MULTISCALE MODEL. SIMUL.
Vol. 5, No. 3, pp. 802–827

AUTOMATED MODEL REDUCTION FOR COMPLEX SYSTEMS EXHIBITING METASTABILITY*

ILLIA HORENKO†, EVELYN DITTMER†, ALEXANDER FISCHER‡, AND CHRISTOF SCHÜTTE†

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 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.

Key words. Bayesian networks, biomolecular conformations, hidden Markov model, maximum likelihood principle, metastability, stochastic differential equations

AMS subject classifications. 05D40, 60H10, 62M05, 60J60

DOI. 10.1137/050623310

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 [15, 18], while upon closer inspection it exhibits a rich temporal multiscale structure [37]. 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 [45, 46]. 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 shorter 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 shorter than the waiting times between hops (and thus much shorter than the time scale 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 [45, 12, 10, 11]. 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 is then based

*Received by the editors January 26, 2005; accepted for publication (in revised form) April 21, 2006; published electronically September 26, 2006. This work was supported by the DFG research center "Mathematics for Key Technologies" (FZT 86) in Berlin.

http://www.siam.org/journals/mms/5-3/62331.html

†Institute of Mathematics II, Free University Berlin, Arnimallee 2–6, 14195 Berlin, Germany (horenko@math.fu-berlin.de, dittmer@math.fu-berlin.de, schuette@math.fu-berlin.de).

‡Courant Institute, New York University, 251 Mercer Street, New York, NY 10012 (fischer@cims.nyu.edu). This author's research was supported by NSF grant DMS02-39625.

on analysis of this transition matrix [46, 13, 11]. 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 approximating 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 $W(t)$ denotes standard Brownian motion, D_x denotes differentiation w.r.t. x, $\Sigma = (\sigma^{(1)}, \ldots, \sigma^{(M)})$ contains noise intensities, and $\mathcal{V} = (V^{(1)}, \ldots, V^{(M)})$ contains 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 shorter 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 $\mathcal{V}$ contains only 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 are 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 are then 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 that is somewhat similar to standard approaches in the parameter estimation of partially observed stochastic processes (see section 1.2).

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 [21, 35] (in our case the "hidden" sequence of metastable states). The hidden Markov model (HMM) [21, 43] is one of the most popular representatives of a Bayesian network. 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.

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 SDEs. Equation (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.

We want to emphasize that the literature on dynamical Bayesian networks and related topics already contains several independent algorithmic approaches which—*under specific assumptions and if properly combined*—would allow an efficient parametrization of metastable time series: One would have to combine specific variants of the EM and Viterbi algorithms that are standard techniques in the field [34, 47] with other standard time series models such as ARIMA processes [6, 36] (e.g., under certain assumptions on linearity and discretization with constant time steps); see section 1.2 for details. We will propose a very direct approach which permits a rather explicit derivation. We claim that the *concept* of combining model reduction of metastable systems with dynamical Bayesian models and our idea for determining the number of necessary metastable states are new and straightforward. Moreover, its realization (almost) requires only assumptions that are natural for the application context considered herein. There is only one step, the application of the Euler scheme for discretizing the SDEs in (1), that is not natural; however, the derivation will show that this discretization step can be avoided completely which leads to a more complicated algorithmic scheme with the extra cost of additional Newton iterations (see [23] for the details of a discretization-free algorithm). In addition, there are very recent results on exact and efficient estimators for discretely observed diffusion processes that exploit the recent advances in simulation methodology for diffusions (exact simulation of diffusion paths) [2]. Furthermore, by additionally applying standard generalizations of the EM scheme we could even handle cases with nonconstant time stepping; one would simply have a slightly more complicated discrete-time model.

After these remarks concerning the background of the statistical techniques used we want to add a few sentences on the dimensionality issue since it always is crucial for complex systems from, e.g., molecular dynamics, climate theory, or materials science: 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 is often 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 section 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 $q(t_j)$ the system occupies at

times t_j are known already, and we have to define the functional w.r.t. which we will then 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:

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

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:

$$\partial_t \rho = \triangle_x V^{(q)}(x) \rho + \nabla_x V^{(q)}(x) \cdot \nabla_x \rho + \frac{1}{2} B^{(q)} \triangle_x \rho, \tag{3}$$

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 self-adjoint 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 [16]) 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. In the one-dimensional case we get

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

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

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

In the 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 in

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

with

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

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

Likelihood function. Whenever we assume the potential is 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, and 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 $t = 0, \ldots, T$. 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 ijth 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:

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

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

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

where the notation $\mathcal{L}(\cdot; O; q)$ and similar notations in the following are used to indicate the *parametric* dependence of the likelihood on the sequences O and q.

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

(1) determine optimal parameters $(\mathcal{T}, \mu^q, D^q, B^q)_{q=1,\dots,M}$ by maximizing the likelihood $\mathcal{L}(\lambda; O; q)$, which 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 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 belong to the same metastable state). The identification of an optimal aggregation from observation of the dynamical behavior is an important algorithmic problem. There are no general solutions to this problem, and the best way to handle this problem is often 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. [35]. 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 [10, 45]. This has led to the construction of an aggregation technique called Perron cluster cluster analysis (PCCA) [12, 13].

