Blind model reduction for high-dimensional time-dependent data

Illia Horenko, Carsten Hartmann

Freie Universität Berlin, Institut für Mathematik
Arnimallee 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 certain dynamical effects are influenced by friction, noise, or systematic drift. 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. Our approach emphasizes the specific hypoelliptic structure of the Langevin equation given high-dimensional observation data, and therefore can be considered as complemetary to the procedure recently proposed in *Horenko et al. (submitted SIAM MMS, 2007)* by one of the authors, or to the problem of incomplete (one-dimensional) observations *Pokern et al. (submitted to JRSSB, 2007).* 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 which it is particularly suited to treat high-dimensional data. That is, we decompose the data into several subsets, each of which gives rise to an appropriate linear Langevin model, where 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),

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

`chartman@math.fu-berlin.de` (Carsten Hartmann).

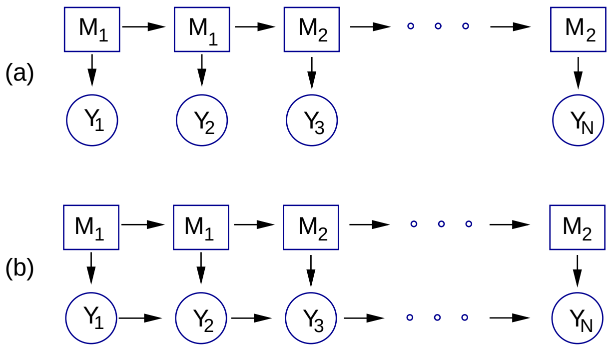

Figure 1. Dynamical Bayesian Networks. Here the arrows denote the casual dependencies, M_i labels the hidden variable or model, Y_i is the observation. In the standard HMM approach the observation is triggered by the sequence of hidden states for a prescribed probability distribution of the output (Figure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

stochastic systems, which means that they are not based on a reasonable 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 multidimensional 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 an observed time series, provided the time steps between successive observations are not too large. In this sense the approach allows for the construction of global physical models for high-dimensional data. A very similar procedure that has been put forward in [14] does not have the limitations regarding the observation time lag, but it fails to preserve the hypoelliptic structure of diffusion model. Finally, the authors of [15] follow a maximum-likelihood strategy for incomplete observations (without momenta). However so far the approach therein is limited to one-dimensional configuration 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 parametrized to Langevin equations on Euclidean configuration space $Q \subseteq \mathbf{R}^n$, which are of the form:

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

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 stemming 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 the mass matrix $M \in \mathbf{R}^{n\times n}$ are symmetric, positive definite matrices, where we do not assume that M is diagonal. The noise amplitude $\sigma \in \mathbf{R}^{n\times n}$ is definite. Exploiting further that some of the involved matrices are symmetric, reduces the number of undetermined parameters from n^2 to $n(n+1)/2$ for the respective matrices matrix.

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

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

Clearly, estimating the parameters in the last equation is hopeless for a general nonlinear potential U. Here we assume that the potential is quadratic, i.e.,

$$U(q) = \frac{1}{2}(q-\mu)^T H(q-u)\,,$$

where $H = D^2U(q) \in \mathbf{R}^{n\times n}$ is symmetric and positive definite. As we will see below this harmonic approximation leads to computationally tractable problems if we embed the parameter estimation procedure into the HMM framework. Moreover harmonic approximations have proven useful on various occasions for elliptic stochastic differential equations [12,13].

Here we do not assume that the time series corresponds to an equilibrium process. Hence we do not require that any kind of fluctuation-dissipation relation between noise and friction coefficients is met. This makes it impossible to estimate the mass matrix explicitly unless we can observe velocities and conjugate momenta independently such that we can take advantage of $p = M\dot{q}$. This degeneracy can be made clear upon introducing mass scaled variables

$q \mapsto M^{1/2}q$ and $p \mapsto M^{-1/2}p$. The latter is clearly is a symplectic transform, and the thus scaled equations read

$$\begin{aligned} \dot{q}(t) &= v(t) \\ \dot{v}(t) &= -H(q(t)-\mu) - \gamma v(t) + \sigma \dot{W}(t)\,, \end{aligned} \tag{2.1}$$

where the coefficients transform according to

$$H \mapsto M^{-1/2}HM^{-1/2}\,, \quad \gamma \mapsto M^{-1/2}\gamma M^{-1/2}\,, \quad \sigma \mapsto M^{-1/2}\sigma$$

The mass scaling amounts to setting $M = \mathbf{1}$ in the Langevin equation, such that we identify tangent space and phase space in the sense that $v = p$. Although we do not assume any kind of fluctuation-dissipation relation, it is important to note that the mass scaling respects this particular relation:

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

Consequently the fluctuation-dissipation relation itself does not provide any additional condition, by means of which the mass matrix in the model could be determined. (The only known possibility employs the covariance matrix of the momentum Maxwell distribution in equilibrium.)

Note that the Langevin equation has some rather specific properties as compared to general hypo-elliptic diffusion equations which are due to its statistical mechanics origin. In particular, q and v (or p) have always the same dimensionality, and the Langevin equation transforms like a Hamiltonian vector field under point transformations, i.e., the Itô-Stratonovich ambiguity disappears. Accordingly the indeterminacy of the mass matrix M amounts to a scaling invariance with respect to the (symplectic) mass scaling transformation.

