Automated Model Reduction for Complex Systems exhibiting Metastability*

Illia Horenko[1] Evelyn Dittmer[1] Alexander Fischer[2]
Christof Schütte[1,†]

January 26, 2005

[1] *Institute of Mathematics II, Free University Berlin, Arnimallee 2–6, 14195 Berlin, Germany*

[2] *Courant Institute, New York University, 251 Mercer Street, New York, 10012 NY, USA*

Abstract

We present a novel method for the identification of the most important metastable states of a system with complicated dynamical behavior from time series information. The novel approach represents the effective dynamics of the full system by a Markov jump process between metastable states, and the dynamics within each of these metastable states by rather simple stochastic differential equations (SDEs). Its algorithmic realization exploits the concept of Hidden Markov Models (HMMs) with output behavior given by SDEs. The numerical effort of the method is linear in the length of the given time series, and quadratic in terms of the number of metastable states. The performance of the resulting method is illustrated by numerical tests and by application to molecular dynamics time series of a trialanine molecule.

Keywords: Bayesian networks; Biomolecular conformations; Hidden Markov model; Maximum likelihood principle; Metastability; Stochastic differential equations

1 Introduction

The macroscopic dynamics of typical biomolecular systems is mainly characterized by the existence of biomolecular conformations which can be understood as metastable geometrical large scale structures, i.e., geometries which are persistent for long periods of time. On the longest time scales biomolecular dynamics is a kind of flipping process between these conformations [9, 12], while on closer

*Supported by the DFG research center "Mathematics for key technologies" (FZT 86) in Berlin. One of the authors (AF) was also supported under NSF grant DMS02-39625.

† *Correspondence to:* Ch. Schütte; E-mail: schuette@math.fu-berlin.de or I. Horenko; E-mail: horenko@math.fu-berlin.de

inspection it exhibits a rich temporal multiscale structure [21]. Biophysical research seems to indicate that typical biomolecular systems possess only few dominant conformations that can be understood as metastable or almost invariant sets in state or configuration space [26, 27]. In other words, the effective or macroscopic dynamics is given by a Markov jump process that hops between the metastable sets while the dynamics within these sets might be mixing on time scales that are smaller than the typical waiting time between the hops. In many applications this Markovian picture is an appropriate description of the dynamics since typical correlation times in the system are sufficiently smaller than the waiting times between hops (and thus much smaller than the timescale the effective description is intended to cover).

The same description of the effective dynamics is true for other complex system including, e.g., climate systems or systems from materials science.

Recently there have been several *set-oriented* approaches to the algorithmic identification of metastable sets of a complex system, and to the computation of the transition probabilities between them [26, 6, 4, 5]. These approaches are based on the construction of a transition matrix that describes transition probabilities between sets in the state space of the system. The identification of metastable sets then is based on analysis of this transition matrix [27, 7, 5]. For higher dimensional systems this always requires coarse graining of state space into sets (a partition of state space in disjoint sets that avoids the curse of dimensionality) that has to be designed carefully since the resulting metastable sets are unions of the sets from the partition.

In this article we will discuss a novel approach to the problem that is no longer purely set-oriented and is not based on (traditional) coarse graining concepts. Instead, we propose to approximate the effective dynamics by stochastic differential equations (SDEs) of the following type for the state $x \in \mathbf{R}^n$ of the system:

$$\begin{aligned} dx(t) &= -D_x V^{(q(t))}(x(t)) + \sigma^{(q(t))}\, dW(t) \\ q(t) &= \text{Markov jump process with states } 1, \ldots, M, \end{aligned} \tag{1}$$

where $W(t)$ is denoting standard Brownian motion, D_x differentiation wrt. x, $\Sigma = (\sigma^{(1)}, \ldots, \sigma^{(M)})$ contains noise intensities, and $\mathcal{V} = (V^{(1)}, \ldots, V^{(M)})$ interaction potentials. The jump process $q(t)$ 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 $M \times M$ rate matrix R. The SDEs (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 $(R, \mathcal{V}, \Sigma)$. In the following we will assume that $\mathcal{V}$ only contains potentials $V^{(q)}$ from a certain family of potentials that is given by a (not too large) tuple of parameters $\theta^{(q)}$ (e.g., polynomial potentials) such that $\mathcal{V}$ is completely determined by the parameters $\Theta = (\theta^{(1)}, \ldots, \theta^{(M)})$.

Consequently, we have to find a procedure that can determine the optimal

model $\lambda = \lambda(R, \Theta, \Sigma)$ for the complex system under consideration. The goal of the algorithmic approach to be presented herein is to identify the optimal model $\lambda(R, \Theta, \Sigma)$ 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 (R, Θ, Σ) simultaneously and iteratively. This procedure will result from some dynamical Bayesian network approach.

Dynamical Bayesian networks generally provide a powerful framework for deriving efficient algorithms to analyze time series (or other observation sequences) that are thought to be governed by some hidden process [14] (in our case the "hidden" sequence of metastable states). The hidden Markov model (HMM) [14, 24] is one of the most popular representatives of a Bayesian network; its graphical representation is shown in Fig. 1 (a) below. In HMMs, an observation sequence is assumed to be a realization of random variables $Y_1, \ldots, Y_N$ that depend on a sequence of hidden states $X = X_1, X_2, \ldots, X_N$ forming a discrete time Markov chain. Associated with HMMs are algorithms for learning and inference, especially the Expectation-Maximization (EM) algorithm for learning parameters of the model, and the Viterbi algorithm for inferring the most probable hidden state space sequence.

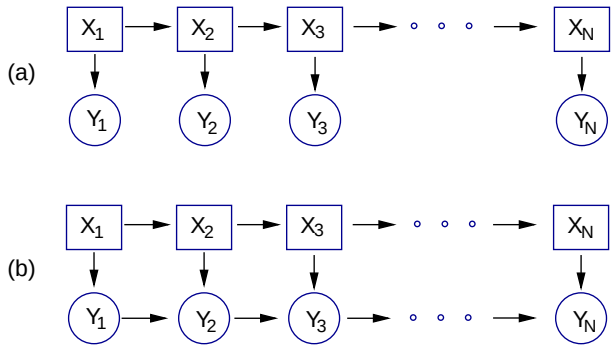

Figure 1: Graphical model of dynamical Bayesian networks for (a) the hidden Markov model and (b) HmmSde . $X_1, \ldots, X_N$ and $Y_1, \ldots, Y_N$ are random variables, rectangles denote hidden (unobservable) random variables, circles observable ones; the arrows specify conditional independence relations.

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. Eq. (1) describes the process in continuous time; the Bayesian network model, however, already reflects the situation that the observation sequence is given in discrete time.

HmmSde has a slightly more complicated conditional independence relationship (see Fig. 1 (b)): The SDE output behavior adds dependence between adjacent observations in its Bayesian network representation, which in general will make learning and inference more complicated. Yet, we will see in the following that the EM and Viterbi algorithms for our specific model are still efficiently computable with the same algorithmic complexity as in the HMM case.

The reader may think that the dimension n of (each state in) the observation sequence has to be identical to the (eventually high) dimension of the state space $\mathbf{X}$ of the complex system under investigation. Yet this is not necessary in most cases, in fact we will be able to have $n \ll dim(\mathbf{X})$. The main requirement is that the observation sequence contains time series of appropriate observables of the system. Here "appropriate" means that the combined information contained in the observation somehow "encodes" the metastability in the system (even if this may be a very cryptic code). For example, it often is possible to identify the main metastable states of a biomolecular system from the observation of some of its torsion or backbone angles. We will illustrate such a procedure in Sect. 3.

We will proceed as follows: First, we will give an outline of the approach together with remarks comparing it to other approaches to related problems. Then, we will present the construction of the identification algorithm, and show how to determine the optimal number of metastable sets. In the last section, we will first illustrate the application of this novel technique to metastable time series from suitable numerical experiments, and finally discuss the applicability of the resulting concept to realistic molecular dynamics problems.

Concept and Commentary

1.1 Concept

Our goal is to identify optimal parameters for our model (1) for given observation data $(O_t)_{t=t_0,\ldots,t_N}$. That is, the states O_{t_j}, the system is in at times t_j, 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 $\mathcal{L}(\lambda|O_t, q_t)$ has to be the probability of output $x(t_j) = O_{t_j}$, $j = 1, \ldots, N$, and the associated sequence of metastable states (q_t) (the state sequence of the Markov jump process at times t_j, $j = 1, \ldots, N$). Thus, in order to construct $\mathcal{L}$ appropriately, we have to know the probability of output of state $x(t_j)$ under the condition of being in metastable state q_{t_j} 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 q_{t_j}.

For the rest of this article let us assume that the potentials $V^{(q)}$ are of

harmonic form:

$$V^{(q)}(x) \quad = \quad \frac{1}{2} D^{(q)} (x - \mu^{(q)})^2 + V_0^{(q)}. \tag{2}$$

This assumption simplifies the derivation of the parametrization 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 later, both assumptions are not necessary.

