Likelihood-Based Estimation of Multidimensional Langevin Models and its Application to Biomolecular Dynamics*

Illia Horenko[**1] and **Christof Schütte**[***1]

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

Key words multidimensional Langevin equation, hypo-elliptic diffusion, model reduction, exponential estimators, parameter estimation, hidden Markov models, maximum–likelihood principle, molecular dynamics, conformations, effective dynamics, model reduction
Subject classification [PACS05.10.Gg,02.50.Ga,05.10.Gg,64.60.My]

We crucially extend the novel method introduced in Horenko et al. (*Multi. Mod. Sim. 2006, 5:802–827*) for the identification of the most important metastable states of a system with complicated dynamical behavior from time series information. The approach represents the effective dynamics of the full system by a Markov jump process between metastable states, and the dynamics within each of these metastable states by diffusions that include the classes called full and overdamped Langevin dynamics in biophysics (local SDEs). Its algorithmic realization combines Hidden Markov Models (HMMs; for the un-observed jump process between the metastable states) with likelihood-based estimation of the parameters in the local linear SDEs based on discrete-time observations of the system. Despite the local linearity the approach is appropriate to handle nonlinear potentials with several dominant wells. Compared to Horenko et al. (*Multi. Mod. Sim. 2006, 5:802–827*), the algorithms proposed herein allow to handle arbitrary dimensions instead of just one-dimensional local SDEs, and to extend the class of diffusions considered. Moreover, the role of the fluctuation-dissipation relation in parameter estimation for Langevin models is discussed in detail. The performance of the algorithms is illustrated by numerical tests and by application to molecular dynamics time series of a 12-alanine molecule with implicit water.

This is a preliminary version. Do not circulate!

Introduction

Modelling and simulation of large molecular systems is a field of world–wide activity with applications ranging from materials science to modelling of highly complex biomolecules like proteins and DNA. Increasing amounts of "raw" simulation data and growing dimensionality of molecular dynamics simulation of processes in molecular systems have led to a persistent demand for modelling approaches which allow to extract physically interpretable information out of large data-sets. Whenever time series are concerned multidimensional statistical analysis may not be enough: whenever a physically appropriate dynamical description of the underlying processes is required one is in need of an approach that allows for *automatized* generation of appropriate physical models out of the (noisy) data. Such *data–driven* approaches should be carefully distinguished from *model–driven* ones (like model reduction, mode elimination, homogenization, or stochastic (re)modelling), which aim at reducing the dimension and/or complexity of an *analytically given model*, for the context of importance herein, we refer to [1, 2, 3, 4].

This article is concerned with a novel data-driven approach to model reduction in molecular dynamics. In order to understand the starting point of this approach we have to be aware that the macroscopic dynamics of typical biomolecular systems is mainly characterized by the existence of biomolecular conformations which can be understood as geometries that are persistent for long periods of time. On

* Supported by the DFG research center MATHEON "Mathematics for key technologies" in Berlin and by Microsoft Research within the project "Conceptionalization of Molecular Dynamics".
** E-mail: horenko@math.fu-berlin.de
*** E-mail: schuette@math.fu-berlin.de

the longest time scales biomolecular dynamics is a kind of flipping process between these conformations [5, 6]. Biophysical research seems to indicate that typical biomolecular systems possess only few dominant conformations that can be understood as metastable or almost invariant sets in state or configuration space [7, 8, 9]. In other words, the effective or macroscopic dynamics is given by a Markov jump process that hops between the metastable sets while the dynamics within these sets might be mixing on time scales that are smaller than the typical waiting time between the hops. In many applications this Markovian picture is an appropriate description of the dynamics since typical correlation times in the system are sufficiently smaller than the waiting times between hops (and thus much smaller than the timescale the effective description is intended to cover). The same description of the effective dynamics is true for other complex system including, e.g., climate systems or systems from materials science.

In this article we will extend an approach to the data-driven construction of reduced models for systems with metastable dynamics that has recently been proposed in [10, 11]. Its basic idea is to (1) construct a finite-state Markov jump process that models the hops *between* the metastable conformations, and (2) for each conformation to parameterize an appropriate stochastic model that allows to approximate the dynamics of the systems as long as it is *within* the respective conformation. In [10, 11] appropriate algorithms have been derived that combine hidden Markov models (for the construction of the unobserved jump process), and optimal, likelihood-based parameterization of the local stochastic models. We will stick to this framework here. However, the results in [10] are limited to (A) specific types of one-dimensional stochastic models (diffusion processes, often also called *overdamped Langevin models*, for a definition see below), (B) time series with rather short lag times (=time between successive discrete observations of the underlying system); in [11] restriction (B) has been overcome but only for one-dimensional models of the type described in (A). The present article will remove both restrictions in the sense that we will demonstrate how to handle (A) diffusions of the type called full and overdamped Langevin-models in biophysics, and (B) arbitrary dimensions (whenever the available time series is long enough).