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 of selecting *in advance* the number M of hidden states. Since our goal is to identify metastable states we can proceed as suggested in [17]: 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 w.r.t. 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) [14]

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

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 [52, 53]. 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 a copious literature on the subject and how to address it; see, for instance, [34]. 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 such as the ones considered herein the precise estimation of drift parameters (or the parameters in the potential) requires rather long observation sequences (cf. section 1.1 of [42]). 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 not only on the time series but also on the details of the algorithmic scheme under consideration. We will herein not address this problem theoretically (see [23] for details) but will give some examples in section 3.

Dimensionality and nonharmonic potentials. In this article we introduce the HmmSde approach for one-dimensional observations and therefore one-dimensional SDEs. Both preconditions are not necessary: 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 [24]. 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 [49].

Why use harmonic potentials? In addition to the above paragraph, it is important to emphasize that the algorithmic concept advocated herein tries to realize the so-called *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 (see Case 3 in section 3.1), 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. However, there are very recent results on exact and efficient estimators for discretely observed diffusion processes [2] that would allow us to even consider nonharmonic potentials within our present algorithmic context; how these would computationally behave in high dimensions will be a question of further investigation.

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

HMMs and time series models. The algorithmic concept has some similarities with other approaches based on the concept of HMMs or hidden Markov processes, in particular the approaches presented in [17, 19, 36]. 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:

$$\text{HmmSde} = \text{SDE parametrization} + \text{HMM metastability analysis.} \tag{8}$$

Concerning *HMM-based metastability analysis* several other approaches exist, all of them with different focus: For example, [17] considers *stationary* output behavior only, and [19] considers global SDE models with hidden data but *without* discrete metastable states.

With *SDE parametrization* we mean the process of identifying the parameters in the jth SDEs in (1) from a piece of the time series that belongs to the jth metastable state. Whenever the time series given is discrete in time with constant sampling time and can be brought into the form of a stationary process (perhaps after a finite repetition of differencing), we enter in the well-known and general framework of autoregression integrable moving average models (ARIMA) [6]. However, one cannot generalize such models to nonharmonic potentials. Moreover, the successful parametrization of an ARIMA model is not the same as "SDE parametrization" since the physical interpretation may be substantially different.

Concerning the *combination* of parametrization and metastability analysis, the literature on switching Markov chains [22, 36] conceptually goes in the direction of the approach intended herein: coupling sudden switches between "trends" or "regimes" with estimation of the parametric models within each "regime" or "trend." However, in the related literature other application contexts are discussed, and the approaches closest to our intentions again use ARIMA parametrization and thus face the same difficulties as commented upon above.

Regarding closeness to our approach, the article that has to be named first is [29]. Obviously this approach has been developed simultaneously to the present approach and also tackles conformational dynamics of single molecules. However, it considers single-molecule fluorescence lifetime experiments instead of molecular dynamics time series, and the coupling between the discrete switch process and the continuous diffusion process is exactly the other way around: The diffusion changes the transition rates in the switch process as a model for fluctuating energy barriers in the molecule instead of as in our case, in which the discrete switches between conformations change the parameters of the internal SDE dynamics within each conformation.

Another link to literature combining diffusions with discrete components is given by [5], which in addition even goes into the direction of avoiding discretization in time. However, in that article, the observations are made on *both* components (diffusion *and* discrete process) so that neither component is completely hidden.

Instead of exploiting the algorithmic approaches mentioned above, we will derive specific variants of the EM and Viterbi algorithms directly for our reduced model (1). Our reasons for this are the following: (A) our HmmSde algorithm allows us to directly implement metastability issues; (B) its derivation is direct and explicit; and (C) this derivation clearly exhibits possible generalizations within the same interpretation context. The last reason may be rather abstract at this point of the manuscript. We therefore add a remark where this is discussed in more detail for the most important generalization; see the end of section 2.

Related approaches based on filtering, smoothing, or SDE parametrizations. There is a huge variety of methods to approach partially observed stochastic processes, for example, in filtering or smoothing (e.g., in [30, 4, 8, 33, 54], mentioning recent works only), robust discretization of filters [26], or control of stochastic processes [9, 54] with a wide spectrum of application backgrounds. Within this variety there also is a rich literature on using HMM techniques including the EM algorithm; c.f., e.g., [33, 26, 19] for recent work. However, we are not aware of other articles on parametrization of stochastic processes with an explicit focus on metastable dynamics in the sense of "formula" (8). There indeed are approaches to parameter estimation for systems of SDEs where not every dimension is observable [20, 27], but then there is no finite state space Markov chain.

In addition to the manifold of techniques mentioned above, another class of typical alternative methods of SDE modeling tries to parametrize a *single* (hence "global") equation to the entire time series (cf. [49, 48]) which could be used for the problem of interest herein by fitting multiwell potentials to the observation time series. The application of such "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: Either (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 (in this case they in fact are given by disjoint metastable sets; see [46]), 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 Figure 1. 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 [25, 44]) with *local* SDE approximations seems to be desirable.

Set-oriented techniques. 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 [45, 13, 10, 11, 39]. 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).