Propagation of probability density. Let us first assume that the jump process in the HMMSDE model (1) is fixed to one state, say $q(t) = q$ for the times t considered. Considering a statistical density function $\rho(x,t)$ of an ensemble of SDE solutions (1) for different realizations of the stochastic process W, we get an equivalent representation of the dynamics in terms of the Fokker-Planck operator:

$$\partial_t \rho = \triangle_x V^{(q)}(x) \rho + \nabla_x V^{(q)}(x) \cdot \nabla_x \rho + B^{(q)} \triangle_x \rho, \tag{3}$$

where $B^{(q)} = (\sigma^{(q)})^2 \in \mathbf{R}^1$ denotes the variance of the white noise (for $\mathbf{R}^d$ it is a positive definite selfadjoint matrix). In the case of harmonic potentials this 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 (3) remains to be a sum of Gaussians whenever the initial probability function $\rho(\cdot, t = 0)$ is. Therefore, let us apply the variational principle (Dirac-Frenkel-MacLachlan principle [10]) to (3) restricted to functions ρ of the form

$$\rho(x,t) = A(t) \exp\left(-(x - x(t))^T \Sigma(t) (x - x(t))^T \right).$$

This leads to the solution of the system of ordinary differential equations

$$\begin{aligned} \dot{x} &= -D^{(q)}(x - \mu^{(q)}), \\ \dot{\Sigma} &= -2B^{(q)} \Sigma^2 + 2D^{(q)} \Sigma, \\ \dot{A} &= \left(D^{(q)} - B^{(q)} \Sigma^l \right) A, \end{aligned} \tag{4}$$

for the time–dependent parameters $\{x, \Sigma, A\}$. The explicit solution of this system of equations on the time-interval $(t, t+\tau)$ where the hidden jump process $q(t)$ is fixed in the state q is:

$$\begin{aligned} x(t+\tau) &= \mu^{(q)} + \exp\left(-D^{(q)} \tau \right) (x(t) - \mu^{(q)}), \\ \Sigma(t+\tau) &= \left(D^{(q)^{-1}} B^{(q)} - \exp\left(-2D^{(q)} \tau \right) \left(D^{(q)^{-1}} B^{(q)} - \Sigma(t)^{-1} \right) \right)^{-1}, \\ A(t+\tau) &= \frac{1}{\sqrt{\pi}} \Sigma(t+\tau)^{1/2}, \end{aligned} \tag{5}$$

In case of initial states that are sums of Gaussians, each Gaussian would move independently according to (5) and we would get the solution of (3) by superposition.

However, in the case considered herein, we are interested in the probability of output $O(t_{j+1})$ in metastable state $q_{t_{j+1}}$ under the condition that the system has been in state O_{t_j} at time t_j. For this, we can now use (5) with $x(t_j) = O_{t_j}$ and $\Sigma(t_j)^{-1} = 0$. Therefore, the output probability distribution results to be

$$\rho(O_{t_{j+1}}|q_{t_j}, O_{t_j}) = A(t_{j+1}) \exp\left(-(O_{t_{j+1}} - x(t_{j+1}))\Sigma(t_{j+1})(O_{t_{j+1}} - x(t_{j+1}))^T\right),$$

with

$$\begin{aligned} x(t_{j+1}) &= \mu^{(q)} + \exp\left(-D^{(q)}\tau\right)(O(t_j) - \mu^{(q)}), \\ \Sigma(t_{j+1}) &= \left(D^{(q)^{-1}}B^{(q)} - \exp\left(-2D^{(q)}\tau\right)D^{(q)^{-1}}B^{(q)}\right)^{-1} \\ &= \left(1 - \exp\left(-2D^{(q)}\tau\right)\right)^{-1} D^{(q)}B^{(q)^{-1}}, \\ A(t_{j+1}) &= \frac{1}{\sqrt{\pi}}\,\Sigma(t_{j+1})^{1/2}, \end{aligned} \tag{6}$$

for metastable state $q = q_{t_{j+1}}$ and with $\tau = t_{j+1} - t_j$.

Likelihood function. Whenever we assume the potential to be harmonic the model is characterized by the parameter tuple $\lambda = (v, R, x, \Sigma, A)$, where v denotes the initial distribution of the Markov Chain, R its transition matrix, and x, Σ, A the parameters of the output distributions due to (6). Suppose that the observed data (O_t) is given with constant time stepping τ, i.e., $t_k = t_{k-1} + \tau$ for all $k = 1, \ldots, N$. Setting $t_0 = 0$ we have $t_k = k\tau$ and especially $T = t_N = N\tau$. For the sake of simplicity of notation we thus may simply write $t = 0, ..., T$. In addition to the observation sequence $O = (O_t)$ we also have the sequence of hidden metastable states $q = (q_t)_{t=0,\ldots,T}$ which herein are given by the M possible states of the Markov jump process, i.e., we have $q_t \in \{1, \ldots, M\}$.

Let R be the rate 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

$$\mathcal{T} = \exp(\tau R).$$

Therefore for given model $\lambda = (v, \mathcal{T}, x, \Sigma, A)$ we have the following joint probability distribution for the observation and hidden state sequences:

$$p(O, q|\lambda) = v(q_0)\rho(O_0|q_0)\prod_{t=1}^{T}\mathcal{T}(q_{t-1}, q_t)\rho(O_t|q_t, O_{t-1}), \tag{7}$$

wherein the probability distributions ρ have the form

$$\rho(O_t|q_t, O_{t-1}) = A^{(q_t)}(t)\exp\left(-(O_t - x^{(q_t)}(t))\Sigma^{(q_t)}(t)(O_t - x^{(q_t)}(t))^T\right),$$

where the superindex refers to the hidden state q_t of the system at time t, and x, Σ and A have to be computed from (6).

Therefore, the joint Likelihood function for the model given the complete data reads:

$$\mathcal{L}(\lambda) = \mathcal{L}(\lambda|O, q) = p(O, q|\lambda).$$

Algorithmic Realization. Our next task will be to construct algorithms that

(1) determine optimal parameters $(\mathcal{T}, \mu^q, D^q, B^q)_{q=1,\ldots,M}$ by maximizing the likelihood $\mathcal{L}(\lambda|O, q)$; this is a nonlinear global optimization problem,

(2) determine an optimal sequence of hidden metastable states (q_t) for given optimal parameters, and

(3) determine the number of important metastable states; up to now we assumed that the number M of hidden states is a priori given - how can we determine an appropriate number?

In Section 2 we will see how to construct appropriate algorithms for problems (1) and (2) based on specifications of the Expectation-Maximization (EM) and Viterbi algorithms.

Determining the number of metastable states. Problem (3) of identifying the number of dominant metastable states can be formulated as the problem of *aggregating* states O_t from the time series into metastable states (i.e., clustering states that belongs to the same metastable state). The identification of an optimal aggregation from observation of the dynamical behavior is an important algorithmic problem. 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 [4, 26]. This has led to the construction of an aggregation technique called "Perron Cluster Cluster Analysis" (PCCA) [6, 7].

We will use PCCA within the HmmSde framework as follows: In the setup of HmmSde for a given observation sequence one is confronted with the task to select *in advance* the number M of hidden states. There are no general solutions to this problem, and the best way to handle this problem often is a mixture of insight and preliminary analysis. However, since our goal is to identify metastable states we can proceed as suggested in [11]: Start the EM algorithm with some sufficient number of hidden states, say M, that should be greater than the expected number of metastable states. After termination of the EM algorithm, take the resulting transition matrix A and aggregate the M hidden states into $M_{\text{meta}} \leq M$ metastable states by means of PCCA. The resulting conformation states will then allow an interpretation of the results in terms of metastable states.

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 T, its dimension n, and the number M of hidden states? The literature on the application of EM and Viterbi algorithms to the parametrization of HMMs demonstrates that *one step* of EM and the entire Viterbi algorithm scale linearly in T and quadratically in M; as we will see below this is still true for the specific HmmSde procedure. The PCCA procedure required to determine the number of dominant metastable states scales like $\mathcal{O}(M^3)$. The scaling wrt. n is not so obvious; it depends mainly on the number of parameters entering the potential. Whenever the potential is harmonic, the scaling will be $\mathcal{O}(n^2)$. Carefully putting all terms together one finds an asymptotic estimate of the form (for harmonic potentials) [8]

$$\mathcal{O}\Big((n^2M + M^2)T + n^3\Big) \cdot \text{number of EM iterations} + \mathcal{O}(M^3).$$

Here, the necessary number of iteration 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 [32, 33]. In this article 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. More precisely, it is known quite generally that for SDEs like the ones considered herein the precise estimation of drift parameters (or the parameters in the potential) requires rather long observation sequences (cf. Sec. 1.1 of [30]). 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. In the following we will call time series with this property "statistically rich enough". Whether this is the case or not will depend on the time series but also on the details of the algorithmic scheme under consideration. We will herein not address this problem theoretically (see [16] for details) but will give some examples in Sect. 3.