Remark 2.1 *For almost all molecular dynamics simulations, e.g., MD trajectories produced with standard numerical integrators like Leapfrog/Verlet [16], the Cartesian data consist of positions q and velocities v. Hence we can trace any position-dependent observable $\phi : \mathbf{R}^n \to \mathbf{R}^k$ (for example, torsion angles) by means of $\phi(t) = \phi(q(t))$; the respective velocities are then given by $\dot{\phi}(t) = D\phi(q(t)) \cdot v(t)$. However the conjugate momenta to ϕ are in general unknown, since the momenta are obtained via Legendre transform of the Lagrange function which is associated with the model and that must be explicitly known as a function of ϕ and $\dot{\phi}$.*

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 2.2 *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. According to Wold's Theorem [17] any time-discrete stationary process $X_t \in \mathbf{R}$ has an infinite moving average (MA) representation, i.e., it can be written in the form*

$$X_t = \mu + \sum_{\tau=0}^{\infty} \psi_\tau W_{t-\tau} \,,$$

where $\mu = \mathbf{E}X_t$ is the constant expectation value of the process X_t, $W_{t-\tau}$ is a realization of white noise at time $(t-\tau)$, and the coefficients ψ_τ satisfy

$$\sum_{\tau=0}^{\infty} |\psi_\tau| < \infty \,.$$

The latter condition guarantees that the process X_t is invertible (i.e., all eigenvalues of the associated characteristic polynomial lie outside the unit circle in the complex plane), in which case the process has an infinite auto-regressive (AR) representation

$$X_t = \mu + \sum_{\tau=1}^{\infty} \beta_\tau X_{t-\tau} + W_t \tag{2.2}$$

with coefficients β_τ that are functions of the ψ_τ and which are called partial autocorrelations. *The β_τ can be computed from the ordinary autocorrelation function*

[1] For example, consider the autocorrelation function of a discretization of the one-dimensional harmonic oscillator, which is clearly periodic (hence non-stationary). But the system is deterministic and defines a Markov process without memory.

$$\rho(t-s) = \frac{\mathbb{E}[X_t X_s]}{\mathbb{E}[X_t^2]}\,, \quad \rho(t-s) = \rho(s-t)$$

by means of the Yule-Walker equations [19]:

$$\begin{aligned}
\rho(1) &= \beta_1\rho(0) + \beta_2\rho(1) + \beta_3\rho(2) + \beta_4\rho(3) + \dots \\
\rho(2) &= \beta_1\rho(1) + \beta_2\rho(0) + \beta_3\rho(1) + \beta_4\rho(2) + \dots \\
\rho(3) &= \beta_1\rho(2) + \beta_2\rho(1) + \beta_3\rho(0) + \beta_4\rho(1) + \dots \\
&\vdots
\end{aligned}$$

As both partial and ordinary autocorrelations will decay as time increases, the infinite system of Yule-Walker equations can be truncated in practice, such that the partial autocorrelations turn out to be the adequate statistical measure for the depth of the non-Markovian memory for stationary time-discrete stochastic processes [18]. In fact, in many interesting cases the autocorrelation function decays rather slowly, whereas the corresponding partial autocorrelation decays several orders of magnitude faster indicating that the memory in the system cannot be estimated just by looking at the autocorrelation function (see the Cyclophane example in the numerics section).

If the partial autocorrelation reveals non-negligible memory at the timescale of interest, the simple Langevin model with linear friction should be replaced by a generalized Langevin equation [20,21] that involves a nontrivial memory kernel, and which can be considered the continuous variant of the auto-regressive model (2.2). For the memory problem we refer to the recent paper [22] by the authors.

3 Optimal model parameters

Given an observation time series, the aim of the current work is to find optimal parameters for the model equations (2.1) by means of some maximum-likelihood principle. To this end it is helpful to rewrite the linear Langevin equation (2.1) as the following first-order system

$$\dot{x}(t) = F\,(x(t) - \nu) + \Sigma \dot{B}(t)$$

with the abbreviations $x = (q, v) \in \mathbf{R}^{2n}$ and

$$F = \begin{pmatrix} \mathbf{0} & \mathbf{1} \\ -H & -\gamma \end{pmatrix}, \quad \nu = \begin{pmatrix} \mathbf{0} \\ \mu \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma \end{pmatrix}, \quad B(t) = \begin{pmatrix} \tilde{W}(t) \\ W(t) \end{pmatrix}$$

Given $h > 0$ the formal solution of the linear Langevin equation becomes

$$x(t+h) = \mu + \exp(hF)(x(t) - \mu) + \int_0^h \exp((h-s)F)\,\Sigma\, dB(s)\,.$$

3.1 Maximum-likelihood principle

Suppose we are given an observation series $X = \{X_1, \ldots X_{M+1}\}$ with $X = (Q, V)$ and equal spacing in time, i.e., $h = t_{k+1} - t_k$. 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

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

with the time-dependent parameters

$$\bar{x}_{k+1} = \mu + \exp(hF)(X_k - \mu)$$

and

$$S = \left(\int_0^h \exp(sF)\,\Sigma\Sigma^T \exp\left(sF^T\right)\right)^{-1}\,.$$