The Langevin equations considered herein are of the following type: (i) The full Langevin equation in form of a hypo-elliptic diffusion

$$M\ddot{q}(t) \quad = \quad -\mathrm{grad}\, U(q(t)) - \gamma\dot{q}(t) + \sigma\dot{W}(t). \tag{1}$$

or, alternatively, as a first order system

$$\begin{aligned} \dot{q}(t) &= M^{-1}p(t) \\ \dot{p}(t) &= -\mathrm{grad}\, U(q(t)) - \gamma M^{-1}p(t) + \sigma\dot{W}(t). \end{aligned} \tag{2}$$

Here $U : R^n \to \mathbf{R}$ denotes the interaction potential, grad stands for differentiation with respect to q, p the momenta associated with q, $W(t)$ is standard n-dimensional Brownian motion, γ the friction matrix, M the mass matrix, and σ the noise intensity matrix. $\gamma \in \mathbf{R}^{n\times n}$, $M \in \mathbf{R}^{n\times n}$, and $\sigma \in \mathbf{R}^{n\times n}$ are positive-definite matrices; we do not assume that M is diagonal. (ii) Whenever γ is sufficiently large or M sufficiently small then the q-dynamics given by (1) is approximately governed by the so-called overdamped Langevin (or diffusive or Smoluchowski) dynamics:

$$\gamma\dot{q}(t) = -\mathrm{grad}\, U(q(t)) + \sigma\dot{W}(t). \tag{3}$$

Langevin models, especially the hypo-elliptic variant, seem to be of major importance here since a series of results from the physical and biophysical literature advocate this type of dynamics or its generalized variants as the reduced models of choice for the effective dynamics of molecular systems [12, 13, 14, 15, 3, 16, 17, 18, 19]. Introducing standard phase-space variables $z = (q, p)$ for positions and momenta, we can rewrite both types of the Langevin equation, hypo-elliptic as well as overdamped, in form of the following first order system

$$\dot{z}(t) \quad = \quad F\left(z(t)\right) + \Sigma\dot{W}(t).$$

Putting these types of local SDEs and the jump process between conformations together, we get reduced models of the form

$$\dot{z}(t) \quad = \quad F^{(i(t))}(z) + \Sigma^{(i(t))}\,\dot{W}(t) \tag{4}$$

$$i(t) \quad = \quad \text{Markov jump process with states } 1, \ldots, L, \tag{5}$$

where $W(t)$ is denoting standard n or $2n$-dimensional Brownian motion, $(\Sigma^{(1)}, \ldots, \Sigma^{(L)})$ noise intensity matrices, and $(F^{(1)}, \ldots, F^{(L)})$ appropriate force functions. We will concentrate on considering *linear* force functions in which case we will use the name HMMSDE for models of form (4)&(5). As will be discussed in detail *nonlinear* forces can be well approximated by HMMSDE models under appropriate circumstances.

The general aim of the present article is to find the optimal model of HMMSDE form for a given time series (i.e., find optimal parameters $(F^{(i)}, \Sigma^{(i)}, \mu^{(i)})_{i=1,\ldots,L}$ and optimal transition matrix of the jump process in a Maximum-Likelihood sense), and then to decide whether the observed dynamics locally is close to full hypo-elliptic Langevin or overdamped Langevin dynamics, or should rather be modelled by other types of processes in the general form of (4); thus in this article we are *not* interested in enforcing the hypo-elliptic form of the Langevin model since we first want to see whether the general form (4) if applied to position and momentum information results in something that is (locally) close to the hypo-elliptic case. The background theory (uniqueness of solutions, sampling) of such stochastic models like (4) is discussed under the title "SDEs with Markovian Switching", e.g., in the recent book [20].

Since the article [10] contains a detailed comparison of the general framework of our approach with other approaches in the literature, we will herein concentrate on commenting on that part of the literature that is directly relevant for our specific approach to parameter-estimation for diffusions of the above type, and related questions: most of the rich existing literature on discrete-time observations does not apply to the case we consider, for example the case closest to the approach presented herein is [21], however there both components, z and j, are observed while here only z is observed. Furthermore, there are alternatives for the estimators considered herein that may allow further generalization but have not been considered for multidimensional applications up to now [22]. In some part of the mathematical literature on estimation of *hypo-elliptic* diffusions, the authors consider the case of *partial observation*, compare [23]: The given time-series contains information on the positions q only, the corresponding velocities $\dot{q}$ or momenta $p = M\dot{q}$ have not been observed. Then, the velocities/momenta have to be estimated in addition to the hidden parameters for stiffness, friction and noise intensity. In contrast, in this article we will mainly consider the case that is typical for data coming from simulations (e.g. in molecular dynamics): the simulation schemes provide information on positions as well as velocities and typically rather long time series; often momenta are not available since the corresponding mass matrix M is not known (see our example in Section 5). We will also shortly address the case in which velocities as well as momenta are *un*-observed; there we will show how to generalize the results of [23].