Generalizations. In this article we introduce the HmmSde approach for one-dimensional observations, therefore one-dimensional SDEs, and harmonic potentials. All these preconditions can be significantly generalized: On the one hand, the proposed algorithm can easily be generalized to higher dimensional observation sequences (this is mainly due to the fact that the above derivation of the solution of the Fokker-Planck equation can be generalized to higher dimensions, at least for harmonic potentials), see our forthcoming article [15]. On the other hand, we can allow for a larger class of potentials. For example, the entire derivation presented herein analogously goes through if the potential is a linear functional of its parameters. This, for example, is true for polynomial potentials; for this case one can even find parameter estimation procedures in the literature [29]. However, it is important to emphasize that the algorithmic concept

advocated herein tries to realize the so-to-say *simplest reduced dynamical model* for metastable complex systems: Markov jump processes and stochastic diffusion governed by harmonic potentials within each metastable state. As we will illustrate below, the potential dynamical substructure within each metastable state can hierarchically be represented in this setting such that the use of nonharmonic potentials may be obsolete. Furthermore, the use of nonharmonic (especially polynomial) potentials may lead to a drastic increase in the number of required parameters; this would lead to a very undesirable explosion in the dimension of the nonlinear parameter optimization problem.

1.2 Commentary

In this section will comment on the similarities and differences between the approach advocated herein and alternative approaches.

Hidden Markov Models. The algorithmic concept has some similarities with the other approaches based on the concept of HMMs or hidden Markov processes, in particular the approaches presented in [11, 13]. However, the fundamental difference is that the HmmSde approach suggested herein combines some discrete hidden process with in general continuous SDE output. That is, the concept behind HmmSde can be expressed shortly in the following way:

$$\text{HmmSde} \;=\; \text{SDE parametrization} \;+\; \text{HMM metastability analysis}, \qquad (8)$$

while [11] is concerned with HMM-based metastability analysis but with *stationary* output behavior only, and [13], e.g., considers global SDE models with hidden data (*without* discrete metastable states). In comparison to (and apart from the general "stationary" versus "dynamical" contrast) HMM-based metastability analysis, HmmSde should have at least one obvious additional feature: it can capture dynamical relaxation behavior within each metastable state.

Other SDE parametrizations. There is a huge variety of methods for parametrizing SDEs based on low-dimensional experimental or observation data, e.g., in (Kalman) filtering, image processing, and other aspects of time series analysis. However, the authors are not aware of approaches that are in depth comparable to the one suggested herein (i.e., mixing discrete and continuous descriptions in the sense of "formula" (8)). Furthermore, typical alternative methods of SDE modelling try to parametrize a single (hence "global") SDE (with or without hidden components) to the entire time series, cf. [13, 29, 28], or the literature on Kalman filtering for examples.

The application of "global" SDE parametrization techniques to the problem of identifying metastable states will in general be troublesome. To understand the reasons for this we have to distinguish two cases: (1) the observation sequence is as highly dimensional as the system under consideration, or (2) the observation sequence is low-dimensional compared to the system's state space. In case (1), the metastable states will be clearly separated from each other

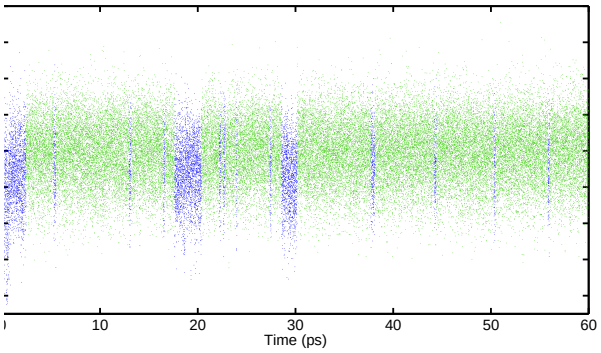

Figure 2: Time series of an intramolecular distance between the end-groups of a trialanine molecule (cf. Fig. 9 below) resulting from a molecular dynamics simulation (see Sect. 3 for further details). Note that the time series clearly indicates the presence of two overlapping metastable states (indicated by dark and light grey coloring of the time series graph).

(in this case they in fact are given by disjoint metastable sets, see [27]) but the complexity of the parametrization process and the statistical requirements on the time series (i.e. on its length) will be enormous. Furthermore, state-dependent noise intensities may be required (since the noise process may differ between the metastable sets). In case (2), due to the nature of low-dimensional projections, different metastable states in general will not be separated clearly but will "overlap", see Fig. 2. This means that their identification within the framework of fitting a *single* global SDE will be problematic. In this direction a combination of statistical model reduction techniques (like those in [17, 25]) with *local* SDE approximations seems to be desirable.

Set-oriented techniques and related approaches. Available set-oriented techniques for the identification of metastable states are based on the assumption that the metastable states are identified as disjoint metastable sets in state space [26, 7, 4, 5, 23]. Therefore, if applied to low-dimensional observations of a high-dimensional system, these techniques face the very same problem that was explained in the last paragraph (overlapping metastable states). The same is true for other related approaches [18, 22].

2 Parameter Estimation

We now construct algorithms that compute (approximate) solutions to the problems (1) and (2) from page 7.

Optimal parameters. To solve problem (1) we will use the *Expectation-Maximization* (EM) algorithm. The EM algorithm is a learning algorithm: it alternately iterates two steps, the expectation step and the maximization

step. Starting with some initial parameter set λ_0 the steps iteratively refine the parameter set, i.e., in step k the present parameter set λ_k is refined to λ_{k+1}. We will work out the details of the EM algorithm for the problem under investigation by following the general framework given in [2, 14]:

The key object of the EM algorithm is the expectation

$$Q(\lambda, \lambda_k) \;=\; \mathbf{E}\Big(\log p(O, q|\lambda) \mid O, \lambda_k \Big) \tag{9}$$

of the complete-data likelihood $\mathcal{L}(\lambda|O, q) = p(O, q|\lambda)$ (in our case given by (7)) wrt. the hidden sequence q given the observation sequence and the current parameter estimate λ_k. One step of the EM algorithm then realizes the following two steps:

- Expectation-step: This step evaluates the expectation value Q based on the given parameter estimate λ_k.
- Maximization-step: This step determines the refined parameter set λ_{k+1} by maximizing the expectation:

$$\lambda_{k+1} = \operatorname*{argmax}_{\lambda} Q(\lambda, \lambda_k). \tag{10}$$

  The maximization guarantees that $\mathcal{L}(\lambda_{k+1}) \geq \mathcal{L}(\lambda_k)$.

While the realization of the E-step for the problem under consideration requires the application of standard techniques only, the M-step requires a detailed analysis which will be performed subsequently.

Optimal sequence of hidden states. Problem (2) from Page 7 can be solved by applying the parameter set obtained from the EM algorithm to the standard Viterbi algorithm [31]. For given λ and O this algorithm computes the most probable hidden path $q^* = (q_1^*, \ldots, q_T^*)$. This path is called the *Viterbi path*. For an efficient computation we define the highest probability along a single path, for the first t observations, ending in the hidden state i at time t,

$$\delta_t(i) = \max_{q_0, q_1, \cdots q_{t-1}} P(q_0, q_1 \ldots q_t = i, O_0, O_1 \ldots O_t|\lambda).$$

This quantity is given by induction as

$$\delta_t(j) = \max_{1 \leq i \leq M} [\delta_{t-1}(i)\mathcal{T}(i, j)]\rho(O_t|q_t, O_{t-1}). \tag{11}$$

In addition, the argument i that maximizes (11) 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 [24].

Further Simplification. The formula (6) 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

$$\begin{aligned} x(t+\tau) &= O_t - D^{(q)}(O_t - \mu^{(q)})\tau && (12)\\ \Sigma(t+\tau) &= \frac{1}{2\tau}\, B^{(q)^{-1}} && (13)\\ A(t+\tau) &= \frac{1}{\sqrt{\pi}}\, \Sigma^{1/2}(t+\tau), && (14) \end{aligned}$$

which is not necessary but simplifies the following steps significantly.

Therefore, for given model parameters λ we have the following joint probability distribution for the observation and hidden state sequences:

$$\begin{aligned} p(O,q|\lambda) &= v(q_0)\rho(O_0|q_0)\prod_{t=1}^{T}\mathcal{T}(q_{t-1},q_t)\rho(O_t|q_t,O_{t-1}) \\ &= v(q_0)A^{(q_0)}(t)\exp\left(-(O_0 - x^{(q_0)}(0))\Sigma^{(q_0)}(0)(O_0 - x^{(q_0)}(0))^T\right) \\ &\quad \prod_{t=1}^{T}\mathcal{T}(q_{t-1},q_t)A^{(q_t)}(t)\exp\left(-(O_t - x^{(q_t)}(t))\Sigma^{(q_t)}(t)(O_t - x^{(q_t)}(t))^T\right). \end{aligned}$$

Due to (12) and (13) the Gaussian observation likelihood reduces to

$$\rho(O_t|q_t, O_{t-1}, ..., O_1) = \rho(O_t|q_t, O_{t-1}) = \frac{1}{(4\pi\tau^2)^{1/4}}\, B^{(q_t)^{-1/2}}\, \exp\left(-(O_t - x^{(q_t)})\frac{1}{2\tau}B^{(q_t)^{-1}}(O_t - x^{(q_t)})^T)\right), \quad (15)$$

with

$$x^{(q_t)} = (O_{t-1} - D^{(q_t)}(O_{t-1} - \mu^{(q_t)})\tau),$$

and the parameter-tuple is $\lambda = (v, \mathcal{T}, \mu, D, B)$.

