Blind model reduction for high–dimensional time–dependent data

Illia Horenko, Carsten Hartmann

Freie Universität Berlin, Institut für Mathematik II
Arnimallee 2-6, 14195 Berlin, Germany

Abstract

We consider the problem of automatically extracting simplified models out of complex high–dimensional and time–dependent data. The simplified model is given by a linear Langevin equation with time–varying coefficients. The reduced model may still be high–dimensional, but it is physically intuitive and much easier to interpret than the original data. In particular we can distinguish whether dynamical effects are influenced by friction, noise, or deterministic motion. The parameters for the reduced model are obtained by a robust and efficient numerical predictor–corrector scheme which relies on analytical solutions to a maximum-likelihood problem provided the time steps between successive observations are not too large. If the data set is very heterogeneous the time series is better described not by a single model, but by a collection of reduced models. This scenario is accounted for by embedding the parameter estimation procedure into the framework of hidden Markov models, i.e., we decompose the data into several subsets, each of which gives rise to an appropriate linear Langevin model. The switching between the local model is done by a Markov jump process. The optimal decomposition into submodels can then be regarded as one global Langevin model with piecewise constant coefficients. We illustrate the performance of the algorithm by means of several examples. Especially we focus on the numerical error as a function of the time step of the observation sequence.

Key words: Langevin equation, model reduction, parameter estimation, hidden Markov models, predictor–corrector scheme, maximum–likelihood principle
PACS: 05.10.Gg, 02.50.Ga, 05.10. Gg, 64.60.My

Email addresses: horenko@math.fu-berlin.de (Illia Horenko), chartman@math.fu-berlin.de (Carsten Hartmann).

1 Introduction

Increasing amount of measurement data and growing complexity of processes in all fields of applied sciences during the last few years has led to a persistent demand for methods that allow for *automatized* extraction of the physically interpretable information out of raw data. Such *data–based modelling approaches* should be able to flexibly incorporate multidimensional statistical models for the observed data, yet they should be simple enough to enable physical understanding of the process under consideration.

Therefore the genuine aim of data–based modelling is to reduce the *complexity* of processes and data; this should be carefully distinguished from analytical approaches like, e.g., spatial decomposition methods such as proper orthogonal decomposition, the Karhunen-Loève expansion, or also averaging techniques. These approaches make the point of *reducing the dimension of a given model*, although the problem of finding a good decomposition may be data–driven as well. See the textbook [1], or the excellent review article [2] for an overview. Compare also [3] for a related approach.

We can distinguish three classes of related approaches for data–based model reduction: (i) Box-Jenkins Model identification strategy, (ii) Bayesian models or neural networks, (iii) and approaches which are based on fitting of the data with a system of differential equations.

The first group of methods (i) is originated in econometrics in the beginning of 1970 and is known under the name *Box-Jenkins technique* or ARIMA (autoregressive integrable models with moving average) [4,5,6]. The main idea of these methods relies on fitting the observed data with a *discrete time stochastic difference scheme*. The Box-Jenkins approach is restricted to the analysis of stochastic processes that can be made *stationary*, i.e., cast into stochastic processes X_t of bounded variation, constant first moment, and second moment $\mathbf{E}(X_t X_s)$ that depends only on $(t-s)$; this can be achieved, e.g., by differencing the time series. Moreover, the resulting autoregressive difference scheme is discrete in time, which implies constant time intervals between single realizations of the process.

The second group (ii) is based on dynamical Bayesian networks, such as hidden Markov models (HMM) [7,8], or neural networks [9,10]. These are *set-oriented approaches*, as they decompose the configuration space into several sets, where the dynamics of the system in each of the domains is described by an independent data model (see Figure 1). The overall dynamics of the process is then governed by a *hidden* process switching between those sets. Most of the approaches that we are aware of are designed in the context of the discrete stochastic systems, which means that they are not based on a reasonable

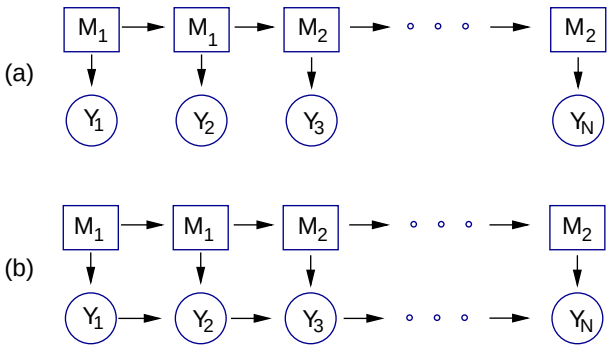

Figure 1. Dynamical Bayesian Networks. Here the arrows denote the casual dependencies, M_i labels the hidden variable or model, Y_i is the observation. In the standard HMM approach the observation is triggered by the sequence of hidden states for a prescribed probability distribution of the output (Figure (a)), whereas in the HmmSde scheme the observation sequence is connected through a physical model, that depends on the hidden states (Figure (b)).

physical model. Moreover the efficient implementation for high–dimensional physical systems is lacking. See Figure 1 for illustration.

The third group of methods (iii) attempts to fit a global physical model, e.g., a Langevin equation, to observed data [11]. Unfortunately the available methods can deal with high–dimensional data only under very specific assumptions (e.g., thermodynamical equilibrium, all matrices are diagonal etc.).

The approach that we develop here is a multi–dimensional extension of the recently proposed HmmSde method (Hidden Markov Models with Stochastic Differential Equations) for the case of Langevin dynamics [12,13]. The method links dynamical Bayesian approaches with local Langevin models that are fitted to observed data. The approach allows for the construction of global physical models for high–dimensional data.

The rest of the article is organized as follows: In Section 2 we introduce the general model, explain the basic method and derive the evolution equations for the time–dependent parameters. The algorithmic strategy for identifying the local Langevin models and to estimate the respective parameters is described in Section 3. Finally we demonstrate the proposed technique by application to some generic examples in Section 4.

2 Reduced model system

We shall restrict the class of models that are to be parameterized to Langevin equations on Euclidean configuration space $Q \subseteq \mathbf{R}^n$, which are of the following type:

$$M\ddot{q}(t) = -\mathrm{grad}\, U(q(t)) - \gamma\dot{q}(t) + \sigma\dot{W}(t) \quad q \in Q\,. \tag{2.1}$$

Here $U : Q \to \mathbf{R}$ denotes the interaction potential, and $\dot{W}(t)$ is the standard Brownian motion. This model can be thought of steming from a separable Hamiltonian including viscous friction and noise; the more general non–separable case will be treated in a forthcoming paper. Here both friction coefficient $\gamma \in \mathbf{R}^{n\times n}$, and noise amplitude $\sigma \in \mathbf{R}^{n\times n}$ are symmetric, positive-definite matrices, and $M \in \mathbf{R}^{n\times n}$ is the constant, positive-definite mass matrix; we do not assume that M is diagonal. Exploiting that the involved matrices are symmetric, reduces the number of undetermined parameters from n^2 to $n(n+1)/2$ for each matrix.

Introducing standard conjugate variables $(q,p) \in T^*Q \simeq \mathbf{R}^n \times \mathbf{R}^n$ for positions and momenta on the cotangent space to Q, we can rewrite the Langevin equation as the following first order system

$$\begin{aligned}
\dot{q}(t) &= M^{-1}p(t)\\
\dot{p}(t) &= -\mathrm{grad}\, U(q(t)) - \zeta p(t) + \sigma\dot{W}(t)
\end{aligned}$$

with the abbreviation $\zeta = \gamma M^{-1}$. It can be seen that the equations of motion above have some scaling invariance: changing variables according to $q \mapsto M^{1/2}q$, $p \mapsto M^{-1/2}p$, which clearly is a symplectic transform, we arrive at the scaled Langevin equation on tangent space TQ

$$\dot{q}(t) = v(t) \tag{2.2}$$

$$\dot{v}(t) = -\mathrm{grad}\, U(q(t)) - \gamma v(t) + \sigma\dot{W}(t)\,. \tag{2.3}$$

Hence we can reasonably identify TQ with T^*Q, just by applying the scaling transform $\gamma \mapsto M^{-1/2}\gamma M^{-1/2}$, $\sigma \mapsto M^{-1/2}\sigma$. This identification amounts to setting $M = \mathbf{1}$ in the original Langevin equation (2.1). Notice that we do not assume that the time series corresponds to an equilibrium process. Hence we do not assume that the fluctuation–dissipation relation is met. Nevertheless it is important to note that the mass scaling respects the fluctuation–dissipation relation, for it is easy to see that

$$\beta\sigma\sigma^T = \gamma \quad \Leftrightarrow \quad \beta M^{-1/2}\sigma\sigma^T M^{-1/2} = M^{-1/2}\gamma M^{-1/2}\,,$$

whenever the inverse temperature $\beta = 1/T$ is well–defined, i.e., in case the system is in thermodynamical equilibrium. Consequently the fluctuation–dissipation relation does not provide an additional condition, by means of which the mass matrix in the model could be determined.

Remark 1 *Apparently we have some freedom in setting up the parameters in the model. However the undeterminacy of the mass matrix M lies deeper,*

for the mass scaling represents a symmetry group of scaling transforms that is present in the Langevin equation (2.1). As we have seen, not even does the fluctuation–dissipation relation provide further information; only in case the given observations contain information about the momenta and *the velocities some knowledge about the mass matrix can be obtained employing the canonical relation* $p = M\dot{q}$. *However for many applications the conjugate momenta will not be available.*

The optimal set of parameters for noise, friction, and the potential function is uniquely determined by a *maximum–likelihood principle*. At a later stage we shall consider parameters which will be only piecewise constant, in the sense that each parameter tuple is optimal only for a specific subsequence of the full time series. As we will show later on we can use the HMM algorithm to switch between these distinct parameter sets; the underlying idea is to decompose a complex time series by means of the *Viterbi algorithm* into several subsequences each of which can be treated again by the maximum–likelihood estimation. Such complex time series may occur in case there is metastability in the system. For examples see [12] and the references therein.