In addition to these considerations one often is keen on certain general properties of the systems under consideration. In molecular dynamics applications, for example, one is interested in the so-called fluctuation-dissipation relation since it guarantees that the invariant distribution has the form of Gibbs or Boltzmann densities. We will consider these additional aspects in detail. We will see that the question of whether the fluctuation-dissipation relation can be assumed valid is deeply related to the unique identifiability of all parameters.

After deriving the algorithms for optimal parametrization of HMMSDE models, and discussing its possible pitfalls and generalizations (Sec. 1-4) we will consider two illustrative examples in Sec. 5: first, a 14-dimensional toy example that is appropriate to demonstrate the performance of the algorithms and, second, a real-world time series originating from molecular dynamics simulations of the oligo-peptide 12-alanine with implicit aqueous solvent.

1 Approach

We consider the following problem: An observation sequence, i.e., time series $Z = \{Z_1, \ldots, Z_T\} \subset \mathbf{R}^n$ of the system under consideration is given where $Z_k = (q(t_k), p(t_k))$ denotes an observation of position, $q(t_k)$, and momenta, $p(t_k)$, at time t_k (or just position in case of the overdamped dynamics), and $\tau = t_{k+1} - t_k$ is the *constant* (wrt. k) time difference between two successive observations. We want to find the optimal parameters $(F^{(i)}, \Sigma^{(i)}, \mu^{(i)})_{i=1,\ldots,L}$ and optimal transition matrix of the jump process in a Maximum-Likelihood sense) of a reduced system of form (4)&(5). Subsequently we want to decide whether the resulting local SDEs are of Langevin form type (2) or (3).

Notation and preparations. We will first consider the local Langevin models (2) or (3). One of the fundamental properties that they have in common is that they exhibit invariant densities of particularly simple form (Gibbs densities),

$$f(q) \propto \exp\left(-\beta U(q)\right), \quad \text{respectively} \quad f(q,p) \propto \exp\left(-\beta(\frac{1}{2}p^\top M^{-1} p + U(q))\right),$$

whenever the fluctuation-dissipation relation $\gamma + \gamma^\top = \beta\sigma\sigma^\top$ holds. Here $\beta > 0$ is some arbitrary constant, the so-called *inverse temperature*. The property that the invariant distribution is given by a Gibbs density is of utmost importance in most biophysical and biochemical applications and will thus play an important role in our considerations.

We are mainly interested in the case of a quadratic potential $U(q) = \frac{1}{2}(q-\mu_{eq})^T D(q-\mu_{eq})$, where D is called the *stiffness matrix*. For physical applications, the *generic case* then is to combine the demand for the fluctuation-dissipation relation to be valid with the requirement that the stiffness matrix D is a symmetric positive definite matrix. In the subsequent, we will use the notation $A > 0$ iff some matrix is positive definite, and $A \geq 0$ iff A is positive-semi-definite. (Note that this does not mean that A is symmetric, cf. Appendix A!)

Whenever the potential U is assumed to be quadratic then the two linear Langevin models to be considered can be written as

$$\begin{array}{rcl} \dot{q}(t) & = & M^{-1}p(t) \\ \dot{p}(t) & = & -D(q(t)-\mu_{eq}) - \gamma M^{-1}p(t) + \sigma\dot{W}(t). \end{array} \tag{6}$$

for the full/hypoelliptic case, and

$$\gamma\dot{q}(t) = -D(q(t)-\mu_{eq}) + \sigma\dot{W}(t), \tag{7}$$

for the overdamped case, respectively.

Introducing $z = (q,p) \in \mathbf{R}^n \times \mathbf{R}^n$ for positions and momenta, we can write the full/hypoelliptic equation as the following first order system

$$\dot{z} = F(z-\mu) + \Sigma\dot{W}, \qquad F = \begin{pmatrix} 0 & M^{-1} \\ -D & -\gamma M^{-1} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 0 & 0 \\ 0 & \sigma \end{pmatrix}, \quad \mu = \begin{pmatrix} \mu_{eq} \\ 0 \end{pmatrix}. \tag{8}$$

For the overdamped dynamics we analogously have

$$\dot{z} = F(z-\mu) + \Sigma\dot{W}, \quad F \;=\; -\gamma^{-1}D, \quad \mu = \mu_{eq}, \quad \Sigma = \gamma^{-1}\sigma. \tag{9}$$

In the subsequent, we will always assume that the spectrum of F associated with the two linear Langevin equations considered above is contained in the negative half plane $\mathbf{C}^- = \{z \in \mathbf{C} | \Re(z) < 0\}$ of the complex plane $\mathbf{C}$, i.e., that $\mathrm{spec}(F) \subset \mathbf{C}^-$, in order to guarantee that the "deterministic part" of the respective SDEs is asymptotically stable.