The last expression is well-defined as can be seen writing down the corresponding Lyapunov equation for the covariance matrix S^{-1} in the conditional probability density (3.1), viz.,

$$FS^{-1} + S^{-1}F^T = A\,, \quad A = \exp(hF)\Sigma\Sigma^T \exp(hF^T) - \Sigma\Sigma^T\,.$$

It follows from the inertia theorem for Lyapunov equations [23] that S (or S^{-1}) is unique and symmetric positive definite, whenever the right hand side of the Lyapunov equation is symmetric negative semidefinite and the matrix

$$[A \;\; FA \;\; F^2A \;\; \ldots \;\; F^{2n-1}A]$$

has maximum rank $2n$. Finally, the positive function

$$\rho_0 = \frac{1}{(2\pi)^n}\sqrt{\det S}$$

normalizes the total probability to one.

We define the log-likelihood function of the observation sequence as

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

where

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

denotes the joint probability distribution of the observation sequence for the parameter set $\lambda = (H, \mu, \gamma, \sigma\sigma^T)$, where we used that $\rho_\lambda(X_{k+1}|X_1, \ldots, X_k) = \rho_\lambda(X_{k+1}|X_k)$ is Markovian. The optimal parameters are those which maximize the log-likelihood function. Inserting the equations (3.1) and (3.3) into (3.2) the log-likelihood becomes

$$\mathcal{L}(\lambda|X) = -\frac{M}{2}\log\det S - \frac{1}{2}\sum_{k=1}^{M}\langle S(X_{k+1} - \bar{x}_{k+1}), X_{k+1} - \bar{x}_{k+1}\rangle\ .$$

3.2 Short-time asymptotics

In order to compute the critical point of the log-likelihood function, we evaluate the necessary condition $\mathbf{d}\mathcal{L} = 0$. For this purpose we have to compute the partial derivatives of the log-likelihood with respect to the parameters $H, \mu, \gamma, \sigma\sigma^T$. Instead of evaluating the solution of the Lyapunov equation for the covariance matrix S^{-1} explicitly, it is convenient to consider its Taylor expansion for sufficiently small observation time lag h. An alternative approach is to consider the exact likelihood, treating the entries of the propagator $\exp(hF)$ as unknowns. Although appealing, we do not follow this route here, since the exact procedure does not preserve the specific block structure of drift matrix F; for the details we refer to [14]. Clearly the resctriction on the time lag has to be verified for each data set. For example, given a sufficiently long time series, we could compute the optimal parameters at different time lags and check if the parameters remain constant. We shall come back to this point later on in the examples section, where we consider the dethreading of Cyclophane.

Doing a Taylor expansion in the time lag h the covariance matrix becomes to lowest order

$$(S^{-1})_0 = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & h\sigma\sigma^T \end{pmatrix} .$$

The fact that the covariance matrix is singular amounts to the hypo-ellipticity of the Langevin system: on short times the equation for the position variables is purely deterministic, whereas the noise immediately "diffuses" the velocity observation. On the other hand the log-likelihood function depends upon the shape matrix S rather than the covariance matrix S^{-1}. The lowest-order expansion for the shape matrix yields

$$S_0 = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\sigma\sigma^T)^{-1}/h \end{pmatrix} ,$$

which is consistent with the non-singular part of the expanded matrix S^{-1} above. The singularity that appears in the lower right block as the time lag h goes to zero describes the fact that the former velocity/momentum observation V_k is sharp (i.e. the conditioning argument in $\rho_\lambda(X_{k+1}|X_k)$). Omitting additive constants, the leading order of $\mathcal{L}(\lambda|X)$ thus reads

$$\mathcal{L}_0(\lambda|X) = \frac{M}{2}\log\det\sigma\sigma^T - \frac{1}{2h}\sum_{k=1}^{M}\left\langle \sigma\sigma^T(V_{k+1}-\bar{v}_{k+1}), V_{k+1}-\bar{v}_{k+1}\right\rangle ,$$

where

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

is obtained by a symmetric second-order discretization of the propagator $\exp(hF)$. Computing the partial derivatives we find for the stiffness matrix

$$\frac{\partial\mathcal{L}_0}{\partial H} = -(\sigma\sigma^T)^{-1}\frac{1}{2}\sum_{k=1}^{M}(Q_k-\mu)\otimes\Delta^v_{k+1} , \tag{3.4}$$

using the abbreviation $\Delta^v_{k+1} = V_{k+1} - \bar{v}_{k+1}$. The tensor product (Kronecker product) is defined by $(X\otimes Y)_{ij} = X_iY_j$, where X, Y are any two vectors from $\mathbf{R}^n$. The derivative with respect to μ is

$$\frac{\partial\mathcal{L}_0}{\partial\mu} = \frac{1}{2}H(\sigma\sigma^T)^{-1}\sum_{k=1}^{M}\Delta^v_{k+1} \tag{3.5}$$

Taking the the derivative with respect to the friction matrix yields the expression

$$\frac{\partial \mathcal{L}_0}{\partial \gamma} = -(\sigma\sigma^T)^{-1} \sum_{k=1}^{M} V_k \otimes \Delta_{k+1}^v \,. \tag{3.6}$$

Last but not least we have

$$\frac{\partial \mathcal{L}_0}{\partial \sigma\sigma^T} = \frac{1}{2h}(\sigma\sigma^T)^{-2} \sum_{k=1}^{M} \Delta_{k+1}^v \otimes \Delta_{k+1}^v - \frac{M}{2}(\sigma\sigma^T)^{-1} \,, \tag{3.7}$$

for the derivative with respect to the covariance matrix of the noise process. The unknown parameters $\lambda = (H, \mu, \gamma, \sigma\sigma^T)$ 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 [29].