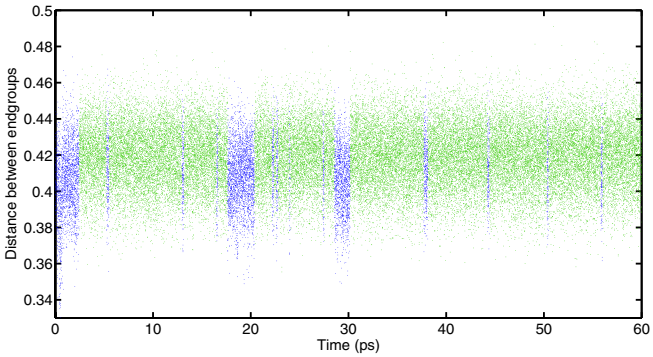

FIG. 1. *Time series of an intramolecular distance between the end-groups of a trialanine molecule (cf. Figure 8) resulting from a molecular dynamics simulation (see section 3 for further details). Note that the time series clearly indicates the presence of two overlapping metastable states (indicated by dark and light coloring of the time series graph).*

Other algorithmic approaches to molecular dynamics using Gaussian metastates. In the physics and chemistry literature one can find several approaches to molecular dynamics also using Gaussian metastates for efficient optimization [40, 41], smoothed dynamics [50], or renormalization of the dynamics [28, 38]. While the former two approaches bear almost no similarity to the approach presented herein (they work in full-dimensional state space and try to simplify optimization/simulation by smoothing the force fields/potentials by means of spatial averages with Gaussian metastates), the latter approach by Shalloway is somewhat similar. Like the set-oriented techniques mentioned above, Shalloway's approach also exploits the hierarchical spectral properties of the transfer operator to design Gaussian metastates that allow us to describe the effective dynamics of the molecular system under investigation. However, it does *not* allow us to deal with reduced dimensions, and it definitely has *nothing* to do with time series analysis, thus being distinctly different from HMMSDE.

2. Parameter estimation. We now construct algorithms that compute (approximate) solutions to the problems (1) and (2) in section 1.1.

Optimal parameters. To solve problem (1) we will use the 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 [3, 21].

The key object of the EM algorithm is the expectation

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

of the complete-data likelihood $\mathcal{L}(\lambda;O;q) = p(O,q|\lambda)$ (in our case given by (7)) w.r.t. 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 expectation step for the problem under consideration requires the application of standard techniques only, the maximization step requires a detailed analysis which will be performed subsequently.

Optimal sequence of hidden states. Problem (2) in section 1.1 can be solved by applying the parameter set obtained from the EM algorithm to the standard Viterbi algorithm [51]. 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,

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

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 [43].

Further simplification. The formula (6) for the parameters of the output distribution can be further simplified by the assumption that we want to know only about the evolution of the system within a short time interval $[t, t+\tau)$. We can then apply an Euler discretization resulting in

$$x(t+\tau) = O_t - D^{(q)}(O_t - \mu^{(q)})\tau, \tag{12}$$

$$\Sigma(t+\tau) = \frac{1}{2\tau} B^{(q)^{-1}}, \tag{13}$$

$$A(t+\tau) = \frac{1}{\sqrt{\pi}} \Sigma^{1/2}(t+\tau), \tag{14}$$

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:

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

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

$$\begin{aligned}\rho(O_t|q_t, O_{t-1}, \dots, O_1) &= \rho(O_t|q_t, O_{t-1}) \\ &= \frac{1}{(4\pi\tau^2)^{1/4}} B^{(q_t)^{-1/2}} \exp\left(-(O_t - x^{(q_t)})\frac{1}{2\tau} B^{(q_t)^{-1}}(O_t - x^{(q_t)})^T)\right),\end{aligned} \tag{15}$$

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 w.r.t. the hidden sequence q. According to [3] (Chap. 4.2) the expectation value Q as defined in (9) can be rewritten as

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

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 [1], among others, 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 w.r.t. $\hat{v}, \hat{\mathcal{T}}, \hat{\mu}, \hat{D}$, and $\hat{B}$. Calculations and representation of these derivatives are made much easier by introducing the so-called forward-backward variables α_t, β_t [1, 43]:

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

These variables are recursively computable with numerical effort growing linearly in T and allow us to compute the derivatives in compact form (see the 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

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

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

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

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 $\mathcal{L}(\lambda; O; q)$. A detailed derivation of the formulas for $\hat{v}_i$ and $\hat{\mathcal{T}}_{ij}$ are to be found in [32]. The remainder is given in the appendix.

Remark. The further simplification (12)–(14) 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 [23]. When using the Euler discretization with constant time step as above, we could also apply ARIMA time series models in combination with an appropriate EM approach.