Obviously, the linear full/hypoelliptic Langevin equations (6) is completely characterized by the 5-tupel $(M, \gamma, D, \sigma, \mu_{eq})$, while the linear overdamped Langevin model (7) will be identified with the quadrupel $(\gamma, D, \sigma, \mu_{eq})$, and the general first order system $\dot{z} = F(z-\mu) + \Sigma\dot{W}$ by the tripel (F, μ, Σ).

Our demand on the spectrum of F leads to the following definition of the classes of models to be considered: The class of linear overdamped Langevin models is

$$\mathbf{OL}(n) = \{(\gamma, D, \sigma, \mu_{eq}) \in \mathrm{Mat}_n(\mathbf{R})^3 \times \mathbf{R}^n \,|\, \gamma, D \text{ regular and } \mathrm{spec}(\gamma^{-1}D) \subset \mathbf{C}^-\},$$

and the class of linear full Langevin models

$$\mathbf{FL}(n) = \{(M, \gamma, D, \sigma, \mu_{eq}) \in \mathrm{Mat}_n(\mathbf{R})^4 \times \mathbf{R}^n \,|\, M, \gamma, D \text{ regular and } \mathrm{spec}(F) \subset \mathbf{C}^-\},$$

where F is defined as in (8). We will say that the linear Langevin models $(M, \gamma, D, \sigma, \mu_{eq})$ or $(\gamma, D, \sigma, \mu_{eq})$ satisfy the fluctuation-dissipation relation iff $\gamma + \gamma^\top = \beta\sigma\sigma^\top$ for some positive constant $\beta > 0$.

Optimal Parameters for the (F, μ, Σ)-model. We will construct an explicit formula for the probability $p(Z|F, \mu, \Sigma)$ that the time series $Z = \{Z_1, \ldots, Z_T\} \subset \mathbf{R}^n$ comes from a solution path of the SDE $\dot{z} = F(z - \mu) + \Sigma\dot{W}$. Then, we consider the likelihood $\mathcal{L}(F, \mu, \Sigma) = p(Z|F, \mu, \Sigma)$ and ask for which parameter set (F, μ, Σ) the likelihood is maximized; this then is understood to be the optimal model wrt. the observation. However, we will see that it is much simpler to consider the maximization of the likelihood in terms of the unknowns $(\exp(\tau F), \mu, \Sigma\Sigma^T)$ instead of (F, μ, Σ). The parameters that maximize $\mathcal{L}$, namely

$$\left(\exp(\tau\hat{F}), \hat{\mu}, \hat{\Sigma}\hat{\Sigma}^T\right) = \mathrm{argmax}_{(\exp(\tau F), \mu, \Sigma\Sigma^T)}\ \mathcal{L}$$

are called the optimal estimators (in the maximal likelihood sense). We will derive explicit analytical expressions for these optimal parameters (see Thm. 2.1 in Section 2.2 below).

Langevin appropriate or not? After identifying the optimal parameters for the (F, μ, Σ)-model, we still have to decide whether the optimal model has the form of either a hypo-elliptic or an overdamped Langevin model, or whether it does not belong to this class of processes. We will illustrate how to distinguish between these cases based on the structure of the optimal parameters $\hat{\mu}, \exp(\tau\hat{F}), \hat{\Sigma}\hat{\Sigma}^T$ (see Secs. 4.3.1 and 5.2). However, for the sake of clarity we want to emphasize again that we do *not* want to *enforce* hypo-elliptic or overdamped Langevin models but that, instead, we want to identify the optimal model in the larger class of (F, μ, Σ)-models in order to then decide whether the process underlying the time series exhibits a Langevin-type structure. This seems crucial for applications to molecular dynamics time series since there it is not clear in advance whether a Langevin structure is appropriate or not.

Non-uniqueness problems. The quite general explicit formula for the optimal parameters of the (F, μ, Σ)-model is one of the key working tools of this article. However, even if the optimal parameters $(\exp(\tau\hat{F}), \hat{\mu}, \hat{\Sigma}\hat{\Sigma}^\top)$ have been computed we still cannot compute all involved parameters uniquely, at least not without further considerations:

(1) *Uniqueness of $\hat{F}$:* Since the matrix logarithm in general is not injective (there are $F_1 \neq F_2$ such that $\exp(F_1) = \exp(F_2)$), how can the optimal estimator $\hat{F}$ be identified from the optimal estimator $\exp(\tau\hat{F})$ in a unique way? We will see that this problem can be solved whenever we have access to $\exp(\tau\hat{F})$ for sufficiently many different τ. We will present preliminary algorithms for this sort of solution in the appendix.

(2) *Identifiability problem:* Whenever the dynamics of interest has the form of the linear overdamped model $(\gamma, D, \sigma, \mu_{eq})$, then the optimal estimators $(\hat{F}, \hat{\mu}, \hat{\Sigma}\hat{\Sigma}^T)$ do not suffice to uniquely identify all three unknown matrices since we do only have the identities $\mu_{eq} = \hat{\mu}$, $\gamma^{-1}D = \hat{F}$ and $\gamma^{-1}\sigma = \hat{\Sigma}$. An analogous problem appears for the full/hypoelliptic Langevin equation as we will see later. How can these identifiability problems be solved or circumvented?

