{(n)} + (t - t_n)\frac{\tilde{P}^{(n+1)} - \tilde{P}^{(n)}}{\tau}.$$

Substituting this into the integral representation (6.21) gives

$$\begin{aligned} R(\alpha) = & \sum_{n=0}^{m-1} \int_{t_n}^{t_{n+1}} \exp(-\alpha s) P(s)\ ds + \int_{t_m}^{\infty} \exp(-\alpha s) P(s)\ ds \\ \approx & \sum_{n=0}^{m-1} \int_{t_n}^{t_{n+1}} \exp(-\alpha s) \left(\tilde{P}^{(n)} + (s - t_n) \frac{\tilde{P}^{(n+1)} - \tilde{P}^{(n)}}{\tau} \right) ds \\ & + \int_{t_m}^{\infty} \exp(-\alpha s) P(t_m)\ ds. \end{aligned} \quad (6.24)$$

Since all integrals in (6.24) can be solved analytically, this yields an approximation $\tilde{R}(\alpha) \approx R(\alpha)$ to the resolvent, and, using equation (6.22), an estimate $\hat{L}^{(\alpha)} = \alpha Id - \tilde{R}^{-1}(\alpha)$ for the generator. Of course the estimate depends on the particular choice of α, but the optimal value of α can be determined by a maximum likelihood approach; see [59] and [58] for details.

It can easily be shown that for any α the entries $\hat{L}_{ii}^{(\alpha)}$ of the estimated generator $\hat{L}^{(\alpha)}$ satisfy the condition $\hat{L}_{ii}^{(\alpha)} = -\sum_{j \neq i} \hat{L}_{ij}^{(\alpha)}$. However, $\hat{L}^{(\alpha)}$ is in general *not* a generator in the sense of (2.7), because $\hat{L}^{(\alpha)}$ can contain negative or even complex off-diagonal elements. This happens if some of the estimated transition matrices $\tilde{P}^{(n)}$ do not belong to set $\mathcal{P}$. The fact that $\hat{L}^{(\alpha)} \notin \mathcal{G}$ is a severe drawback of the resolvent method. This method is rather designed to estimate a generator that best reflects the given discrete propagators, which belong indeed (albeit to different) transition semigroups. It is applied for instance to the molecular sequence analysis and the reconstruction of phylogenetic trees. However, this method ties up a resolvent function between several estimators of a transition semi-group but it is not the method of choice if a generator for a more general transition matrix is sought-after.

6.4.2 Quadratic Optimization Method

In contrast to the resolvent method, the approach introduced by Crommelin and Vanden-Eijnden [18] provides an estimate that *does* belong to the set $\mathcal{G}$. As in the resolvent approach, first an approximative propagator $\tilde{P}^{(1)} \approx P(t_1) = P(\tau)$ is computed by Equation (6.23). Now suppose an eigenvalue decomposition

$$\tilde{P}^{(1)} = U \Lambda U^{-1}$$

with a diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ containing the eigenvalues exists, and that $\lambda_k \neq 0$ for all k. Then, the matrix

$$\tilde{L} = U Z U^{-1} \quad \text{with} \quad Z = \mathrm{diag}(z_1, \ldots, z_d), \quad z_k = \frac{\log[\lambda_k]}{\tau} \quad (6.25)$$

can be defined, and the approximative propagator can be expressed in terms of the matrix exponential

$$\exp(\tau\tilde{L}) = \exp\left(U\log[\Lambda]U^{-1}\right) = U\Lambda U^{-1} = \tilde{P}^{(1)}.$$

In spite of this relation, $\tilde{L}$ cannot be considered as a reasonable estimate for the generator because $\tilde{L} \notin \mathcal{G}$ in many cases. In order to find an estimate with the correct structural properties, Crommelin and Vanden-Eijnden propose to compute the generator $\hat{L} \in \mathcal{G}$ which is in maximal accordance with the eigenvalue decomposition (6.25). This is motivated by the fact that many properties of a continuous-time Markov chain (such as, e.g., its stationary distribution) depend strongly on the eigenvalues and eigenvectors of its generator. Therefore, in [18] the generator is estimated by solving the quadratic minimization problem

$$\begin{aligned}\hat{L} = \mathrm{argmin}_{L\in\mathcal{G}} \sum_{k=1}^{d} \big(\alpha_k \, |U_k^{-1}L - z_k U_k^{-1}|^2 \quad &+ \quad \beta_k \, |LU_k - z_k U_k|^2 \\ &+ \quad \gamma_k \, |U_k^{-1}LU_k - z_k|^2\big)\end{aligned} \tag{6.26}$$

U_k denotes here the k-th column of U, U_k^{-1} is the k-th row of U^{-1}, and α_k, β_k and γ_k are normalization constants with suitably chosen coefficients a_k, b_k, c_k: $\alpha_k = a_k|z_k U_k^{-1}|^{-2}$, $\beta_k = b_k|z_k U_k|^{-2}$ and $\gamma_k = c_k|z_k|^{-2}$. The problem (6.26) can be solved by quadratic programming (QP). A standard quadratic optimizer is implemented by the matlab `quadprog` command after reformulating (6.26) as

$$\hat{L} = \mathrm{argmin}_{L\in\mathcal{G}} \frac{1}{2}\langle L, HL\rangle + \langle F, L\rangle + E_0$$

with a tensor $H \in \mathbb{R}^{d\times d\times d\times d}$ and a matrix $F \in \mathbb{R}^{d\times d}$; see [18] for details.

6.4.3 Pros and Cons

In order to compare these three approaches, we first of all take the computational complexity into consideration.

The resolvent method requires as input several estimators of the transition matrices at different time points $P(t_k), k = 1, ..., K$. Their computation requires an effort of $O(K(M + d))$, where M is the length of the time series and d the number of states. The summation over the analytically solved integrals requires $O(MN^2)$ and finally the computation of L from $R(\alpha)$ needs a matrix inversion and costs thus $O(d^3)$. Altogether the effort amounts to $O(TM + d^3 + MN^2)$.

For the QP approach suffices one estimated transition matrix $O(M+d)$. Its eigenvalue decomposition is carried out in $O(d^3)$. The more costly part is the quadratic minimization. The minimization problem has $d^2 - d$ degrees

Resolvent	QP	MLE
$O(TM + d^3 + MN^2)$	$O(M + d^3)$ + $O((d^2 - d)^3)$ per iteration	$O(M + d^3)$ + $O(d^6)$ per iteration

Table 6.1: Complexity of the three estimation methods.

of freedom and hence is the Hessian a $(d^2 - d) \times (d^2 - d)$ matrix. For the optimization can be used standard algorithms. The matlab routine `quadprog` for instance uses *active sets methods*. Though the speed of convergence for this algorithm is exponential in the worst case, in the average case it behaves linearly in the number of degrees of freedom. Another class of QP algorithms, the *interior point methods*, has in the worst case polynomial complexity and needs even in the average case less iterations than the active sets methods. But the computational cost in each iteration is significantly higher. The dominant cost of an iteration in the active sets approach is the solution of a system of linear equations, which is cubic in the number of degrees of freedom $O((d^2 - d)^3)$. A short overview about the complexity of optimization problem is to be found in [35].

Eventually, in the MLE approach first the frequency matrix in $O(M)$ and the eigenvalue decomposition of L_0 in $O(d^3)$ is computed. The costly part is again the iteration procedure. In each iteration step have to be computed the expectations for R_i^{kl} in $(O(d^3))$ and N_{ij}^{kl} in $(O(d^2(d^2 - d)))$ as prescribed in (6.16) via (6.17), which takes also $(O(d^2))$. Altogether it amounts to $O(d^2(d^2(d^2 - d) + d^3)) = O(d^6)$. Finally the reestimation for L as prescribed in (6.19) by (6.13) needs $O(d^4)$ operations. This does not alter the effort per iteration of $O(d^6)$. The speed of convergence depends on the initial value L_0 and on the likelihood landscape. We have little knowledge about its convergence behavior, but some examples are discussed in [56].

For the sake of lucidity we summarize the complexity in the Table 6.1. The resolvent method has clearly the lowest complexity, but its severe drawback is the fact that $\hat{L} \in \mathcal{G}$ is not guaranteed. The complexity of the MLE and the QP method are both $O(d^6)$ per iteration. In most observations the QP method seemed to have a better convergence rate. However, for a high-dimensional state space the matlab routine `quadprog` becomes intractable since a $(d^2 - d) \times (d^2 - d)$ matrix is required as input unlike the MLE method, where only $d \times d$ matrices have to be handled. The accuracy of the estimators obtained by either method are comparable. For the application to HMMs, of course the MLE approach is the method of choice. In the EM algorithm the maximum likelihood estimation of the discrete transition matrix is simply replaced by the maximum likelihood estimation of the generator. The increase of the likelihood is still ensured, this is not the case in the QP method. In practice the estimators of MLE and QP correspond at large, but it is still conceivable, that the quadratic optimization yields an

estimator that decreases the likelihood of the entire HMM.

All of the presented methods provide an estimator for a Markov process. If it is applied to a time series, which is not a realization of a Markov process or which is not Markovian at all, the quality of the estimator becomes worse. It is up to the user to choose the time increment appropriately, such that non-Markovian effects are avoided.

6.4.4 Perturbation Theory

We have discussed several approaches to estimate a generator matrix L from a given time series. In this section we will focus on the question how good a generator can be approximated based on a time series of finite length.

Let $Y_0, ..., Y_M$ be a realization from a Markov process $P(\tau) = \exp(\tau L)$. For the MLE of the transition matrix $\hat{P}(\tau)$, which is obtained from the relative frequencies (cf. 6.3), the asymptotic behavior is well-known [18, 2]:

$$\sqrt{M}(\hat{P}(\tau) - P(\tau)) \to Q(\tau).$$

$Q(\tau)$ is normal distributed with mean 0 and covariance matrix

$$\mathbb{E}\left[Q_{x,y}(\tau)Q_{x',y'}(\tau)\right] = \frac{P_{xy}(\tau)}{\pi(x)}\big(\delta(y,y') - P_{x'y'}(\tau)\big)\delta(x,x').$$

This way we can determine the distribution of the perturbation error of the propagator depending on the length of the time series M. If the time series is long enough and thus the error

$$\epsilon = P(\tau) - \hat{P}(\tau)$$

is small enough, the condition of Proposition 2.1

$$\tau^{-1}\|P^{-1}(\tau)\|\|\epsilon\| \leq \min_{i,j\in S}|L_{ij}|$$

is fulfilled and hence, the perturbed generator is imbeddable as well. The error of the generator matrix is bounded by 2.12

$$\left\|\tau(\hat{L} - L)\right\| \leq \|\epsilon\| \left\|P^{-1}(\tau)\right\|.$$

6.5 Numerical Examples

6.5.1 Generator Estimation under Perturbation

In this example, we consider the transition matrix $P(\tau)$ with $\tau = 0.2$ which is generated by the matrix

$$L = \begin{pmatrix} -4.293 & 0.678 & 0.301 & 0.819 & 0.592 & 0.149 & 0.543 & 0.411 & 0.774 & 0.023 \\ 0.033 & -3.833 & 0.633 & 0.260 & 0.636 & 0.878 & 0.485 & 0.527 & 0.147 & 0.231 \\ 0.857 & 0.995 & -5.466 & 0.704 & 0.532 & 0.021 & 0.441 & 0.920 & 0.148 & 0.845 \\ 0.682 & 0.499 & 0.005 & -4.691 & 0.208 & 0.923 & 0.626 & 0.379 & 0.639 & 0.726 \\ 0.801 & 0.430 & 0.816 & 0.082 & -4.268 & 0.632 & 0.077 & 0.638 & 0.093 & 0.694 \\ 0.917 & 0.829 & 0.690 & 0.875 & 0.241 & -5.584 & 0.544 & 0.173 & 0.928 & 0.383 \\ 0.388 & 0.116 & 0.981 & 0.077 & 0.720 & 0.632 & -4.667 & 0.785 & 0.485 & 0.479 \\ 0.472 & 0.598 & 0.069 & 0.741 & 0.400 & 0.753 & 0.270 & -4.435 & 0.163 & 0.967 \\ 0.088 & 0.221 & 0.045 & 0.125 & 0.394 & 0.769 & 0.291 & 0.776 & -3.495 & 0.783 \\ 0.925 & 0.398 & 0.740 & 0.443 & 0.411 & 0.808 & 0.822 & 0.342 & 0.131 & -5.022 \end{pmatrix}.$$

In order to investigate the impact of perturbations due to, e.g., sampling from a time series, we estimate a generator based on a perturbed transition matrix

$$P_\epsilon(\tau) = \exp(\tau L) + k\epsilon, \quad k = 0, ..., 19,$$

where ϵ is the perturbation matrix

$$\epsilon = 10^{-5} \cdot \begin{pmatrix} 4.055 & -3.552 & 1.754 & 0.805 & -4.090 & -3.519 & 4.719 & 0.047 & 0.696 & -0.917 \\ 3.104 & -3.508 & -1.609 & 2.874 & 1.319 & -0.671 & 2.020 & 1.459 & 1.272 & -6.261 \\ -3.22 & -0.978 & -2.611 & 5.673 & -3.653 & 2.386 & 5.726 & -2.478 & 0.154 & -0.993 \\ 4.467 & -1.238 & -5.225 & 1.944 & -1.021 & -3.496 & 2.433 & -2.047 & 2.687 & 1.497 \\ 4.698 & -4.188 & -1.271 & 1.949 & -4.191 & -0.450 & -0.850 & 3.649 & -4.336 & 4.991 \\ 4.376 & -2.336 & -1.603 & 3.415 & 1.556 & 1.850 & -4.529 & -2.277 & 4.355 & -4.808 \\ 1.200 & -2.234 & 5.509 & -4.121 & -1.151 & -0.133 & -3.341 & -3.631 & 4.118 & 3.785 \\ 2.836 & -1.009 & 2.731 & -3.009 & -1.067 & -4.559 & 2.699 & 2.614 & 3.194 & -4.432 \\ -1.478 & 4.040 & -0.318 & -3.722 & -0.412 & 1.249 & 0.450 & -2.992 & -2.153 & 5.336 \\ -1.460 & -1.569 & 5.235 & -0.772 & -2.618 & 4.252 & -2.006 & -0.251 & 0.705 & -1.514 \end{pmatrix}.$$

The upper panel of Figure 6.2 shows the deviation of the estimated generators from the unperturbed generator as a function of the perturbation factor k. The QP-method performs slightly better but both errors $\|L - \tilde{L}_{QP}\|$ and $\|L - \tilde{L}_{MLE}\|$ are of the same order of magnitude. Furthermore, the errors scale linearly with the perturbation factor k. This observation is plausible since for small perturbations the logarithm $\log[P + \epsilon]$ can be approximated by $\log[P] + P^{-1}\epsilon + o(\epsilon)$ as stated in equation (2.12). The lower panel of Figure 6.2 illustrates the behavior of the errors of the estimated transition matrices $\exp(\tau\tilde{L}_{QP})$ and $\exp(\tau\tilde{L}_{MLE})$ respectively. A similar reasoning as above explains the linear scaling.

Finally, we consider the error of the estimated transition matrices $\exp(\tau\tilde{L}_{QP})$ and $\exp(\tau\tilde{L}_{MLE})$ with respect to the perturbed transition matrix $P_\epsilon(\tau) = \exp(\tau L) + k\epsilon$, depicted in Figure 6.3. Notice that the error $\|P_\epsilon(\tau) - \exp(\tau\tilde{L})\|$ is bounded from above, namely

$$\|P_\epsilon(\tau) - \exp(\tau\tilde{L})\| \leq \|\exp(\tau L) - \exp(\tau\tilde{L})\| + k\|\epsilon\|.$$

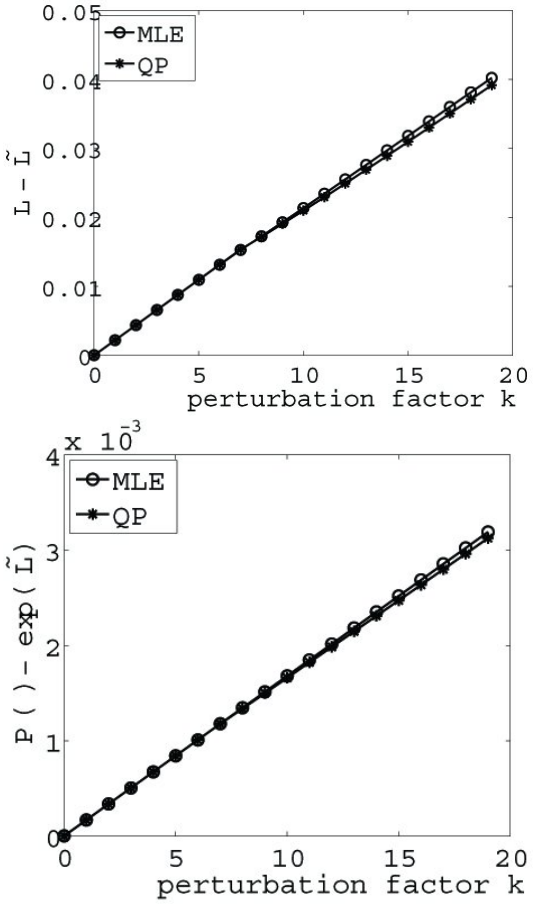

Figure 6.2: Upper: Approximation error of the generator estimates $\tilde{L}_{QP}$ and $\tilde{L}_{MLE}$ with respect to the unperturbed generator as a function of the perturbation factor k. Lower: Error of the estimated transition matrices $\exp(\tau\tilde{L}_{QP})$ and $\exp(\tau\tilde{L}_{MLE})$ with respect to the unperturbed transition matrix $\exp(\tau L)$ as a function of the perturbation factor k. Results for $\tau = 0.2$.