Optimal parameters via the Maximum Likelihood Principle. We aim to estimate the parameters that maximize the expectation Q of the log-likelihood $\log \mathcal{L}(O, q|\lambda)$ of the complete data wrt. the hidden sequence q. According to [2] (Chap. 4.2) the expectation value Q as defined in (9) can be rewritten as

$$Q(\lambda|\lambda_k) = \sum_{q=(q_t)\in S^{T+1}} p(O,q|\lambda_k)\log\left(p(O,q|\lambda)\right), \quad (16)$$

where S denotes the state space of the hidden states. As we will see below this form will allow us to find very efficient maximizers. Due to Baum et al. [1] the function $Q(\cdot|\lambda_k)$ exhibits a unique maximum such that the new parameter iterate λ_{k+1} is uniquely determined by (10). To simplify notation we will use the notation $\lambda = (v, \mathcal{T}, \mu, D, B) = \lambda_k$ and $\hat{\lambda} = \lambda_{k+1}$ for the old and new parameter iterate, respectively.

In order to identify $\hat{\lambda} = (\hat{v}, \hat{\mathcal{T}}, \hat{\mu}, \hat{D}, \hat{B})$ we have to find the zeros of the partial derivatives of Q wrt. $\hat{v}, \hat{\mathcal{T}}, \hat{\mu}, \hat{D}$, and $\hat{B}$. Calculations and representation of these derivatives is made much easier by introducing the so-called forward-backward variables α_t, β_t [1, 24]:

$$\begin{aligned} \alpha_t(i) &= P(O_0, ..., O_t, q_t = i|\bar{\lambda}), \\ \beta_t(i) &= P(O_{t+1}, ..., O_T|q_t = i, O_t, \bar{\lambda}). \end{aligned}$$

These variables are recursively computable with numerical effort growing linearly in T and allow to compute the derivatives in compact form (see Appendix). Given these, we find $\hat{\lambda} = \operatorname{argmax} Q(\cdot|\lambda)$ to be uniquely given by

$$\hat{v}_i = \frac{\alpha_1(i)\beta_1(i)}{\sum_{i=1}^{M} \alpha_1(i)\beta_1(i)}, \tag{17}$$

$$\hat{\mathcal{T}}_{ij} = \frac{\sum_{t=0}^{T-1} \alpha_t(i)\bar{\mathcal{T}}(i,j)\rho(O_t|q_t, O_{t-1}, \bar{\lambda})\beta_{t+1}(j)}{\sum_{t=0}^{T-1} \alpha_t(i)\beta_t(i)} \tag{18}$$

and

$$\hat{\mu}^{(i)} = \frac{X_1X_2 - X_3X_4}{X_1X_4 - X_3X_5}, \tag{19}$$

$$\hat{D}^{(i)} = \frac{\sum_{t=2}^{T} \alpha_t(i)\beta_t(i)(O_t - O_{t-1})(O_{t-1} - \hat{\mu}^{(i)})}{-\tau \sum_{t=2}^{T} \alpha_t(i)\beta_t(i)(O_{t-1} - \hat{\mu}^{(i)})^2}, \tag{20}$$

$$\hat{B}^{(i)} = \frac{\sum_{t=2}^{T} \alpha_t(i)\beta_t(i)(-O_t + O_{t-1} - D^i(O_{t-1} - \hat{\mu}^{(i)})\tau)^2}{\tau \sum_{t=0}^{T} \alpha_t(i)\beta_t(i)} \tag{21}$$

with

$$X_1 = \sum_{t=1}^{T} \alpha_t(i)\beta_t(i)(O_t - O_{t-1}), \qquad X_2 = \sum_{t=1}^{T} \alpha_t(i)\beta_t(i)O_{t-1}^2,$$

$$X_3 = \sum_{t=1}^{T} \alpha_t(i)\beta_t(i)(O_t - O_{t-1})O_{t-1}, \quad X_4 = \sum_{t=1}^{T} \alpha_t(i)\beta_t(i)O_{t-1},$$

$$X_5 = \sum_{t=1}^{T} \alpha_t(i)\beta_t(i).$$

These formulas are applicable iteratively to approximate a maximum (not necessarily global) of $\mathcal{L}(O, q|\lambda)$. A detailed derivation of the formulas for $\hat{v}_i$ and $\hat{\mathcal{T}}_{ij}$ are to be found in [20]. The remainder is given in the Appendix.

Remark: The further simplification by means of Euler discretization 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; details will be published elsewhere [16].
Furthermore, we could also use alternative methods [28] for the parametrization of the SDEs in question; this may speed-up convergence of the parameter estimation and thus reduce the requirements on the length of the time series.

3 Results and Discussion

In this section we first want to illustrate the fundamental features and performance of the algorithm, its nice properties and problematic scenarios. Then, we will demonstrate how it can be applied to time series derived from molecular dynamics simulations of trialanine.

3.1 Illustrative Examples

In order to generate data sets for illustrative means, we generated time series from direct realizations of given models. For the first two test cases we used direct realizations of models of type (1) (parameters $(v, \mathcal{T}, \mu, D, B)$ known). Based on the output sequence of such realizations we will try to re-identify the parameters by application of the HmmSde identification algorithm based on the results of the last section. For the third test case we used a realization of a diffusive motion in a multi-well potential. Again, we will try to identify optimal parameters for systems of type (1).

In all numerical experiments the initial parameter guesses were based on the same procedure: The initial $M \times M$ transition matrix was chosen to be a stochastic matrix with off-diagonal entries 0.001 and identical diagonal entries. The model parameters were obtained by the re-estimation formulas (17)-(21), where $\alpha_t(i)$ and $\beta_t(i)$ were computed via the forward-backward algorithm based on randomized determination of the probabilities $P(O_t|q_t, O_{t-1})$ (they were chosen uniformly distributed on $[0, 1]$).

Case 1: Two coupled metastable sets. For the first test case we compute a realization of (1) for $M = 2$ states of the jump process with transition matrix

$$R = \begin{pmatrix} 0.997 & 0.003 \\ 0.003 & 0.997 \end{pmatrix}.$$

The parameters $(\mu^{(q)}, D^{(q)}, \sigma^{(q)} = \sqrt{B^{(q)}})$ of the two associated SDEs can be found in Table 1 in the column entitled "model". The time series resulted from discretization of the SDEs by the Euler-Maruyama scheme with stepsize $dt = 0.1$; each 10th step entered the time series ($\tau = 1.0$).

This table also displays the results of the re-estimation procedure for output sequences of length $N = 1.000$ and $N = 10.000$. We also included error estimates for the estimated parameters. These error estimates are computed by the following procedure: we simply repeat the re-estimation procedure $L = 1.000$ times (every time with new realizations of the observation sequence) and compute the means and standard deviation of the resulting ensemble of estimated parameters. The results nicely demonstrate that the re-estimation procedure converges to the correct values and that the error decays with the length of the available observation sequence. The estimated transition matrix comes out to be:

$$R_{N=10^4} = \begin{pmatrix} 0.9968 & 0.003 \\ 0.003 & 0.9969 \end{pmatrix} \pm \begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix} \cdot 10^{-4}.$$

First SDE	model	$T = 10^3$	$T = 10^4$
$\mu^{(1)}$	1	0.99 ± 0.18	1.0 ± 0.03
$D^{(1)}$	0.01	0.03 ± 0.025	0.01 ± 0.003
$\sigma^{(1)}$	0.02	0.025 ± 0.01	$0.02 \pm 2 \cdot 10^{-4}$
Second SDE	model		
$\mu^{(2)}$	1.1	1.11 ± 0.03	1.094 ± 0.008
$D^{(2)}$	0.1	0.1 ± 0.07	0.1 ± 0.005
$\sigma^{(2)}$	0.05	0.04 ± 0.01	$0.05 \pm 4 \cdot 10^{-4}$

Table 1: Parameters of the SDEs for test case 1: original data ("model") and re-estimated data (last two columns). Re-estimation was based on observation sequences of length 1.000 (middle) and 10.000 (right). The computation of the error of the re-estimation results are explained in the text below.

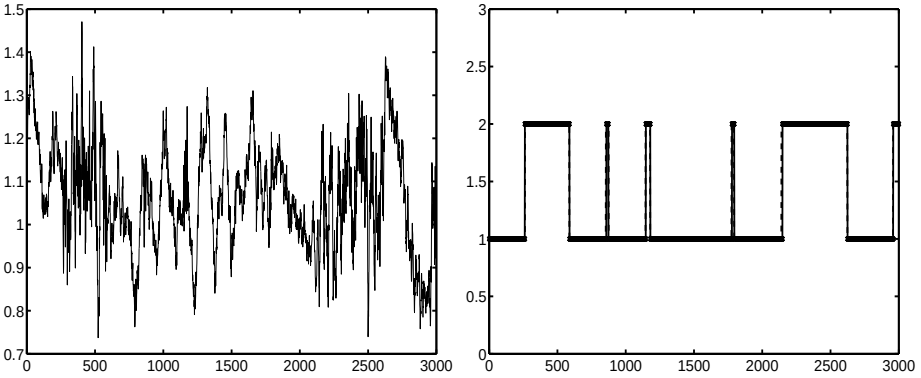

Figure 3: Realization of HmmSde model for two metastable sets with two different $D^{(1)} = 0.01, D^{(2)} = 0.1$ (left) against time and the comparison of the exact path of the jump process (dashed) with the result of the Viterbi algorithm (solid, right).