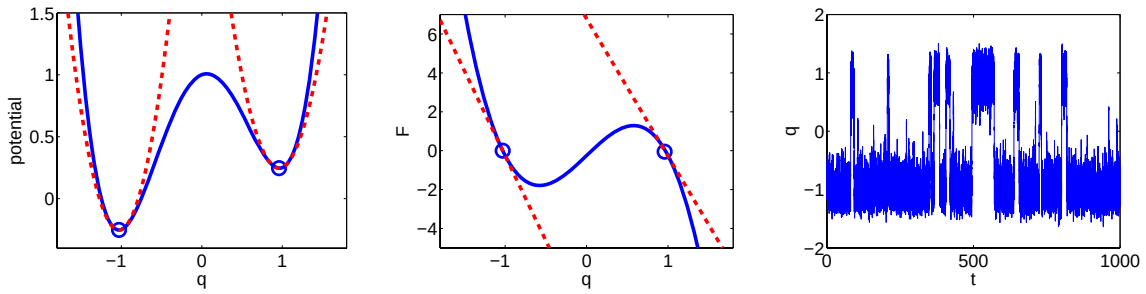

Fig. 1 Left: Two-well potential $U = U(q)$ and the two harmonic approximations around its minima. Middle: Related nonlinear force $F(q) = -U'(q)$ and the associated linearizations around its zeros. On both sides the circles indicate the minima of U and the equilibrium points of F, respectively. Left: Typical solution path of $\ddot{q} = -U'(q) - \dot{q} + \sigma\dot{W}$ with $\sigma = 0.75$.

Approximation of nonlinearity. In general a linear Langevin model, whether overdamped or full, will not be appropriate for modelling the dynamics of a molecular system in a multi-well energy landscape. One would like to consider nonlinear Langevin models like (1), where the potential energy function U allows for incorporation of the multi-well dynamics. Thus, parameter estimation procedures that allow for functions U of more general form seem attractive. However, if U is more general than a harmonic function then its specification will need more parameters. However, the quality of the parameter estimation will reduce with increasing number of parameters. Because of these problems we intend to represent the potential function U as the *best possible combination of the local harmonic models*. For example, consider the situation shown in Fig. 1 and assume that the timeseries in fact comes from the nonlinear Langevin model (1) with potential U as shown on the left of Fig. 1 and noise intensity is small compared to the energy barrier between the two minima in the potential. Then the resulting dynamics exhibits rare jumps between the two wells and, while remaining in one of the wells, it can be approximated by a linear Langevin model given by the harmonic approximation on the potential around the respective minimum. But then, a model that describes rare jumps between two otherwise linear models is desirable. Therefore, the following type of model seems appropriate:

$$\dot{z}(t) \quad = \quad F^{(i(t))}\left(z(t) - \mu^{(i(t))}\right) + \Sigma^{(i(t))}\dot{W}(t) \tag{10}$$

where $i(t)$ is a continuous time Markov process on $\{1, \ldots, L\}$ that is generated by a $n \times n$ rate matrix $\mathcal{R}$. This leads us to the key problem of this article:

(HMMSDE) Can we generalize the above results for the (F, μ, Σ)-model to this case of *local* (F, μ, Σ)-models even if only the time series $Z = \{Z_1, \ldots, Z_T\} \subset \mathbf{R}^n$ of observations of the z-component is given while the jump process i remains hidden, i.e., is not observed?

We will solve this problem by constructing the associated likelihood and then identifying the respective optimal estimators $(\exp(\tau\hat{F}^{(i)}), \hat{\mu}^{(i)}, (\hat{\Sigma}\hat{\Sigma}^\top)^{(i)}, \hat{\mathcal{R}})$; this time we will not have an explicit expression for the estimators but we will generalize the so-called *expectation-maximization algorithm* in order to iteratively compute/approximate them. These results (generalization and algorithmic approximation) are main new findings of this article.

We will proceed as follows: We will first present the analysis of the (F, μ, Σ)-model in Section 2. Then we will solve problem (HMMSDE) in Section 3. Section 4 contains the solutions to the non-uniqueness problems (1) and (2) while Section 5 will present our numerical experiments, and illustrate the performance of the algorithm. In particular, it will also illustrate the procedure of deciding whether a Langevin model is appropriate or not.

2 Optimal Parameters for the (F, μ, Σ)-Model

Suppose again that a discrete time series $Z = \{Z_1, \ldots, Z_T\}$ is given, where Z_k denotes an observation of the state of the system at time t_k and $\tau = t_{k+1} - t_k$ is the *constant* (wrt. k) time difference between two successive observations.