Remark 2 *The reader may argue that the considered Langevin model with linear friction does not capture memory effects, which may be important, e.g., for the dynamics of biomolecules. This objection is typically formulated in terms of slowly decaying velocity autocorrelations in the data. However it is often ignored that these "global" autocorrelation functions, i.e., autocorrelation functions that are estimated over the full time series, are meaningful only for stationary time series; for non–equilibrium processes the autocorrelation function may be totally misleading.*[1] *Furthermore the autocorrelation is no reliable measure for the memory in the system as it known from the theory of time series analysis [4], even for stationary time series. Instead the* partial *autocorrelation, which can be computed from the ordinary autocorrelation function, is an exact statistical measure for the depth of the Markovian memory (provided the data is generated by a generalized time–discrete Markov process [14]. In many interesting cases the (velocity) autocorrelation function decays rather slowly, whereas the corresponding partial autocorrelation decays several orders of magnitude faster. See the Cyclophane example in the numerics section. Hence the decay time of the partial autocorrelation indicates on which time scales the linear friction model makes sense at all.*

[1] For example, consider the autocorrelation function of a discretization of the one–dimensional harmonic oscillator, which is clearly periodic. But as the system is deterministic, this is a Markov process without memory.

Given a time series, the aim of the current work is to find optimal parameters for the model equations (2.2)–(2.3) by means of some maximum–likelihood principle. To this end we define the Kolmogorov forward equation

$$\partial u(q,v,t) = \mathcal{A}u(q,v,t)\,, \quad u(q,v,0) = u_0(q,v,t_0)$$

that is associated with our model system, where $\mathcal{A} : L^1(\nu) \to L^1(\nu)$ is the forward generator of the time evolution considered to act on functions on the function space space $L^1(\nu)$, where $\nu = \mathrm{d}q\,\mathrm{d}v$ denotes the ordinary Liouville measure. Letting $\partial/\partial v$ denote the (directional) derivative with respect to the vector $v = \dot{q}(t)$, the forward operator is defined as

$$\mathcal{A} = \sum_{i,j} \frac{\sigma_{ij}^2}{2} \frac{\partial^2}{\partial v_i \partial v_j} + \sum_i \left(\gamma v + \operatorname{grad} U\right)_i \frac{\partial}{\partial v_i} - \sum_k v_k \frac{\partial}{\partial q^k} + \operatorname{tr}(\gamma)\,.$$

In order to determine the parameters of our model we approximate the solution of the Kolmogorov forward equation by localized Gaussians. This is for two reasons: Assuming that the initial observation is sharp, the density will have a Gaussian shape after a short time, provided the coefficients in the evolution equation are sufficiently smooth [15,16]. Hence we study (i) the local evolution of a Dirac–like density between two successive observations which are close in time. Then (ii), we can compute the solution of the forward equation analytically which has proven useful for an efficient maximum–likelihood estimation of the parameters [12]. We introduce the Gaussian probability density of the random variable $x = (q,v) \in \mathbf{R}^n \times \mathbf{R}^n$ centered at $\bar{x}(t)$ at time t

$$\rho(x,t) = \rho_0(t) \exp\left(-\frac{1}{2}\left\langle \Sigma(t)(x-\bar{x}(t)), x-\bar{x}(t)\right\rangle\right)\,, \quad x = (q,v)\,,$$

where $\langle\cdot,\cdot\rangle$ stands for the inner product in either $\mathbf{R}^n$ or in $\mathbf{R}^{2n}$, and $\Sigma \in \mathbf{R}^{2n\times 2n}$ denotes the symmetric, positive-definite shape matrix

$$\Sigma(t) = \begin{pmatrix} A(t) & B(t) \\ B(t) & C(t) \end{pmatrix}\,.$$

Here the block matrices A, B, C correspond to the variables q and v in the obvious way. Plugging the Gaussian ansatz functions into the forward equation, and equating powers of $(q-\bar{q}(t))(v-\bar{v}(t))$ we end up with a system of (symmetrized) ordinary differential equations for the time–dependent parameters

$$\ddot{\bar{q}}(t) = -\mathrm{grad}\, U(\bar{q}(t)) - \gamma\dot{\bar{q}}(t) \tag{2.4}$$

$$\dot{\rho}_0(t) = \left(\mathrm{tr}\,(\gamma) - \frac{1}{2}\mathrm{tr}\,(\sigma^2 C(t))\right)\rho_0(t) \tag{2.5}$$

$$\dot{A}(t) = B(t)H(\bar{q}(t)) + H(\bar{q}(t))B(t) - B(t)\sigma^2 B(t) \tag{2.6}$$

$$\begin{aligned}\dot{B}(t) = &\frac{1}{2}\left(B(t)\gamma + \gamma B(t) - B(t)\sigma^2 C(t) - C(t)\sigma^2 B(t)\right) \\ &+\frac{1}{2}\left(C(t)H(\bar{q}(t)) + H(\bar{q}(t))C(t)\right) - A(t)\end{aligned} \tag{2.7}$$

$$\dot{C}(t) = C(t)\gamma + \gamma C(t) - 2B(t) - C(t)\sigma^2 C(t)\,, \tag{2.8}$$

where $H(\bar{q}) = D^2U(\bar{q}) \in \mathbf{R}^{n\times n}$ denotes the Hessian matrix of the potential function $U(q)$ evaluated at $q = \bar{q}$ (see the Remark below).

Note that the equation for the center $\bar{q}(t)$ is decoupled from the remaining equations and vice versa; thus, in order to solve the equations, we can choose convenient discretization schemes for each of the equations. Clearly, depending on which equation is how disretized we will obtain different convergence properties in the numerics for the respective parameters. We will come back to this issue later on in the examples section. We stress once again that the derivation here does not require that the observation series is in thermodynamic equilibrium, i.e., that the parameters γ and σ obey the fluctuation–dissipation relation. Rather we consider the data to describe some inherent non–equilibrium process.

2.2 Short–time asymptotics

On condition that the initial density at time $t_0 = t$ is sharply peaked around the observed value $(q, v) = (Q(t), V(t))$ we may derive an asymptotic expression for the density at a small time step $t \mapsto t + h$ with $h = \mathcal{O}(\epsilon)$, where $\epsilon \ll 1$ is a small parameter. To put this differently, the probability distribution can be regarded as an infinitely narrow Gaussian density, and we shall try to solve parameter equations of motion for these specific initial conditions. After time $t + h$ we expect the shape matrix to be of the form

$$\Sigma(t+h) = \begin{pmatrix} \mathcal{O}(\epsilon^{-1}) & \mathcal{O}(\epsilon) \\ \mathcal{O}(\epsilon) & \mathcal{O}(\epsilon^{-1}) \end{pmatrix}.$$

For sufficiently small ϵ we can use some formal arguments from singular perturbation theory: First of all we expand the shape matrices in powers of ϵ:

$$
\begin{aligned}
A^\epsilon(s) &= \epsilon^{-1}A_{-1}(s) + A_0(s) + \epsilon A_1(s) + \dots \\
B^\epsilon(s) &= \epsilon^{-1}B_{-1}(s) + B_0(s) + \epsilon B_1(s) + \dots \\
C^\epsilon(s) &= \epsilon^{-1}C_{-1}(s) + C_0(s) + \epsilon C_1(s) + \dots \,.
\end{aligned}
$$

We assume that all coefficients A_k, B_k, C_k are uniformly bounded and sufficiently smooth, which we also require for the only remaining time–dependent coefficient $H(s) = H(\bar{q}(s))$. Defining the microscopic time scale by $s = \epsilon t$, we can rescale the equations of motion (2.6)–(2.8), which then become

$$
\begin{aligned}
\epsilon^{-1}\dot{A}^\epsilon(s) &= B^\epsilon(s)H(s) + H(s)B^\epsilon(s) - B^\epsilon(s)\sigma^2 B^\epsilon(s) \\
\epsilon^{-1}\dot{B}^\epsilon(s) &= \frac{1}{2}\left(B^\epsilon(s)\gamma + \gamma B^\epsilon(s) - B^\epsilon(s)\sigma^2 C^\epsilon(s) - C^\epsilon(s)\sigma^2 B^\epsilon(s)\right) \\
&\quad + \frac{1}{2}\left(C^\epsilon(s)H(s) + H(s)C^\epsilon(s)\right) - A^\epsilon(s) \\
\epsilon^{-1}\dot{C}^\epsilon(s) &= C^\epsilon(s)\gamma + \gamma C^\epsilon(s) - 2B^\epsilon(s) - C^\epsilon(s)\sigma^2 C^\epsilon(s)\,.
\end{aligned}
$$

We can plug in the expansions of $A^\epsilon, B^\epsilon, C^\epsilon$, and equate powers of ϵ. From this we obtain a hierarchy of equations, the lowest order of which is

$$
\dot{A}_{-1}(s) = -B_{-1}(s)\sigma^2 B_{-1}(s) \tag{2.9}
$$

$$
\dot{B}_{-1}(s) = -\frac{1}{2}\left(B_{-1}(s)\sigma^2 C_{-1}(s) + C_{-1}(s)\sigma^2 B_{-1}(s)\right) \tag{2.10}
$$

$$
\dot{C}_{-1}(s) = -C_{-1}(s)\sigma^2 C_{-1}(s)\,. \tag{2.11}
$$

So far we have not said too much about initial conditions. Clearly the expansion is supposed to still hold at $s = 0$. As we have stated the initial density is sharply peaked around (q, v); this means that the shape matrix is diagonal–dominant with $A(0), C(0) \gg 1$, and $B(0) = 0$. Hence we can infer from the smoothness properties of the coefficients that B_{-1} identically vanishes in a neighbourhood of the origin. In particular this can be seen by first solving the equation for $C_{-1}(s)$, which is independent of all the other shape matrices, and then plugging the expression into the equation for $B_{-1}(s)$; since we do not want the solution to blow up at $s = 0$, we demand $\dot{B}_{-1} = B_{-1} = 0$.