Figure 3 shows the realization itself: we observe that from the given data one cannot directly see metastability. It also displays the original (hidden) path of the Markov jump process and the one computed via the Viterbi algorithm. Although the jumps between metastable sets are not clear from the observation sequence the algorithm almost perfectly identified the hidden path.

Case 2: Three coupled states with two metastable subsets. Next we want to check whether the approach is able to detect a hierarchy of metastable states of different importance. Therefore we will analyse the case where the jump process has more states than metastable aggregates. To this end we compute a realization of (1) for $M = 3$ states of the jump process with transition matrix

$$R = \begin{pmatrix} 0.995 & 0.005 & 0 \\ 0 & 0.8 & 0.2 \\ 0.01 & 0.19 & 0.8 \end{pmatrix}.$$

This jump process obviously has 2 metastable aggregates: Ag_1 = state 1, and Ag_2 = states 2+3. The parameters $(\mu^{(q)}, D^{(q)}, \sigma^{(q)} = \sqrt{B^{(q)}})$ of the three associated SDEs can be found in Table 2. The time series again resulted from discretization of the SDEs by the Euler-Maruyama scheme with stepsize $dt = 0.1$; again each 10th step entered the time series ($\tau = 1.0$).

First SDE	model	$T = 10^4$	$T = 5 \cdot 10^4$
$\mu^{(1)}$	1	0.99 ± 0.1	1.0 ± 0.001
$D^{(1)}$	0.05	0.05 ± 0.01	0.05 ± 0.005
$\sigma^{(1)}$	0.03	0.03 ± 0.008	$0.03 \pm 4 \cdot 10^{-3}$
Second SDE	model		
$\mu^{(2)}$	1.2	1.2 ± 0.002	1.2 ± 0.001
$D^{(2)}$	0.08	0.08 ± 0.01	$0.08 \pm 1 \cdot 10^{-3}$
$\sigma^{(2)}$	0.005	$0.005 \pm 0.7 \cdot 10^{-3}$	$0.005 \pm 3 \cdot 10^{-5}$
Third SDE	model		
$\mu^{(3)}$	1.4	1.4 ± 0.01	1.4 ± 0.008
$D^{(3)}$	0.1	0.1 ± 0.002	$0.1 \pm 8 \cdot 10^{-4}$
$\sigma^{(3)}$	0.005	$0.005 \pm 0.7 \cdot 10^{-3}$	$0.005 \pm 2 \cdot 10^{-5}$

Table 2: Parameters of the SDEs for test case 2: original data ("model") and reestimated data (last two columns). Re-estimation was based on observation sequences of length 10.000 (middle) and 50.000 (right). The error estimates are computed as described above.

This table also displays the results of the re-estimation procedure for output sequences of length $T = 10.000$ and $T = 50.000$. We again included error estimates for the estimated parameters. The results again show that the re-estimation procedure converges to the correct values and that the error decays with increasing length of the available observation sequence. The estimated transition matrix comes out to be:

$$R_{N=10^4} = \begin{pmatrix} 0.985 \pm 0.02 & 0.004 \pm 0.0001 & 0.01 \pm 0.01 \\ 0.02 \pm 0.02 & 0.8 \pm 0.05 & 0.17 \pm 0.02 \\ 0.01 \pm 0.0004 & 0.15 \pm 0.05 & 0.84 \pm 0.05 \end{pmatrix}.$$

Figure 4 shows the original realization itself and displays the original (hidden) path of the Markov jump process aggregated wrt. Ag_1 and Ag_2 and the one computed via the Viterbi algorithm. Again, the algorithm almost perfectly identified the hidden path.

Case 3: Diffusive motion in multi-well potential. In this case we produce the output sequence by a realization of the diffusive motion given by

$$dx = -D_x V(x)\, dt + \eta dW, \tag{22}$$

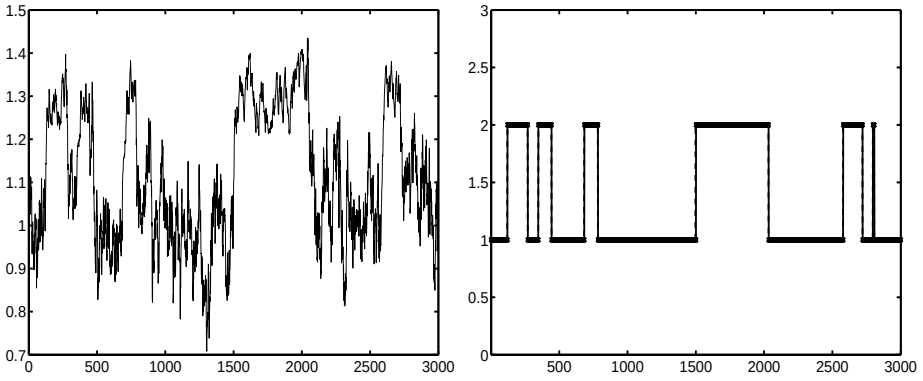

Figure 4: Realization of HMMSDE model with three coupled SDEs with two metastable aggregates Ag_1 and Ag_2 (left) versus time, and the comparison of the exact path of the Markov jump process between its metastable components (i.e., aggregates wrt. Ag_1 and Ag_2; dashed) with the result of the Viterbi algorithm (also aggregated; solid, right).

with potential

$$\begin{aligned} V(x) &= p(x) + \alpha \sin(\beta x), \\ p(x) &= \sum_{k=0}^{11} a_k x^k \\ a &= (0.300, 0.000, 5.187, -7.194, -25.507, 23.585, \\ &\qquad 59.049, -15.538, -63.822, -10.904, 24.794, 10.050) \\ (\alpha, \beta) &= (0.020, 50.000) \end{aligned}$$

This system will exhibit metastability between its wells whenever the noise amplitude η is small enough; see the illustration of the potential in Fig. 5. We will use $\eta = 0.85$ which clearly leads to metastable behavior, as we can see from the typical realization of the dynamics illustrated in Figure 5. Throughout this example the observation sequence comes from numerical simulation of (22) with the Euler-Maruyama scheme with discretization time step $dt = 0.005$; every second step entered the observation sequence, so the observation time step is $\tau = 0.01$.

We used this time series to train our HMMSDE model, and considered three cases with $M = 2, 3, 4$ hidden states for the jump process. In order to judge the quality of the results we may visit Figs. 6 and 7 that illustrate the harmonic potentials resulting from the optimal parametrization. We observe that for $M = 2$ the algorithm identifies the two most important metastable states (left and right of the main energy barrier). For $M = 3$ the algorithm further resolves the internal structure of the right metastable states; it identifies the deepest well on the RHS of the main barrier and approximates the middle part appropriately. For $M = 4$ the algorithm further decomposes the middle state into its slightly metastable parts. Fig. 7 moreover illustrates how the quality of the identification

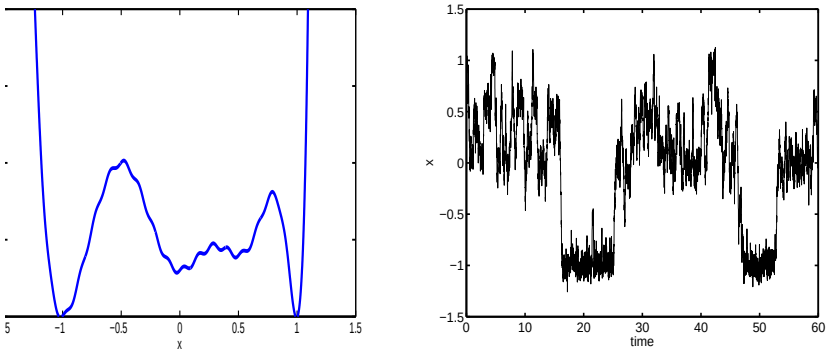

Figure 5: Right: Multi-well potential $V = V(x)$ as defined in the text. Left: Typical realization of the dynamics given by the SDE (22) with noise intensity $\eta = 0.85$. The time series is generated as described in the text, has length 6.000 and is plotted against time.

results deteriorates with decreasing length of the observation time series.

In order to demonstrate the quality of the transition matrix approximation by means of the HMMSDE model wrt. *metastable behavior* we calculate the first five eigenvalues λ_{HMMSDE} of the transition matrix for $M = 7$. We compare them with the corresponding five first eigenvalues λ_{box} of the HMMSDE transition matrix resulting from box-discretization into 50 equidistant boxes of the transfer operator $P_t = \exp(tA)$, $A = \frac{\eta^2}{2}\Delta - DV(x) \cdot D_x$ associated with the SDE (22) (see [27] for details).

$$\lambda_{\text{box}} = \begin{pmatrix} 1.000 \\ 0.997 \\ 0.986 \\ 0.943 \\ 0.883 \end{pmatrix}, \quad \lambda_{\text{HMMSDE}} = \begin{pmatrix} 1.000 \\ 0.997 \\ 0.982 \\ 0.938 \\ 0.890 \end{pmatrix} \pm \begin{pmatrix} 0 \\ 0.001 \\ 0.004 \\ 0.006 \\ 0.007 \end{pmatrix}. \tag{23}$$

where the standard variation of the eigenvalues results from 100 realizations of the HMMSDE procedure.