3. Results and discussion. In this section we first want to illustrate the fundamental features and performance of the algorithm, its nice properties, and its 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 reidentify 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 multiwell 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 reestimation 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

TABLE 1

Parameters of the SDEs for Case 1: *original data ("model") and reestimated data (last two columns). Reestimation was based on observation sequences of length* 1.000 *(middle) and* 10.000 *(right). The computation of the error of the reestimation results are explained in the text below.*

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	$0.05 \pm 4 \cdot 10^{-4}$

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 reestimation 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 reestimation 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 reestimation 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

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

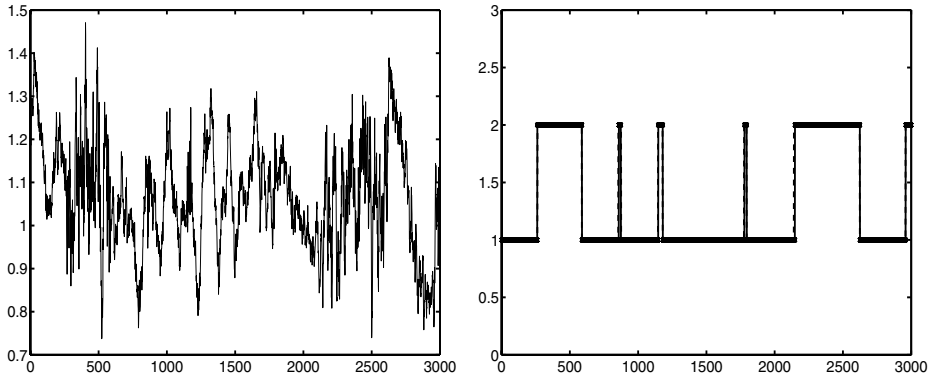

FIG. 2. *Realization of* HMMSDE *model for two metastable sets with two different* $D^{(1)} = 0.01, D^{(2)} = 0.1$ *(left) against time and the comparison of the exact path of the jump process (dashed) with the result of the Viterbi algorithm (solid) (right).*

Figure 2 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 with different importance. Therefore we will analyze 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 two metastable aggregates: $\mathrm{Ag}_1 =$ state 1 and $\mathrm{Ag}_2 =$ 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 ($\tau = 1.0$).

TABLE 2

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

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	$0.005 \pm 0.7 \cdot 10^{-3}$	$0.005 \pm 2 \cdot 10^{-5}$

This table also displays the results of the reestimation 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 reestimation 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 3 shows the original realization itself and displays the original (hidden) path of the Markov jump process aggregated w.r.t. 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 multiwell potential.* In this case we produce the output sequence by a realization of the diffusive motion given by

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

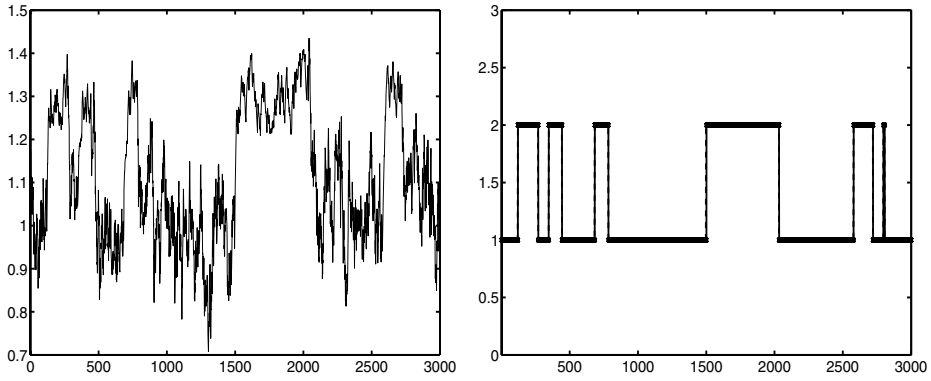

FIG. 3. *Realization of* HMMSDE *model with three coupled SDEs with two metastable aggregates* Ag_1 *and* Ag_2 *(left) versus time and the comparison of the exact path of the Markov jump process between its metastable components (i.e., aggregates w.r.t.* Ag_1 *and* Ag_2*; dashed) with the result of the 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 Figure 4. 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 4. 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 Figures 5 and 6, which 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 right-hand side 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. Figure 6, moreover, illustrates how the quality of the identification 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 w.r.t. *metastable behavior* we calculate the first five eigenvalues $\lambda_{\mathrm{HMMSDE}}$ of the transition matrix for $M = 7$. We compare them with the corresponding first five eigenvalues λ_{box} of the HMMSDE transition matrix re-

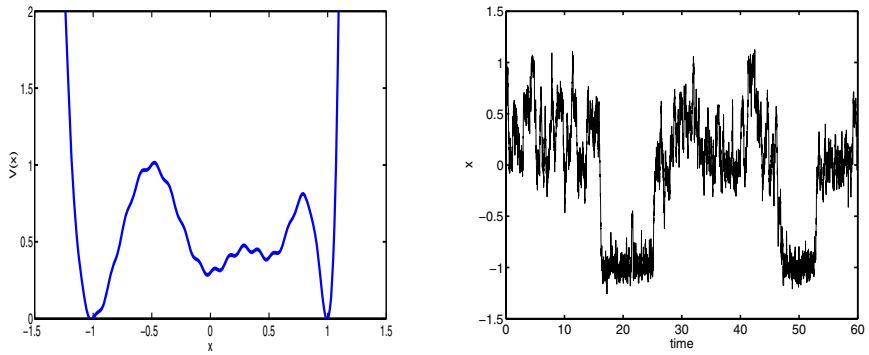

FIG. 4. *Right: multiwell potential $V = V(x)$ as defined in the text. Left: typical realization of the dynamics given by the SDE* (22) *with noise intensity $\eta = 0.85$. The time series is generated as described in the text, has length* 6.000*, and is plotted against time.*