Hence the system of equations (2.9)(2.11) can be solved by analytic means, provided that $B_{-1} = 0$. Scaling back to the original time scale $t = s/\epsilon$ we obtain the first–order result at time $t + h$

$$
\Sigma(t+h) = \begin{pmatrix} A(t) & \mathbf{0} \\ \mathbf{0} & C(t)(h\sigma^2)^{-1} \end{pmatrix}, \tag{2.12}
$$

where the initial values satisfy $A(t), C(t) \gg 1$ and $B(t) = 0$. By performing a

second-order accurate, symmetric, and explicit discretization of the equations for $\bar{q}$, and $\bar{v}$ we arrive at [17]

$$\bar{q}(t+h) = Q(t) + \frac{h}{2} V(t) \tag{2.13}$$

$$\bar{v}(t+h) = V(t) - h\left(\operatorname{grad} U(\bar{q}(t+h)) + \gamma V(t)\right) \tag{2.14}$$

Clearly the time evolution of the amplitude $\rho_0(t+h)$ is determined by the condition that the probability density $\rho(\cdot, t)$ stays normalized:

$$\rho_0(t+h) = \frac{1}{h\pi^n} \sqrt{\frac{\det K}{\det \sigma^2}}\,. \tag{2.15}$$

We can check the shape matrix result for consistency by taking the intermediate result to the next order in the small parameter; in doing so we find that $A_0 = \text{const.}, B_0 \propto h$ and $C_0 \propto \log h$. Clearly the asymptotic derivation of the shape matrix' time evolution is purely formal; nevertheless it can be supported by numerically solving the equations of motion, where it turns out that the asymptotic result is valid over relatively long times; in the last section we will quantify the validity of the asymptotic expansion for increasing step size h.

Remark 3 *It is necessary to make an arrangement for the generic case that the potential $U(q)$ in the Langevin equation is unknown. Then we assume that the Hessian matrix $H(\bar{q}) = D^2U(\bar{q})$ is constant, which leads to a quadratic potential function of the form*

$$U(q) = \frac{1}{2} \left\langle H(q-\mu), q-\mu \right\rangle\,.$$

Thus H, μ are the unknown parameters, where μ denotes the center of the harmonic potential, which should not be confused with the time-dependent center $\bar{q}(t)$ of the Gaussian density. In praxi *the harmonic approximation leads to computationally tractable problems at all, but moreover it follows by consistency with the choice of the Gaussian ansatz functions that led to evolution equations for the shape matrices, where no higher–order derivatives of $U(\bar{q})$ appeared; determination of higher–order terms would require higher-order moments in the ansatz density $\rho(x,t)$, in case of which the emerging parameter equations cannot be solved for high–dimensional problems.*

3 Optimal model parameters

3.1 Maximum–likelihood principle

For the moment we assume that we are dealing with only one, possibly high–dimensional, Langevin model without hidden states, that has to be parameterized.

The procedure then works as follows: Suppose we are given a discrete time series $X = \{X_1, \ldots, X_{M+1}\}$, where $X_k = (Q(t_k), V(t_k))$ denotes a position and velocity observation, and $h = t_{k+1} - t_k$ is the constant time difference between two successive observations. We are aiming at maximizing the probability density of the output X_{k+1} that is evolved according to the Langevin model, starting from the observed datum X_k. The corresponding conditional probability density is given by the expression

$$\rho_\lambda(X_{k+1}|X_k) = \rho_0 \exp\left(-\frac{1}{2}\left\langle \Sigma(X_{k+1} - \bar{x}_{k+1}), X_{k+1} - \bar{x}_{k+1}\right\rangle\right) . \tag{3.1}$$

with the time–dependent parameters $\rho_0, \Sigma, \bar{x}$ evaluated at $t + h$ as determined by the equations (2.12)–(2.15). Of course the time–*dependent* parameters are functions of the observations and the time–*independent* parameters $\lambda = (\gamma, \sigma^2, H, \mu)$, that ought to be determined. In particular $\bar{x}_{k+1}$ is a function of the former observation X_k. We define the log–likelihood function of the observation sequence as

$$\mathcal{L}(\lambda|X) = \log p(X|\lambda) \tag{3.2}$$

where $p(X|\lambda)$ denotes the joint probability distribution of the observation sequence

$$p(X|\lambda) = \prod_{k=1}^{M} \rho_\lambda(X_{k+1}|X_k)\,, \tag{3.3}$$

that satisfies the Markov property $\rho(X_{k+1}|X_1, \ldots, X_k) = \rho(X_{k+1}|X_k)$. The optimal parameters λ are those which maximize the log–likehood function, which reads upon inserting the equations (3.1) and (3.3) into (3.2)

$$\begin{aligned}\mathcal{L}(\lambda|X) &= \sum_{k=1}^{M} \log \rho_\lambda(X_{k+1}|X_k) \\ &= M \log \rho_0 - \frac{1}{2}\sum_{k=1}^{M} \langle \Sigma(X_{k+1} - \bar{x}_{k+1}), X_{k+1} - \bar{x}_{k+1} \rangle \\ &= C - \frac{M}{2} \log \det \sigma^2 \\ &\quad -\frac{1}{2h}\sum_{k=1}^{M} \left\langle \sigma^{-2}(V_{k+1} - \bar{v}_{k+1}), V_{k+1} - \bar{v}_{k+1} \right\rangle .\end{aligned}$$

Here $C < 0$ denotes a constant that collects all terms that do not depend on the undetermined parameters $\lambda = (\gamma, \sigma^2, H, \mu)$, and $\bar{v}_{k+1}$ is defined according to (2.13)–(2.14)

$$\bar{v}_{k+1} = V_k - h\left(H(Q_k + \frac{h}{2}V_k - \mu) + \gamma V_k\right) .$$

In order to compute the critical point of the log–likelihood function, we evaluate the necessary condition $\mathbf{d}\mathcal{L} = 0$. To this end we compute the individual partial derivatives of the log–likelihood: for the friction coefficients

$$\frac{\partial \mathcal{L}}{\partial \gamma} = -\sigma^{-2}\sum_{k=1}^{M} V_k \otimes \Delta_{k+1}^v , \tag{3.4}$$

the noise covariance matrix

$$\frac{\partial \mathcal{L}}{\partial \sigma^2} = \frac{1}{2h}\sigma^{-4}\sum_{k=1}^{M} \Delta_{k+1}^v \otimes \Delta_{k+1}^v - \frac{M}{2}\sigma^{-2} , \tag{3.5}$$

the Hessian of the potential function

$$\frac{\partial \mathcal{L}}{\partial H} = -\sigma^{-2}\frac{1}{2}\sum_{k=1}^{M} (Q_k - \mu) \otimes \Delta_{k+1}^v , \tag{3.6}$$

and last but not least we compute the derivative with respect to the centre of the potential

$$\frac{\partial \mathcal{L}}{\partial \mu} = \frac{1}{2}H\sigma^{-2}\sum_{k=1}^{M} \Delta_{k+1}^v \tag{3.7}$$

using the abbreviation $\Delta_{k+1}^v = V_{k+1} - \bar{v}_{k+1}$, and exploiting some basic properties of the tensor product, that is defined as $(X \otimes Y)_{ij} = X_i Y_j$, where X, Y are any two vectors from $\mathbf{R}^n$.

The unknown parameters $\lambda = (\gamma, \sigma^2, H, \mu)$ are determined by solving the nonlinear system of equations (3.4)–(3.7) for a given observation sequence $X = \{X_1, \ldots, X_M\}$. If either the configuration space is one–dimensional or all degrees of freedom are decoupled from each other we can solve this system analytically. This explicit solution may then serve as a predictor in solving the fully coupled high–dimensional system numerically. The numerical scheme therefore can be considered as predictor–corrector method, where the corrector step is performed using a standard Newton iteration [18].

3.2 Hidden Markov model and expectation-maximization algorithm

Up to now we have considered a single, possibly high–dimensional global model, which approximates the whole time series in the *maximum–likelihood* sense. Alternatively we could imagine that different segments of the time series correspond to different *local* Langevin models, each of which is characterized by a particular set of constant parameters $\lambda_i = (\gamma_i, \sigma_i^2, H_i, \mu_i)$. Switching back and forth between these local parameter sets can then be understood as one *global* model with parameters that are piecewise constant in time.

We shall consider the problem of estimating optimal parameters within the framework of hidden Markov models (HMM): For a prescribed number L of local parameter sets $\lambda_i, i = 1, \ldots L$, we use the expectation–maximization algorithm [7,19,20]. Hence we assume that the switching between the different parameter sets is governed by a Markov jump process. For example, one may think that the configuration space has a metastable decomposition; then every instance t in the time series is assigned to a metastable set $i(t)$. Thus the model consists of two related stochastic processes $X(t)$ and $i(t)$, where the latter is not directly observed (hidden) and fullfils the Markov property. On the other hand the observation sequence is a stochastic process $X(t) = (X|i)(t)$ conditional on the hidden state $i(t)$ at time t.