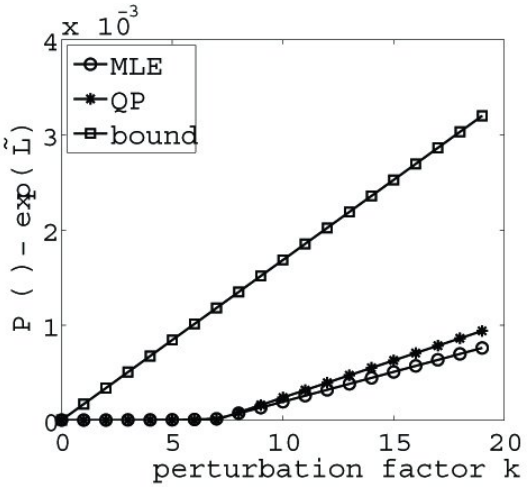

Figure 6.3: Error of the estimated transition matrices $\exp(\tau\tilde{L}_{QP})$ and $\exp(\tau\tilde{L}_{MLE})$ with respect to the perturbed transition matrix $P_\epsilon(\tau) = \exp(\tau L) + k\epsilon$ as a function of the perturbation factor k. The upper bound was computed via $\tilde{L}_{MLE}$.

Indeed Figure 6.3 shows that both errors obey that bound. For the perturbation factors up to $k = 5$, the matrix logarithm of P_ϵ is still a generator whereas for $k = 6, \ldots, 19$ the perturbation is apparently high enough to destroy the generator structure of the matrix logarithm of P_ϵ. Indeed the condition of Corollary 2.1

$$\tau^{-1}\|P^{-1}(\tau)\|\|\epsilon\| \leq \min_{i,j\in S}|L_{ij}|$$

is fulfilled up to $k = 5$. However, the accuracy of both methods is again of the same order of magnitude.

6.5.2 A Metastable Generator

In a second example we investigate a generator matrix with a strong metastable character. The matrix L consists of two metastable blocks.

$$L = \left(\begin{array}{ccc|ccc} -0.9426 & 0.4860 & 0.4565 & 0.0001 & 0 & 0 \\ 0.2311 & -0.2497 & 0.0185 & 0 & 0.0001 & 0 \\ 0.6068 & 0.7621 & -1.3690 & 0 & 0 & 0.0001 \\ \hline 0.0001 & 0 & 0 & -1.3276 & 0.9218 & 0.4057 \\ 0 & 0.0001 & 0 & 0.6154 & -1.5510 & 0.9355 \\ 0 & 0 & 0.0001 & 0.7919 & 0.1763 & -0.9683 \end{array}\right)$$

For the sake of simplicity has been chosen a small 6×6-matrix. Typical realizations of molecular dynamics time series require of course a larger number of discretization boxes. However, a better insight into the quality of the estimated parameters provide small test examples.

The off-block-diagonal entries are close to zero. A perturbation of these can violate easily the generator condition. And indeed we obtain from a realization as in Figure 6.4 by counting a transition matrix $\tilde{P}$, which is not in $\mathcal{P}$.

Besides from that for sufficiently small τ quasi-zero entries occur also in the transition matrix $P(\tau)$. During the estimation procedure $P_{ij}(\tau)$ has to be inverted for each pair i, j, for which a state transition was observed. If an entry $P_{ij}(\tau)$ is too close to machine accuracy the quality of the estimator could be perturbed. For an actual zero entry no $i \to j$ transition was observed and hence no inversion is necessary. Consequently, zero entries are no problem, though quasi-zeros are.

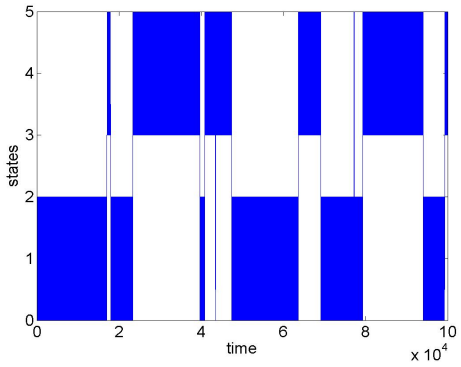

Figure 6.4: Realization from the generator L stated above. The time lag was set constantly to 1.

$$\tilde{P} = \begin{pmatrix} 0.4716 & 0.3705 & 0.1577 & 0.0002 & 0.0001 & 0 \\ 0.1423 & 0.8245 & 0.0331 & 0.0001 & 0 & 0.0001 \\ 0.2483 & 0.4400 & 0.3114 & 0 & 0.0002 & 0 \\ 0.0001 & 0.0002 & 0 & 0.4357 & 0.2760 & 0.2880 \\ 0 & 0.0001 & 0 & 0.2891 & 0.3387 & 0.3722 \\ 0 & 0.0001 & 0 & 0.3180 & 0.1698 & 0.5120 \end{pmatrix}$$

$$\log[\tilde{P}] = \begin{pmatrix} -0.9482 & 0.4974 & 0.4505 & 0.0005 & \boxed{\text{-0.0001}} & \boxed{\text{-0.0001}} \\ 0.2360 & -0.2538 & 0.0178 & 0.0000 & \boxed{\text{-0.0000}} & 0.0001 \\ 0.6041 & 0.7389 & -1.3433 & \boxed{\text{-0.0004}} & 0.0010 & \boxed{\text{-0.0004}} \\ 0.0002 & 0.0003 & \boxed{\text{-0.0001}} & -1.2850 & 0.8901 & 0.3945 \\ \boxed{\text{-0.0000}} & 0.0000 & 0.0000 & 0.5793 & -1.5392 & 0.9599 \\ \boxed{\text{-0.0001}} & 0.0001 & 0.0000 & 0.7739 & 0.1763 & -0.9502 \end{pmatrix}$$

The generator estimation by the MLE and the QP method show rather slight differences. The distance to the original generator is measured by $\|L - \hat{L}\|$, in Table 6.2 are listed the distances as well as the likelihood of

	MLE	QP
$\|L - \hat{L}\|$	0.0723	0.0720
Likelihood	-85902	-85903

Table 6.2: Comparison of the distance to the original generator and the likelihood.

Original L	-1.9234 + 0.3052i	-1.9234 - 0.3052i	-1.7001	-0.8611	-0.0002	0
$\hat{L}_{MLE}$	-1.8869 + 0.3309i	-1.8869 - 0.3309i	-1.6813	-0.8637	-0.0003	0
$\hat{L}_{QP}$	-1.8871 + 0.3309i	-1.8871 - 0.3309i	-1.6810	-0.8629	-0.0003	0

Table 6.3: Eigenvalues of the generator matrices.

the data. While the distance to the original generator is a little smaller for the QP-estimator the likelihood is slightly higher for the MLE-estimator. Table 6.3 contains the eigenvalues that are recovered satisfactory by both methods.

7 HMM with Generator Estimation

We will call the combination of the HMM framework with an Markov jump process **HMM-MJP**. A Markov jump process can be integrated into an HMM in two ways: as output process and as hidden process. In this chapter we will discuss both by means of several examples. In both cases we will recover an HMM-MJP from time series that were generated themselves by an HMM-MJP.

Furthermore, HMMs with Markov jump output process will be applied to metastable time series where each metastable set is expressed by a separate Markov jump process. This approach will be compared to the estimation of one large metastable generator matrix in a second example. Eventually we will compare the HMM-MJP estimation with the HMMSDE approach in a final example.

The combination of a hidden Markov jump process with different output processes is possible as well. We will discuss this topic in Section 7.1.2.

There are models in literature that exhibit some similarities to the HMM-MJP with hidden Markov jump process. For example in [13, 12] a more general case of an infinite-dimensional state space is presented. They approach the problem in a more technical way without specifying an implementation. The model described in [57] additionally requires the jump times to be known. For the HMM-MJP presented here, this assumption is not necessary. Also in [62] an HMM is defined, which they call continuous time Bayesian network. But the design differs from ours as follows: The observables are discrete points of a Markov jump process and the hidden variables are the continuous time intervals between the observables. This way a single Markov process only can be modeled. The generator estimation is similar to our approach with slight differences in specifying the likelihood function.

7.1 Hidden Markov Jump Process

First, let the Markov jump process be a hidden process. While so far we confined ourselves to estimating the transition matrix of the hidden process, we next aim at estimating a generator as parameter of a hidden Markov model. This allows for modeling time-continuity. The observables are random variables $O_0, ..., O_M$. Each O_m only depends on the hidden state Y_m. However, $Y = Y_0, ..., Y_M$ is no longer a realization of a time-discrete Markov chain but consists of discrete data points of a time-continuous Markov process realization (cf. Figure 7.1). To estimate a generator for the discrete transitions between the hidden states Y we have to face the problem that even the discrete data points $Y_0, ..., Y_M$ are not observable. We need a twofold expectation step:

- First, estimate the occupancy probabilities for each time point and each state. This results in an estimated frequency matrix (or several

in case of varying time steps respectively).

- Second, use the estimated frequency matrices to maximize (6.12).

The first step is realized by a well-known HMM technique: the forward-backward algorithm (3.2). For the second step an expectation value for the frequency matrix is needed. Instead of observable states we only have occupancy probabilities. That is, the characteristic function $\mathbb{1}_{\{Y_m=i,Y_{m+1}=j\}}$ from (6.2) is converted to the probability $\mathbb{P}(O_0,...,O_M,Y_m=i,Y_{m+1}=j)$, where $O_0,...,O_M$ denote the observables. Thus we obtain

$$\hat{C}_{ij} \quad = \quad \sum_{m=1}^{M-1} \mathbb{P}(O_0,...,O_M,Y_m=i,Y_{m+1}=j) = \mathbb{E}\left[C_{ij}\right]. \qquad (7.1)$$

This proceeding leads to a nested EM algorithm: in each estimation step the estimators for the observable distributions as well as the estimator for the hidden Markov process with respect to (7.1) are computed. For the latter estimator another EM algorithm is applied.

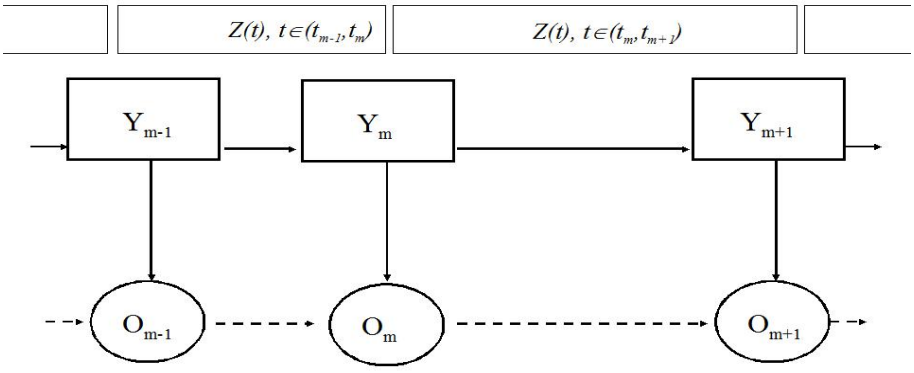

Figure 7.1: Hidden Markov model with hidden Markov jump process. The observable data points are $O_0...,O_{m-1},O_m,O_{m+1},...,O_M$. The hidden Markov chain is time-continuous. In the nested EM algorithm, first the discrete hidden states Y_m at the time points t_m are estimated, and based on this, the entire hidden continuous process $Z \cup Y$.

7.1.1 Concept

Now we focus on a more detailed description of both steps: First, we recapitulate the standard expected likelihood

$$Q_1(\lambda,\lambda_0) = \mathbb{E}\big[\log[\mathcal{L}(O,Y|\lambda)]|\lambda_0\big].$$

For the sake of simplicity a Gaussian is used as output distribution in the model design. However, the output distribution is exchangeable. The expected likelihood can be transformed to

$$\mathbb{E}\big[\log[\mathcal{L}(O|Y,(\mu_k)_{k\in S},(\Sigma_k)_{k\in S})]|\lambda_0\big] + \mathbb{E}\big[\log[\mathcal{L}(Y|L)]|\lambda_0\big], \tag{7.2}$$

where $\lambda = ((\mu_k)_{k\in S},(\Sigma_k)_{k\in S},L)$. The parameters $(\mu_k)_{k\in S}$ and $(\Sigma_k)_{k\in S}$ denote the mean and the covariance matrix of the output process and L the generator matrix determining the hidden process. For the standard HMM with discrete transition matrix P the maximum likelihood estimator is

$$P_{ij} = \frac{\hat{C}_{ij}}{\sum_j \hat{C}_{ij}},$$

where $\hat{C}$ is the expected frequency matrix as given in (7.1). This is exactly the reestimation formula for P as stated in the Baum-Welch formulas (3.10). With the expected frequency matrix the second step is straightforward. As pointed out in [56], $\mathcal{L}(Y|\lambda)$ can also be formulated as $\mathcal{L}(C|\lambda)$. We restate the second term of (7.2) as

$$\mathbb{E}\big[\log[\mathcal{L}(Y|L)]|\lambda_0\big] = \log[\mathcal{L}(\hat{C}|L)]. \tag{7.3}$$

$\log[\mathcal{L}(\hat{C}|L)]$ is, analogous to (6.1), the discrete likelihood and the marginal distribution of the continuous likelihood (6.10). The maximization of (7.3) is now carried out with a second EM algorithm, where in each iteration

$$L_{k+1} = \max_L Q_2(L, L_k)$$

for

$$Q_2(L,L_k) = \mathbb{E}\big[\log[\mathcal{L}(Y,Z|L)]|\hat{C},L_k\big] = \mathbb{E}\big[\log[\mathcal{L}(Y,Z|L)]|Y,L_k\big]$$

is determined. Here Z denotes the continuous process between the discrete data points $t_0,...,t_M$, while $Y = Y_0,...,Y_M$ denotes the discrete process at the specified time points. The expected continuous likelihood Q_2 is conditioned on the expected frequency matrix $\hat{C}$ and the current parameter guess L_k. The expected log-likelihoods Q_1 and Q_2 differ in the conditions: while Q_1 only depends on the parameters λ, Q_2 depends on the parameters L as well as on the observables $\hat{C}$. However, the well-known relation from EM-theory [5, 22]

$$Q(\lambda_{k+1},\lambda_k) \geq Q(\lambda_k,\lambda_k) \Rightarrow \mathcal{L}(\lambda_{k+1}) \geq \mathcal{L}(\lambda_k) \tag{7.4}$$

holds in both cases. For the standard HMM case this was shown in [5] by the following calculation:

$$\begin{aligned}
& \log\left[\frac{\mathbb{P}(O|\lambda_{k+1})}{\mathbb{P}(O|\lambda_k)}\right] \\
= \; & \log\left[\frac{\int \mathbb{P}(O,Y|\lambda_{k+1})\; dq}{\mathbb{P}(O|\lambda_k)}\right] \\
= \; & \log\left[\int \frac{\mathbb{P}(O,Y|\lambda_k)}{\mathbb{P}(O|\lambda_k)}\; \frac{\mathbb{P}(O,Y|\lambda_{k+1})}{\mathbb{P}(O,Y|\lambda_k)}\; dq\right] \\
\geq \; & \int \frac{\mathbb{P}(O,Y|\lambda_k)}{\mathbb{P}(O|\lambda_k)} \log\left[\frac{\mathbb{P}(O,Y|\lambda_{k+1}}{\mathbb{P}(O,Y|\lambda_k)}\right] dq \\
= \; & \frac{1}{\mathbb{P}(O|\lambda_k)} \int \mathbb{P}(O,Y|\lambda_k)\big[\log[\mathbb{P}(O,Y|\lambda_{k+1})] - \log[\mathbb{P}(O,Y|\lambda_k)]\big] dq \\
= \; & \frac{1}{\mathbb{P}(O|\lambda_k)} \left[Q_1(\lambda_{k+1},\lambda_k) - Q_1(\lambda_k,\lambda_k)\right].
\end{aligned}$$

In line 4 Jensen's inequality is applied. Thus, if $Q_1(\lambda_{k+1},\lambda_k) - Q_1(\lambda_k,\lambda_k) \geq 0$, $\log\left[\frac{\mathbb{P}(O|\lambda_{k+1})}{\mathbb{P}(O|\lambda_k)}\right] \geq 0$ follows immediately. This is equivalent to $\mathbb{P}(O|\lambda_{k+1}) \geq \mathbb{P}(O|\lambda_k)$. Now we perform the same calculation for the expected likelihood stated in Q_2.