3.3 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 [7,30,31]. 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 fulfills 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 [32,33]. 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 lowest-order likelihood $\mathcal{L}_0(\theta|X, i)$; 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

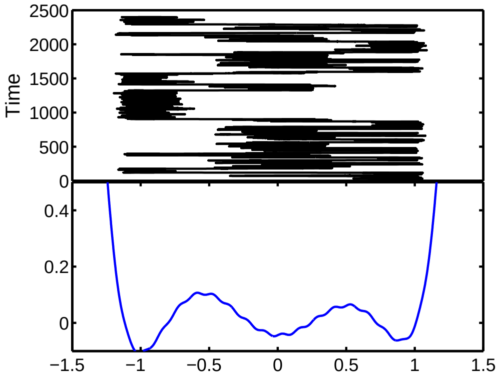

Figure 2. Lower panel: Multi-well potential $U = U(q)$ as defined in the text. Upper panel: Typical realization of the dynamics given by the Langevin equation (4.1) with noise intensity $\eta^2 = 0.1$. The time series has total length of 60.000.

identified by iteratively maximizing the entropy

$$\mathcal{S}(X) = \max_{\theta} \sum_{i} \mathcal{L}_0(\theta|X, i) \log \mathcal{L}_0(\theta|X, i)\,.$$

For the identification of the optimal sequence of hidden metastable states the Viterbi algorithm [34] is used, which exploits dynamic programming techniques to resolve the optimization problem

$$\max_{i} \mathcal{L}_0(\theta|X, i)$$

in a recursive manner. For the details see [35] 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 [36,37]. 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 [12,38,39]. 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 from a one-dimensional Langevin equation whose hidden states are implicitly defined by the metastable sets of a perturbed three-well potential, demonstrating the data-based decomposition of the dynamics

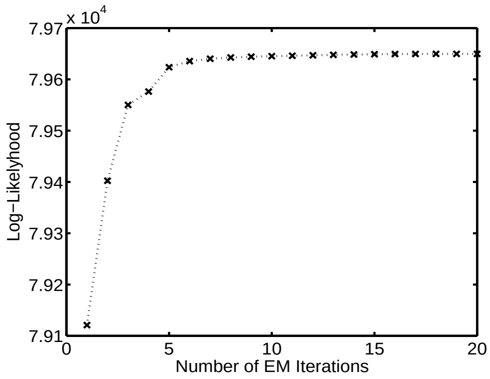

Figure 3. Log-likelihood maximization with the EM algorithm. The separation into linearized models and the estimation of the optimal parameters converges after approximately ten iterations.

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 studying drift, friction and noise parameters for each local model we obtain moreover information about the dominant dynamical effects in the metastable regions. 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.

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

$$\ddot{q}(t) = -\nabla U(q(t)) - \gamma \dot{q}(t) + \eta \dot{W}(t) \tag{4.1}$$

with the potential defined by

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

where the parameters are

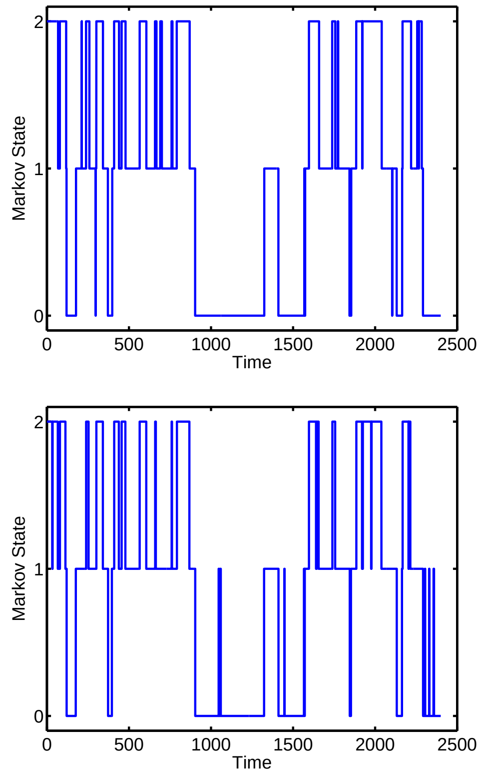

Figure 4. Jumps between the three dominant metastable states $i \in \{1, 2, 3\}$ versus time t. Left: As computed from the original time series with the perturbed three-well potential (state 1 = $\{x < -0.5\}$, state 2 =$\{-0.5 \geq x \geq 0.5\}$, state 3 = $\{x > 0.5\}$). Right: Viterbi path computed for $L = 3$.