Overall a HMM is fully specified by an initial distribution π of hidden states, a transition matrix T of the hidden Markov chain $i(t)$, and by the parameters of the output process λ_i for each state i. If the rate matrix of the jump process is denoted by $R \in \mathbf{R}^{L\times L}$, then the transition probability to jump from state $i(t_k) = m$ to state $i(t_{k+1}) = n$ within time h is given by the respective entry of the transition matrix

$$T(m, n) = (\exp(hR))_{mn} \ .$$

In the standard version of HMM the observables $X(t)$ are identical and independent random variables [21,22]. Here instead we consider random variables

that are the output of the Langevin equation (2.1) for the current hidden state $i = i(t)$, that is,

$$\begin{aligned} \dot{q}(t) &= v(t) \\ \dot{v}(t) &= -H_i(q(t) - \mu_i) - \gamma_i v(t) + \sigma_i \dot{W}(t) \\ i &: \mathbf{R} \to \{1, 2, ..., L\} . \end{aligned} \tag{3.8}$$

Now embedding the problem of estimating optimal parameters for the model (3.8) into the context of HMM, the joint probability distribution (3.3) of the observation sequence reads

$$r(X|\lambda) = \prod_{k=1}^{M} T(i_k, i_{k+1}) \varrho_\lambda(X_{k+1}|i_{k+1}, X_k) , \tag{3.9}$$

where the conditional probability $\varrho_\lambda(\cdot|\cdot)$ is defined as $\rho_\lambda(\cdot|\cdot)$ before except that the parameters now depend on the hidden state $i_{k+1} = i(t_{k+1})$. The algorithm for the identification of parameters conditional on the hidden (metastable) states comprises the following three steps:

(1) Determine the optimal parameters $\theta = (\pi, A, \lambda_i)$ for all states $i = 1, \ldots, L$ by maximizing the likelihood $\mathcal{L}(\theta|X, i)$; in general this is a nonlinear global optimization problem.
(2) Determine the optimal sequence of hidden metastable states $\{i_k\} := \{i(t_k)\}$ for given optimal parameters.
(3) Determine the number of important metastable states (up to now we have simply assumed that the number L of hidden states is given a priori).

The first two problems can be addressed by standard HMM algorithms. The parameter estimation on the partially observed data is carried out using the expectation–maximization (EM) algorithm. The optimal parameters θ are identified by iteratively maximizing the entropy

$$\mathcal{S}(X) = \max_\theta \sum_i \mathcal{L}(\theta|X, i) \log \mathcal{L}(\theta|X, i) .$$

For the identification of the optimal sequence of hidden metastable states the Viterbi algorithm [23] is used, which exploits dynamic programming techniques to resolve the optimization problem

$$\max_i \mathcal{L}(\theta|X, i)$$

in a recursive manner. For the details see [24] and the references therein.

Addressing the first two problems (1) and (2) requires the specification of a number L of hidden states, which is unknown *a priori*. A practical way to

handle this problem is to assume a sufficiently large number of hidden states and then aggregate the resulting transition matrix, which gives the minimum number of hidden states which are necessary to resolve the metastable sets [25,26]. The aggregation is performed by the Perron cluster cluster analysis (PCCA), exploiting the spectral properties of the transition matrix T to transform it to a matrix with quasi–block structure [12,27,28]. These blocks then correspond to the existing metastable states.

4 Numerical examples

In this section we present different types of numerical examples for the proposed method. We start with a one–dimensional HMM–Langevin model of the type (3.8) with two hidden states, producing an ensemble of the realizations. From the realizations we then estimate the model parameters and show that the sequence of the hidden states can be completely recovered. Subsequently, we apply the procedure to a one–dimensional Langevin equation whose hidden states are implicitly defined by the metastability arising from a perturbed three–well potential, demonstrating the data–based decomposition of the dynamics into locally harmonic Langevin models that are connected by a Markov jump process.

As a slightly more challenging task, we apply the reduction algorithm to a multidimensional problem with known parameters. In the parameter estimation we especially focus on the quantitatively correct reconstruction of the flipping dynamics between metastable sets by coupling several local Langevin models; moreover we can reproduce the interaction between the different spatial dimensions of the original model and assign them to different dynamical properties, such as deterministic motion, friction or noise. We show that the approach, in contrast to simple correlation analysis of a time series, maintains the physical structure of the underlying dynamics; it is therefore possible to reconstruct physical processes by means of incomplete observations.

In the last example we apply the method to a molecular dynamics simulation of Cyclophane, demonstrating the ability of also estimating parameters of inherent non–equilibrium processes, only from short fragments of the MD simulation. We also perform a numerical investigation of the time step length influence on the quality of the parameter estimation.

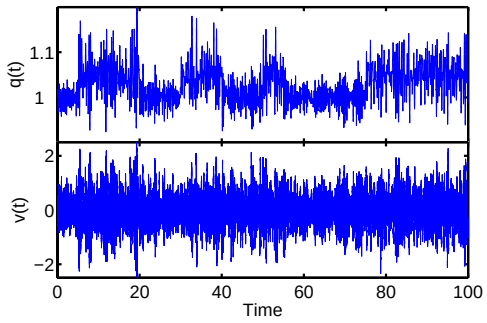

Figure 2. Typical realization $(q(t), \dot{q}(t))$ of the Langevin equation (4.1). The time series has a length $M = 200.000$.

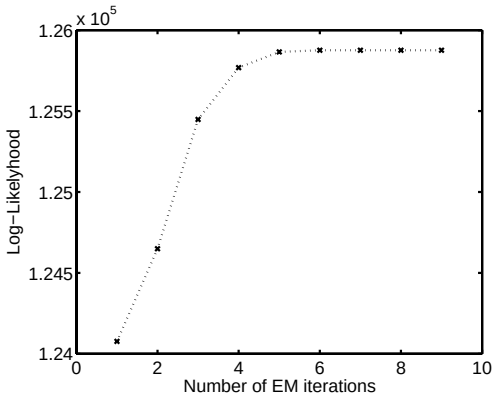

Figure 3. Log–likelihood maximization with EM algorithm. The separation into linearized models and the estimation of the optimal parameters converges after approximately five iterations.

4.1 One–dimensional examples

Case 1: HMM–Langevin model with two states. In the first example we produce the output sequence by realizations of the following Langevin equation

$$\ddot{q}(t) = -H^{i(t)}(q(t) - \mu^{i(t)}) - \gamma^{i(t)}\dot{q}(t) + \sigma^{i(t)}\dot{W}(t)\,. \tag{4.1}$$

The parameters are given in Table 1 below. Figure 2 shows one realization of this process with initial values $(q(0), \dot{q}(0)) = (1, 0)$.

We start the EM algorithm for two hidden states; alternatively one can start with more states and cluster the obtained transition matrices, e.g., with PCCA resulting in two metastable hidden states. On average, the EM algorithm converges after approximately five iterations, where the average is taken over

Table 1
Parameters of the two local Langevin models (4.1) used for generating the test sequences. The corresponding Viterbi path is shown in the left panel of the Figure 4.

state i	μ^i	H^i	γ^i	$(\sigma^i)^2$
$i=1$	1.00	600	6	3.69
$i=2$	1.05	300	6	5.76

Table 2
Estimated parameters for the Langevin models averaged over 100 different realizations of the given HMM-Langevin process. Each realization has a length of $M = 200.000$ with $h = 5 \cdot 10^{-4}$.

state i	μ^i	H^i	γ^i	$(\sigma^i)^2$
$i=1$	$1.00 \pm 2.4 \cdot 10^{-7}$	599.9 ± 40	$6.06 \pm 1 \cdot 10^{-1}$	$3.69 \pm 7.1 \cdot 10^{-5}$
$i=2$	$1.05 \pm 1.3 \cdot 10^{-6}$	301.4 ± 26	$6.05 \pm 1 \cdot 10^{-1}$	$5.76 \pm 1.4 \cdot 10^{-4}$

100 different realizations (see Figure 3). If we compare the original model parameters with their estimates from the ensemble of 100 realizations we can see, that the algorithm more or less exactly recovers the original values. See Table 2 and Figure 5 below. Figure 4 shows a comparison of the "true" Viterbi path $i(t)$ with the estimated one.

The numerical error of the parameter estimation as a function of the time step h and the length M of the time series is plotted in the Figures 6 and 7, respectively. In accordance with the law of the large numbers the relative error of the method scales with the inverse square root of the total length M of the (equilibrium) trajectory. On the other hand we observe a rather unequal behaviour of the numerical error as a function of the step size h: the errors of the Hessian and the friction coefficient scale quadratically, whereas the error of the noise intensity scales linearly. This behavior can clearly be explained by the different discretization schemes (2.12)–(2.15) that were used for the various parameters.

Case 2: Diffusive motion in a perturbed three-well potential. As a second example we consider realizations of the Langevin equation

$$\ddot{q}(t) = -\mathrm{grad}\, U(q(t)) - \gamma \dot{q}(t) + \eta \dot{W}(t) \tag{4.2}$$

with the potential defined by

$$U(q) = p(q) + \alpha \, \sin(\beta q)\,, \quad p(q) = \sum_{k=0}^{6} a_k q^k\,,$$

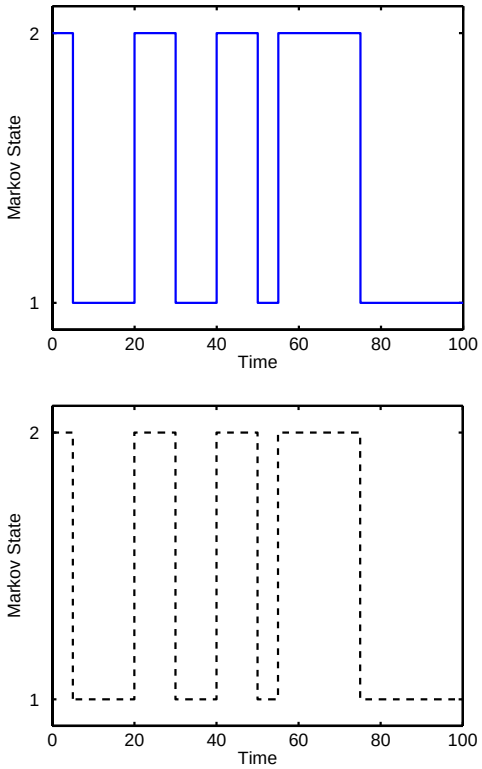

Figure 4. Jumps between the two local Langevin models $i = 1$ and $i = 2$ versus time t. Left: Original sequence used to generate the time series. Right: Computed Viterbi path for $L = 2$.