2.1 Likelihood of Linear Langevin Models

We are aiming at maximizing the observation probability of the sequence Z wrt. the parameters of the linear Langevin model (8). The solution to (8) on the time interval $[t, t+\tau]$ is given by

$$z(t+\tau) = \mu + e^{\tau F}(z(t) - \mu) + \int_0^\tau e^{(\tau - s)F} \Sigma dW(s). \tag{11}$$

Thus, the probability density $\rho_\lambda(Z_{k+1}|Z_k)$ of observation Z_{k+1} at time $k+1$ under the condition of observation Z_k at k is proportional to a Gaussian:

$$\rho_\lambda(Z_{k+1}|Z_k) \propto \exp\left[-\frac{1}{2}\xi_k^\top R^{-1}(\tau)\,\xi_k\right], \tag{12}$$

where

$$\xi_k = Z_{k+1} - \mu - e^{\tau F}(Z_k - \mu), \tag{13}$$

$$R(\tau) = \int_0^\tau e^{sF} \Sigma\Sigma^\top e^{sF^\top} ds. \tag{14}$$

Thus, we can express the joint conditional probability density of the observation of the time series Z from our model by

$$P(Z|\lambda) = \prod_{k=1}^{T-1} \rho_\lambda(Z_{k+1}|Z_k) = \prod_{k=1}^{T-1} \rho_0(\tau) \exp\left[-\frac{1}{2}\xi_k^\top R^{-1}(\tau)\,\xi_k\right] \tag{15}$$

$$\rho_0(\tau) = (2\pi)^{-n/2} / \sqrt{\det(R(\tau))}. \tag{16}$$

As it can be seen from (15), for fixed τ any Langevin model of type (4) or (9) can be uniquely defined by the set of *solution parameters* λ of the form

$$\lambda = (\mu, \exp(\tau F), R(\tau)).$$

We use these parameter set instead of the perhaps more natural one, (μ, F, Σ) since it makes derivation of explicit expressions for the optimal estimators much easier.

The integral in (14) can be solved by partial integration resulting in the following linear matrix equation

$$R(\tau)F^\top + FR(\tau) = e^{\tau F}\Sigma\Sigma^\top e^{\tau F^\top} - \Sigma\Sigma^\top \tag{17}$$

We define the log–likelihood function of the observation sequence as (compare [21])

$$\mathcal{L}(\lambda|Z) = \log P(Z|\lambda) \tag{18}$$

The optimal parameters λ are those which maximize the log–likelihood function

$$\begin{aligned}\mathcal{L}(\lambda|Z) &= \sum_{k=1}^{T-1} \log \rho_\lambda(Z_{k+1}|Z_k) \\ &= C - \frac{T-1}{2}\log\det R(\tau) \\ &\quad -\frac{1}{2}\sum_{k=1}^{T-1}\left(Z_{k+1} - \bar\mu - e^{\tau F}(Z_k - \bar\mu)\right)^\top R^{-1}(\tau)\left(Z_{k+1} - \bar\mu - e^{\tau F}(Z_k - \bar\mu)\right).\end{aligned} \tag{19}$$

Here $C < 0$ denotes a constant that collects all terms that do not depend on the undetermined parameters λ.

2.2 Maximization of Likelihood

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 matrix $\exp(\tau F)$

$$\frac{\partial \mathcal{L}}{\partial \exp(\tau F)} = 0 \quad \Longrightarrow \quad \exp(\tau F) = \left(A_1(\mu) A_2^{-1}(\mu)\right) \tag{20}$$

$$A_1(\mu) = \sum_{k=1}^{T-1} \left(Z_{k+1} - \mu\right) \left(Z_k - \mu\right)^\top$$

$$A_2(\mu) = \sum_{k=1}^{T-1} \left(Z_k - \mu\right) \left(Z_k - \mu\right)^\top,$$

for the center of the potential

$$\frac{\partial \mathcal{L}}{\partial \mu} = 0 \quad \Longrightarrow \quad \mu = \frac{1}{T-1} (I - e^{\tau F})^{-1} \sum_{k=1}^{T-1} \left(Z_{k+1} - e^{\tau F} Z_k\right), \tag{21}$$

and for the derivative with respect to the covariance matrix $R(\tau)$

$$\frac{\partial \mathcal{L}}{\partial R} = 0 \quad \Longrightarrow \quad R(\tau) = \frac{1}{T-1} \sum_{k=1}^{T-1} d_k d_k^\top \tag{22}$$

$$d_k = \left(Z_{k+1} - \bar{\mu} - e^{\tau F} \left(Z_k - \bar{\mu}\right)\right).$$

The optimal parameters $(\exp(\tau \hat{F}), \hat{\Sigma}\hat{\Sigma}^\top, \hat{\mu})$ are determined by solving the nonlinear system of equations (20)–(22) for a given observation sequence $Z = \{Z_1, \ldots, Z_T\}$. Note the difference to the approach presented in [24]: there the parameter $\hat{\mu}$ is simply assumed being zero, this clearly leads to a decoupled system of estimator equations which can be explicitly solved providing the closed expressions for $(\exp(\tau \hat{F}), \hat{\Sigma}\hat{\Sigma}^\top)$. The presence of the unknown parameter $\hat{\mu}$ in our case makes the task of calculating the optimal estimators from the coupled system of equations (20)–(22) not so straightforward as in [24] (as we will demonstrate in the appendix, also the expression for the optimal estimator of $\hat{\mu}$ is not just the expectation value of the time series but gets an additional term responsible for the relaxation behavior). We can use the fact that the equations (20)-(21) are independent of the variables $R(\tau)$ and $\Sigma\Sigma^{\mathcal{T}}$. This results in the explicit expressions for the unique solution that are given in Theorem 2.1 below which is a direct consequence of Lemma 6.3 and Lemma 6.4 of Appendix C.