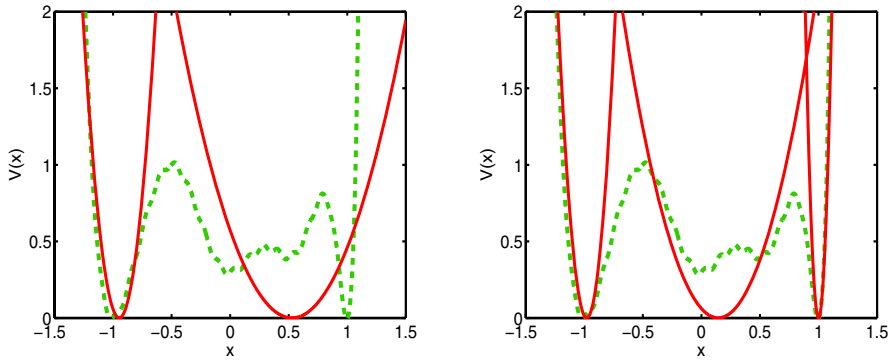

FIG. 5. *Multiwell potential (dashed) and harmonic potentials (solid) resulting from* HMMSDE *parametrization with $M = 2$ (left) and $M = 3$ (right) hidden states based on the observation sequence of length* 30.000 *(see text). Remark: The energy at the minimum of the harmonic potentials is a free parameter (only the derivative of the potential enters the SDEs) and has been set to $V = 0$.*

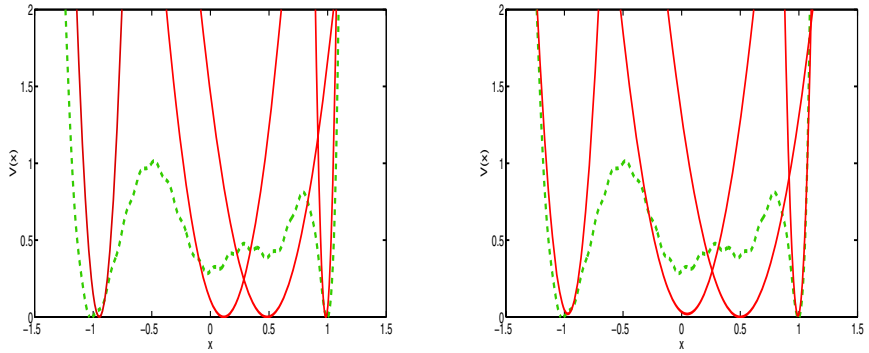

FIG. 6. *Multiwell potential (dashed) and harmonic potentials (solid) resulting from* HMMSDE *parametrization with $M = 4$ hidden states based on observation time series of length* 3.000 *(left) and* 30.000 *(right). One observes that the identification based on the longer sequence results in very good approximations, while the results of the identification based on the shorter sequence exhibit significant deviation from the expected outcome.*

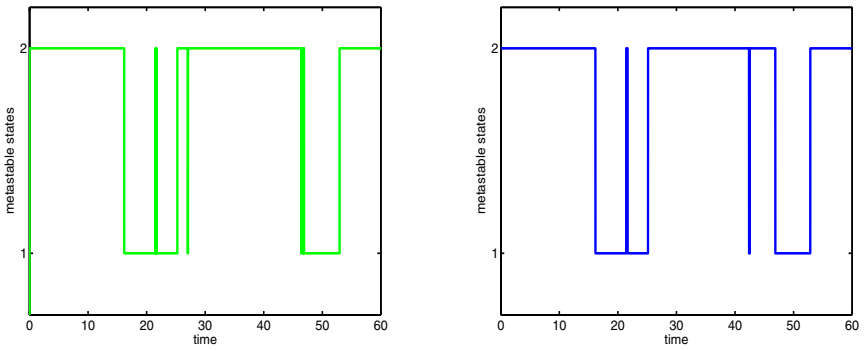

FIG. 7. *Jumps between the two dominant metastable states (indicated by* 1 *and* 2*) versus time* $t \in [0, 60]$ *w.r.t. the time series (length* 60.000*) shown in Figure* 4*. Left: as computed from the original time series (state* $1 = (x < x_0)$*, state* $2 = (x \geq x_0)$*, where* $x_0 = -0.5$ *is the position of the energy barrier separating the two states). Right: Viterbi path as computed based on* HMMSDE *parametrization for* $M = 2$*.*

sulting 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 [46] 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 the hopping behavior identified by means of the Viterbi algorithm after HMMSDE parametrization with $M = 2$; see Figure 7. 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 Figure 4 and cf., 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. 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 [31], in which trialanine is represented by 21 extended atoms. The structural and dynamical properties of this molecule normally are mainly determined

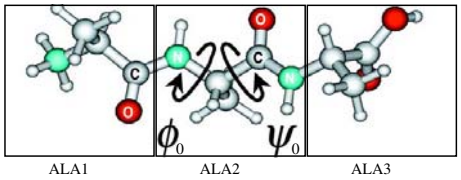

FIG. 8. *Illustration of the trialanine molecule and indication of the three alanine groups of which the molecule is composed.*