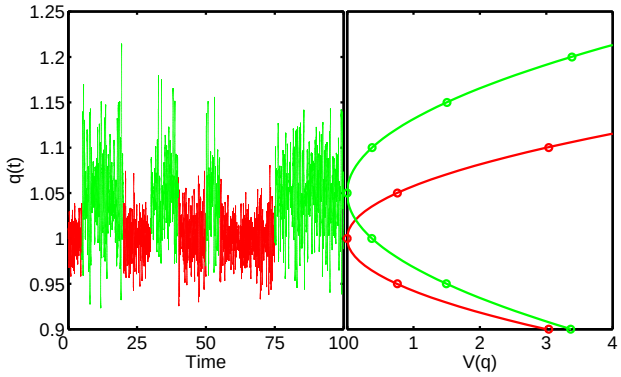

Figure 5. Colouring of the time series according to the optimal decomposition into linearized models. Right panel: two local Langevin potential wells (solid), and harmonic potentials calculated for one realization by means of the EM algorithm with $L = 2$ hidden states (circles). Left panel: original time series coloured according to the HMM states from Figure 4.

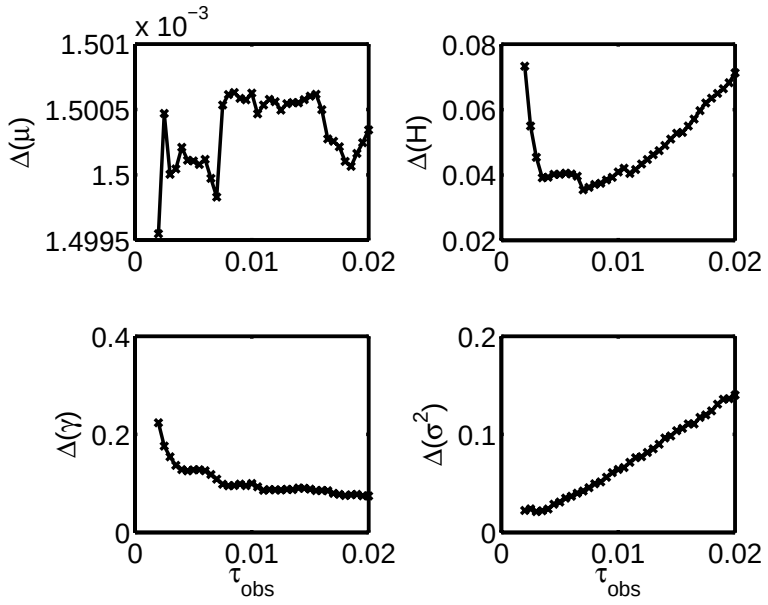

Figure 6. Mean relative error (left) and its variance (right) as functions of time step h for the fixed length of the time series $M = 5.000$. The plot shows the mean value of 100 different realizations.

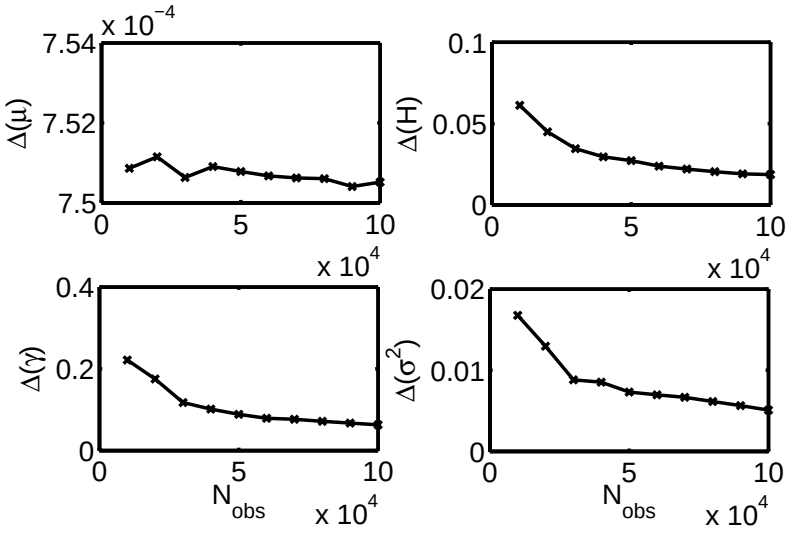

Figure 7. Mean relative error (left) and its variance (right) as functions of the observation time M for fixed time step $h = 5 \cdot 10^{-4}$. The values were averaged over 100 different realizations.

where the parameters are

$$a = (1.3515, 0.2104, -2.3786, -0.1462, 1.0123, -0.0168, -0.0438)$$
$$(\alpha, \beta) = (0.005, 50.000)\,.$$

Table 3
Parameters of the Langevin models (4.2).

	1st Langevin model	2nd Langevin model	3rd Langevin model
μ	-0.97	0.05	0.88
H	7.77	0.44	6.38
γ	1.02	1.00	1.09
η^2	0.109	0.104	0.11

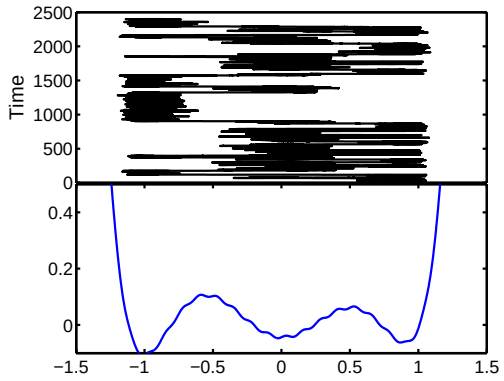

Figure 8. Lower panel: Multi–well potential $U = U(q)$ as defined in the text. Upper panel: Typical realization of the dynamics given by the Langevin equation (4.2) with noise intensity $\eta^2 = 0.1$. The time series has total length of 60.000.

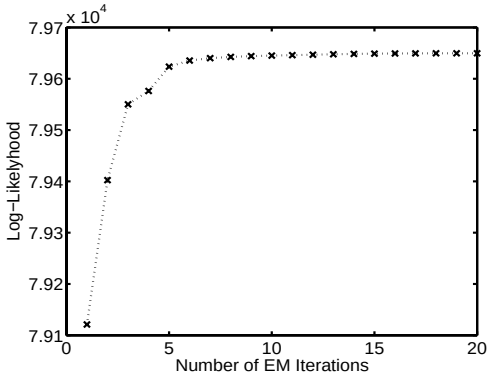

Figure 9. Log–likelihood maximization with the EM algorithm. The separation into linearized models and the estimation of the optimal parameters converges after approximately ten iterations.

This system exhibits metastable transitions between its three wells, if the noise amplitude η is reasonably small; the potential is shown in Figure 8. We set $\eta^2 = 0.1, \gamma = 1$ which leads to metastability, as we can see from the realization shown in Figure 8. The observation sequence is generated by numerical integration of (4.2) using the Euler-Maruyama [29] scheme with time step $\tau = 0.02$. Only every second step enters the observation sequence, thus the observation time step is $h = 0.04$.

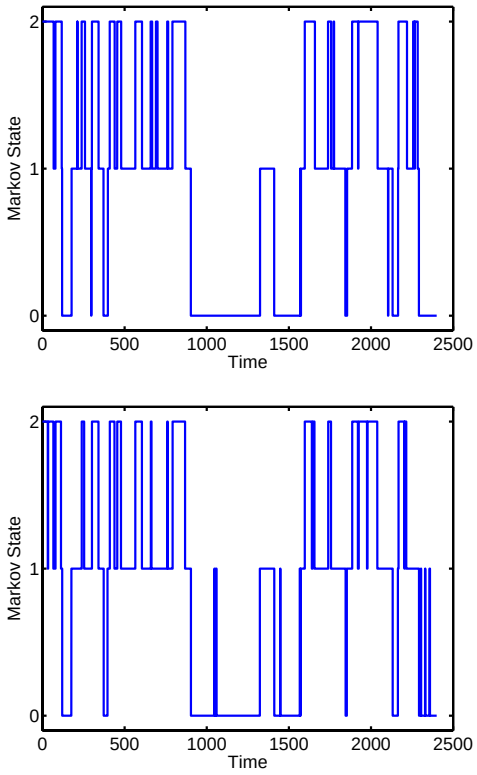

Figure 10. Jumps between the three dominant metastable states $i \in \{1, 2, 3\}$ versus time t. Left: As computed from the original time series with the perturbed three–well potential (state 1 = $\{x < -0.5\}$, state 2 =$\{-0.5 \geq x \geq 0.5\}$, state 3 = $\{x > 0.5\}$). Right: Viterbi path computed for $L = 3$.