$$\begin{aligned}
& \log\left[\frac{\mathbb{P}(Y|L_{k+1})}{\mathbb{P}(Y|L_k)}\right] \\
= \; & \log\left[\frac{\int \mathbb{P}(Y,Z|L_{k+1})\; dZ}{\mathbb{P}(Y|L_k)}\right] \\
= \; & \log\left[\int \frac{\mathbb{P}(Y,Z|L_k)}{\mathbb{P}(Y|L_k)}\; \frac{\mathbb{P}(Y,Z|L_{k+1})}{\mathbb{P}(Y,Z|L_k)}\; dZ\right] \\
\geq \; & \int \frac{\mathbb{P}(Y,Z|L_k)}{\mathbb{P}(Y|L_k)} \log\left[\frac{\mathbb{P}(Y,Z|L_{k+1}}{\mathbb{P}(Y,Z|L_k)}\right] dZ \\
= \; & \int \mathbb{P}(Z|Y,L_k)\big[\log[\mathbb{P}(Y,Z|L_{k+1})] - \log[\mathbb{P}(Y,Z|L_k)]\big] dZ \\
= \; & \left[Q_2(L_{k+1},L_k) - Q_2(L_k,L_k)\right].
\end{aligned}$$

The only difference occurs in line 5, where the Bayesian formula was applied and hence the likelihood measure in the Q-functional becomes conditional on Y. For both calculations holds: equality is achieved only in a critical point of $\mathcal{L}(\lambda)$ (resp. $\mathcal{L}(L)$), which is a fixed point of the EM-iteration. In particular, from

$$Q_2(L_{k+1},L_k) \geq Q_2(L_k,L_k) \Rightarrow \mathcal{L}(\hat{C}|L_{k+1}) \geq \mathcal{L}(\hat{C}|L_k) \tag{7.5}$$

also follows an increase of the right term in (7.2). Furthermore, we get a maximization of the left term of (7.2) by the standard reestimation of the

output parameters. Altogether, we get

$$\begin{aligned} & Q_1(\lambda, \lambda_0) \\ = \; & \mathbb{E}\big[\log[\mathcal{L}(O|Y, (\mu_k)_{k\in S}, (\Sigma_k)_{k\in S})]|\lambda_0\big] + \mathbb{E}\big[\log[\mathcal{L}(Y|L)]|\lambda_0\big] \\ \geq \; & \mathbb{E}\big[\log[\mathcal{L}(O|Y, (\mu_k^0)_{k\in S}, (\Sigma_k^0)_{k\in S})]|\lambda_0\big] + \mathbb{E}\big[\log[\mathcal{L}(Y|L_0)]|\lambda_0\big] \\ = \; & Q_1(\lambda_0, \lambda_0), \end{aligned}$$

and with (7.4), $\mathcal{L}(O|\lambda) \geq \mathcal{L}(O|\lambda_0)$ follows. This shows that both EM-algorithms can be linked without affecting the correctness of the algorithm. In the next section we will state the implementation for an HMM with a time-continuous hidden Markov process.

7.1.2 Parameter Estimation

Eventually, we will point out the reestimation procedure following the scheme as introduced in Section 3.1.

Likelihood. First, we state the joint likelihood of the observable process O and the hidden continuous process consisting of the discrete time series $Y = Y_0, ..., Y_M$ and the continuous part $Z = Z(t),\ t \in [t_0, t_M]\backslash\{t_0, ..., t_M\}$ between the time points t_m

$$\begin{aligned} \mathcal{L}(O, Y, Z|\lambda) \;\; = \;\; & \mathbb{P}(Y_0|\lambda)\ \psi(O_0|Y_0, \lambda) \\ & \prod_{m=1}^{M} \mathbb{P}(Z(t_{m-1}, t_m), Y_m|Y_{m-1}, \lambda)\ \psi(O_m|Y_m, \lambda). \end{aligned}$$

We will denote the output distribution, which can stand for any probability distribution by $\psi(O_m|Y_m)$. Here we take a Gaussian. Without loss of generality $\psi(O_m|Y_m)$ can also be substituted by $\rho(O_m|O_{m-1}, Y_m)$ from Section 5.2.1 or $\varphi(O_m|O_{m-1}, Y_m)$ from Section 7.2.2.

To make the parameter estimation feasible instead of the time-discrete transition probability the time-continuous variant from (6.10) will be used.

$$\begin{aligned} \mathbb{P}(Z(t_{m-1}, t_m), Y_m|Y_{m-1} = i, \lambda) \;\; = \;\; & \prod_{k,l\in S, l\neq k} ((L_i)_{kl})^{N_{i_{kl}}(t_m - t_{m-1})} \\ & \exp\Big(- ((L_i)_{kl})\ R_{i_k}(t_m - t_{m-1})\Big), \end{aligned}$$

with N and R as specified in (6.7) and (6.8).

Expectation step. Since amongst the random variables O, Y, Z only O is observable, we have to build the expectation value over Y and Z. We obtain

the expected likelihood as combination of Q_1 and Q_2 from the foregoing Section 7.1.1.

$$\begin{aligned}
Q(\lambda,\lambda_0) &= \mathbb{E}_Y\left[\mathbb{E}_Z\left[\log[\mathcal{L}(O,Y,Z|\lambda)]\middle|Y,L_0\right]\middle|\lambda_0\right] \\
&= \mathbb{E}_Y\left[\log[\mathcal{L}(O|Y,\lambda)]\middle|\lambda_0\right] + \mathbb{E}_Y\left[\mathbb{E}_Z\left[\log[\mathcal{L}(Y,Z|\lambda)]\middle|Y,L_0\right]\middle|\lambda_0\right] \\
&= \sum_{i\in S}\Big(\left(\log[\pi_i] + \log[\varphi_0(O_0|Y_0)]\right)\,\alpha_i(t_0)\beta_i(t_0) + \\
&\quad \sum_{m=1}^{M}\log[\psi(O_m|Y_m=i)]\;\alpha_i(t_m)\beta_i(t_m)\Big) + \\
&\quad \sum_{i,j\in S}\sum_{m=1}^{M-1}\Big(\mathbb{E}_Z\left[\log[\mathbb{P}(Z(t_{m-1},t_m),Y_m|Y_{m-1}=i,\lambda)]\middle|Y,L_0\right] \\
&\quad \alpha_i(t_m)\;(P_0(t_{m+1}-t_m))_{ij}\;\psi(O_{m+1}|Y_{m+1}=j)\;\beta_i(t_m)\Big).
\end{aligned}$$

Note that the transition probability, which also enters the forward backward variables, is no longer independent from the time step. This is indicated by the time increment in the brackets $(P_0(t_{m+1}-t_m))_{ij} = \mathrm{expm}((t_{m+1}-t_m)L)_{ij}$. The adapted forward-backward recursion is given by

$$\begin{aligned}
\alpha_i(t_m) &= \sum_{Y_{m-1}\in S}\alpha_{Y_{m-1}}(t_{m-1})\;P_{Y_{m-1}i}(t_m-t_{m-1})\;\psi(O_m|Y_m=i) \qquad (7.6)\\
\beta_i(t_m) &= \sum_{Y_{m+1}\in S}P_{iY_{t_{m+1}}}(t_{m+1}-t_m)\;\psi(O_{m+1}|Y_{m+1}=j)\;\beta_{Y_{m+1}}(t_{m+1}).
\end{aligned}$$

Maximization step. The estimators of the output distribution are obtained by the reestimation formulas derived above (e.g. μ and Σ from 3.10). We will derive now the reestimation formula for L. The part of Q depending on L is

$$\begin{aligned}
&\sum_{i,j\in S}\sum_{m=1}^{M-1}\mathbb{E}_Z\left[\log[\mathbb{P}(Z(t_{m-1},t_m),Y_m|Y_m=i,Y_{m-1}=i,\lambda)]|Y,L_0\right] \\
&\qquad \alpha_i(t_m)\;(P_0(t_{m+1}-t_m))_{ij}\;\psi_d(O_{m+1}|Y_{m+1}=j)\;\beta_i(t_m).
\end{aligned}$$

Using the frequency matrix $\hat{C}$ from (7.1)

$$\hat{C}_{ij} = \sum_{m=1}^{M}\alpha_i(t_m)\;(P_0(t_{m+1}-t_m))_{ij}\;\psi_d(O_{m+1}|Y_{m+1}=j)\;\beta_i(t_m),$$

the expression above simplifies to

$$\sum_{i,j\in S}\left(\sum_{\substack{k,l\,\in\,S\\ l\,\neq\,k}}\log[L_{kl}]\mathbb{E}\left[N_{kl}(\tau)|Y,L_0\right]-L_{kl}\mathbb{E}\left[R_k(\tau)|Y,L_0\right]\right)\hat{C}_{ij}.$$

By applying the frequency matrix a constant time increment τ is assumed. This is not necessary, but in case of varying time steps a frequency matrix C for each time increment is required, which is computed by

$$\begin{aligned}\hat{C}_{ij}(\tau_n)=\sum_{m=1}^{M}\quad & \alpha_i(t_m)\ (P_0(t_{m+1}-t_m))_{ij}\ \psi_d(O_{m+1}|Y_{m+1}=j)\\ & \beta_i(t_m)\ \mathbb{1}_{\{t_{m+1}-t_m=\tau_n\}}.\end{aligned}$$

The root of the partial derivative with respect to L_{ij} leads to the reestimation formula

$$L_{ij}=\frac{\sum_{x,y\in S}\mathbb{E}\left[N_{ij}(\tau)|x,y,L_0\right]C_{xy}}{\sum_{x,y\in S}\mathbb{E}\left[R_i(\tau)|x,y,L_0\right]C_{xy}}.$$

We conclude this section with a summary of the EM algorithm in pseudocode notation (see Algorithm 7.1). The output process, here a Gaussian, is replaceable by several other processes like a Markov jump process (Section 7.2) or an Ornstein-Uhlenbeck process (Section 5). However, not every process can handle varying time lags. The multi-dimensional HMMSDE as presented in [41] for example requires a time series with a constant time lag. Thus it can be combined with a hidden Markov jump process, however, only with equidistant time steps.

7.1.3 Numerical Examples

In the following examples we investigate an HMM with four hidden states that has a continuous hidden process and Gaussian output. The output distributions are overlapping considerably. The models in both examples differ only in the transition matrix from each other. From either model a time series with 500 000 steps was generated. But we do not take the entire time series into consideration but only certain time points, such that the time lag between two observables is one, two or three. The time lags were chosen at random with probability 0,6 for time lag 1 and 0,2 for time lag 2 and 3, respectively.

Model with generator. As a first example we consider a model with a transition matrix that has an underlying generator. The output distributions reveal a noticeable overlap (see Fig. 7.2).

Algorithm 7.1 EM algorithm for an HMM with hidden Markov jump process.

Require: Time series $O_0, \ldots, O_M$,
initial guess of parameters $\lambda_0 = (\pi, L, (\mu_i)_{i\in S}, (\Sigma_i)_{i\in S})$, accuracy ε.
(1) Set $\tilde{\lambda} := \lambda_0$.
(2) Compute observation likelihood $\mathcal{N}(O_m, \mu_i, \Sigma_i)$ and forward-backward variables $\alpha_i(t_m), \beta_i(t_m)$, for $i \in S, m = 1, .., M$ via (7.6).
(3) Compute for each τ_n a frequency matrix

$$\hat{C}_{kl}(\tau_n) = \sum_{m=1}^{M-1} \Big(\quad \alpha_k(t_m)\ P_{kl}(\tau_m)\ \mathcal{N}(O_{m+1}, \mu_l, \Sigma_l) \\ \beta_l(t_{m+1}) \mathbb{1}_{\{t_{m+1}-t_m=\tau_n\}} \Big)$$

(4) M-Step: Reestimate model parameters:
Estimate $\tilde{L}_{ij}$ via Algorithm 6.1.
Estimate output parameters via (3.10).
if $\mathcal{L}(\tilde{\lambda}) - \mathcal{L}(\lambda_0) > \varepsilon$ **then**
$\lambda_0 = \tilde{\lambda}$
Go to Step (2).
else
return $\tilde{\lambda}$.
end if

Original model parameters

$$L = \begin{pmatrix} -1.7930 & 1.4312 & 0.1865 & 0.1753 \\ 1.4397 & -1.4519 & 0.0072 & 0.0050 \\ 0.0271 & 0.0517 & -0.1083 & 0.0295 \\ 0.0123 & 0.0095 & 0.0097 & -0.0315 \end{pmatrix}$$

$$\mu_1 = \begin{pmatrix} 1.0 \\ 2.0 \end{pmatrix} \mu_2 = \begin{pmatrix} 2.0 \\ 2.0 \end{pmatrix}$$

$$\mu_3 = \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix} \mu_4 = \begin{pmatrix} 2.0 \\ 1.0 \end{pmatrix}$$

$$\Sigma_1 = \begin{pmatrix} 0.5 & -0.3 \\ -0.3 & 0.5 \end{pmatrix} \Sigma_2 = \begin{pmatrix} 0.5 & 0.3 \\ 0.3 & 0.5 \end{pmatrix}$$

$$\Sigma_3 = \begin{pmatrix} 0.5 & 0.0 \\ 0.0 & 0.5 \end{pmatrix} \Sigma_4 = \begin{pmatrix} 0.5 & 0.0 \\ 0.0 & 0.5 \end{pmatrix}$$

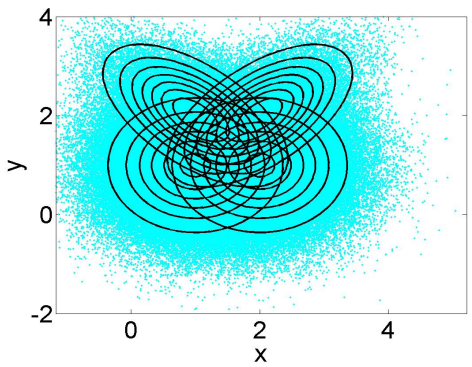

Figure 7.2: Original distributions. Two-dimensional data points generated from a hidden Markov model together with the output distributions for each of the four states as contour plots.

The EM algorithm requires as input the number of hidden states and an initial guess of the model parameters. The number of states – four – was assumed to be known, the initial guess for the HMM estimator was generated from the data, by drawing the forward-backward variables at random and run the estimation step once. After 174 iterations the EM converged with an accuracy of 1e-2. The estimated model is given below (see Fig. 7.3).

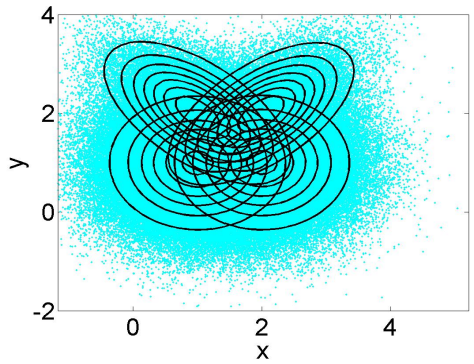

Figure 7.3: Estimated distributions. Two-dimensional data points generated from a hidden Markov model together with the estimated output distributions for each of the four states as contour plots.

Estimated model parameters

$$\hat{L} = \begin{pmatrix} -1.8540 & 1.4908 & 0.1836 & 0.1796 \\ 1.4947 & -1.5041 & 0.0094 & 0.0000 \\ 0.0276 & 0.0533 & -0.1081 & 0.0272 \\ 0.0138 & 0.0077 & 0.0101 & -0.0315 \end{pmatrix}$$

$$\hat{\mu}_1 = \begin{pmatrix} 1.0191 \\ 1.9736 \end{pmatrix} \hat{\mu}_2 = \begin{pmatrix} 1.9892 \\ 1.9904 \end{pmatrix}$$

$$\hat{\mu}_3 = \begin{pmatrix} 1.0043 \\ 1.0029 \end{pmatrix} \hat{\mu}_4 = \begin{pmatrix} 2.0024 \\ 1.0043 \end{pmatrix}$$

$$\hat{\Sigma}_1 = \begin{pmatrix} 0.5227 & -0.3195 \\ -0.3195 & 0.5291 \end{pmatrix} \hat{\Sigma}_2 = \begin{pmatrix} 0.5084 & 0.3105 \\ 0.3105 & 0.5058 \end{pmatrix}$$

$$\hat{\Sigma}_3 = \begin{pmatrix} 0.5069 & -0.0037 \\ -0.0037 & 0.5032 \end{pmatrix} \hat{\Sigma}_4 = \begin{pmatrix} 0.4998 & 6.0E-4 \\ 6.0E-4 & 0.5029 \end{pmatrix}$$

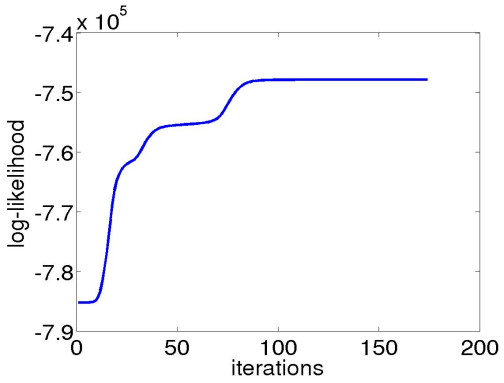

Figure 7.4: Evolution of the log-likelihood during the course of the EM-iterations.