In order to be able to judge the quality of the assignment of states to the metastable states we compare the "true" hopping behavior of the original dynamics between the two main metastable states with that one identified by means of the Viterbi algorithm after HMMSDE parametrization with $M = 2$, see Fig. 8. We observe that the agreement is pretty good; small deviations between the two paths result from very short recrossing events of the barrier (cf. the graph of the time series in Fig. 5 and compare, e.g., the behavior around time $t = 28$).

3.2 Application to Molecular Data

Time series from molecular dynamics and Metropolis Monte Carlo simulations of biomolecules are often analyzed in terms of some essential torsion angles.

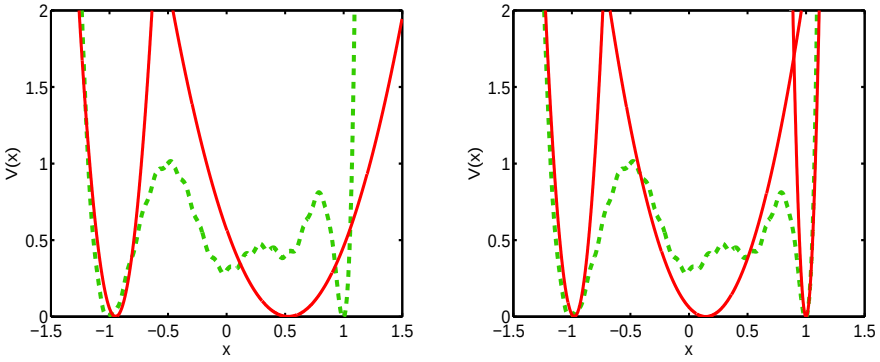

Figure 6: Multi-well potential (light grey, dashed), and harmonic potentials (solid) resulting from HmmSde parametrization with $M = 2$ (left) and $M = 3$ (right) hidden states based on the observation sequence of length 30.000 (see text). Remark: The energy at the minimum of the harmonic potentials is a free parameter (only the derivative of the potential enters the SDEs) and has been set to $V = 0$.

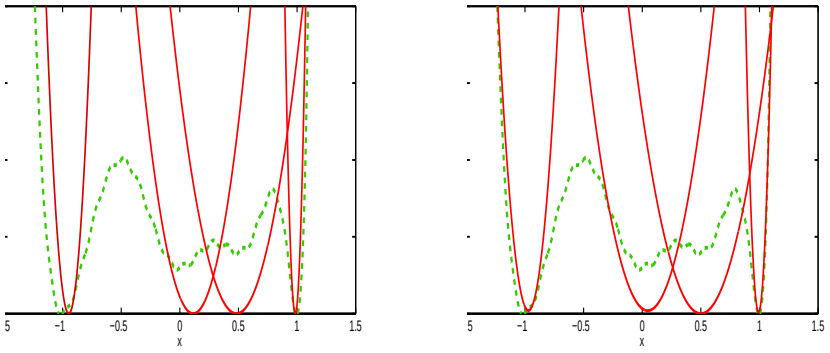

Figure 7: Multi-well potential (light grey, dashed), and harmonic potentials (solid) resulting from HmmSde parametrization with $M = 4$ hidden states based on observation time series of length 3.000 (left) and 30.000 (right). One observes that the identification based on the longer sequence results in very good approximations while the results of the identification based on the shorter sequence exhibit significant deviation from the expected outcome.

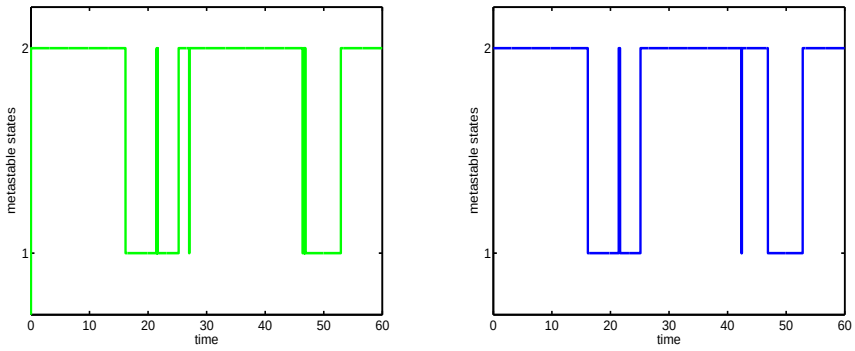

Figure 8: Jumps between the two dominant metastable states (indicated 1 and 2) versus time $t \in [0, 60]$ wrt. the time series (length 60.000) shown in Fig. 5. Left: As computed from the original time series (state 1 = $(x < x_0)$, state 2 =$(x \geq x_0)$, where $x_0 = -0.5$ is the position of the energy barrier separating the two states). Right: Viterbi path as computed based on HmmSde parametrization for $M = 2$.

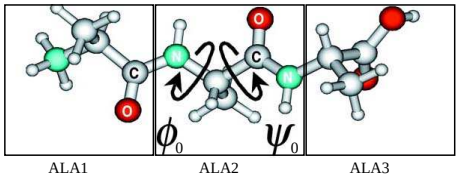

Figure 9: Illustration of the trialanine molecule and indication of the three alanine groups the molecule is composed of.

We illustrate that distances between the mass centers of functional groups also contain the information about metastable sets and can be used in conformational analysis of molecular dynamics data. In the following we will present the application of the proposed HmmSde algorithm to trialanine, a small peptide composed of three alanine amino acid residues. We will consider observation sequences of the one-dimensional distance between mass centers of the two amino-acid end-groups.

For the simulation of trialanine we used the Gromacs implementation of the Gromacs force field [19], in which trialanine is represented by 21 extended atoms. The structural and dynamical properties of this molecule normally are mainly determined by two central peptide backbone angles Φ_0 and Ψ_0 (see Fig. 9). In addition, at very high temperatures the other (otherwise planar) peptide bond angles (especially the Ω angles) may also undergo some conformational transition.

The time series of 544555 steps has been generated by means of Hybrid Monte Carlo (HMC) [3] at a temperature of 750 K. Each sub-trajectory for a HMC proposal step has a length of 0.1 ps and was computed with the Verlet integration scheme based on 1 fs time steps. This yielded an acceptance rate of

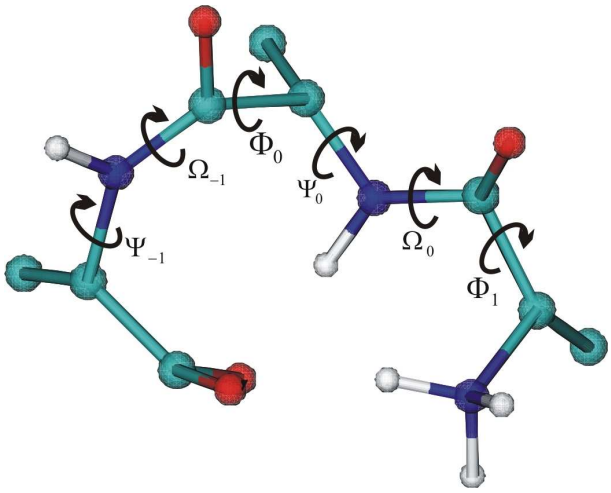

Figure 10: Illustration of the trialanine molecule and indication of the six torsion angles used for PCCA analysis as described in the text.

about 99 per cent.

By analysis of the time series of the six main torsion/peptide angles (see Fig. 10) using PCCA [6, 7] and some fine grid discretization we found three metastable states.

In order to get a useful illustration of the power of the HmmSde approach, we further reduced the information that is given to the algorithm: we constructed a one-dimensional observation sequence by simply computing the modulus of the distance vector between the centers of mass of the ALA1 and ALA3 end-groups (see Fig. 9). This observation time series is shown in Fig. 11. As illustrated in this figure the HmmSde approach (for $M = 4$) also identifies three (partially overlapping) metastable states. Figure 11 also contains information on the three SDEs within each single metastable state resulting from the HmmSde procedure; the resulting parameters are collected in Table 3 below. Comparison of the spread of the distance data within each of the metastable states with the distributions generated by the SDEs shows good agreement.

state j	$\mu^{(j)}$	$D^{(j)}$	$B^{(j)} = \sigma^{(j)^2}$
$j = 1$	0.384	0.93	0.0003
$j = 2$	0.409	0.96	0.0003
$j = 3$	0.419	0.98	0.0003

Table 3: Parameters of the SDEs from the HmmSde fitting of the ALA1-ALA3 distance in trialanine; the invariant distributions of the SDEs resulting from these values are illustrated in Fig. 11.