The HMM–Langevin model (3.8) is trained on this time series employing the expectation–maximization algorithm for $L = 6$ hidden states (more than we actually expect), the subsequent clustering of the transition matrix results in $L = 3$ hidden states for the jump process. As can be seen from Figure 9 the algorithm quickly converges towards a local maximum of the log–likelihood function. The estimated optimal parameters of three linearized Langevin models are given in the Table 3.

In order to evaluate the quality of the assignment of states to three locally linearized Langevin models, we compare the jump sequence between the three metastable states produced by the original dynamics with that identified by the Viterbi algorithm for $L = 3$. Figure 10 shows that the two pathways are in good agreement. Small deviations between the two paths may result from rare recrossings of the barrier (cf. the time series Figure 8, in particular around $t = 1400$). The shape of the corresponding harmonic potentials in the estimated model is illustrated in Figure 11. Notice that the algorithm resolves the internal structure of the metastable states; both the centers μ^i and the stiffnesses H^i of the harmonic potentials approximate the mean Hessians of the metastable sets quite well.

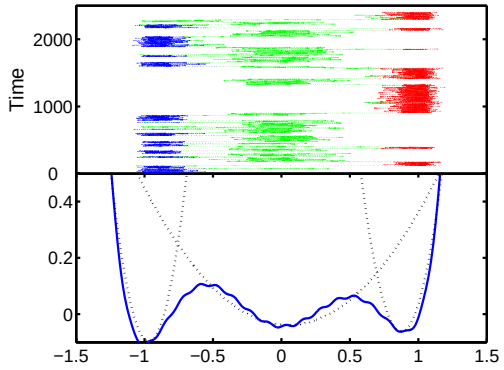

Figure 11. Upper panel: Colouring of the time series according to the optimal decomposition into linearized models. Lower panel: Multi-well potential (solid), and harmonic approximations with $L = 3$ hidden states (dashed).

Table 4
Parameters of the three–hole potential. The corresponding Viterbi path is shown in Figure 13

l	a_l	μ_l	δ_0	k
$l = 1$	3.00	$(0, 1/3)$	0.05	3.00
$l = 2$	-3.00	$(0, 5/3)$	–	–
$l = 3$	-5.00	$(1, 0)$	–	–
$l = 4$	-5.00	$(-1, 0)$	–	–

4.2 High–dimensional examples

Nonlinear potential coupled to a harmonic bath. We consider realizations of the Langevin equation

$$\ddot{q}(t) = -\mathrm{grad}\, U(q(t)) - \gamma \dot{q}(t) + \sigma \dot{W}(t) \tag{4.3}$$

with $q = (x, y) \in \mathbf{R}^2 \times \mathbf{R}^{10}$ and the three–hole potential defined by

$$\begin{aligned} U(x,y) = \sum_{l=1}^{4} a_l \exp(-\langle x - \mu_l, x - \mu_l \rangle + \frac{1}{2} \langle Hy, y \rangle \\ + \delta_0 \left(\cos(2\pi k(x_1 + x_2)) + \cos(2\pi k(x_1 - x_2))\right) , \end{aligned}$$

where $\delta_0 \ll 1$ is a perturbation parameter, and x labels those degrees of freedom which are attached to the three–hole potential; the harmonic bath variables are denoted by y. The parameters of the three–hole potential are given in Table 4 below.

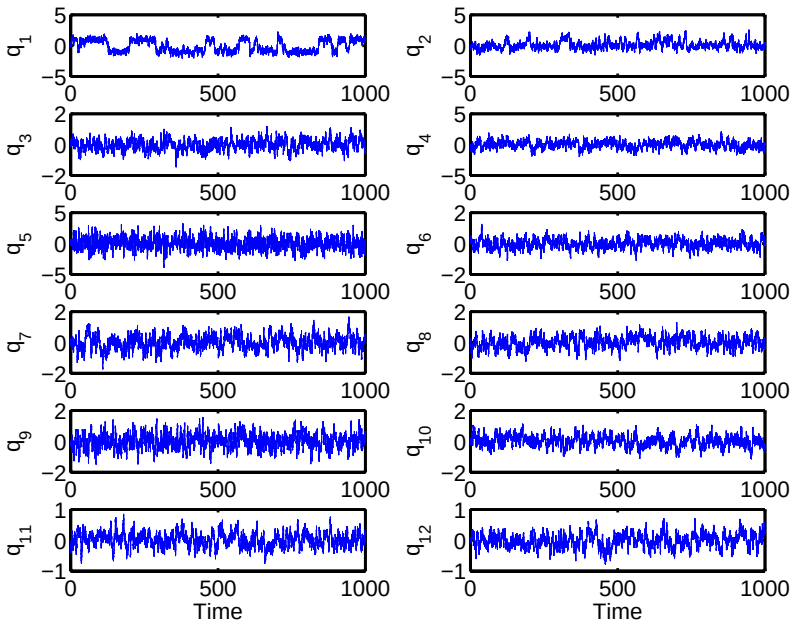

Figure 12. Realization of (4.3) with 60.000 observations and timestep $h = 0.01$.

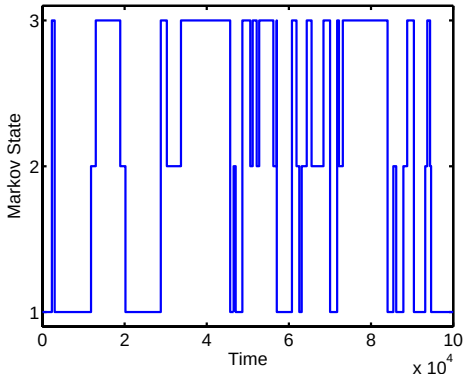

Figure 13. Viterbi path for the three–hole problem.

As a test we generate a realization of the Langevin model 4.3 with 60.000 observations and a time step $h = 0.01$. As the potential energy function in this example has three local wells, the model reduction produces three locally harmonic 12–dimensional models with a Markov chain switching between them. The corresponding Viterbi path produced by the EM algorithm is shown in Figure 13, which should be compared to the projection of the time series onto first two degrees of freedom (see Figure 14). The colouring is due to the computed Viterbi path, and it can be seen that the states of the hidden Markov chain coincide with the respective local minima of the potential energy function.

Additionally we test the quality of estimated parameters by comparing them

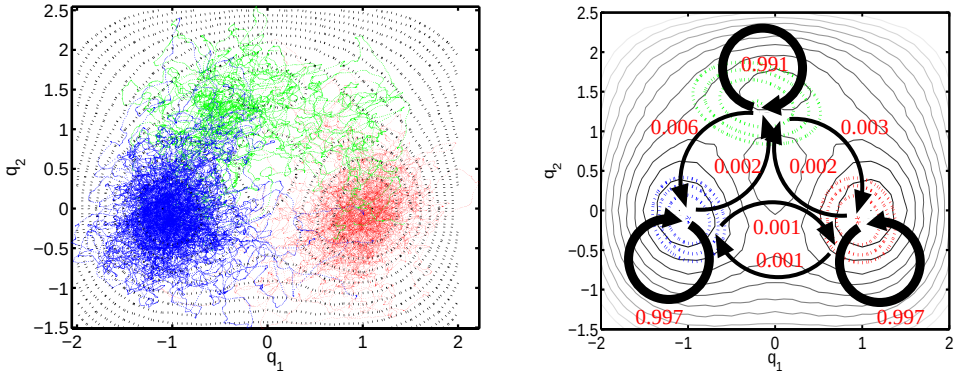

Figure 14. Left: Projection of the 12–dimensional time series onto the 2–dimensional subspace of the three–hole potential. The projected time series is coloured according to the Viterbi path in Figure 13. Right: Comparison of the contour lines of the three–hole potential (solid lines) with the contour plots of three locally harmonic models as obtained from the EM algorithm. The arrows graphically represent transitions and the corresponding rates between the hidden states.

to the exact model parameters that have been used generating the time series (cf. Figure 15). Apparently the estimated parameters are in good agreement with the exact ones, and it is even possible to resolve the fine off–diagonal structure of the parameter matrices, that is responsible for the coupling between different degrees of freedom.

4.3 Dethreading of Cyclophane

In previous examples the performance of the numerical scheme was tested on artificial models with known parameters. In this section we shall apply the technique to a real molecular system whose underlying physical model is *a priori* unknown. To this end we consider a time series of a *Cyclophane dethreading* process that has been provided by Alessandro Laio and Michele Parinello at ETHZ [30]. The system represents a complex of tetracationic Cyclophane and a 1,5-Dihydroxynaphtalene solvated in Acetonitrile as illustrated in Figure 17.

One of the basic insights in the work [30] is that the essential dynamics of the system is well represented by two internal coordinates: q_1 is the distance between the centroids of the Cyclophane and the Naphtalene molecules, and q_2 labels the coordination number of the Naphtalene with the molecules of the solvent. The two–dimensional time series comes as a $7ns$ observation sequence with a time stepping of $h = 2fs$.