The transition matrix has not been estimated, it is computed from the generator matrix as

$$\hat{P}(1) = \text{expm}(\hat{L}).$$

The proximity to the transition matrix based on the original generator is evident:

$$P(1) = \begin{pmatrix} 0.398 & 0.395 & 0.1041 & 0.1029 \\ 0.3951 & 0.4902 & 0.0585 & 0.0562 \\ 0.0298 & 0.0396 & 0.9003 & 0.0302 \\ 0.01 & 0.0099 & 0.01 & 0.9701 \end{pmatrix}$$

$$\hat{P}(1) = \text{expm}(\hat{L}) = \begin{pmatrix} 0.3962 & 0.3985 & 0.1024 & 0.1029 \\ 0.3974 & 0.4878 & 0.0598 & 0.0550 \\ 0.0307 & 0.0406 & 0.9006 & 0.0281 \\ 0.0102 & 0.0092 & 0.0104 & 0.9701 \end{pmatrix}.$$

Figure 7.4 almost reveals a saddle point in the likelihood. Due to this behavior it is important to choose the convergence threshold small enough. However, a comparison of the estimator with the original model shows a good correspondence. Hence the EM algorithm seems to have converged in the right (global) maximum of the likelihood. The output distributions are nearly identical (see Fig. 7.5). The estimators exhibit a good approximation

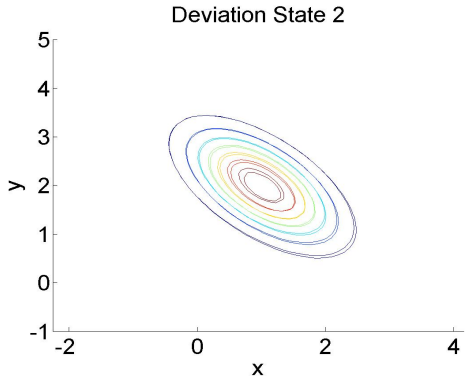

Figure 7.5: Deviation of the estimator from the original distribution – exemplarily for the second state.

to the original output parameters and the transition matrix as Table 7.1 shows. Alone the generator matrix reveals a noticeable deviation. Yet a closer look to the entrywise distances suggests a weighting of the estimators with the stationary distribution of the several states.

$$|L - \hat{L}| = \begin{pmatrix} 0.0610 & 0.0596 & 0.0029 & 0.0043 \\ 0.0550 & 0.0522 & 0.0022 & 0.0050 \\ 0.0005 & 0.0016 & 0.0002 & 0.0023 \\ 0.0015 & 0.0018 & 0.0004 & 0 \end{pmatrix}$$

$\|L - \hat{L}\|$	$\|P(1) - \hat{P}(1)\|$	$\|P(2) - \hat{P}(2)\|$	$\|P(3) - \hat{P}(3)\|$
0.1144	0.0056	0.0051	0.007
$\|\mu_1 - \hat{\mu}_1\|$	$\|\mu_2 - \hat{\mu}_2\|$	$\|\mu_3 - \hat{\mu}_3\|$	$\|\mu_4 - \hat{\mu}_4\|$
0.0066	0.0157	0.0058	0.0037
$\|\Sigma_1 - \hat{\Sigma}_1\|$	$\|\Sigma_2 - \hat{\Sigma}_2\|$	$\|\Sigma_3 - \hat{\Sigma}_3\|$	$\|\Sigma_4 - \hat{\Sigma}_4\|$
0.0133	0.0034	0.0077	9.7027e-04

Table 7.1: Unweighted distance of the estimated parameters from the original parameters in 2-norm.

$\|\rho(L - \hat{L})\|$	$\|\rho(P(1) - \hat{P}(1))\|$	$\|\rho(P(2) - \hat{P}(2))\|$	$\|\rho(P(3) - \hat{P}(3))\|$
0.0015	7.1703e-04	0.0012	0.0017
$\|\rho_1(\mu_1 - \hat{\mu}_1)\|$	$\|\rho_2(\mu_2 - \hat{\mu}_2)\|$	$\|\rho_3(\mu_3 - \hat{\mu}_3)\|$	$\|\rho_4(\mu_4 - \hat{\mu}_4)\|$
5.1886e-04	0.0014	0.0012	0.0024
$\|\rho_1(\Sigma_1 - \hat{\Sigma}_1)\|$	$\|\rho_2(\Sigma_2 - \hat{\Sigma}_2)\|$	$\|\rho_3(\Sigma_3 - \hat{\Sigma}_3)\|$	$\|\rho_4(\Sigma_4 - \hat{\Sigma}_4)\|$
0.0010	2.9915e-04	0.0015	6.1652e-04

Table 7.2: Distance in 2-norm of the estimated parameters from the original parameters weighted by the stationary distribution ρ of the Markov chain.

The larger deviations arise in rarely visited states. The stationary distribution of the Markov chain is

$$\rho = (0.0784, 0.0885, 0.1978, 0.6354).$$

The weighting with the stationary distribution is reasonable since for more frequently visited states more accurate estimators are available. Table 7.2 shows the weighted distances between the original and the estimated parameters. For comparison, we also compute the generator estimator from the path of the Markov chain states directly. With 312 EM iterations, an accuracy of $1e-2$ in the likelihood was achieved. The distance between estimator and original generator are similar to the HMM estimation (see Table 7.3).

Model without generator. In the following example we proceed just as described above but with a different original model. The difference to the foregoing example is that the transition matrix has one negative eigenvalue

	$\|L - \hat{L}\|$	$\|\rho(L - \hat{L})\|$
HMM Estimation	0.1144	0.0015
Observable Markov Chain	0.1054	0.0028

Table 7.3: Comparison between the generator estimator obtained by HMM estimation and by direct estimation from the state path.

and therefore has no underlying generator (cf. Corollary 2.1). The model is given below, the generated output data is shown in Figure 7.6.

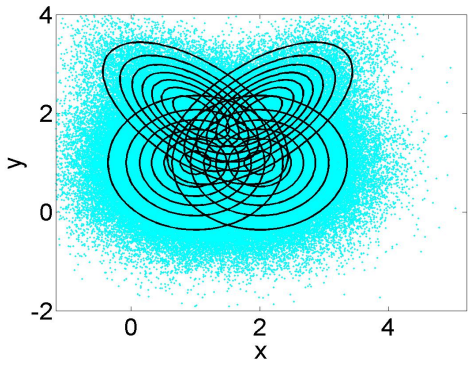

Figure 7.6: Original distributions. Two-dimensional data points generated from a hidden Markov model together with the output distributions for each of the four states as contour plots.

Original model parameters

$$P(1) = \begin{pmatrix} 0.0 & 0.4 & 0.6 & 0.0 \\ 0.5 & 0.0 & 0.0 & 0.5 \\ 0.03 & 0.07 & 0.9 & 0.0 \\ 0.08 & 0.02 & 0.0 & 0.9 \end{pmatrix}$$

$$\mu_1 = \begin{pmatrix} 1.0 \\ 2.0 \end{pmatrix} \mu_2 = \begin{pmatrix} 2.0 \\ 2.0 \end{pmatrix}$$

$$\mu_3 = \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix} \mu_4 = \begin{pmatrix} 2.0 \\ 1.0 \end{pmatrix}$$

$$\Sigma_1 = \begin{pmatrix} 0.5 & -0.3 \\ -0.3 & 0.5 \end{pmatrix} \Sigma_2 = \begin{pmatrix} 0.5 & 0.3 \\ 0.3 & 0.5 \end{pmatrix}$$

$$\Sigma_3 = \begin{pmatrix} 0.5 & 0.0 \\ 0.0 & 0.5 \end{pmatrix} \Sigma_4 = \begin{pmatrix} 0.5 & 0.0 \\ 0.0 & 0.5 \end{pmatrix}$$

After 272 EM iterations the likelihood has converged up to an accuracy of $1e-2$. Again a quasi saddle point was overcome as Figure 7.8 shows. But even if we iterate the EM algorithm up to an accuracy of $1e-7$, there is no big difference in the results. Thus we can assume that the EM algorithm has converged. The estimated model is given below (see also Fig. 7.7).

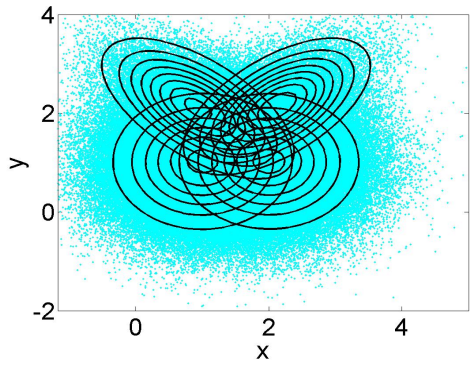

Figure 7.7: Estimated distributions. Two-dimensional data points generated from a hidden Markov model together with the estimated output distributions for each of the four states as contour plots.

Estimated model parameters

$$\hat{L} = \begin{pmatrix} -1.9476 & 1.0969 & 0.8506 & 0.0000 \\ 1.4270 & -2.1983 & 0.0000 & 0.7713 \\ 0.0003 & 0.1028 & -0.1031 & 0.0000 \\ 0.1094 & 0.0000 & 0.0000 & -0.1094 \end{pmatrix}$$

$$\hat{\mu}_1 = \begin{pmatrix} 0.9351 \\ 2.0973 \end{pmatrix} \hat{\mu}_2 = \begin{pmatrix} 2.1210 \\ 2.1376 \end{pmatrix}$$

$$\hat{\mu}_3 = \begin{pmatrix} 1.0047 \\ 1.0265 \end{pmatrix} \hat{\mu}_4 = \begin{pmatrix} 2.0038 \\ 1.0271 \end{pmatrix}$$

$$\hat{\Sigma}_1 = \begin{pmatrix} 0.4914 & -0.3007 \\ -0.3007 & 0.4767 \end{pmatrix} \hat{\Sigma}_2 = \begin{pmatrix} 0.4545 & 0.2652 \\ 0.2652 & 0.4418 \end{pmatrix}$$

$$\hat{\Sigma}_3 = \begin{pmatrix} 0.4899 & 0.0038 \\ 0.0038 & 0.5134 \end{pmatrix} \hat{\Sigma}_4 = \begin{pmatrix} 0.4917 & -0.0034 \\ -0.0034 & 0.5073 \end{pmatrix}$$

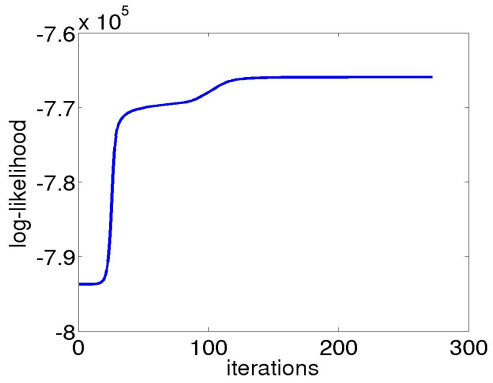

Figure 7.8: Evolution of the log-likelihood in the course of the EM-iterations.

$\|\rho(P(1)) - \hat{P}(1))\|$	$\|\rho(\mu_1 - \hat{\mu}_1)\|$	$\|\rho(\mu_2 - \hat{\mu}_2)\|$	$\|\rho(\mu_3 - \hat{\mu}_3)\|$	$\|\rho(\mu_4 - \hat{\mu}_4)\|$
0.0231	0.0094	0.0134	0.013	0.01
$\|\rho(\Sigma_1 - \hat{\Sigma}_1)\|$	$\|\rho(\Sigma_2 - \hat{\Sigma}_2)\|$	$\|\rho(\Sigma_3 - \hat{\Sigma}_3)\|$	$\|\rho(\Sigma_4 - \hat{\Sigma}_4)\|$	
0.0019	0.0064	0.0067	0.0033	

Table 7.4: Distance in 2-norm of the estimated parameters from the original parameters weighted by the stationary distribution ρ of the Markov chain.

	$\|\rho(P(1)) - \text{expm}(\hat{L}))\|$	$\|\rho(\tilde{L} - \hat{L})\|$
HMM Estimation	0.0231	0.0258
Observable Markov Chain	0.0053	0.0039

Table 7.5: Comparison between the generator estimator obtained from the HMM estimation and from direct estimation from the state path. For comparison was used the transition matrix and the nearest existing generator to the spectrum of $\text{logm}(P(1))$. This generator is denoted by $\tilde{L}$.

In this example the distance between the original transition matrix and the estimated one is per construction greater than that in the previous example since here for the original model does not even exist a generator. The estimated transition matrix

$$\hat{P}(1) = \text{expm}(\hat{L}) = \begin{pmatrix} 0.2689 & 0.1927 & 0.4071 & 0.1312 \\ 0.2462 & 0.2279 & 0.1871 & 0.3388 \\ 0.0228 & 0.0454 & 0.9104 & 0.0215 \\ 0.0522 & 0.0186 & 0.0267 & 0.9025 \end{pmatrix}$$

has an underlying generator and therefore necessarily another structure. But the changes mainly affect the zeros in the original matrix and hence states with a minor part of the stationary distribution. Thus, in Table 7.4 we only compare the weighted distances.

Finally we compare the generator estimate from the HMM approach with that one estimated from the observable state path. Since no original generator is available, we compute following Crommelin and Vanden-Eijnden [18] that one, which is closest in a spectral sense to the logarithm of the transition matrix. This generator will be denoted with $\tilde{L}$.

In Table 7.5 it becomes apparent that the generator that was estimated directly from the state path approximates $\tilde{L}$, as expected, better than the HMM estimator. The influence of the twofold estimation step accounts for a decimal place in accuracy. But the main structure of the transition matrix (two metastable states and two conductive states) is still reproduced correctly.

7.2 Markov Jump Output Process

At last we will define an HMM with a kinetic output process, which is a Markov jump process. In contrast to HMM with Gaussian output, we define here an HMM with kinetic patterns that are encoded also in the output parameters. The structure is similar to HMMSDE, but here the output process has a discrete state space. To apply this on continuous (possibly even multivariate) data, at first a box-discretization as described in several publications [23, 24, 21, 67] is carried out: The state space is divided in boxes and each (multi-dimensional) data point is assigned to a box and the resulting time series reflects the transition behavior between the particular boxes. Denote the discrete observation space by B, the Markov jump process takes values from B

$$X(t) \in B, \forall t \in [0, T].$$

The discrete observable data points we denote by $O_m = X(t_m) \in B$, for $m = 0, ..., M$, the states of the hidden Markov chain by Y_m, for $m = 0, ..., M$. The model structure is illustrated in Figure 7.9. The EM algorithm is introduced

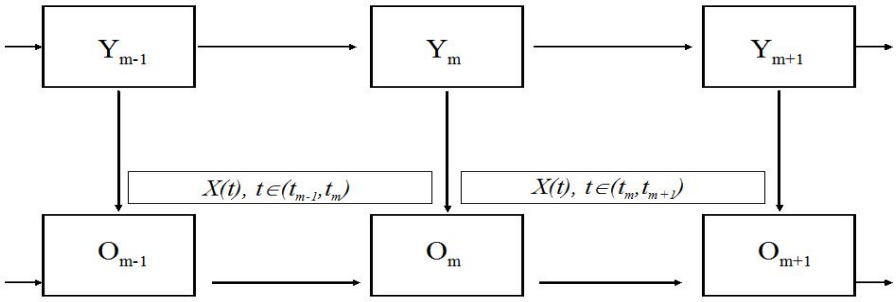

Figure 7.9: Hidden Markov model with Markov jump output process. The observable data points are $O_0..., O_{m-1}, O_m, O_{m+1}, ..., O_M$. The continuous process between these data points $X(t)$ is hidden as well as the states $Y_0, ..., Y_{m-1}, Y_m, Y_{m-1}, ..., Y_M$.

again following the scheme from Section 3.1.

Later it is applied in several examples. In a first example we consider a time series generated by an HMM with varying time lags. In a second and third example we restrict ourselves to constant time lags and investigate time series exhibiting strong metastability. In the second example we reconsider the generator matrix from Section 6.5.2 and compare the direct generator estimation to the HMM-MJP approach. In the third example we estimate an HMM-MJP from a time series generated by Smoluchowski dynamics like in Section 5.5. Finally, the results are compared to the HMMSDE approach.

7.2.1 Model Design

For the application of an HMM with Gaussian output or HMMSDE we have to make certain assumptions on the data, i.e. the data is supposed to be normal distributed. If this is not the case, miscellaneous problems can occur during the estimation procedure. The advantage of HMMs with Markov jump output (HMM-MJP) is that we do not have to specify a distribution of the observed data. Only the transition behavior between the discretization boxes is relevant for the likelihood.