Theorem 2.1 *Let the running average, the covariance matrix and normalized autocorrelation of the time series be defined by*

$$\begin{aligned}
\bar{Z} &= \frac{1}{T-1} \sum_{k=1}^{T-1} Z_k, \\
Cov(Z) &= \frac{1}{T-1} \sum_{k=1}^{T-1} (Z_k - \bar{Z})(Z_k - \bar{Z})^\top, \\
Cor(Z) &= \frac{1}{T-1} \sum_{k=1}^{T-1} (Z_{k+1} - \bar{Z})(Z_k - \bar{Z})^\top \cdot Cov(Z)^{-1}.
\end{aligned}$$

Suppose that $Cov(Z)$ is positive definite. Then, the optimal estimators $\exp(\tau \hat{F})$ and $\hat{\mu}$ are given by

$$\begin{aligned}
\exp(\tau \hat{F}) &= Cor(Z), \\
\hat{\mu} &= \bar{Z} - (\mathrm{Id} - Cor(Z))^{-1} \delta.
\end{aligned}$$

where $\delta = (Z_T - Z_1)/(T-1)$. *The estimator* $\hat{F}$ *inherits the property* $\mathrm{spec}(\hat{F}) \subset \mathbf{C}^-$ *iff* $\|Cor(Z)\|_2 < 1$*; as outlined above this is the generic situation for the cases considered as long as* $\tau > 0$.

In addition, the optimal noise intensity matrix estimator $\hat{\Sigma}\hat{\Sigma}^\top$ *is then determined by*

$$\hat{\Sigma}\hat{\Sigma}^\top = -\Big((Cov(Z)+E)\hat{F}^\top + \hat{F}(Cov(Z)+E)\Big), \tag{23}$$

where E *is a symmetric matrix that satisfies the Sylvester equation*

$$Cor(Z)E\,Cor(Z)^\top - E \quad = \quad f(Z_1, Z_T, \bar{Z}), \tag{24}$$

where f *is just the following symmetric-matrix-valued, quadratic function*

$$f(Z_1, Z_T, \bar{Z}) = -\delta\delta^\top + \frac{1}{T-1}\Big((Z_T-\bar{Z})(Z_T-\bar{Z})^\top - (Z_1-\bar{Z})(Z_1-\bar{Z})^\top\Big).$$

Whenever $\mathrm{spec}(\hat{F}) \subset \mathbf{C}^-$ *(24) has a unique symmetric solution* E *.*

Let us assume that the time series comes from $\dot{z} = F(z-\mu) + \Sigma\dot{W}$ and that the underlying process is ergodic. Then, the optimal estimators are *consistent* (see Appendix C). That is, for $T \to \infty$ they converge to the values of $\mu, \exp(\tau F), \Sigma\Sigma^T$ that were used for generation of the time series.

3 HMMSDE: Switching between (F, Σ, μ)-Models

Up to now we have considered a single linear Langevin model, which approximates the entire 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 $\lambda^{(i)} = (F^{(i)}, \Sigma^{(i)}, \mu^{(i)})$. Switching back and forth between these local parameter sets can be understood as one *global* model with parameters that are piecewise constant in time.

To this end 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 $\lambda_i, i = 1, \dots L$, we assume that the switching between the different parameter sets is governed by a Markov jump process. Thus the model consists of two related stochastic processes $z(t)$ and $i(t)$, where the latter is not directly observed (hidden) (note the difference to [21]) and satisfies the Markov property. That is, our dynamical model now has the form

$$\dot{z}(t) \quad = \quad F^{(i(t))}\left(z(t) - \mu^{(i(t))}\right) + \Sigma^{(i(t))}\dot{W}(t) \tag{25}$$

where $i(t)$ is a continuous time Markov process on $\{1, \dots, L\}$ that is generated by the $n \times n$ rate matrix $\mathcal{R}$.