In order to estimate if fitting of the linear friction model to the given data is reasonable, we compute the partial autocorrelation function for the velocities v_1 and v_2 out of the differenced time series for q_1 and q_2. We assume that

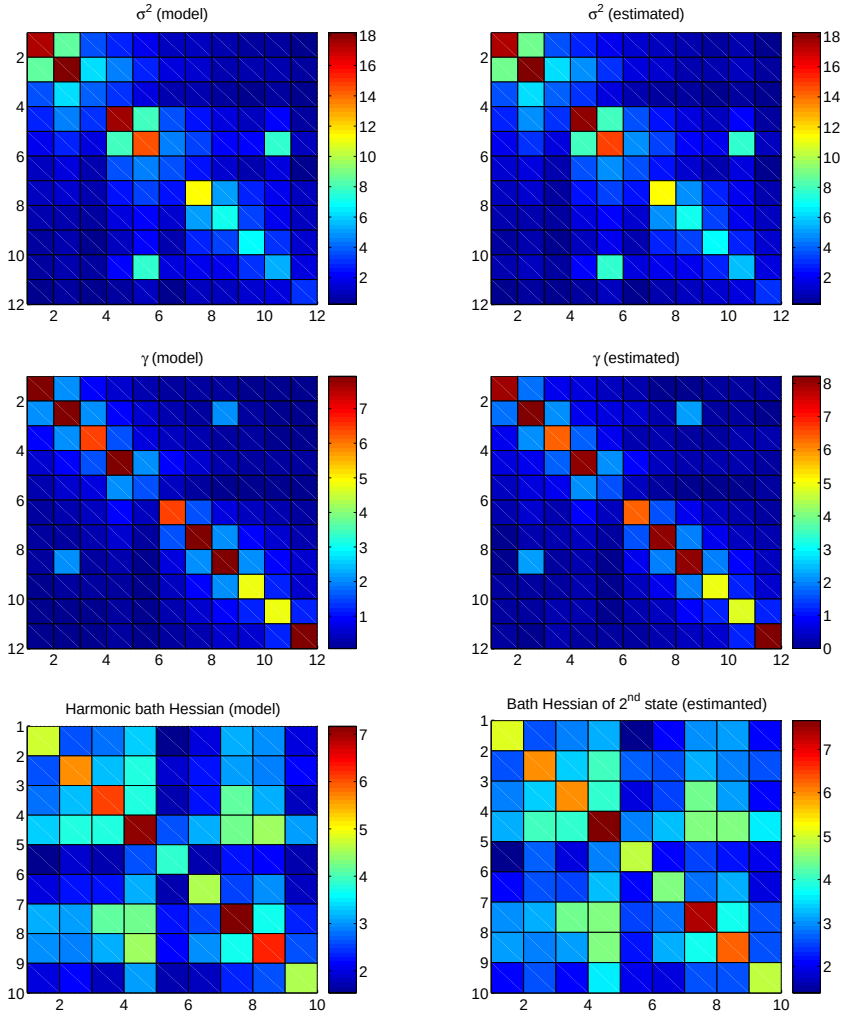

Figure 15. Comparison of the original noise, friction and Hessian matrices (left column) with the parameters estimated by the EM algorithm (right column). The difference between the real and estimated parameters in matrix 2–norm is of the order of magnitude 10^{-2}.

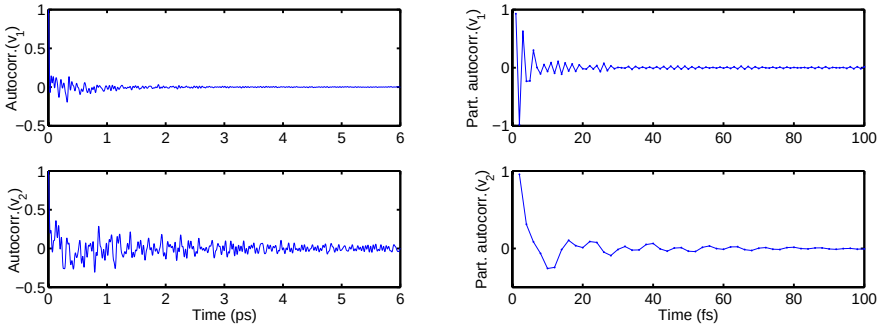

Figure 16. Autocorrelation and partial autocorrelation functions for the velocities v_1, v_2. Note the different time scales on the time–axes of autocorrelation (left) and partial autocorrelation (right).

v_1, v_2 can be considered as a realization of a generalized time–discrete Markov process. Then, as it can be seen from the comparison of autocorrelation and partial autocorrelation functions in Figure 16, the partial autocorrelation of v_1, v_2 decays after about $20fs$, and so does the memory of the process. The ordinary velocity autocorrelation function however tells a different story: here the autocorrelations decay on time scales which are far beyond picoseconds, hence it is misleading regarding memory effects in the system. In Figure 19 it can be seen that the parameters hardly change, while the time step is changed from $2fs$ to $24fs$. Therefore we are confident that linear viscous friction is an appropriate description for the friction in the system.

The free energy landscape computed with respect to the two essential coordinates is anharmonic (see Figure 18). Application of the estimation procedure with one hidden state however produces a meaningful harmonic approximation of the free landscape around the minimum. Incorporating further hidden states in the model clearly gives a better approximation of the free energy landscape and results in a global Langevin model which consists of several locally harmonic Langevin models that are connected by a rapidly mixing Markov chain; note that no severe metastability is observable in the essential subspace.

The convergence of the model parameters as a function of observation sequence length is shown in Figure 19. As it can be clearly seen, even relatively short trajectories (split nanosecond) give reliable estimates of the Langevin parameters, which is in accordance with the lack of metastability. Figure 19 shows the estimated parameters as functions of the observation time step for a fixed–length observation sequence (500.000 points). Finally, Figure 20 illustrates that the numerical effort of the parameter estimation scales linearly with the length of the analyzed time series.

5 Conclusions

The algorithm introduced here allows for the parametrization of reduced models for high–dimensional time series. The proposed Langevin models are simple enough to provide physical insight into complicated data, yet flexible enough, so as to capture a variety of dynamical phenomena. The algorithm does neither require stationarity of the time series, nor thermodynamical equilibrium (fluctuation–dissipation relation). The numerical effort of the method scales linearly with the total length of the time series, quadratic in the dimensionality and the number of hidden states, i.e., in the number of local models (cf. [12]); nevertheless the method works quite well even for high–dimensional data, although estimating the parameters for the Langevin equation is a global nonlinear optimization problem. Moreover the method reveals information about

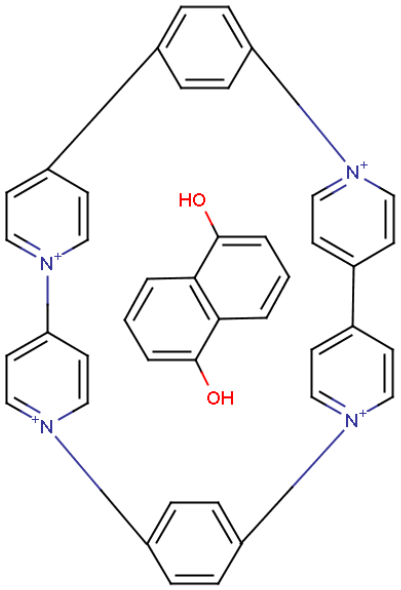

Figure 17. Chemical structure of Cyclophane (left) and the 1,5-Dihydroxynaphtalene (right)

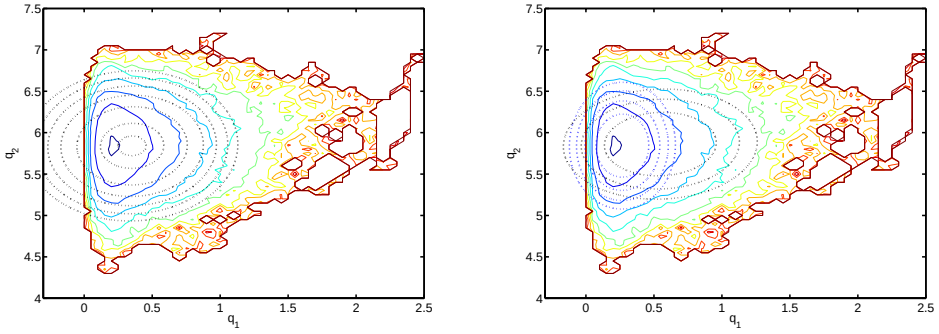

Figure 18. Comparison of the free energy surface (solid) of the reduced time series with the locally harmonic potentials of the Langevin models (dashed) with one hidden state (left) or two hidden states (right).

the interaction and coupling of certain degrees of freedom, that is, it allows to decide whether certain dynamical effects are induced by friction, deterministic motion or noise.

The parameter estimation for the reduced model is based on a predictor–corrector scheme exploiting an analytical solution to the corresponding maximum–likelihood problem. We have shown in the examples section by means of several model problems that the numerics successfully recovers the original parameters of the used model, whenever the time stepping between successive observations is not too large. The time stepping issue reveals the main difficulty for the algorithm: what does a small step size mean? Unfortunately there is no *a priori* criterion at hand in order to decide whether a given time