The output process is completely parameterized by its generator matrix L^{out} and an initial distribution φ_0 for each hidden state. The entire HMM-MJP is hence specified by the model parameters

$$\lambda = (\pi, P, (\varphi_{0_i})_{i \in S}, (L_i^{out})_{i \in S}). \tag{7.7}$$

7.2.2 Likelihood

Let $\varphi(O_m | O_{m-1}, Y_m = i)$ denote the likelihood of the observed process. The discrete likelihood has the simple form

$$\bar{\varphi}(O_m | O_{m-1}, Y_m = i) = \operatorname{expm}((t_m - t_{m-1}) L_i^{out})_{O_{m-1} O_m}. \tag{7.8}$$

Yet for aforementioned reasons, the discrete likelihood is not feasible for maximization (cf. Section 6.1). Therefore the continuous likelihood (6.10) will be used

$$\begin{aligned} &\varphi(O([t_{m-1}, t_m]) | O_{m-1}, Y_m = i) = \\ &\quad \prod_{\substack{k,l \in S \\ l \neq k}} ((L_i^{out})_{kl})^{N_{kl}(t_m - t_{m-1})} \exp\left(-((L_i^{out})_{kl}) R_k (t_m - t_{m-1})\right). \end{aligned} \tag{7.9}$$

Here $X([0,T])$ denotes the continuous process which is observed only at discrete time points $O_m = X(t_m), m = 0, ..., M$. Now we can express the joint likelihood of the continuous process X (observed at discrete points) and the hidden data Y by the model parameters defined in (7.7):

$$\mathbb{P}(X, Y | \lambda) \quad = \quad \pi(Y_0) \varphi_0(O_0 | Y_0) \prod_{m=1}^{M} P(Y_{m-1}, Y_m) \varphi(O([t_{m-1}, t_m]) | O_{m-1}, Y_m),$$

where φ_0 denotes an initial distribution which for each $Y_0 \in S$ simply is

$$\varphi_0(x) = \begin{cases} 1, & x = O_0 \\ 0 & \text{else} \end{cases}, \ x \in B.$$

7.2.3 Partial Observability

Expectation step. Additional to the hidden process Y most of the continuous process X is not observable. Thus we have to build the expectation over Y and X. As Figure 7.9 illustrates the data points O_m solely are observable, the continuous data X in between are hidden as well as the hidden states Y are hidden. For this purpose we combine the results from (3.9) and (6.12) and obtain the expected likelihood

$$\begin{aligned} Q(\lambda, \lambda_0) = \sum_{i \in S} & \Big((\log[\pi_i] + \log[\varphi_0(O_0|Y_0)]) \ \alpha_i(t_0)\beta_i(t_0) + \\ \sum_{m=1}^{M} & \log\Big[\mathbb{E}[\varphi(O([t_{m-1}, t_m])|Y_m = i)|O_{m-1}, O_m, L_{0_i}^{out}]\Big] \ \alpha_i(t_m)\beta_i(t_m)\Big) + \\ \sum_{i,j \in S} & \sum_{m=1}^{M-1} \log[P_{ij}] \ \alpha_i(t_m) \ (P_0)_{ij} \ \bar{\varphi}(O_{m+1}|O_m, Y_{m+1} = j) \ \beta_i(t_m), \end{aligned}$$

with

$$\begin{aligned} & \log\Big[\mathbb{E}\left[\varphi(O([t_{m-1}, t_m])|Y_m = i)|O_{m-1}, O_m, L_{0_i}^{out}\right]\Big] \\ = & \sum_{k=1}^{d} \sum_{l \neq k} \log[L_{i_{kl}}^{out}] \mathbb{E}\left[N_{i_{kl}}([t_{m-1}, t_m])|O_{m-1}, O_m, L_{0_i}^{out}\right] \\ - & \sum_{k=1}^{d} \sum_{l \neq k} L_{i_{kl}}^{out} \mathbb{E}\left[R_{i_k}([t_{m-1}, t_m])|O_{m-1}, O_m, L_{0_i}^{out}\right]. \end{aligned} \tag{7.10}$$

Maximization step. Again the maximum likelihood estimators can be obtained by the partial derivatives with respect to the parameters π, P and L_i^{out}. The initial distribution of the output process φ_0 is negligible since it is already determined by O_0. While π and P follow from the standard Baum-Welch formulas (3.10), we still have to derive L_i^{out}. The relevant term in Q is

$$\sum_{m=1}^{M} \log\Big[\mathbb{E}[\varphi(O([t_{m-1}, t_m])|Y_m = i)|O_{m-1}, O_m, L_{0_i}^{out}]\Big] \alpha_i(t_m)\beta_i(t_m).$$

Note that by homogeneity of the Markov process, the expected likelihood does not depend on $O([t_{m-1}, t_m])$, since O_{m-1} and O_m are given, but only on the length of the time interval $t_m - t_{m-1}$. In case of equidistant time steps $t_m - t_{m-1} = \tau$ this term simplifies to

$$\sum_{x,y \in S} \log\Big[\mathbb{E}[\varphi(\tau, i)|x, y, L_{0_i}^{out}]\Big] C_i^{xy}, \tag{7.11}$$

where C_i is the frequency matrix

$$C_i^{xy} = \sum_{m=1}^{M} \mathbb{1}_{\{O_m=x, O_{m+1}=y\}} \alpha_i(t_m)\beta_i(t_m).$$

For non-equidistant time steps a frequency matrix for each different time-lag τ_k has to be computed separately and instead of (7.11) we use the expression

$$\sum_{k=1}^{n} \sum_{x,y \in S} \log \left[\mathbb{E}[\varphi(\tau_k, i)|x, y, L_{0_i}^{out}] \right] C_i^{xy}(\tau_k).$$

Remark 7.1. To handle a time series with non-equidistant time steps we modify the hidden process as described in Section 7.1 in more detail.

Using (7.10) in (7.11) and deriving it with respect to $L_{i_{kl}}^{out},\ k \neq l$ yields the reestimation formula:

$$L_{i_{kl}}^{out} = \frac{\sum_{x,y \in S} \mathbb{E}\left[N_{i_{kl}}([\tau])|x, y, L_{0_i}^{out}\right] C_i^{xy}}{\sum_{x,y \in S} \mathbb{E}\left[R_{i_k}([\tau])|x, y, L_{0_i}^{out}\right] C_i^{xy}}.$$

The expectations of N and R can be computed according to (6.18). The diagonal entries of $L_{i_{kl}}^{out}$ are determined by the generator property (2.7). Finally, we state the Algorithm 7.2 in pseudocode.

7.2.4 Example: Recovering an HMM-MJP from a Realization with Varying Time Lag

In this example we consider model parameters estimated from a time series which in turn has been generated from an HMM. We investigate two cases. In the first case the observation space can be separated clearly in two different regions. In the second case no spatial separation is possible. For both hidden states the output process stays in the same region.

Case 1: separate regions. The original model has two hidden states and a 10-dimensional observation state space. As in the previous example from Section 7.1.3 the time lag was chosen randomly: $\tau = 0.1$ with probability 0.6 and $\tau = 0.2$ respectively $\tau = 0.3$ either with probability 0.2. The generator of the hidden process was chosen as

$$L^{hidden} = \begin{pmatrix} -0.01 & 0.01 \\ 0.02 & -0.02 \end{pmatrix}.$$

The generators of the output Markov process have the shape

$$L_1^{out} = \begin{pmatrix} L1 & A \\ B & C \end{pmatrix} \qquad L_2^{out} = \begin{pmatrix} C & B \\ A & L2 \end{pmatrix}$$

Algorithm 7.2 EM algorithm for an HMM with Markov jump output process.

Require: Time series $O_0, \ldots, O_M$,
initial guess of parameters $\lambda_0 = (\pi, P, (L_i^{out})_{i \in S})$, accuracy ε.
(1) Set $\tilde{\lambda} := \lambda_0$.
(2) Compute observation likelihood $\bar{\varphi}(O_m | O_{m-1}, Y_m = i)$ via (7.8) and forward-backward variables $\alpha_i(t_m), \beta_i(t_m)$, for $i \in S, m = 1, .., M$ via (3.2).
(3) Compute for each state i a frequency matrix
$C_i^{xy} = \sum_{m=1}^{M} \mathbb{1}_{\{O_m = x, O_{m+1} = y\}} \alpha_i(t_m) \beta_i(t_m)$
(4) M-Step: Reestimate model parameters:
Estimate π, P via (3.10).
Estimate output parameters via L_i^{out} via Algorithm 6.1 with input C_i,
for each $i \in S$.
if $\mathcal{L}(\tilde{\lambda}) - \mathcal{L}(\lambda_0) > \varepsilon$ **then**
$\lambda_0 = \tilde{\lambda}$
Go to Step (2).
else
return $\tilde{\lambda}$.
end if

with

$$A = \begin{pmatrix} 0 & \ldots & 0 \\ \vdots & & \vdots \\ 0 & \ldots & 0 \end{pmatrix} B = \begin{pmatrix} 0.1 & \ldots & 0.1 \\ \vdots & & \vdots \\ 0.1 & \ldots & 0.1 \end{pmatrix}$$

$$C = \begin{pmatrix} -0.5 & 0 & 0 & 0 & 0 \\ 0.1 & -0.6 & 0 & 0 & 0 \\ 0.1 & 0.1 & -0.7 & 0 & 0 \\ 0.1 & 0.1 & 0.1 & -0.8 & 0 \\ 0.1 & 0.1 & 0.1 & 0.1 & -0.9 \end{pmatrix}$$

$$L1 = \begin{pmatrix} -13 & 5 & 2 & 2 & 4 \\ 2 & -11 & 2 & 4 & 3 \\ 2 & 16 & -23 & 2 & 3 \\ 5 & 2 & 2 & -11 & 2 \\ 3 & 4 & 3 & 3 & -13 \end{pmatrix}$$

$$L2 = \begin{pmatrix} -59 & 16 & 16 & 3 & 24 \\ 3 & -10 & 3 & 2 & 2 \\ 3 & 3 & -10 & 2 & 2 \\ 2 & 4 & 2 & -10 & 2 \\ 9 & 7 & 2 & 3 & -21 \end{pmatrix}.$$

Figure 7.10 shows the resulting time series as realization of the described model. In each state a different region of the observation space was visited. Next, we estimate from this time series model parameters by means of the

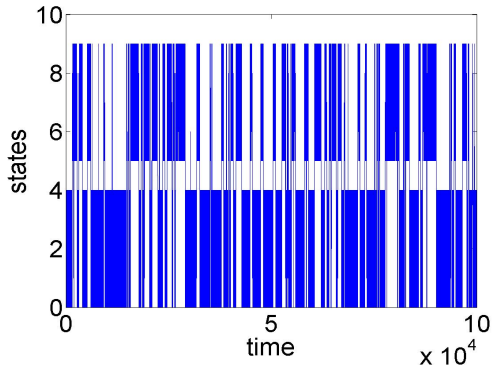

Figure 7.10: Realization of the original model.

EM algorithm. The initial parameters were generated at random. After 552 iterations the EM algorithm converged up to an accuracy of $1e-3$. The resulting estimator for the generator of the hidden process is

$$\hat{L}_{hidden} = \begin{pmatrix} -0.0126 & 0.0126 \\ 0.0199 & -0.0199 \end{pmatrix}.$$

The submatrices of the output generator for the first state

$$\hat{L}_1^{out} = \begin{pmatrix} \hat{L1} & \hat{A} \\ \hat{B} & \hat{C} \end{pmatrix}$$

were estimated as follows:

$$\hat{L1} = \begin{pmatrix} -13.3711 & 5.0817 & 2.0166 & 2.0615 & 4.2114 \\ 1.7908 & -10.9387 & 2.1680 & 4.1706 & 2.8093 \\ 2.6818 & 16.6118 & -23.0632 & 1.6711 & 2.0985 \\ 5.3415 & 2.2331 & 1.6901 & -11.2154 & 1.9506 \\ 3.1645 & 3.4948 & 2.7751 & 2.8445 & -12.2789 \end{pmatrix},$$

the zero matrix A was recovered correctly

$$\hat{A} = \begin{pmatrix} 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{pmatrix},$$

whereas $\hat{B}$ and $\hat{C}$ exhibit higher deviations. However, the lower triangular structure of C was mostly reflected, except from line 2. The matrices $\hat{B}$

and $\hat{C}$ express the transition behavior between two metastable sets only. Accordingly for these states the least statistical data is available.

$$\hat{B} = \begin{pmatrix} 0.1808 & 0.0908 & 0 & 0.5621 & 0 \\ 0 & 0.6027 & 0 & 0 & 0.0877 \\ 0 & 0 & 0.1256 & 0.0283 & 0.2886 \\ 0 & 0.4463 & 0 & 0 & 0 \\ 0.0424 & 0 & 0.0039 & 0.1919 & 0 \end{pmatrix}$$

$$\hat{C} = \begin{pmatrix} -0.8337 & 0 & 0 & 0 & 0 \\ 0.1190 & -0.8786 & \boxed{0.0379} & \boxed{0.0313} & 0 \\ 0.1156 & 0.1425 & -0.7007 & 0 & 0 \\ 0.1377 & 0 & 0.2120 & -0.7960 & 0 \\ 0.0801 & 0 & 0.1666 & 0 & -0.4849 \end{pmatrix}.$$

For the second state the following parameters were estimated:

$$\hat{L}_2^{out} = \begin{pmatrix} \hat{L2} & \hat{A} \\ \hat{B} & \hat{C} \end{pmatrix}$$

with

$$\hat{L2} = \begin{pmatrix} -39.3496 & 9.2831 & 11.5246 & 5.2381 & 13.3038 \\ 2.4133 & -9.7114 & 2.9157 & 2.1413 & 2.2410 \\ 1.7389 & 3.3495 & -9.4933 & 2.0471 & 2.3577 \\ 1.0729 & 4.1116 & 2.4152 & -10.4174 & 2.8178 \\ 6.1767 & 8.6161 & 2.6222 & 2.0193 & -19.4344 \end{pmatrix}$$

$$\hat{A} = \begin{pmatrix} 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{pmatrix}$$

$$\hat{C} = \begin{pmatrix} -0.4843 & 0 & 0 & 0 & 0 \\ 0.0799 & -0.5949 & 0 & 0 & \boxed{0.0086} \\ 0 & 0.3489 & -0.9198 & 0 & 0 \\ 0.1530 & 0.1368 & 0.2046 & -0.9887 & 0 \\ 0.1867 & 0.2791 & 0.0739 & 0 & -1.0912 \end{pmatrix}$$

$$\hat{B} = \begin{pmatrix} 0 & 0.3214 & 0.0963 & 0.0667 & 0 \\ 0 & 0 & 0.2451 & 0.0027 & 0.2586 \\ 0 & 0.5709 & 0 & 0 & 0 \\ 0 & 0.3245 & 0 & 0 & 0.1698 \\ 0 & 0.1106 & 0.3591 & 0.0818 & 0 \end{pmatrix}.$$

The zero matrix A was recovered correctly again, the matching of $\hat{B}$ and $\hat{C}$ is also comparable to the first state.

The lower triangular structure of C was almost preserved, all entries upside the diagonal but the last entry of the second row are zero. Altogether

the estimated parameters reflect the original parameters satisfactory. The essential dynamical behavior within a metastable state is expressed in the blocks L1 and L2. We compare the generator matrices for both states as

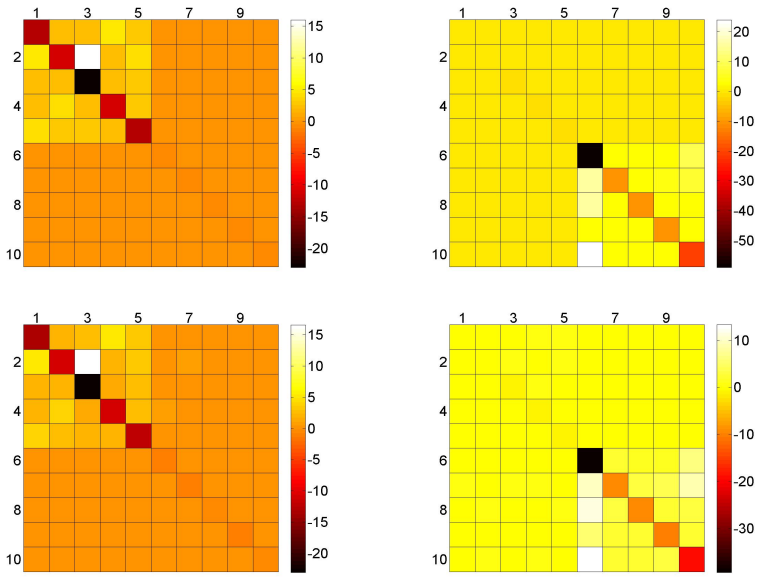

Figure 7.11: Original generator matrix for state 1 of the output process (left top) for state 2 of the output process(right top). Estimated generator matrices for state 1 (left bottom) and for state 2 (right bottom).

boxplots in Figure 7.11, which mainly reflect the occupation structure of L1 resp. L2. The highest distance from the original parameters is exhibited by the first row of $\hat{L}2$. A closer look on the histogram explains the bad approximation since state 5 was visited least. However, the propagator

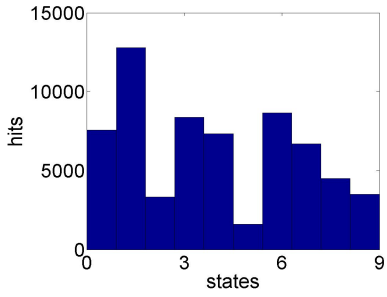

Figure 7.12: Histogram of the time series from Figure 7.10. State 5 (the first state of the second metastable set) exhibits the fewest hits.

matrices are still fitted well to the data. Let denote

$$P_i(\tau) = \text{expm}(\tau L_i^{out})$$

τ	0.1	0.2	0.3
$\|P_1(\tau) - \hat{P}_1(\tau)\|$	0.0458	0.0816	0.1143
$\|P_2(\tau) - \hat{P}_2(\tau)\|$	0.0497	0.0710	0.0947
$\max_{kl}\{\|P_1(\tau)_{kl} - \hat{P}_1(\tau)_{kl}\|\}$	0.0387	0.0723	0.1012
$\max_{kl}\{\|P_2(\tau)_{kl} - \hat{P}_2(\tau)_{kl}\|\}$	0.0235	0.0424	0.0587

Table 7.6: Distance of the estimated propagator from the original component-wise and norm-wise.

and

$$\hat{P}_i(\tau) = \text{expm}(\tau \hat{L}_i^{out}).$$

Table 7.6 compares the distances in terms of the Euclidian norm $\|P_i(\tau) - \hat{P}_i(\tau)\|$ and maximal component-wise distance $\max_{kl}\{|P_i(\tau)_{kl} - \hat{P}_i(\tau)_{kl}|\}$. Although the second generator matrix deviates most from the original parameters, the components with high deviance do not carry much weight. The distances of the propagator have the same order of magnitude. The second propagator was even slightly better approximated than the first one.