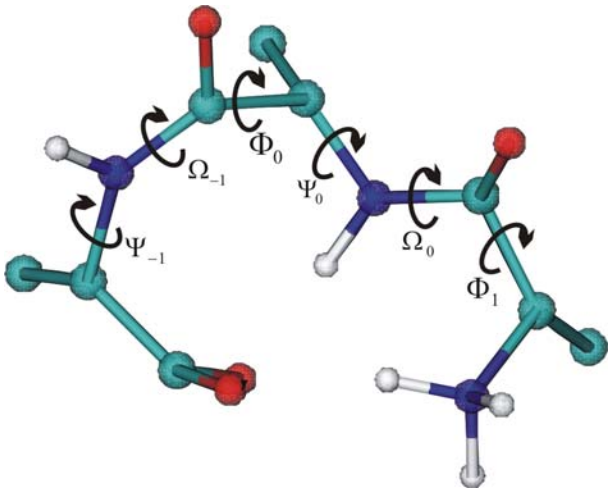

FIG. 9. *Illustration of the trialanine molecule and indication of the six torsion angles used for PCCA analysis as described in the text.*

by two central peptide backbone angles Φ_0 and Ψ_0 (see Figure 8). 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) [7] at a temperature of 750 K. Each subtrajectory 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 about 99 percent.

By analysis of the time series of the six main torsion/peptide angles (see Figure 9) using PCCA [12, 13] 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 Figure 8). This observation time series is shown in Figure 10. As illustrated in this figure the HMMSDE approach (for $M = 4$) also identifies three (partially overlapping) metastable states. Figure 10 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. Comparison of the spread of the distance data within each of the metastable states with the distributions generated by the SDEs shows good agreement.

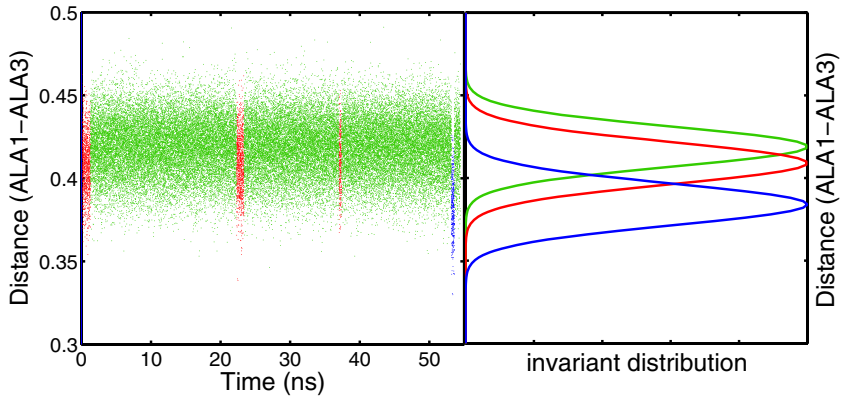

FIG. 10. *Left: time series of the intramolecular ALA1–ALA3 distance in trialanine as introduced in the text. Different colorings (indicated by dark and light coloring of the time series graph) of phases of the time series according to the three metastable states determined by* HMMSDE *. Right: invariant distributions of the three one-dimensional SDEs resulting from the* HMMSDE *algorithm versus the distance (y-axis).*

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 Figure* 10.

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

The resulting Viterbi path is illustrated in Figure 11(left) in comparison to the jump process between metastable states identified from the PCCA analysis of the full torsion angle time series Figure 11(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 HMMs 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 multidimensional generalization of the technique essentially poses no problem [24], we will have to ask whether the time series is "statistically rich enough" (cf. the discussion in section 1.1) to reliably identify all the parameters of the multidimensional 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 forthcoming paper [24]. However, we are interested in an approach that allows us 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 such as molecular dynamics or climate theory time series are usually extremely rich.

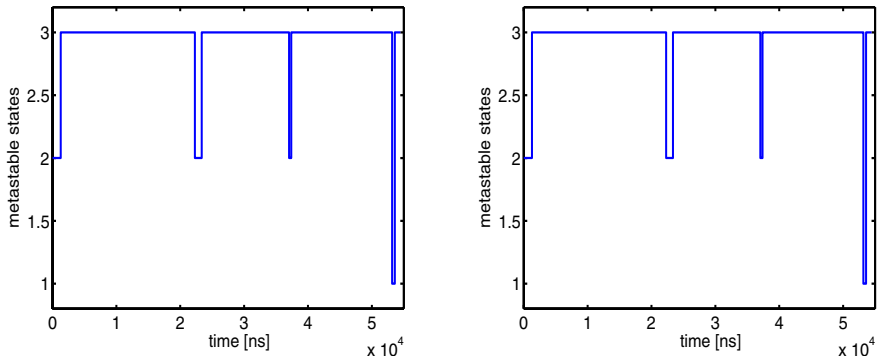

FIG. 11. *Jumps between the three dominant metastable states (indicated by* 1, 2, *and* 3*) versus time w.r.t. the time series (length* 544555 *steps) shown in Figure* 10. *Right: as computed from PCCA decomposition of the full six-dimensional torsion angle state space. Left: Viterbi path as computed based on* HMMSDE *parametrization for the one-dimensional distance observation sequence for* $M = 4$.

We discussed in detail the manifold of somewhat comparable approaches. The authors hope that research combining other approaches with the one presented herein might allow us to construct error and convergence indicators.

Furthermore, we herein assumed that the observation stepsize τ is small 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. This is worked out in detail in [23] based on the results presented herein .