The resulting Viterbi path is illustrated in Fig. 12 (left) in comparison to the jump process between metastable states identified from the PCCA analy-

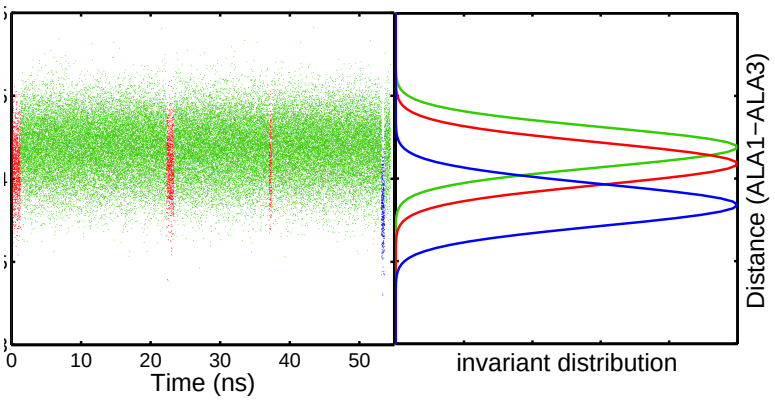

Figure 11: Left: Time series of the intramolecular ALA1-ALA3 distance in trialanine as introduced in the text. Different colorings (indicated by dark and light grey coloring of the time series graph) of phases of the time series according to the three metastable states determined by HmmSde . Right: Invariant distributions of the three one-dimensional SDEs resulting from the HmmSde algorithm versus the distance (y-axis).

sis of the full torsion angle time series (right). The agreement in fact is very convincing.

Conclusion

We introduced a novel approach to the identification of metastable states and the stochastic dynamics within each state for a complex (eventually highly dimensional) system from low-dimensional time series. The approach combines Hidden Markov models with optimally parametrized SDE output within each hidden state (to cope with temporal correlations within the states); its novelty mainly lies in the fact that it can serve as a tool for *automated* model reduction. We gave a detailed derivation for the case of one-dimensional time series, and demonstrated the successful application to different test cases.

However, we considered the one-dimensional case only. While the multi-dimensional generalization of the technique essentially poses no problem [15], we will have to ask whether the time series is "statistically rich enough" (cf. discussion on page 8) to reliably identify all the parameters of the multi-dimensional potentials (the more details the SDEs within each metastable state are supposed to represent correctly, the more data will in general be necessary to reliably parametrize). This problem will be investigated in a forth-coming paper [15]. However, we are interested in an approach that allows to identify metastability based on *low*-dimensional observations such that the problem of dimensionality might not be that problematic. In addition, in the fields of application like molecular dynamics or climate theory time series usually are extremely rich.

Furthermore, we herein assumed that the observation stepsize τ is small

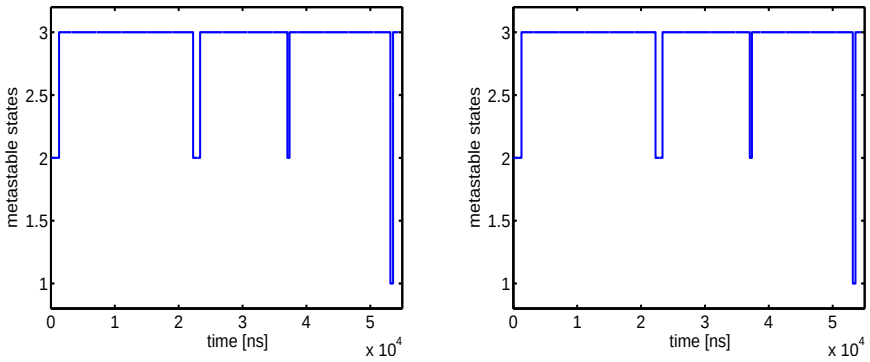

Figure 12: Jumps between the three dominant metastable states (indicated 1, 2 and 3) versus time wrt. the time series (length 544555 steps) shown in Fig. 11. Right: As computed from PCCA decomposition of the full six-dimensional torsion angle state space. Left: Viterbi path as computed based on HmmSde parametrization for the one-dimensional distance observation sequence for $M = 4$.

compared to the typical waiting time between hops of the jump process such that the sequence of hidden states will contain long subsequences in which that state does not change. In this case, the time series should in general be statistically rich enough to allow for a reliable parameter fit for the SDE parameters. Whether the proposed method is also able to successfully handle time series for which this assumption is not valid (e.g., the hops between the hidden states are frequent but the length of the time within each hidden state still is long enough) will also have to be discussed in later works; it will definitely be necessary to avoid the Euler discretization used herein (see [16]).

References

[1] L. E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities*, 3:1–8, 1972.

[2] J. A. Bilmes. A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and Hidden Markov Models. Technical report, International Computer Science Institute, Berkeley, 1998.

[3] A. Brass, B. J. Pendleton, Y. Chen, and B. Robson. Hybrid Monte Carlo simulations theory and initial comparison with molecular dynamics. *Biopolymers*, 33:1307–1315, 1993.

[4] M. Dellnitz and O. Junge. On the approximation of complicated dynamical behavior. *SIAM J. Num. Anal.*, 36(2):491–515, 1999.

[5] M. Dellnitz and R. Preis. Congestion and almost invariant sets in dynamical systems. In *Proceedings of Symbolic and Numerical Scientific Computation (SNSC'01)*, LNCS 2630, pages 183–209, 2003.

[6] P. Deuflhard, W. Huisinga, A. Fischer, and C. Schütte. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Lin. Alg. Appl.*, 315:39–59, 2000.

[7] P. Deuflhard and M. Weber. Robust Perron cluster analysis in conformation dynamics. ZIB-Report 03-19, Zuse Institute Berlin, 2003.

[8] E. Dittmer. Projizierte Hidden-Markov-Modelle in der Metastabilitätsanalyse hochdimensionaler Zeitreihen. Diploma thesis, Department of Mathematics and Computer Science, Free University Berlin, 2004.

[9] R. Elber and M. Karplus. Multiple conformational states of proteins: A molecular dynamics analysis of Myoglobin. *Science*, 235:318–321, 1987.

[10] E. Faou and C. Lubich. A Poisson integrator for Gaussian wavepacket dynamics. *Submitted to Comput. Visual. Sci.*, 2004.

[11] A. Fischer, S. Waldhausen, and C. Schütte. Identification of biomolecular conformations from incomplete torsion angle observations by Hidden Markov Models. *Submitted to J. Comput. Phys.*, 2004.

[12] H. Frauenfelder, P. J. Steinbach, and R. D. Young. Conformational relaxation in proteins. *Chem. Soc.*, 29A:145–150, 1989.

[13] J. Frydman and P. Lakner. Maximum likelihood estimation of hidden Markov processes. *Ann. Appl. Prob.*, 13(4):1296–1312, 2003.

[14] Z. Ghahramani. An introduction to hidden Markov models and Bayesian networks. *Int. J. Pattern Recognition and Artificial Intelligence*, 15(1):9–42, 2001.

[15] I. Horenko, E. Dittmer, and C. Schütte. Reduced SDE models for complex systems: The problem of dimensionality. *manuscript in preparation*, 2005.

[16] I. Horenko, E. Dittmer, and C. Schütte. Reduced stochastic models for complex molecular systems. *manuscript in preparation*, 2005.

[17] M. Hristache, A. Juditski, J. Polzehl, and V. Spokoiny. Structure adaptive approach for dimension reduction. *Annals of Statistics*, 29(6):1537–1566, 2001.

[18] D. Korenblum and D. Shalloway. Macrostate data clustering. *Phys. Rev. E*, 67, 2003.

[19] E. Lindahl, B. Hess, and D. der Spoel. Gromacs 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Mod.*, 7:306–317, 2001.

[20] L. A. Liporace. Maximum likelihood estimation for multivariate observations of Markov sources. *IEEE Trans. Informat. Theory*, 28(5):729–734, 1982.

[21] G. U. Nienhaus, J. R. Mourant, and H. Frauenfelder. Spectroscopic evidence for conformational relaxation in Myoglobin. *PNAS*, 89:2902–2906, 1992.

[22] M. Oresic and D. Shalloway. Hierarchical characterization of energy landscapes using Gaussian packet states. *J. Chem. Phys.*, 101:9844–9857, 1994.

[23] K. Padberg, R. Preis, and M. Dellnitz. Integrating multilevel graph partitioning with hierarchical set oriented methods for the analysis of dynamical systems. *Preprint 152 of the DFG Schwerpunktprogramm 1095*, 2004.

[24] L. R. Rabiner. A tutorial on Hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, 1989.

[25] A. Samarov, V. Spokoiny, and C. Vial. Component identification and estimation in nonlinear high-dimensional regression models. *Preprint 828, WIAS Berlin (www.wias-berlin.de)*, 2003.

[26] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard. A direct approach to conformational dynamics based on hybrid Monte Carlo. *J. Comput. Phys.*, 151:146–168, 1999.

[27] C. Schütte and W. Huisinga. Biomolecular conformations can be identified as metastable sets of molecular dynamics. In P. G. Ciaret and J.-L. Lions, editors, *Handbook of Numerical Analysis*, volume Computational Chemistry. North-Holland, 2003.