Finally, we compare the assignment of the hidden states based on the estimated model parameters – the Viterbi path – with the hidden Markov chain of the realization from the original parameters in Figure 7.13. The Viterbi path mostly corresponds to the original Markov chain, which is a further indicator of the reliability of the estimated parameters. But after

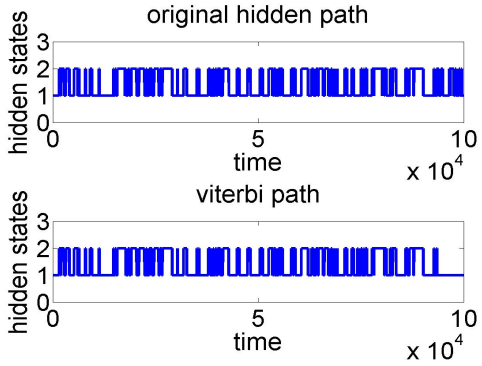

Figure 7.13: Realization of the hidden Markov process based on the original model and Viterbi path based on the estimated model.

all the occurrence of zero-entries in the generator matrix can easily lead to numerical difficulties as pointed out in Section 6.5.2.

Case 2: overlapping regions. In the previous example was considered a time series, in which for both states different regions of the observation space $\{0, ..., 4\}$ respectively $\{5, ..., 9\}$ have been visited. But even if these regions

τ	0.1	0.2	0.3
$\|P_1(\tau) - \hat{P}_1(\tau)\|$	0.0184	0.0117	0.0114
$\|P_2(\tau) - \hat{P}_2(\tau)\|$	0.0293	0.0153	0.0161
$\max_{kl}\{\|P_1(\tau)_{kl} - \hat{P}_1(\tau)_{kl}\|\}$	0.0143	0.0068	0.0050
$\max_{kl}\{\|P_2(\tau)_{kl} - \hat{P}_2(\tau)_{kl}\|\}$	0.0183	0.0093	0.0075

Table 7.7: Distance of the estimated propagator from the original component-wise and norm-wise.

are totally overlapping, similar results have been achieved. Consider the same model as above, with the modified generators for the output process

$$L_1^{out} = (L1) \qquad L_2^{out} = (L2)\,,$$

with the matrices $L1$, $L2$ as defined above. Here, the effective dynamics for both states happen in the whole observation state space $B = \{0, ..., 4\}$. However, the HMM-MJP analysis produced similar results as in the first case:

$$\begin{aligned}
\hat{L}_{hidden} &= \begin{pmatrix} -0.0115 & 0.0115 \\ 0.0215 & -0.0215 \end{pmatrix} \\
\hat{L}1 &= \begin{pmatrix} -12.7709 & 5.0287 & 1.9135 & 2.0687 & 3.7599 \\ 1.9081 & -10.9242 & 1.7867 & 3.9065 & 3.3228 \\ 2.2971 & 14.5782 & -21.8340 & 2.3274 & 2.6312 \\ 5.3409 & 2.2011 & 1.8514 & -11.2902 & 1.8969 \\ 2.7348 & 4.1294 & 2.9082 & 3.2056 & -12.9779 \end{pmatrix} \\
\hat{L}2 &= \begin{pmatrix} -37.4506 & 11.9836 & 10.5426 & 2.3710 & 12.5535 \\ 2.1157 & -9.3554 & 3.1659 & 2.0046 & 2.0691 \\ 1.8149 & 3.5233 & -9.8443 & 1.9437 & 2.5624 \\ 1.8365 & 3.8118 & 1.7382 & -9.7991 & 2.4125 \\ 4.3896 & 6.6814 & 3.4167 & 3.3046 & -17.7921 \end{pmatrix}.
\end{aligned}$$

Again the occupation structure of the matrices $\hat{L}1$ and $\hat{L}2$ are comparable to the first example and the poorest fit was achieved at the first row of matrix $\hat{L}2$. But the EM algorithm took 452 iterations only to converge with an accuracy of $1e-3$ and recurrent executions showed it was less vulnerable to numerical difficulties since the generator matrices did not contain entries that are close to zero. Also the algorithm did not come up with local maxima, which happened in the previous example several times.

We again compare the distances of the original and the estimated matrices $L1$ and $L2$ in Table 7.7 and observe that the match is clearly better than in the first example. Also for increasing τ the distance remains small since the propagator is already close to the equilibrium, albeit far enough to fulfill $\det(P_i(\tau) > 0)$.

Eventually we come to the conclusion: in case of strong metastability in the hidden process, parameters can be identified even if the state space for the particular metastable states are largely overlapping. At last the comparison of the original hidden path with the estimated Viterbi path ratifies the quality of the model. Except for a few differences they reveal a high agreement (cf. Figure 7.14).

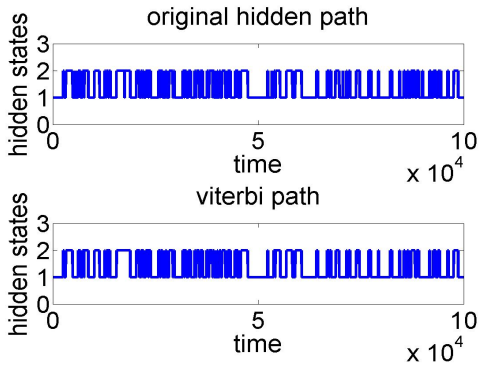

Figure 7.14: Realization of the hidden Markov process based on the original model with overlapping metastable regions and Viterbi path based on the respective estimated model.

7.2.5 Example: A Metastable Generator Revisited

For the next example we reconsider the generator from Section 6.5.2. The quasi zero entries beyond the metastable blocks led to numerical difficulties (c.f. Section 6.5.2). At this point we wish to design a more stable model by the means of HMMs. The output parameters consist of smaller generator matrices corresponding to the metastable blocks and the transition behavior between the metastable blocks is encoded in the hidden process. However, the problem of this model design is the dependency of the output likelihood on the previous time step

$$\varphi(O_m|O_{m-1}, Y_m = i).$$

If a transition between two metastable sets say $M_1 = \{1, 2, 3\}$ and $M_2 = \{4, 5, 6\}$ occurs, the generator matrices describing the output process have to be defined on the entire state space rather than only on one metastable set.

Our initial assumption was that an HMM that describes the data from

example 6.5.2 best, would have the following shape:

$$
\begin{aligned}
L^{hidden} &= \begin{pmatrix} -0.0001 & 0.0001 \\ 0.0001 & -0.0001 \end{pmatrix} \\
L_1^{out} &= \begin{pmatrix} -0.9426 & 0.4860 & 0.4565 & 0.0001 & 0 & 0 \\ 0.2311 & -0.2497 & 0.0185 & 0 & 0.0001 & 0 \\ 0.6068 & 0.7621 & -1.3690 & 0 & 0 & 0.0001 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \\
L_2^{out} &= \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0.0001 & 0 & 0 & -1.3276 & 0.9218 & 0.4057 \\ 0 & 0.0001 & 0 & 0.6154 & -1.5510 & 0.9355 \\ 0 & 0 & 0.0001 & 0.7919 & 0.1763 & -0.9683 \end{pmatrix}.
\end{aligned}
$$

But the EM algorithm with an a priori model initialized at random yields the following model:

$$
\begin{aligned}
L^{hidden} &= \begin{pmatrix} -0.3529 & 0.3529 \\ 0.5393 & -0.5393 \end{pmatrix} \\
L_1^{out} &= \begin{pmatrix} -1.1788 & 0.6330 & 0.5452 & 0.0006 & 0 & 0 \\ 0.2206 & -0.2397 & 0.0191 & 0 & 0 & 0 \\ 0.1737 & 1.2769 & -1.4511 & 0 & 0.0005 & 0 \\ 0 & 0.0005 & 0 & -1.2662 & 0.9016 & 0.3641 \\ 0 & 0.0000 & 0 & 0.7469 & -1.7590 & 1.0122 \\ 0 & 0.0001 & 0 & 1.2459 & 0.2522 & -1.4982 \end{pmatrix} \\
L_2^{out} &= \begin{pmatrix} -0.7174 & 0.3517 & 0.3657 & 0 & 0 & 0 \\ 0.2764 & -0.2884 & 0.0118 & 0.0001 & 0 & 0.0002 \\ 1.0210 & 0.1501 & -1.1711 & 0 & 0 & 0 \\ 0.0002 & 0 & 0 & -1.4572 & 0.9249 & 0.5322 \\ 0 & 0 & 0 & 0.2180 & -1.2489 & 1.0309 \\ 0 & 0 & 0 & 0.4417 & 0.1100 & -0.5517 \end{pmatrix}.
\end{aligned}
$$

The generator of the hidden process is highly mixing, and the output generators are a rather arbitrary split of the original generator matrix with $\frac{1}{2}(L_1^{out} + L_2^{out}) \approx L$. Figure 7.15 illustrates the distance of the arithmetic mean of both output generators from the original generator.

But what about the likelihood? It turns out that the likelihood of the second model is indeed higher than the one of the expected model. To point it out we perform another EM algorithm with our expectation as a priori model parameter. After 37 iterations the EM converged up to an accuracy of $1e-3$ and the likelihood was still below the likelihood of the second

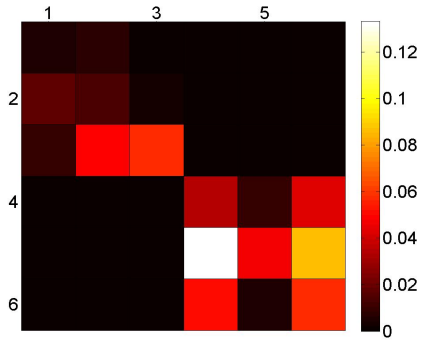

Figure 7.15: Distance between the original generator and the arithmetic mean of the HMM output generators $|L - \frac{1}{2}(L_1^{out} + L_2^{out})|$.

model. During the 37 iterations, only a slight adjustment of the a priori parameters happened. As a third test we execute a further EM algorithm, this time with initial parameters that were computed from the time series with a predetermined hidden path. This path is easy to determine in our example since both metastable regions are completely disjoint.

$$Y(m) = \begin{cases} 1 & \text{if } O(m) \leq 3 \\ 2 & \text{else} \end{cases}$$

We finally set the weights $\alpha_i(m)\beta_i(m)$ to $\mathbb{1}_{\{Y_m=i\}}$ and compute the reestimation formulas according to step (3) and (4) in algorithm (7.2). With the resulting initial parameters we obtain another model, that recognizes the particular metastable sets as different hidden states:

$$L^{hidden} = 1.0e-003 * \begin{pmatrix} -0.1560 & 0.1560 \\ 0.1452 & -0.1452 \end{pmatrix}$$

$$L_1^{out} = \begin{pmatrix} -1.8470 & 0 & 0 & 1.8470 & 0.0000 & 0 \\ 0 & -4.8759 & 0 & 0.8003 & 0 & 4.0756 \\ 0 & 0 & -2.7548 & 0 & 2.7548 & 0 \\ 0 & 0 & 0 & -1.2848 & 0.8931 & 0.3917 \\ 0 & 0 & 0 & 0.5789 & -1.5414 & 0.9625 \\ 0 & 0 & 0 & 0.7742 & 0.1750 & -0.9493 \end{pmatrix}$$

$$L_2^{out} = \begin{pmatrix} -0.9480 & 0.4975 & 0.4505 & 0 & 0 & 0 \\ 0.2359 & -0.2537 & 0.0178 & 0 & 0 & 0 \\ 0.6041 & 0.7389 & -1.3430 & 0 & 0 & 0 \\ 1.0282 & 5.2101 & 0 & -6.2383 & 0 & 0 \\ 0 & 0.5528 & 0 & 0 & -0.5528 & 0 \\ 0 & 5.1975 & 0 & 0 & 0 & -5.1975 \end{pmatrix}.$$

The metastability is recovered by the generator of the hidden process and the metastable blocks enter each in a different output generator. A comparison

	A priori model	Likelihood	iterations
λ_1	expected parameters	-85957.968	37
λ_2	random	-85894.674	1712
λ_3	predetermined hidden path	-85904.455	115

Table 7.8: Likelihood and speed of convergence of the models obtained by EM algorithm with different a priori parameters.

of the likelihood is to be found in Table 7.8. The likelihood is maximized by the second model. Both other models converged much sooner, but this is due to the fact, that we have put some knowledge into the initial parameters. However, each of the three models is a maximum in the likelihood landscape. To illustrate this we consider a one-dimensional projection of the parameter space along the straight line between two parameter tuples. Figure 7.16 shows the likelihood conditioned on the parameters along the line between two EM results λ_i and λ_j at the points

$$\lambda_i + k\frac{\lambda_j - \lambda_i}{n},$$

for $k = -20, ..., n + 100$ and $n = 100$. The kinks in the figures are due to the generator condition that has to be obeyed. In some cases negative off-diagonal entries for $k < 0$ or $k > n$ occurred, those have been set to zero. So every kink in the likelihood landscape corresponds to a kink in the parameter space. Overall, the likelihood has a complicated shape. The EM

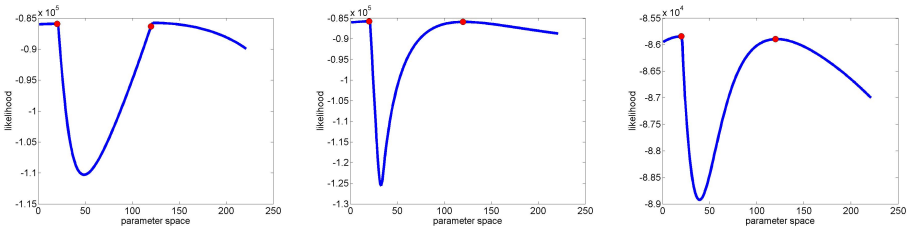

Figure 7.16: Likelihood landscape projected on the line $\lambda_2+\eta(\lambda_3-\lambda_2)$ (left), $\lambda_1+\eta(\lambda_3-\lambda_1)$ (middle) and $\lambda_1+\eta(\lambda_2-\lambda_1)$ (right), with $\eta \in [-0.2, 2]$. The red dots mark the likelihood at the points λ_1, λ_2 and λ_3.

algorithm tends to return local maxima. This phenomenon is not surprising as the specification of the output process in terms of generator matrices implies a large number of parameters. Yet, the more structure goes into the model, the more complex becomes the likelihood structure and hence the number of extrema grows.

However, the resulting parameters λ_3 with the predetermined hidden path are able to express the main features of the time series. It reflects the dynamics within the particular metastable states as well as the transition behavior between these states in terms of L^{hidden} and L^{out}. Note, that the

columns of L^{out} beyond the metastable blocks are zero (except from the diagonal entries). This structure ensures that after a hidden state change the time series moves towards the appropriate metastable set.