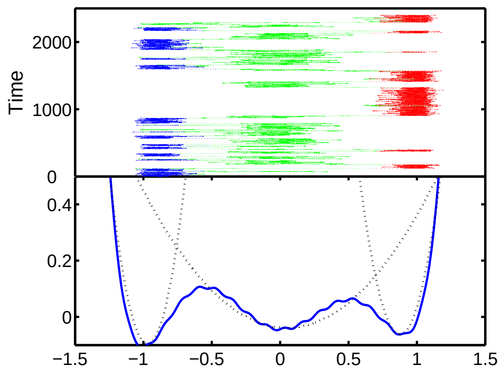

Figure 5. Upper panel: Colouring of the time series according to the optimal decomposition into linearized models. Lower panel: Multi-well potential (solid), and harmonic approximations with $L = 3$ hidden states (dashed).

$$
\begin{aligned}
a &= (1.3515, 0.2104, -2.3786, -0.1462, 1.0123, -0.0168, -0.0438) \\
(\alpha, \beta) &= (0.005, 50.000)\,.
\end{aligned}
$$

This system exhibits metastable transitions between its three wells, if the noise amplitude η is reasonably small; the potential is shown in Figure 2.

Table 1
Parameters of the Langevin models (4.1).

	1^{st} Langevin model	2^{nd} Langevin model	3^{rd} 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

Table 2
Parameters of the three-hole potential. The corresponding Viterbi path is shown in Figure 7

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	$(-1, 0)$	–	–

We set $\eta^2 = 0.1, \gamma = 1$ which leads to metastability, as we can see from the realization shown in Figure 2. The observation sequence is generated by numerical integration of (4.1) using the Euler-Maruyama [40] scheme with time step $\tau = 0.02$. Only every second step enters the observation sequence, thus the observation time step is $h = 0.04$.

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 3 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 1.

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 4 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 2, in particular around $t = 1400$). The shape of the corresponding harmonic potentials in the estimated model is illustrated in Figure 5. 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.

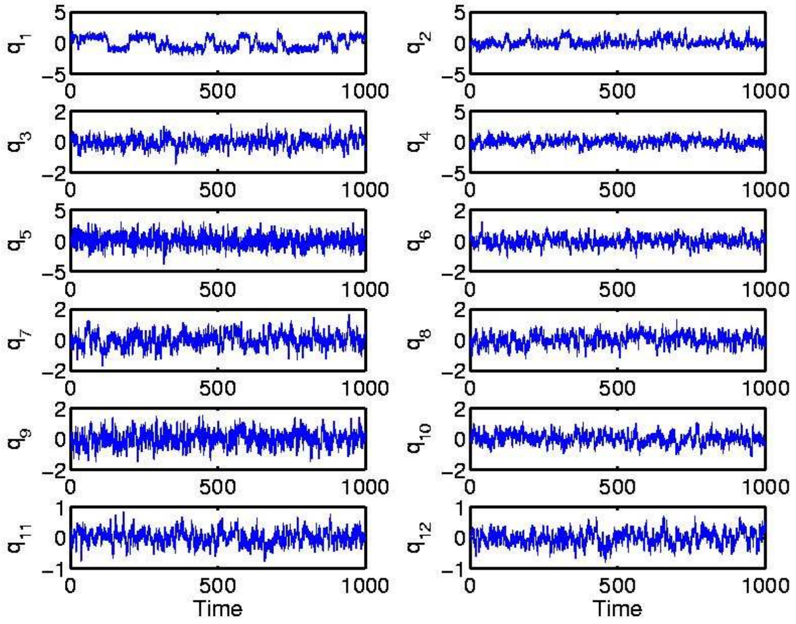

Figure 6. Realization of (4.2) with 60.000 observations and time step $h = 0.01$.

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

$$\ddot{q}(t) = -\nabla U(q(t)) - \gamma \dot{q}(t) + \sigma \dot{W}(t) \tag{4.2}$$

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

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

where $\delta_0 \ll 1$ is a perturbation parameter, and $x = (x_1, x_2)$ labels those degrees of freedom of the three-hole potential, where the wells (holes) are located at $(-1, 0)$, $(1, 0)$ and $(0, 5/3)$; the harmonic bath variables are denoted by y. The parameters of the three-hole potential are given in Table 2 below.

As a test we generate a realization of the Langevin model 4.2 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 7, which should be compared to the projection of the time series onto first two degrees of freedom (see Figure 8). The colouring is due to the com-

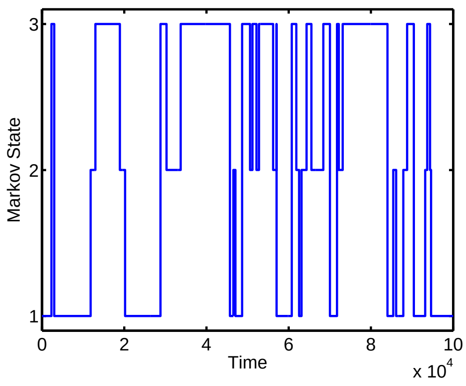

Figure 7. Viterbi path for the three-hole problem.