Last but not least, we should ask, When might the proposed algorithms fail or produce misleading results? So far our experiences show that (assuming "long enough" time series and a reliable upper bound on the possible number of important metastable states) the most important source of problems is by far the convergence of the EM algorithm: It is hard to judge whether it really has converged and, if converged, whether the results might be only a local maximum or really the global one. However, this really seems to be a general problem of the field whose solution would be of general interest far beyond the framework of the present article.

Appendix.

Forward-backward variables. The forward-backward variables are defined via

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

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

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

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

and directly compute

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

For β we get similarly

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

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

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

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

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

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

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

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

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

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

Acknowledgment. We are indebted to an anonymous referee for useful comments on earlier drafts of the manuscript, especially for pointing out the literature on switching Markov chains.

REFERENCES

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

[2] A. Beskos, O. Papaspiliopoulos, G. Roberts, and P. Fearnhead, *Exact and computational efficient likelihood-based estimation for discretely observed diffusion processes*, J. R. Statist. Soc. B, 68 (2006), pp. 1–29.

[3] J. A. Bilmes, *A Gentle Tutorial of the EM Algorithm and Its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models*, Tech. report, International Computer Science Institute, Berkeley, CA, 1998.

[4] J. P. N. Bishwal, *Maximum likelihood estimation in partially observed stochastic differential. system driven by a fractional Brownian motion*, Stochastic Anal. Appl., 21 (2003), pp. 995–1007.

[5] P. G. Blackwell, *Bayesian inference for Markov processes with diffusion and discrete components*, Biometrica, 90 (2003), pp. 613–627.

[6] G. E. P. Box and G. M. Jenkins, *Time Series Analysis: Forecasting and Control*, Holden–Day, San Francisco, 1976.

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

[8] T. E. Dabbous, *Filtering of linear partially observe stochastic systems: The fuzzy logic approach*, Dynam. Control, 11 (2001), pp. 315–331.

[9] T. E. Dabbous, *Fuzzy optimal control for bilinear stochastic systems with fuzzy parameters*, Dynam. Control, 11 (2001), pp. 243–259.

[10] M. Dellnitz and O. Junge, *On the approximation of complicated dynamical behavior*, SIAM J. Numer. Anal., 36 (1999), pp. 491–515.

[11] M. Dellnitz and R. Preis, *Congestion and almost invariant sets in dynamical systems*, in Proceedings of Symbolic and Numerical Scientific Computation (SNSC'01), Lecture Notes in Comput. Sci. 2630, Springer, Berlin, 2003, pp. 183–209.

[12] P. Deuflhard, W. Huisinga, A. Fischer, and C. Schütte, *Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains*, Linear Algebra Appl., 315 (2000), pp. 39–59.

[13] P. Deuflhard and M. Weber, *Robust Perron cluster analysis in conformation dynamics*, Linear Algebra Appl., 398 (2005), pp. 161–184.

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

[15] R. Elber and M. Karplus, *Multiple conformational states of proteins: A molecular dynamics analysis of myoglobin*, Science, 235 (1987), pp. 318–321.

[16] E. Faou and C. Lubich, *A Poisson integrator for Gaussian wavepacket dynamics*, Comput. Vis. Sci., 9 (2006), pp. 45–55.

[17] A. Fischer, S. Waldhausen, and C. Schütte, *Identification of biomolecular conformations from incomplete torsion angle observations by hidden Markov Models*, J. Comput. Phys., submitted.