State space restriction. The success of an EM parameter estimation depends on the nature of the likelihood landscape. The problems in the present example are due to a number of local maxima on the same altitude. It is not trivial to decide, which model provides the best description of the time series. Surely the dynamical behavior and the metastable sets should be expressed by the estimated model. This is not taken into account by the EM algorithm, if the likelihood does not reflect the dynamics properly. In this case the parameters λ_2, which do not reflect the metastability at all, have nearly the same (or even a slightly higher) probability as the models λ_1 and λ_3 with inherent metastability. Now, if we know some metastable regions of the process a priori – or recognize it from the time series by eye – it can be reasonable to restrict the observation space from the beginning. That is, certain columns will be set to zero (the off-diagonal entries only). We apply the state space restriction to the EM algorithm. The initial parameters of the prior model are generated at random, the convergence threshold is set to $1e-3$ as above. If we proceed this way in the present example, we obtain after 119 iterations the following model:

$$
\begin{aligned}
L^{hidden} &= 1.0e-003 * \begin{pmatrix} -0.1452 & 0.1452 \\ 0.1560 & -0.1560 \end{pmatrix} \\
L_1^{out} &= \begin{pmatrix} -0.9480 & 0.4975 & 0.4505 & 0 & 0 & 0 \\ 0.2359 & -0.2537 & 0.0178 & 0 & 0 & 0 \\ 0.6041 & 0.7389 & -1.3430 & 0 & 0 & 0 \\ 1.0295 & 5.2187 & 0 & -6.2482 & 0 & 0 \\ 0 & 0.5531 & 0 & 0 & -0.5531 & 0 \\ 0 & 5.1975 & 0 & 0 & 0 & -5.1975 \end{pmatrix} \\
L_2^{out} &= \begin{pmatrix} -1.8479 & 0 & 0 & 1.8474 & 0.0006 & 0 \\ 0 & -4.7317 & 0 & 0.8546 & 0 & 3.8771 \\ 0 & 0 & -2.7569 & 0 & 2.7569 & 0 \\ 0 & 0 & 0 & -1.2831 & 0.8873 & 0.3958 \\ 0 & 0 & 0 & 0.5795 & -1.5386 & 0.9591 \\ 0 & 0 & 0 & 0.7723 & 0.1784 & -0.9507 \end{pmatrix}.
\end{aligned}
$$

The resulting parameter estimator is nearly identical to the third estimated model λ_3. Especially the likelihood -85904.457 matches up to two decimal places. Thus, if the metastable regions are known, we do not need any knowledge about the hidden state path to achieve suitable results. The restriction of the observation space will also be applied to the next example. We conclude this example with a comparison of the estimated metastable

	metastable set $\{1,2,3\}$	metastable set $\{4,5,6\}$
generator estimation without HMMs	0.0370	0.0723
HMM-MJP with predetermined hidden path	0.0370	0.0724
HMM-MJP with restricted state space	0.0370	0.0742

Table 7.9: Deviation from the metastable blocks of the estimators to the metastable block of the original generator given by the norm $\|B - \hat{B}\|$. B denotes the metastable block of the original generator $\hat{B}$ the metastable block of the particular estimators.

blocks for the parameters obtained by the EM algorithm with predetermined path, the observation space restriction approach, and the generator estimation without HMMs as carried out in Section 6.5.2. In Table 7.9 we compare the normwise distance of both metastable blocks $\|B-\hat{B}\|$ from the estimators to the original generator.

The metastable blocks of the resulting estimators are indeed similar, but the off-block diagonal entries of the generator without HMM contain more entries that are close to zero. Thus the HMM-MJP approach is the method of choice to avoid numerical difficulties.

7.2.6 Example: A Discrete Generator for an Smoluchowski Process

In a last example we will compare the HMM-MJP with the HMMSDE estimation for a time series that was generated by propagation of Smoluchowski dynamics given by the SDE

$$\dot{X}(t) \quad = \quad -\nabla_X V(X(t)) \, + \, \sigma \, \dot{W}(t). \tag{7.12}$$

In simulations of biomolecules the potential describes a molecular force field. However, here we use for illustrative reasons a small double well potential with minima at 1 and -1

$$V(x) = (x^2 - 1)^2$$

as test example. If the potential was harmonic, the SDE above would describe an Ornstein-Uhlenbeck process (cf. Chapter 4).

The generator of the process is given by the Fokker-Planck equation

$$L = \triangle_x V(x) + \nabla_x V(x) \cdot \nabla_x + \frac{1}{2} \nabla_x \cdot B \nabla_x.$$

Its discretization can be found at [55]. However, in our case we use the trivial Dirichlet boundary conditions $L_{ij} = 0$ for $i, j < 0$, or $i, j > N$ respectively,

where N denotes the number of boxes. We obtain the discrete generator

$$\begin{aligned} L_{00} &= -\frac{\beta^{-1}}{h^2} - \frac{V'(x_0)}{2h} \\ L_{NN} &= -\frac{\beta^{-1}}{h^2} + \frac{V'(x_N)}{2h} \\ L_{ij} &= \begin{cases} \frac{\beta^{-1}}{h^2} + \frac{V'(x_i)}{2h}, & \text{if } i = j-1, \quad 0 < i \leq N \\ -2\frac{\beta^{-1}}{h^2}, & \text{if } i = j, \qquad 0 < i, j < N \\ \frac{\beta^{-1}}{h^2} - \frac{V'(x_i)}{2h}, & \text{if } i = j+1, \quad 0 \leq j < N \\ 0 & \text{else.} \end{cases} \end{aligned} \tag{7.13}$$

The parameter β is defined by the noise intensity σ according to the relation

$$\beta = \frac{1}{\sigma^2}. \tag{7.14}$$

The time series propagated by (7.12) is shown in Figure 7.17. The noise intensity σ was set to $\frac{1}{\sqrt{2}}$.

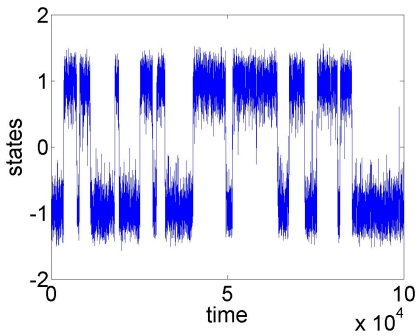

Figure 7.17: Time series generated by Smoluchowski dynamics (7.12) with noise intensity $\sigma = \frac{1}{\sqrt{2}}$ and step size $\tau = 1e-2$ (for details see below).

Now we estimate an HMM with MJP output process and two hidden states. For each state we obtain an generator that describes an output process. As will turn out, it is reasonable to assume the associated potentials being harmonic. Thus the described process is Ornstein-Uhlenbeck. Before we compare the resulting estimates with the HMMSDE estimators we will shortly describe the proceeding of the HMM-MJP approach:

At first we estimate an HMM-MJP with a Markov jump output process. Since we have chosen a constant time lag, we take for the hidden process a discrete Markov chain.

The output process stays in different subregions depending on the hidden state. If we restrict the observation space for the particular states on subregions, the effort reduces clearly. A number of columns is set to zero, thus no transition out of the subregion is possible without an hidden

state change. The data was discretized rather coarse grained into 32 boxes. The observations space was restricted a priori to $\{1, ..., 22\}$ for the first and $\{15, ..., 32\}$ for the second state. This way a considerable reduction of effort was achieved and simultaneously the observation spaces for both states are still overlapping, so that the transition behavior between both hidden states can be expressed properly.

After 51 iterations the EM algorithm converged up to an accuracy of $1e-3$. The estimated generators have tridiagonal structure. We estimate the potential parameters for the resulting generator not directly as specified in (7.13) since the discretization has to be very small to ensure, that the generator condition is fulfilled. Here we used a discretization step $h = 0.1$. With this boxsize the generator condition in (7.13) is violated as there exist some negative entries L_{ij}. That is why we estimate the potential parameters rather by means of the free energy. The exact proceeding will be pointed out in the following:

The stationary distribution of a process like (7.12) is given by the Boltzmann distribution

$$\pi(x) = \frac{1}{Z} \exp(-\beta V(x)).$$

The constant Z is a normalization constant

$$Z = \int_{x \in \mathbb{R}} \exp(-\beta V(x)) \; dx,$$

such that $\pi(x)$ becomes a probability measure. Its discretization with respect to the boxes (or intervals) $\mathcal{B}_i$ is defined as

$$\bar{\pi}_i = \frac{1}{Z} \int_{x \in \mathcal{B}_i} \exp(-\beta V(x)) \; dx.$$

The discretized stationary distribution $\bar{\pi}$ then specifies the discretized free energy

$$\bar{F}_i = -\beta^{-1} \log[\bar{\pi}_i].$$

In the continuous case the free energy

$$F(x) = -\beta^{-1} \log[\pi(x)] = V(x) - \log[Z]$$

is exactly the potential energy itself up to a constant.

This way we can compute the continuous and discrete stationary distribution π and $\bar{\pi}$ as well as the continuous and discrete free energy F and $\bar{F}$. The discretized parameters we will compare with our estimators. From an estimated generator $\hat{L}$ we obtain the free energy as the solution of

$$\begin{aligned} \hat{\pi}' \hat{L} &= 0 \\ \hat{F} &= -\beta^{-1} \log(\hat{\pi}). \end{aligned}$$

In our example we have the HMM parameters $\hat{L}_1$ and $\hat{L}_2$ – since we have assumed two hidden states – and based on these we obtain $\hat{\pi}_1$, $\hat{\pi}_2$, $\hat{F}_1$ and $\hat{F}_2$.

Each of the estimated free energy vectors $\hat{F}_1$ and $\hat{F}_2$ reflect one well of the double well potential as Figure 7.18 shows. Accordingly, the estimated stationary distributions $\hat{\pi}_1$ and $\hat{\pi}_2$ approximate one peak of the stationary distribution π each.

Finally, we can fit to the estimated free energy $\hat{F}_1$ and $\hat{F}_2$ a harmonic potential by the least squares method

$$\text{argmin}_{\mu,D\in\mathbb{R}} \sqrt{\sum_i \left| \left(\frac{1}{2}\hat{D}(x_i - \hat{\mu})^2 - \hat{F}_k(i) \right) \hat{\pi}_k(i) \right|^2}.$$

The sum index i here indicates the boxes, and the points x_i denote the box-centers. The Euclidian norm was weighted with the estimated stationary distribution, to achieve a better fit to the highly frequented states. Furthermore, to avoid some asymmetric effects we take only the states $\{1, ..., 16\}$ of $\hat{F}_1$ and the states $\{17, ..., 32\}$ of $\hat{F}_2$ into account. Although the state space restriction was set to $\{1, ..., 22\}$ for the first and $\{15, ..., 32\}$ for the second state to ensure a certain overlap.

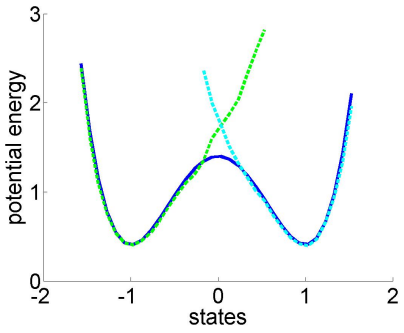

Figure 7.18: Discretized free energy $\bar{F}$ (solid line) and estimators $\hat{F}_1$ and $\hat{F}_2$ for both hidden states (dashed line).

This way we have transformed the model parameters from $\lambda = (A, (L_k^{out})_{k\in S})$ to $\lambda = (A, (\mu_k, D_k, \beta_k)_{k\in S})$, which correspond exactly to the HMMSDE parameters (5.2). The only difference are the parameters β in the HMM-MJP and B in the HMMSDE approach. Both arise from the noise intensity σ of the SDE (7.12) as $\beta = \frac{1}{\sigma^2}$ and $B = \sigma^2$.

We could not estimate the parameter β with the proceeding described above. It was set to $\beta = 4$. We made use of the a priori knowledge of the noise intensity since we have set it to $\frac{1}{\sqrt{2}}$ propagating the SDE (7.12). The parameter B on the other hand was estimated by HMMSDE as $\hat{B}_1 =$

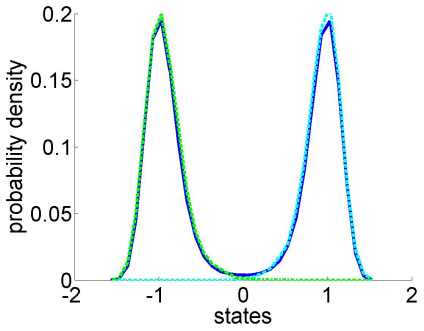

Figure 7.19: Discretized stationary distribution $\bar{\pi}$ (solid line) and estimators $\hat{\pi}_1$ and $\hat{\pi}_2$ for both hidden states (dashed line).

	$\hat{\mu}_1$	$\hat{D}_1$
HMM-MJP	-0.9656	5.5756
HMMSDE	-0.9318	5.2720

	$\hat{\mu}_2$	$\hat{D}_2$
HMM-MJP	0.9688	5.8889
HMMSDE	0.9312	5.3035

Table 7.10: Comparison of the potential parameters and noise intensities estimated by the HMM-MJP and the HMMSDE approach for state 1 at the top and state 2 at the bottom.

0.5024 and $\hat{B}_2 = 0.5066$, which approximates the square of the a priori noise intensity $B = \sigma^2 = 0.5$ up to an accuracy of two decimal places.

The other resulting parameters, which were estimated in both approaches are listed in the Table 7.10. The EM algorithm in the HMMSDE approach took 41 iterations to achieve a convergence threshold of $1e-3$. The potential parameters are similar to the estimates resulting from the HMM-MJP approach. However, the HMM-MJP approximated the double well potential slightly better. The minima $\hat{\mu}_1$ and $\hat{\mu}_2$ are closer to -1 and 1 (cf. Figure 7.20).

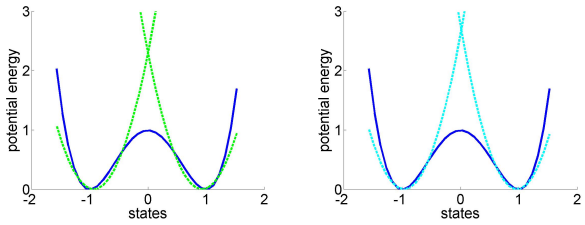

Figure 7.20: Discretized free energy and estimated harmonic potentials based on the HMMSDE approach (left) and on the HMM-MJP approach (right).

Altogether both models were able to describe the dynamical behavior within the metastable states. HMM-MJP provides a tool to recover the

local dynamics even for coarse-grained discretizations. On the other hand it requires more effort since more parameters have to be estimated. HMMSDE requires the estimation of three parameters (μ, D, B) for each hidden state, HMM-MJP of an $M \times N$ – matrix per each state, where N is the number of boxes in the entire and M the number of boxes in the restricted state space. In this case the matrix was tridiagonal. Thus we have to estimate just $2(M-1)$ matrix entries.

Another possibility would be the estimation of one large generator matrix containing metastable blocks instead of hidden states. But this variant implies numerical difficulties as already discussed in the previous example.

To get rid of the small transition rates, we considered generators that describe the transition behavior within a metastable subregion and swapped the transition behavior between the metastable sets out to the hidden process.

7.3 Alternative Approaches to HMM Variants

There exist articles focussing on HMMs combined with Markov jump processes [13, 12] but they consider a more general case of an infinite-dimensional state space. However, the modification of HMMs presented herein provides a directly implementable algorithm for a finite-dimensional state space, which is not the case by existing approaches in aforementioned papers. Further, in contrast to [57] no additional assumption about the jump times are necessary in the presented expectation maximization (EM) algorithm. In the article [62] was designed a dynamical Bayesian network with continuous time. Dynamical Bayesian networks are actually hidden Markov models, but the authors designed the model such that the observables are discrete points of the Markov process and the hidden variables are the continuous time intervals between the observables. The continuous time Bayesian network describes therefore only a Markov process not – as in our case – also an associated output process. However, the generator estimation follows the same procedure as we use in Chapter 6, but the authors take a slightly different likelihood function into account with generator estimators as specified in [3, 36], whereas we keep with [11]. These approaches consider HMMs with purely geometrical output parameters. In this work are designed HMMs with kinetic output processes focussing on the transition behavior of the observables. In [75] kinetic models have been applied for a better description of MD systems. However, the combination of HMMs kinetic models is new.

We call the combination of HMMs with a Markov jump process HMM-MJP and with Ornstein-Uhlenbeck processes HMMSDE. The algorithmic concept of HMMSDE has some similarities with other approaches based on the concept of HMMs or hidden Markov processes, in particular approaches presented in [29, 32, 60]. However, the fundamental difference is that the HMMSDE approach suggested herein combines some discrete hidden process

with generally continuous stochastic differential equation (SDE) output.

That is, the concept behind HMMSDE and HMM-MJP can be expressed shortly in the following way:

$$\left.\begin{array}{l}\text{HMMSDE}\\ \text{HMM-MJP}\end{array}\right\} = \text{ kinetic parameterization + HMM metastability analysis.}$$

Concerning *HMM-based metastability analysis* several other approaches exist, all of them with different focus: For example, [29] considers *stationary* output behavior only, and [32] global SDE models with hidden data but *without* discrete metastable states.

8 Summary

In this thesis a set of procedures for the analysis of time series was developed. The models introduced here are based on the concept of Hidden Markov models. A hidden Markov model (HMM) consists of two stochastic processes. However, only one of these is observable. The HMM variants presented herein have been developed with regard to a prospective application to biomolecular time series.