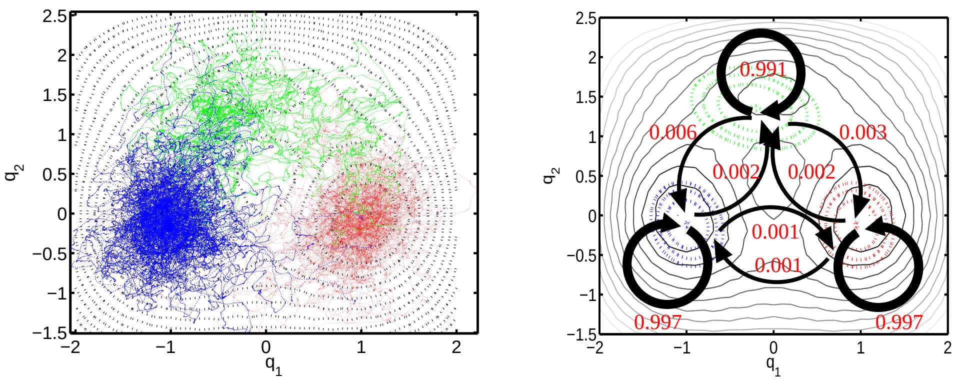

Figure 8. Left: Projection of the 12-dimensional time series onto the 2-dimensional subspace of the three-hole potential. The projected time series is coloured according to the Viterbi path in Figure 7. Right: Comparison of the contour lines of the three-hole potential (solid lines) with the contour 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.

puted 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 to the exact model parameters that have been used generating the time series (cf. Figure 9). 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.

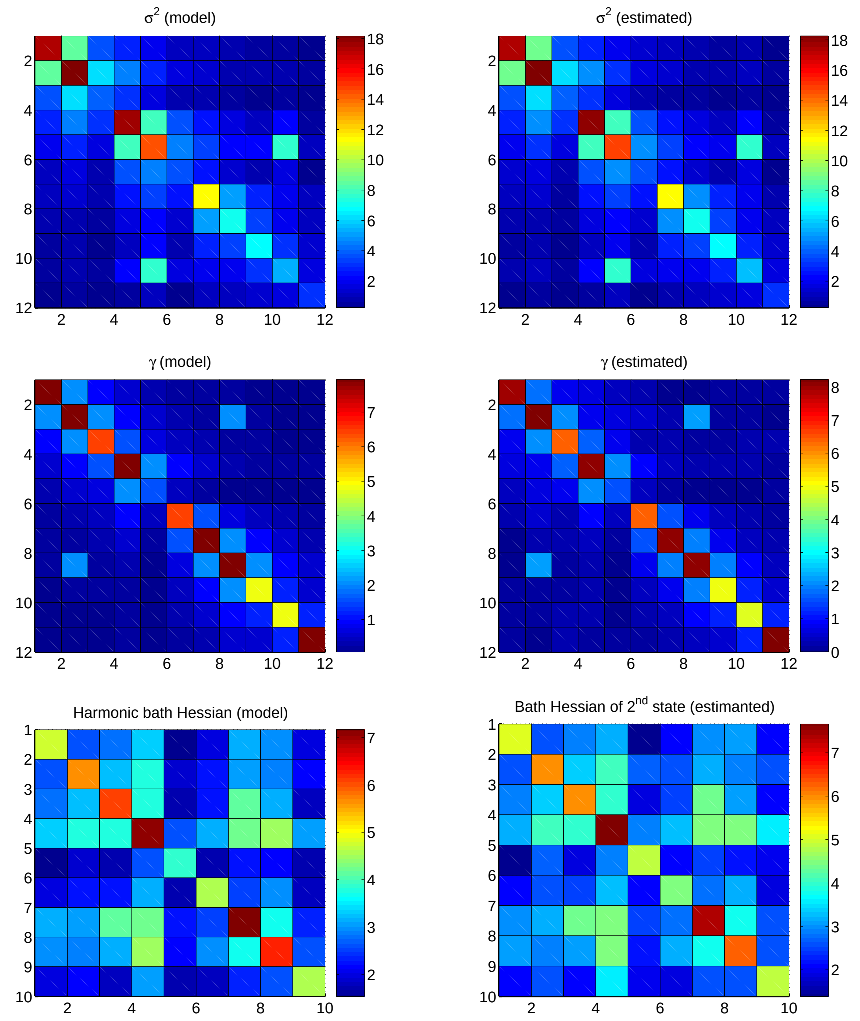

Figure 9. Comparison of the original noise, friction and Hessian matrices (left column) with the parameters estimated by the EM algorithm (right column). The difference between the real and estimated parameters in matrix 2-norm is of the order of magnitude 10^{-2}.

4.1 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 [41]. The system represents a complex of tetracationic Cyclophane and a 1,5-Dihydroxynaphtalene solvated in Acetonitrile as illustrated in Figure 10.

One of the basic insights in the work [41] is that the essential dynamics of the system is well represented by two internal coordinates: q_1 is the distance

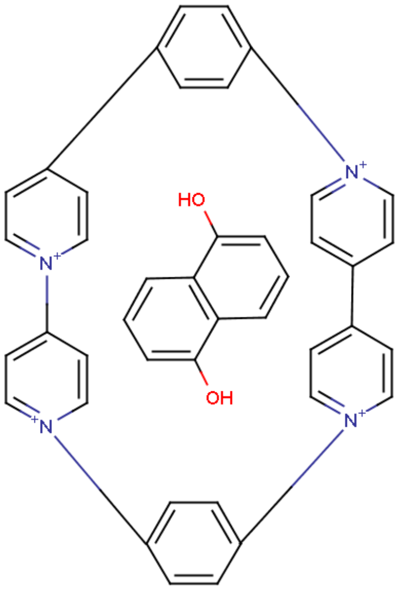

Figure 10. Chemical structure of Cyclophane (left) and the 1,5-Dihydroxynaphtalene (right)