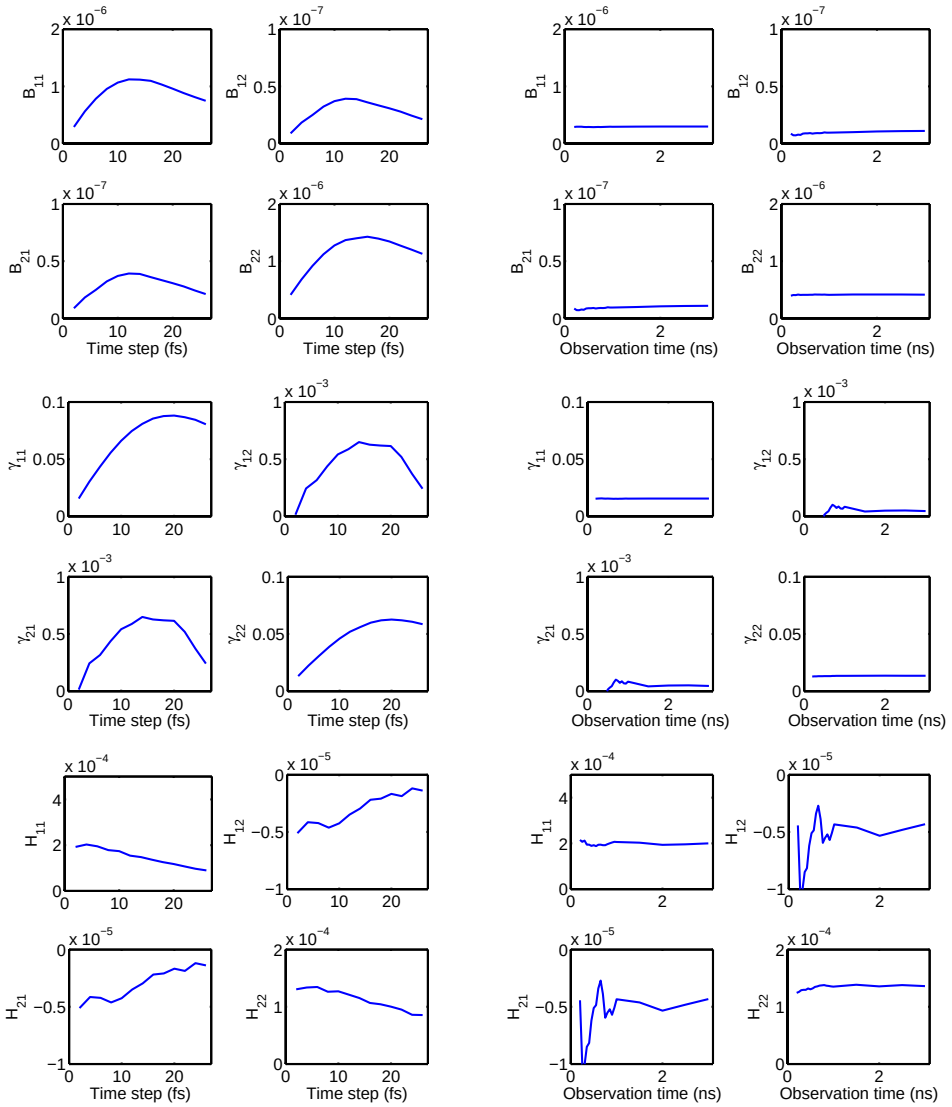

Figure 19. Convergence of the 2–dimensional parameter estimation for different time steps h (left column), and with increasing length of the observation time M for a single hidden state (right column).

series is fine enough or not. However the parameter estimation can be performed, checking *a posteriori* whether the truncated terms in the short–time asymptotics are negligible indeed. Alternatively we could also solve the exact equations of motion for the parameters numerically, i.e., without any approximations, and then use this result maximizing the log-likelihood by means of Newtons method with an appropriate damping scheme. However we have decided to stick to the analytical expressions that are available from the lowest–order perturbative expansion, since this has proven quite efficient, and it lets the parameter estimation be remarkably robust.

A second restriction concerns the linearity of the Langevin equation: neither do we consider memory effects, nor do we treat Langevin equations that orig-

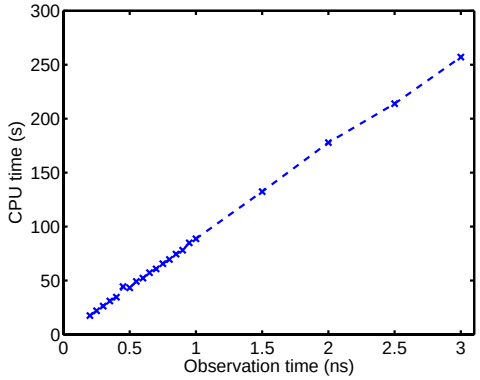

Figure 20. Numerical performance (CPU time in seconds) of the algorithm as a function of the time series' length (in nanoseconds).

inate from non–separable Hamiltonians. Memory effects play a crucial role on time scales, where partial correlations in the system have not been decayed yet. Although the correlation times of the "global" autocorrelation functions are far beyond the short time intervals between the individual observations [31], partial autocorrelations, which are a measure for the memory in the system, often decay much faster. In this case it seems reasonable to parameterize the linear Langevin model, as was shown in the last example. The limitation further pertains data that arise, e.g., in rigid body motion or in coarse–grained modelling of DNA [32], for such systems usually have non–separable Hamiltonians. An extension of the present method to non–separable systems is work in progress.

Acknowledgements

We would like to thank John H. Maddocks and Christof Schütte for stimulating discussions concerning the solution of the Langevin equation. Moreover we are indebted to Alessandro Laio for providing the Cyclophane data. The work of IH is supported by the DFG-SFB 450 "Analysis and Control of Ultrafast Photoinduced Reactions", CH is supported by the DFG Research Center Matheon "Mathematics for Key Technologies" in Berlin. Finally, Frank Noe is acknowledged for carefully reading this manuscript.

References

[1] P. Holmes, J.L. Lumley, and G. Berkooz. *Turbulence, Coherent Structures, Dynamical Systems and Symmetry.* Cambridge University Press, 1996.

[2] D. Givon, R. Kupferman, and A. Stuart. Extracting macroscopic dynamics: Model problems and algorithms. *Nonlinearity*, 17:R55–R127, 2004.

[3] R. Kupferman and A.M. Stuart. Fitting sde models to nonlinear kac-zwanzig heat bath models. *Physica D*, 199:279–316, 2004.

[4] G. Box and G. Jenkins. *Time Series Analysis, Forecasting, and Control*. Holden–Day, 1976.

[5] S. Makridakis, S.C. Wheelwright, and R.J. Hyndman. *Forecasting: methods and applications*. John Wiley & Sons, New York, 1998.

[6] A. Pankratz. *Forecasting with univariate Box-Jenkins models: concepts and cases*. John Wiley & Sons, New York, 1983.

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

[8] A. Fischer, S. Waldhausen, I. Horenko, E. Meerbach, and C. Schütte. Identification of biomolecular conformations from incomplete torsion angle observations by hidden Markov models. *J. Chem. Phys.*, 2005. submitted.

[9] V. Schultheis, T. Hirschberger, H. Carstens, and P. Tavan. Extracting markov models of peptide conformational dynamics from simulation data. *J. Chem. Theory Comput.*, 1:515–526, 2005.

[10] A.H. Monahan. Nonlinear principal component analysis by neural networks: Theory and application to the lorenz system. *J. Climate*, 13:821–835, 2000.

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

[12] I. Horenko, E. Dittmer, and A. Fischer C. Schütte. Automated model reduction for complex systems exhibiting metastability. *SIAM Multiscale Modeling and Simulation*, 2005. submitted.

[13] I. Horenko, E. Dittmer, F. Lankas, J. Maddocks, P. Metzner, and Ch. Schütte. Macroscopic dynamics of complex metastable systems: Theory, algorithms, and application to b-DNA. *J. Appl. Dyn. Syst.*, 2005. submitted.

[14] P.J. Brockwell and R.A. Davis. *Introduction to Time Series and Forecasting*. Springer, Berlin, 2002.

[15] H Risken. *The Fokker-Planck Equation. Methods of Solution and Applications*. Springer, Berlin, 1992.

[16] C.W. Gardiner. *Handbook of Stochastic Methods for Physics, Chemistry and the Natural Sciences*. Springer, Berlin, 2004.

[17] W. Wang and R.D. Skeel. Analysis of a few numerical integration methods for the langevin equation. *Mol. Phys.*, 101(14):2149–2156, 2003.

[18] P. Deuflhard. *Newton Methods for Nonlinear Problems. Affine Invariance and Adaptive Algorithms*, volume 35 of *Computational Mathematics*. Springer, Heidelberg, 2004.

[19] J.A. Bilmes. *A Gentle Tutorial of the EM Algorithm and its Applications to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Thechnical Report*. International Computer Science Institute, Berkeley, 1998.

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

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

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

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

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

[25] C. Schütte and W. Huisinga. On conformational dynamics induced by Langevin processes. In B. Fiedler, K. Gröger, and J. Sprekels, editors, *Equadiff 99*, volume 2 of *Proceedings of the International Conference on Differential Equations*, pages 1247–1262. World Scientific, 2000.

[26] P. Deuflhard, M. Dellnitz, O. Junge, and Ch. Schütte. Computation of essential molecular dynamics by subdivision techniques. In P. Deuflhard, J. Hermans, B. Leimkuhler, A. E. Marks, S. Reich, and R. D. Skeel, editors, *Computational Molecular Dynamics: Challenges, Methods, Ideas*, volume 4 of *Lecture Notes in Computational Science and Engineering*, pages 98–115. Springer, Heidelberg, 1999.

[27] W. Huisinga and B. Schmidt. Metastability and dominant eigenvalues of transfer operators. In C. Chipot, R. Elber, A. Laaksonen, B. Leimkuhler, A. Mark, T. Schlick, C. Schütte, and R. Skeel, editors, *New Algorithms for Macromolecular Simulation*, volume 49 of *Lecture Notes in Computational Science and Engineering*, pages 167–182. Springer, 2005.

[28] M. Weber. Clustering by using a simplex structure. *ZIB–Report*, 03:1–22, 2004.

[29] P.E. Kloeden and E. Platen. *Numerical Solution of Stochastic Differential Equations*. Springer, Berlin, 1999.

[30] A. Laio, A. Rodriguez-Fortea, F.L. Gervasio, M. Ceccarelli, and M. Parinello. Assessing the accuracy of metadynamics. *J. Phys. Chem. B*, 109(14):6714 – 6721, 2005.

[31] W. Min, G. Luo, B.J. Cherayil, S.C. Kou, and X.S. Xie. Observation of a power-law memory kernel for fluctuations within a single protein molecule. *Phys. Rev. Lett.*, 94:198302, 2005.

[32] O. Gonzalez and J.H. Maddocks. Extracting parameters for base-pair level models of dna from molecular dynamics simulations. *Theoretical Chemistry Accounts*, 106:76–82, 2001.