[18] H. Frauenfelder, P. J. Steinbach, and R. D. Young, *Conformational relaxation in proteins*, Chem. Soc., 29 (1989), pp. 145–150.
[19] H. Frydman and P. Lakner, *Maximum likelihood estimation of hidden Markov processes*, Ann. Appl. Probab., 13 (2003), pp. 1296–1312.
[20] A. R. Gallant and G. Tauchen, *Reprojecting partially observed systems with application to interest rate diffusions*, J. Amer. Statist. Assoc., 93 (1998), pp. 10–24.
[21] Z. Ghahramani, *An introduction to hidden Markov models and Bayesian networks*, Int. J. Pattern Recognition and Artificial Intelligence, 15 (2001), pp. 9–42.
[22] J. D. Hamilton, *A new approach to the economic analysis of nonstationary time series and the business cycle*, Econometrica, 57 (1989), pp. 357–384.
[23] I. Horenko, E. Dittmer, and C. Schütte, *Reduced stochastic models for complex molecular systems*, Comput. Vis. Sci., 9 (2006), pp. 89–102.
[24] I. Horenko, E. Dittmer, and C. Schütte, *Stochastic Model Reduction for Complex Metastable Systems: The Problem of Dimensionality*, manuscript.
[25] M. Hristache, A. Juditski, J. Polzehl, and V. Spokoiny, *Structure adaptive approach for dimension reduction*, Ann. Statist., 29 (2001), pp. 1537–1566.
[26] M. R. James, V. Krishnamurthy, and F. LeGland, *Time discretization of continuous-time filters and smoothers for hmm parameter estimation*, IEEE Trans. Inform. Theory, 42 (1996), pp. 593–605.
[27] T. N. Kaseke and M. E. Thompson, *Estimation for rainfall-runoff modeled as a partially observed Markov process*, Stochastic Hydrology and Hydraulics, 11 (1997), pp. 1–16.
[28] D. Korenblum and D. Shalloway, *Macrostate data clustering*, Phys. Rev. E(3), 67 (2003), 056704.
[29] S. C. Kou, X. S. Xie, and J. S. Liu, *Bayesian analysis of single-molecule experimental data*, Appl. Statist., 54 (2005), pp. 469–506.
[30] N. R. Kristensen, H. Madsen, and S. B. Jorgensen, *Parameter estimation in stochastic grey-box models*, Automatica, 40 (2004), pp. 225–237.
[31] E. Lindahl, B. Hess, and D. van der Spoel, *Gromacs* 3.0: *A package for molecular simulation and trajectory analysis*, J. Mol. Mod., 7 (2001), pp. 306–317.
[32] L. A. Liporace, *Maximum likelihood estimation for multivariate observations of Markov sources*, IEEE Trans. Inform. Theory, 28 (1982), pp. 729–734.
[33] A. Logothetis and V. Krishnamurthy, *Expectation maximization algorithms for map estimation of jump markov linear systems*, IEEE Trans. Signal Process., 47 (1999), pp. 2139–2156.
[34] G. McLachlan, *The EM Algorithm and Extensions*, 2nd ed., Wiley, New York, 2005.
[35] G. McLachlan and D. Peel, *Finite Mixture Models*, Wiley, New York, 2000.
[36] C. R. Nelson and C. J. Kim, *State-space models with regime switching: Classical and gibbs-sampling approaches with applications*, MIT Press, Cambridge, MA, 1999.
[37] G. U. Nienhaus, J. R. Mourant, and H. Frauenfelder, *Spectroscopic evidence for conformational relaxation in myoglobin*, Proc. Natl. Acad. Sci. USA, 89 (1992), pp. 2902–2906.
[38] M. Oresic and D. Shalloway, *Hierarchical characterization of energy landscapes using Gaussian packet states*, J. Chem. Phys., 101 (1994), pp. 9844–9857.
[39] K. Padberg, R. Preis, and M. Dellnitz, *Integrating Multilevel Graph Partitioning with Hierarchical Set Oriented Methods for the Analysis of Dynamical Systems*, Preprint 152, DFG Schwerpunktprogramm 1095, DFG, Bonn, Germany, 2004.
[40] J. Pillardy and L. Piela, *Molecular dynamics on deformed energy hypersurfaces*, J. Phys. Chem., 99 (1995), pp. 11805–11812.
[41] J. Pillardy and L. Piela, *Smoothing techniques of global optimization: Distance scaling method in searches for most stable Lennard-Jones atomic clusters*, J. Phys. Chem., 18 (1997), pp. 2040–2049.
[42] Y. Pokern, A. M. Stuart, and P. Wiberg, *Parameter Estimation for Partially Observed Hypoelliptic Diffusion*, http://www.maths.warwick.ac.uk/∼stuart (2004), submitted.
[43] L. R. Rabiner, *A tutorial on hidden Markov models and selected applications in speech recognition*, Proc. IEEE, 77 (1989), pp. 257–286.
[44] A. Samarov, V. Spokoiny, and C. Vial, *Component identification and estimation in nonlinear high-dimensional regression models by structural adaptation*, J. Amer. Statist. Assoc., 100 (2005), pp. 429–445.
[45] C. Schütte, A. Fischer, W. Huisinga, and P. Deuflhard, *A direct approach to conformational dynamics based on hybrid Monte Carlo*, J. Comput. Phys., 151 (1999), pp. 146–168.
[46] C. Schütte and W. Huisinga, *Biomolecular conformations can be identified as metastable sets of molecular dynamics*, in Handbook of Numerical Analysis, P. G. Ciarlet and J.-L. Lions, eds., North–Holland, Amsterdam, 2003.

[47] S. L. Scott, *Bayesian methods for hidden Markov models: Recursive computing in the* 21*st century*, J. Amer. Statist. Assoc., 97 (2002), pp. 337–352.
[48] I. Shoji and T. Ozaki, *Comparative study of estimation methods for continuous time stochastic processes*, J. Time Ser. Anal., 18 (1997), pp. 485–506.
[49] V. N. Smelyanskiy, D. A. Timucin, A. Brandrivskyy, and D. G. Luchinsky, *Model reconstruction of nonlinear dynamical systems driven by noise*, Phys. Rev. Lett., submitted.
[50] J. E. Straub, J. Ma, and P. Amara, *Simulated annealing using coarse grained classical dynamics: Smoluchowski dynamics in the Gaussian density approximation*, J. Phys. Chem., 103 (1995), pp. 1574–1581.
[51] A. J. Viterbi, *Error bounds for convolutional codes and an asymptotically optimum decoding algorithm*, IEEE Trans. Inform. Theory, 13 (1967), pp. 260–269.
[52] C. F. J. Wu, *On the convergence properties of the EM algorithm*, Ann. Statist., 11 (1983), pp. 95–103.
[53] L. Wu and M. I. Jordan, *On the convergence properties of the EM algorithm for Gaussian mixtures*, Neural Comput., 8 (1996), pp. 129–151.
[54] Q. Zhang, *Nonlinear filtering and control of a switching diffusion with small observation noise*, SIAM J. Control Optim., 36 (1998), pp. 1638–1668.