[28] I. Shoji and T. Ozaki. Comparative study of estimation methods for continuous time stochastic processes. *J. Time Ser. Anal.*, 18(5):485–506, 1997.

[29] V. N. Smelyanskiy, D. A. Timucin, A. Brandrivskyy, and D. G. Luchinsky. Model reconstruction of nonlinear dynamical systems driven by noise. *Submitted to Phys. Rev. Lett.*, 2004.

[30] A. Stuart and P. Wiberg. Parameter estimation for partially observed hyopelliptic diffusion. *submitted (available via www.maths.warwick.ac.uk/∼stuart)*, 2004.

[31] A. J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Informat. Theory*, IT-13:260–269, 1967.

[32] C. Wu. On the convergence properties of the EM algorithm. *Ann. Stat.*, 11(1):95–103, 1983.

[33] L. Wu and M. Jordan. On the convergence properties of the EM algorithm for Gaussian mixtures. *Neural Comput.*, 8:129–151, 1996.

Appendix

Forward-Backward Variables

The forward-backward variables are defined via

$$\begin{aligned} \alpha_t(i) &= P(O_0, ..., O_t, q_t = i) \\ \beta_t(i) &= P(O_{t+1}, ..., O_T | q_t = i, O_t) \end{aligned}$$

Together with

$$\mathcal{T}(i,j) = P(q_t = j | q_{t-1} = i)$$

we get the following recursion

$$\alpha_t(i) = \sum_{j=1}^{M} \alpha_{t-1}(j) \mathcal{T}(j,i) P(O_t | q_t = i, O_{t-1})$$

In order to derive this recursion, use the independencies on the far past (O_t depends on O_{t-1} only) to see that

$$P(O_t | O_0, q_0, ..., O_{t-1}, q_{t-1}, q_t = i) = P(O_t | O_{t-1}, q_t) \quad (24)$$

$$P(q_t = i | O_0, q_0, ...O_{t-1}, q_{t-1}) = P(q_t = i | q_{t-1}), \quad (25)$$

and directly compute:

$$\begin{aligned}
\alpha_t(i) &= P(O_0,...,O_t,q_t=i)\\
&= \sum_{q_0,\dots,q_{t-1}} P(q_0,O_0,...,q_{t-1},O_{t-1},q_t=i,O_t)\\
&= \sum_{q_0,\dots,q_{t-1}} P(q_0,O_0,...,q_{t-1},O_{t-1},q_t=i)P(O_t|O_0,q_0,...,O_{t-1},q_{t-1},q_t=i)\\
&=^{(24)} \sum_{q_0,\dots,q_{t-1}} P(q_0,O_0,...,q_{t-1},O_{t-1})P(q_t=i|q_0,O_0,...,q_{t-1},O_{t-1})P(O_t|O_{t-1},q_t)\\
&=^{(25)} \sum_{q_0,\dots,q_{t-1}} P(q_0,O_0,...,q_{t-1},O_{t-1})P(q_t=i|q_{t-1})P(O_t|O_{t-1},q_{t-1})\\
&= \sum_{q_0,\dots,q_{t-2}}\sum_{j=1}^{M} P(q_0,O_0,...,q_{t-1}=j,O_{t-1})P(q_t=i|q_{t-1}=j)P(O_t|O_{t-1},q_{t-1}=j)\\
&= \sum_{j=1}^{M} P(O_0,...,O_{t-1},q_{t-1}=j)\mathcal{T}(i,j)P(O_t|q_t=j,O_{t-1})\\
&= \sum_{j=1}^{M} \alpha_{t-1}(j)\mathcal{T}(i,j)P(O_t|q_t=j,O_{t-1}),
\end{aligned}$$

For β we get similarly

$$\begin{aligned}
\beta_t(i) &= P(O_{t+1},...,O_T|q_t=i,O_t)\\
&= \sum_{q_{t+1},\dots,q_T} P(q_{t+1},O_{t+1},...,q_T,O_T|q_t=i,O_t)\\
&= \sum_{q_{t+1},\dots,q_T} P(O_{t+1}|q_{t+1},O_t)P(q_{t+1}|q_t=i)P(q_{t+2},O_{t+2},...,q_T,O_T|q_{t+1},O_{t+1})\\
&= \sum_{q_{t+2},\dots,q_T}\sum_{j=1}^{M} P(O_{t+1}|q_{t+1}=j,O_t)P(q_{t+1}=j|q_t=i)P(q_{t+2},O_{t+2},...,q_T,O_T|q_{t+1}=j,O_{t+1})\\
&= \sum_{j=1}^{M} \mathcal{T}(i,j)P(O_{t+1}|q_t=j,O_t)\beta_{t+1}(j).
\end{aligned}$$

Derivatives of Likelihood Q

We derive partial derivatives of the likelihood Q defined in (16), which then are used to compute the unique maximum of $Q(\cdot,\bar{\lambda})$.

Using the forward-backward variables defined above, direct calculations yield:

$$\begin{aligned}
\frac{\partial Q}{\partial \mu^{(i)}} &= \sum_{q\in S^T} \mathcal{L}(\bar{\lambda}|O,q) \sum_{t=1,q_t=i}^{T} (O_t - O_{t-1} + D^i(O_{t-1}-\mu^i)\tau)B^{(i)^{-1}}D^i/2\Big) \\
&= \sum_{t=1}^{T} \underbrace{\sum_{q\in S^T} \mathcal{L}(\bar{\lambda}|O,q_0,...,q_t=i,...,q_T)}_{=\alpha_t(i)\beta_t(i)} (O_t - O_{t-1} + D^i(O_{t-1}-\mu^i)\tau)B^{(i)^{-1}}D^i/2 \\
\frac{\partial Q}{\partial D^{(i)}} &= \sum_{t=1}^{T} \alpha_t(i)\beta_t(i)(-O_t + O_{t-1} - D^i(O_{t-1}-\mu^i)\tau)B^{(i)^{-1}}(O_{t-1}-\mu^{(i)})/2 \\
\frac{\partial Q}{\partial B^{(i)}} &= \frac{1}{2}\Big[\sum_{t=0}^{T} -\alpha_t(i)\beta_t(i)B^{(i)^{-1}} + \\
&\quad \sum_{t=1}^{T} \alpha_t(i)\beta_t(i)(-O_t + O_{t-1} - D^i(O_{t-1}-\mu^i)\tau)^2/\tau\Big]
\end{aligned}$$

The maximum is computed via the zeros of this derivatives as follows:

$$\begin{aligned}
\frac{\partial Q}{\partial \mu^{(i)}} &= 0 \\
\Leftrightarrow \mu^i &= \frac{\sum_{t=1}^{T} \alpha_t(i)\beta_t(i)(O_t - O_{t-1} + D^i O_{t-1}\tau)}{D^i\tau\sum_{t=1}^{T}\alpha_t(i)\beta_t(i)}, \qquad (26) \\
\frac{\partial Q}{\partial D^{(i)}} &= 0 \\
\Leftrightarrow D^i &= \frac{\sum_{t=1}^{T} \alpha_t(i)\beta_t(i)(O_t - O_{t-1})(O_{t-1}-\mu^{(i)})}{-\tau\sum_{t=1}^{T}\alpha_t(i)\beta_t(i)(O_{t-1}-\mu^{(i)})^2}, \qquad (27) \\
\frac{\partial Q}{\partial B^{(i)}} &= 0 \\
\Leftrightarrow B^{(i)} &= \frac{\sum_{t=1}^{T}\alpha_t(i)\beta_t(i)}{\tau\sum_{t=0}^{T}\alpha_t(i)\beta_t(i)}(-O_t + O_{t-1} - D^{(i)}(O_{t-1}-\mu^{(i)})\tau)^2.
\end{aligned}$$

The parameters $\mu^{(i)}$ and $D^{(i)}$ are independent of $B^{(i)}$. Hence we have to solve (26) and (27):

$$\Rightarrow^{(26),(27)} \frac{\sum_{t=1}^{T} \alpha_t(i)\beta_t(i)(O_t - O_{t-1})(O_{t-1}-\mu^{(i)})}{-\tau\sum_{t=1}^{T}\alpha_t(i)\beta_t(i)(O_{t-1}-\mu^{(i)})^2} = \frac{\sum_{t=1}^{T} \alpha_t(i)\beta_t(i)(O_t - O_{t-1})}{\tau\sum_{t=1}^{T}\alpha_t(i)\beta_t(i)(O_{t-1}-\mu^{(i)})}$$

$$\Leftrightarrow \qquad \mu^i = \frac{X_1X_2 - X_3X_4}{X_1X_4 - X_3X_5}$$

with

$$X_1 = \sum_{t=1}^{T}\alpha_t(i)\beta_t(i)(O_t - O_{t-1}), \qquad X_2 = \sum_{t=1}^{T}\alpha_t(i)\beta_t(i)(O_{t-1}^2),$$

$$X_3 = \sum_{t=1}^{T}\alpha_t(i)\beta_t(i)(O_t - O_{t-1})O_{t-1}, \quad X_4 = \sum_{t=1}^{T}\alpha_t(i)\beta_t(i)O_{t-1},$$

$$X_5 = \sum_{t=1}^{T}\alpha_t(i)\beta_t(i).$$