Therefore, the investigated time series are realizations of processes that can be characterized as follows: They jump between metastable states. The correlation times are small with respect to the exit times from a metastable state. On a time scale, chosen in such a way that the process is Markovian, these jumps occur instantaneously. That is: *Transitions between metastable states in equilibrium take place in one or very few time steps.*

By means of the presented methods in particular, the question how metastable states can be distinguished by kinetic patterns is addressed. The local dynamics are modeled either by a space-discrete Markov jump process or by a continuous diffusion process.

Both concepts are discussed by means of several examples. Particularly we focussed on the issue of generator estimation. The combination of the generator estimation with the concept of the hidden Markov model is new. It allows for analyzing time-continuous processes with standard HMM techniques. Especially the time-continuity of the model makes the analysis of time series with varying time lags possible. Furthermore, a model with Markov jump output process has the advantage that the process is determined by the generator matrix only. Hence no additional assumption about the distribution of the data is required. Beyond this in the examples we observed that even if the box-discretization is rather coarse-grained, the local dynamics still can be expressed satisfactorily by an HMM-MJP.

In the scope of this thesis HMM-MJP was applied to small systems, generated by Smoluchowski dynamics, by a discrete generator or by an HMM itself. However, the algorithms are applicable to the high-dimensional case. How HMM-MJP performs in the application to larger systems – such as the simulation of biomolecules – has to be clarified in further investigations. Difficulties arising with larger systems are on the one hand computational costs and on the other hand cumulations of small entries in the generator matrix. Too many small entries close to zero can lead to numerical instabilities. One approach to handle these problems is the restriction of the state space as described in the examples 7.2.5 and 7.2.6.

9 Zusammenfassung

In dieser Arbeit wurde ein Verfahrenskatalog zur Zeitreihenanalyse metastabiler Systeme entwickelt. Die hier eingeführten Modelle basieren auf dem Konzept der Hidden Markov Modelle. Ein Hidden Markov Modell (HMM) besteht aus zwei stochastischen Prozessen, von denen nur einer beobachtbar ist. Die HMM-Varianten in dieser Arbeit wurden im Hinblick auf die spätere Anwendung auf Biomolekülzeitreihen entwickelt.

Deshalb sind die zu analysierenden Zeitreihen Realisierungen von Prozessen, die sich wie folgt charakterisieren lassen: Sie springen zwischen metastabilen Zuständen. Die Korrelationsszeiten sind im Vergleich zu den Austrittszeiten aus einem metastabilen Zustand klein. Auf einer Zeitskala, die so gewählt ist, dass der Prozess Markovsch ist, passieren diese Sprünge "plötzlich". Das heißt: *Übergänge zwischen metastabilen Zuständen im Gleichgewicht passieren in einem oder in sehr wenigen Zeitschritten.*

Mit den vorgestellten Verfahren lässt sich insbesondere die Frage, wie metastabile Zustände anhand kinetischer Muster zu unsterscheiden sind, behandeln. Die lokale Dynamik wird entweder duch einen räumlich diskreten Markovschen Sprungprozess oder aber durch einen kontinuierlichen Diffusionsprozess modelliert.

Diese beiden Konzepte wurden anhand von Beispielen diskutiert. Insbesondere die Frage der Generatorschätzung wurde in dieser Arbeit eingehend behandelt. Die Kombination der Generatorschätzung mit dem Konzept des Hidden Markov Modells ist neu. Sie ermöglicht die Analyse zeit-kontinuierlicher Prozesse mit bewährten HMM-Techniken. Die Darstellung durch zeit-kontinuierliche Modelle erlaubt insbesondere die Analyse von Zeitreihen mit unterschiedlicher Zeitschrittweite. Desweiteren hat ein HMM mit beobachtbarem Markovschen Sprungprozess den Vorteil, dass der Prozess allein durch die Generatormatrix bestimmt wird und keine zusätzliche Annahme über die Verteilung der Daten notwendig ist. Darüber hinaus hat sich in den Bespielen gezeigt, dass sich selbst bei groben Boxdiskretisierungen die lokale Dynamik durch ein HMM-MJP noch gut beschreiben lässt.

Im Rahmen dieser Arbeit wurde das HMM-MJP auf kleine Systeme angewandt, die von einer Smoluchowski Dynamik, einem diskreten Generator oder einem HMM generiert wurden. Die Algorithmen sind jedoch auf den hochdimensionalen Fall übertragbar. Wie sich das HMM-MJP in der Anwendung auf größere Systeme – etwa der Simulation von Biomolekülen – bewährt, ist noch in weiterführenden Arbeiten zu untersuchen. Die Schwierigkeiten, die größere Systeme mit sich bringen, sind zum einen der Rechenaufwand und zum anderen die Häufung sehr kleiner Einträge in der Generatormatix, die zu numerischer Instabilität führen können. Ein Ansatz, diese Probleme zu bewältigen, ist die Einschränkung des Zustandsraumes, wie in den Beispielen 7.2.5 und 7.2.6 beschrieben.

Abbreviations

MD	molecular dynamics
HMM	hidden Markov model
MJP	Markov jump process
EM	expectation maximization (algorithm)
SDE	stochastic differential equation
HMMSDE	HMM with an output process that is determined by a SDE
HMM-MJP	HMM with a Markov jump process either as output or as hidden process
MLE	maximum likelihood estimation
a.s.	almost sure
pdf	probability density functions
FPE	Fokker-Planck equation
PCCA	Perron cluster analysis
ODE	ordinary differential equation
QP	quadratic programming

References

[1] M. Allen and D. Tildesley. *Computer Simulations of Liquids*. Clarendon Press, Oxford, 1990.

[2] T. W. Anderson and L. A. Goodman. Statistical inference about Markov chains. *Ann. Math. Stat.*, 28:89–109, 1957.

[3] S. Asmussen, O. Nerman, and M. Olsson. Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, 23:419–441, 1996.

[4] Pierre Baldi and Soeren Brunak. *Bioinformatics : the machine learning approach*. MIT Press, 1998.

[5] L. Baum. An inequality and associated maximization technique in statistical estimation for probalistic functions of Markov processes. *Inequalities*, 3:1–8, 1972.

[6] L. Baum and J. Eagon. An inequality with applications to statistical estimation for probalistic functions of Markov processes and to a model for ecology. *American Mathematical Society Bulletin*, 73:360–363, 1967.

[7] L. Baum and T. Petrie. Statistical inference for probabilistic functions of finite state Markov chains. *Annals of Mathematical Statistics*, 37:1554–1563, 1966.

[8] L.E. Baum and G.R. Sell. Growth transformations for functions on manifolds. *Pac. J. of Math.*, 27, 1968.

[9] Leonard E. Baum, Ted Petrie, George Soules, and Norman Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.*, 41:164–171.

[10] J.A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models, 1998. Technical Report TR-97-021, International Computer Science Institute, Berkeley CA.

[11] M. Bladt and M. Sorensen. Statistical inference for discretely observed Markov jump processes. *J.R.Statist. Soc. B*, 67(3):395–410, 2005.

[12] A.V. Borisov. Analysis and estimation of the states of special jump Markov processes. I. martingale representation. *Automation and Remote Control*, 65(1):44–57, 2004.

[13] A.V. Borisov and A.I. Stefanovich. Optimal filtering for HMM governed by special jump processes. *Decision and Control, CDC-ECC '05.*, pages 5935– 5940, 2005.

[14] Petros Boufounos, Sameh El-Difrawy, and Dan Ehrlich. Hidden Markov models for DNA sequencing. citeseer.ist.psu.edu/571136.html.

[15] Pierre Brémaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer, New York, 1999.

[16] K. L. Chung. *Markov chains with stationary transition probabilities*. Springer, New York, 1967.

[17] Peter Clote and R. Backofen. *Computational Molecular Biology: An Introduction*. John Wiley & Sons, Ltd, Aug 2000.

[18] D. T. Crommelin and E. Vanden-Eijnden. Fitting timeseries by continuous-time Markov chains: A quadratic programming approach. *J. Comp. Phys.*, 217:782–805, 2006.

[19] James R. Cuthbert. On uniqueness of the logarithm for Markov semigroups. *J. London Math. Soc.*, s2-4:623–630, 1972.

[20] James R. Cuthbert. The logarithm function for finite-state Markov semi-groups. *J. London Math. Soc.*, s2-6:524–532, 1973.

[21] Michael Dellnitz and Oliver Junge. On the approximation of complicated dynamical behavior. *SIAM J. Num. Anal.*, 36(2):491–515, 1999.

[22] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[23] Peter Deuflhard, Wilhelm Huisinga, Alexander Fischer, and Christof Schütte. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Lin. Alg. Appl.*, 315:39–59, 2000.

[24] Peter Deuflhard and Marcus Weber. Robust Perron cluster analysis in conformation dynamics. Special Issue on Matrices and Mathematical Biology, 2005.

[25] Evelyn Dittmer. Projizierte Hidden-Markov-Modelle in der Metastabilitätsanalyse hochdimensionaler Zeitreihen. Diploma thesis, Department of Mathematics and Computer Science, Free University Berlin, 2004.

[26] G. Elfving. Zur theorie der markoffschen ketten. *Acta Social Science Fennicae n.*, series A.2(no. 8):1–17, 1937.

[27] Erwan Faou and Christian Lubich. A Poisson integrator for Gaussian wavepacket dynamics. *Comput. Visual. Sci.*, 9(no. 2):45–55, 2006.

[28] Alexander Fischer. *An Uncoupling-Coupling Method for Markov Chain Monte Carlo Simulations with an Application to Biomolecules*. PhD thesis, Free University Berlin, 2003.

[29] Alexander Fischer, Sonja Waldhausen, and Christof Schütte. Identification of biomolecular conformations from incomplete torsion angle observations by Hidden Markov Models. *J. Comp. Chem.*, 28:2453–2464, 2007.

[30] C. Franzke, D. Crommelin, A. Fischer, and A.J. Majda. A hidden Markov model perspective on regimes and metastability in atmospheric flows. *Journal of Climate*, 21:1740–1757, 2008.

[31] D. Frenkel and B. Smit. *Understanding Molecular Dynamics: From Algorithms to Applications*. Academic Press, London, 2002.

[32] J. Frydman and P. Lakner. Maximum likelihood estimation of hidden Markov processes. *Ann. Appl. Prob.*, 13(4):1296–1312, 2003.

[33] C. W. Gardiner. *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*. Springer, Berlin, 2004.

[34] G. S. Goodman. An intrinsic time for non-stationary finite Markov chains. *Z. Wahrscheinlichkeitsth.*, 16:165–180, 1970.

[35] Nicholas I.M. Gould. Some reflections on the current state of active-set and interior-point methods for constrained optimization. 2003.

[36] I. Holmes and G. M. Rubin. An expectation maximization algorithm for training hidden substitution models. *J Mol Biol*, 317(5):753–764, April 2002.

[37] I. Horenko, R. Klein, S. Dolaptchiev, and Ch. Schütte. Automated generation of reduced stochastic weather models i: Simultaneous dimension and model reduction for time series analysis. *Mult. Mod. Sim.*, 6(4):1125–1145, 2008.

[38] Illia Horenko, Evelyn Dittmer, Alexander Fischer, and Christof Schütte. Automated model reduction for complex systems exhibiting metastability. *Mult. Mod. Sim.*, 5(3):802–827, 2006.

[39] Illia Horenko, Evelyn Dittmer, Filip Lankas, John Maddocks, Philipp Metzner, and Christof Schütte. Macroscopic dynamics of complex metastable systems: Theory, algorithms, and application to B-DNA. *SIAM J. Appl. Dyn. Syst.*, 7(2):532–560, 2008.

[40] Illia Horenko, Evelyn Dittmer, and Christof Schütte. Reduced stochastic models for complex molecular systems. *Comp. Vis. Sci.*, 9(2):89–102, 2005.

[41] Illia Horenko and Christof Schütte. Likelihood-based estimation of multidimensional langevin models and its application to biomolecular dynamics. *To appear in Mult. Mod. Sim.*, 2007.

[42] W. Huisinga, S. Meyn, and C. Schütte. Phase transitions & metastability in Markovian and molecular systems, 2004.

[43] Wilhelm Huisinga. *Metastability of Markovian systems. A transfer operator based approach in application to molecular dynamics*. PhD thesis, Free University Berlin, 2001.

[44] Wilhelm Huisinga and Eike Meerbach. Markov processes for everybody, 2005.

[45] R. B. Israel, J. S. Rosenthal, and J. Z. Wei. Finding generators for Markov chains via empirical transition matrices, with applications to credit ranking. *Mathematical finance*, 11(2):245–265, 2001.

[46] S. Johansen. Some results on the embedding problem for finite Markov chains. *J. London Math. Soc.*, 8:191–195, 1974.

[47] J. F. C. Kingman. The imbedding problem for finite Markov chains. *Z. Wahrscheinlichkeitstheorie*, 1:14–24, 1962.

[48] P. E. Kloeden, E. Platen, and H. Schurz. *Numerical Solution of SDE Through Computer Experiments*. Springer, Berlin, 1994.

[49] Andrzej Lasota and Michael C. Mackey. *Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics*. Springer, New York, 1994.

[50] L. A. Liporace. Maximum likelihood estimation for multivariate observations of Markov sources, 1989.

[51] G. McLachlan. *The EM Algorithm and Extensions*. Wiley, New York, 2005. Second Edition.

[52] G. McLachlan and D. Peel. *Finite mixture models*. Wiley, New York, 2000.

[53] E. Meerbach, E. Dittmer, I. Horenko, and Ch. Schütte. Multiscale modelling in molecular dynamics: Biomolecular conformations as metastable states. *Lecture Notes in Physics*, 703:457–497, 2006.

[54] E. Meerbach, Ch. Schütte, and A. Fischer. Eigenvalue bounds on restrictions of reversible nearly uncoupled Markov chains. *Lin. Alg. Appl.*, 398:141–160, 2005.

[55] P. Metzner. *Transition Path Theory for Markov Processes, Application to Molecular Dynamics*. PhD thesis, Free University Berlin, 2007.

[56] Ph. Metzner, E. Dittmer, T. Jahnke, and Ch. Schütte. Generator estimation of Markov jump processes. *J. Comp. Phys.*, 227(1):353–375, 2007.

[57] B. M. Miller and W. J. Runggaldier. Kalman filtering for linear systems with coefficients driven by a hidden Markov jump process. *Systems and Control Letters*, 31(2):93– 102(10), 1997.

[58] T. Müller. *Modellierung von Proteinevolution.* PhD thesis, Heidelberg, 2001.

[59] T. Müller, R. Spang, and M. Vingron. Estimating amino acid substitution models: A comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. *Mol Biol Evol*, 19:8–13, 2002.

[60] C. R. Nelson and C. J. Kim. State-space models with regime switching: Classical and gibbs-sampling approaches with applications. *MIT Press*, 1999.

[61] M. F. Neuts. *Algorithmic Probability: a Collection of Problems.* Chapman and Hall, London, 1995.

[62] U. Nodelman, C. R. Shelton, and D. Koller. Expectation maximization and complex duration distributions for continuous time Bayesian networks. In *Proceedings of the Twenty-first Conference on Uncertainty in AI (UAI)*, pages 421–430, Edinburgh, Scottland, UK, July 2005.

[63] A. Quarteroni and F. Saleri. *Wissenschaftliches Rechnen mit MATLAB.* Springer, New York, 2006.

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

[65] H. Risken. *The Fokker-Planck-Equation. Methods of Solution and Applications.* Springer, Berlin, 1989.

[66] C. Schütte and W. Huisinga. *Biomolecular Conformations can be Identified as Metastable Sets of Molecular Dynamics.* North–Holland, 2003.

[67] Christof Schütte, Alexander Fischer, Wilhelm Huisinga, and Peter Deuflhard. A direct approach to conformational dynamics based on hybrid Monte Carlo. *J. Comput. Phys.*, 151:146–168, 1999.

[68] E. Seneta. *Non-negative matrices and Markov chains.* Springer, New York, 1981.

[69] Petar Todorovic. *An Introduction to Stochastic Processes and Their Applications*. Springer, New York, 1992.

[70] Andrew J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inform. Theory*, 13:260–269, 1967.

[71] Karl-Heinz Waldmann and Ulrike M. Stocker. *Stochastische Modelle*. Springer, Berlin, 2003.

[72] Darren J. Wilkinson. *Stochastic Modelling for Systems Biology (Mathematical and Computational Biology)*. Chapman & Hall/CRC, April 2006.

[73] C.F.J. Wu. On the convergence properties of the EM algorithm. *Ann. Stat.*, 11(1):95–103, 1983.

[74] L. Wu and M.I. Jordan. On the convergence properties of the EM algorithm for Gaussian mixtures. *Neural Comput.*, 8:129–151, 1996.

[75] Jianhua Xing, Hongyun Wang, and George Oster. From continuum fokker-planck models to discrete kinetic models. *Biophys. J.*, 89(3):1551–1563, September 2005.

[76] Huan-Xiang Zhou, Stanislaw T. Wlodek, and J. Andrew McCammon. Conformation gating as a mechanism for enzyme specificity. *Proc. Natl. Acad. Sci.*, 95:9280 – 9283, August 1998.