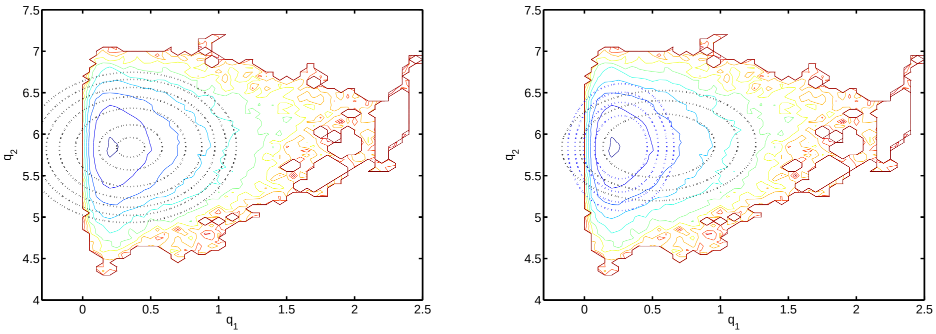

Figure 11. Comparison of the free energy surface (solid) of the reduced time series with the locally harmonic potentials of the Langevin models (dashed) with one hidden state (left) or two hidden states (right).

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 lag time of $h = 2$fs between successive obervations.

The free energy landscape computed with respect to the two essential coordinates is anharmonic (see Figure 11). 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

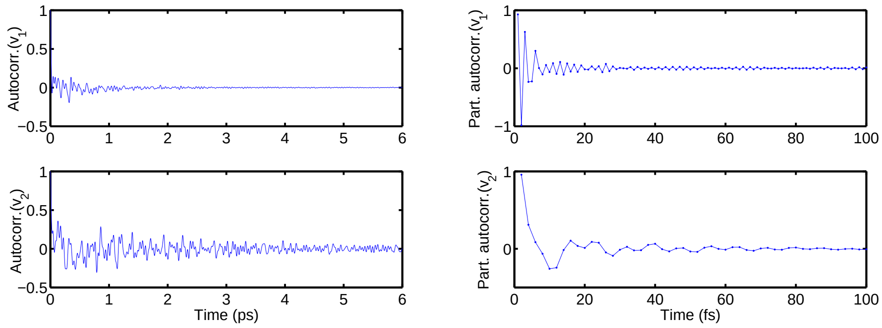

Figure 12. Autocorrelation and partial autocorrelation functions for the velocities v_1, v_2. Note the different time scales on the time axes of autocorrelation (left) and partial autocorrelation (right).

Markov chain.

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. We assume that 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 12, 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. Indeed, as Figure 13 shows (left panel), some of the estimated parameters change about one order of magnitude, while the observation lag time is varied from 2fs to 24fs. This could be either because of a too large lag time or because of memory effects that prohibit the use of a Markovian model below a lag time of 20fs. Finally, the convergence of the model parameters for a as a function of observation sequence length for fixed lag time is illustrated in Figure 13 (right panel).

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,

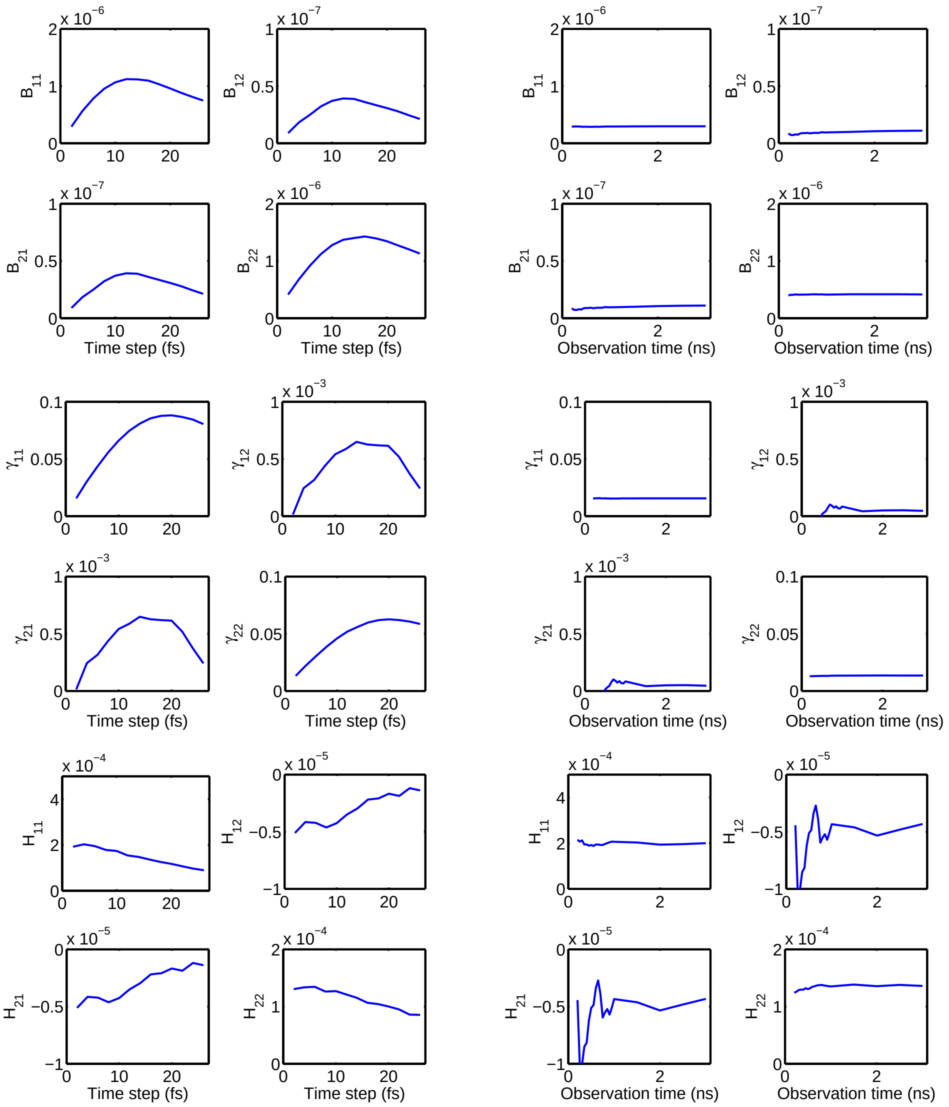

Figure 13. Convergence of the 2-dimensional parameter estimation for different time steps h (left column), and with increasing length of the observation time M for a single hidden state (right column).

although estimating the parameters for the Langevin equation is a global nonlinear optimization problem. Moreover the method reveals information about the interaction and coupling among certain degrees of freedom or regions in phase space. In addition to that we gain some knowledge about the dominant dynamical effects in the metastable regions, by means of which one could, for example, explain why different molecular conformations have different flexibilities.

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

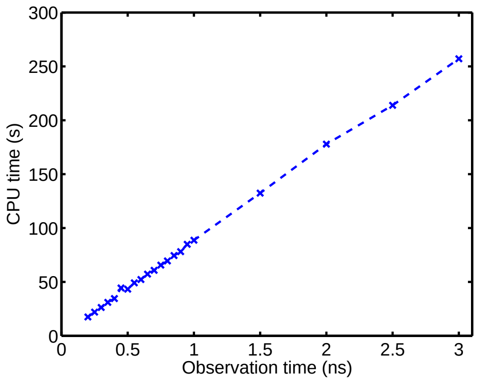

Figure 14. Numerical performance (CPU time in seconds) of the algorithm as a function of the time series' length (in nanoseconds).

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 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 originate from non-separable Hamiltonians. The second limitation pertains data that arise, e.g., in rigid body motion or in coarse-grained modelling of DNA [43], for such systems are usually described by 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 [42], partial autocorrelations, which are a measure for the memory in the system, often decay much faster. One such instance is the cyclophane example where we show that the observation time lag restriction may compete with the requirement of choosing a time lag that is larger than the maximum decay time of the partial autocorrelations. The problem of finding appropriate parameters for non-Markovian systems exhibiting memory is addressed in the recent work [22] by the authors.

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] I. Horenko and Ch. Schtte. Likelihood-Based Estimation of Multidimensional Langevin Models and its Application to Biomolecular Dynamics. *submitted to SIAM Mult. Mod. Sim.*, 2007.

[15] Y. Pokern, A. Stuart, and P. Wiberg. Parameter estimation for partially observed hypoelliptic diffusions. *submitted to Journal of the Royal Statistical Society of Britain*, 2007.

[16] M. Allen and D. Tildesley. *Computer Simulations of Liquids*. Clarendon Press, Oxford, 1990.

[17] H. Wold. *A Study in the Analysis of Stationary Time Series*. Stockholm: Almqvist and Wiksel, 1938.

[18] S.M. Kay. Vector space solution to the multi dimensional yule-walker equations. In *IEEE International Conference on Acoustics, Speech and Signal Proceedings.*, volume 3, pages 289–292, 2003.

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

[20] A.J. Chorin, O.H. Hald, and R. Kupferman. Optimal prediction and the Mori-Zwanzig representation of irreversible processes. *Proc. Natl. Acad. Sci.*, 97(7):2969–2973, 2000.

[21] A.J. Chorin, O.H. Hald, and R. Kupferman. Optimal prediction with memory. *Physica D*, 166:239–257, 2002.

[22] I. Horenko, F. Noe, C. Hartmann, and Ch. Schtte. Data-based parameter estimation of generalized multidimensional Langevin processes. *to appear in Phys. Rev. E*, 2007.

[23] R. Loewy. An inertia theorem for Lyapunov's equation and dimension of a controllability space. *Lin. Alg. Appl.*, 260:1–7,1997

[24] E. J. Heller. Phase space interpretation of semiclassical theory. *J. Chem. Phys.*, 67(7):3339–3351, 1977.

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

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

[27] I. Horenko, S. Lorenz, Ch. Schütte, and W. Huisinga. Adaptive approach for non-linear sensitivity analysis of reaction kinetics. *J. Comp. Chem.*, 26(9):941–948, 2005.

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

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

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

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

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

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

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

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

[36] 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.

[37] 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.

[38] 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.

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

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

[41] 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.

[42] 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.

[43] 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.