In general, a HMM is fully specified by an initial distribution π of hidden states, the rate matrix $\mathcal{R}$ of the hidden Markov chain $i(t)$, and the parameters of the Langevin output process $\lambda^{(i)} = (F^{(i)}, \Sigma^{(i)}, \mu^{(i)})$ for each state i. For given rate matrix $\mathcal{R}$, the transition probability to jump from state $i(t_k) = m$ to state $i(t_{k+1}) = j$ within time τ is given by the respective entry of the transition matrix

$$T(m, j) = \left(\exp(\tau\mathcal{R})\right)_{mj} .$$

Now we have to embed the problem of estimating optimal parameters for the model (25) into the context of standard HMM. Therefore, we start with the joint probability distribution of the observation sequence that here reads

$$\mathcal{L}(\lambda|Z, i) = P(Z, i|\lambda) = \prod_{k=1}^{T-1} T(i_k, i_{k+1})\varrho_\lambda(Z_{k+1}|i_{k+1}, Z_k)\,, \tag{26}$$

where the conditional probability $\varrho_\lambda(\cdot|\cdot)$ is defined exactly as $\rho_\lambda(\cdot|\cdot)$ in equation (12) above, except that the parameters now depend on the hidden state $i_{k+1} = i(t_{k+1})$. The algorithm for the identification of optimal parameters conditional on the hidden (metastable) states comprises the following three steps:

(1) Determine the optimal parameters $\theta = (\pi, T, \lambda_1, \ldots, \lambda_L)$ by maximizing the likelihood $\mathcal{L}(\theta|Z, i)$ associated with (26); in general this is a nonlinear global optimization problem.

(2) Determine the optimal sequence of hidden 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 [25, 26, 27, 28, 29] in a specification that we will outline below. Generally, in the EM algorithm the optimal parameters λ are identified by iteratively maximizing the entropy

$$\mathcal{S}(Z) = \max_{\lambda} \sum_{i} \mathcal{L}(\lambda|Z, i) \log \mathcal{L}(\lambda|Z, i)\,. \tag{27}$$

For the identification of the optimal sequence of hidden metastable states the Viterbi algorithm [30] is used, which exploits dynamic programming techniques to resolve the optimization problem

$$\max_{i} \mathcal{L}(\lambda|Z, i)$$

in a recursive manner. For the details see [31] and the references therein.

It is worthwhile to mention that the above algorithmic steps all scale linearly in the length of the time series considered.

For any algorithmic realization of the first two problems (1) and (2) one selects a specific number L of hidden states *in advance*. Problem (3) means that one has to find the *optimal* number of hidden states in the sense that each hidden state finally corresponds to one of the dominant metastable sets of the system under consideration. A practical way to handle this problem is to start with a sufficiently large number of hidden states L and then aggregate the resulting transition matrix, which gives the minimum number of hidden states that are necessary to resolve the metastable sets [32, 33]. The aggregation is performed by Perron cluster cluster analysis (PCCA), which exploits the spectral properties of the transition matrix T in order to transform it to a matrix with quasi–block structure [10, 34, 35]. These blocks then correspond to the existing metastable states.

3.1 Specification of the Expectation-Maximization Algorithm

In order to solve (1) for a predefined number of hidden states, we suggest to specify the standard Expectation-Maximization algorithm as follows: As usual, in the Expectation step (E-step) the occupation and transition probabilities for a hidden Markov–chain are calculated for the actual value of model parameters λ and T, subsequently in the Maximization step (M-step) the values of the model parameters are updated via the *re-estimation* formulas. The M-step guarantees that the likelihood does not decrease in each single iteration. Whereas the E-step in our case is identical to a standard procedure described in [10], the *re-estimation* formulas have to be modified: Let the parameters $\nu_k(i)$ be the probabilities to observe the hidden process in the state i in the time k (as computed in the E-step and fixed given for the M-step). In order to obtain the estimator formulas in the case of L hidden states, in analogy to (20)-(22) we derive the individual partial derivatives of the log–likelihood in each hidden state i while taking into account the probability $\nu_k(i)$ to be in this state at time k: for the matrix $B^{(i)}(\tau) = \exp(\tau F^{(i)})$

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial B^{(i)}} = 0 \quad \Longrightarrow \quad B^{(i)} &= \left(A_1^{(i)}(\mu^{(i)}) A_2^{(i)}(\mu^{(i)}) \right)^{-1} \\ A_1^{(i)}(\mu) &= \sum_{k=1}^{T-1} \nu_{k+1}(i) \left(Z_{k+1} - \mu^{(i)} \right) \left(Z_k - \mu^{(i)} \right)^{\top} \\ A_2^{(i)}(\mu) &= \sum_{k=1}^{T-1} \nu_{k+1}(i) \left(Z_k - \mu^{(i)} \right) \left(Z_k - \mu^{(i)} \right)^{\top}, \end{aligned} \tag{28}$$

for the center of the potential

$$\frac{\partial \mathcal{L}}{\partial \mu^{(i)}} = 0 \quad \Longrightarrow$$

$$\mu^{(i)} = \frac{1}{\sum_k^{T-1} \nu_{k+1}(i)} (I - B^{(i)})^{-1} \sum_{k=1}^{T-1} \nu_{k+1}(i) \left(Z_{k+1} - B^{(i)} Z_k \right), \tag{29}$$

and for the derivative with respect to the covariance matrix $R^{(i)}(\tau) = \int_0^\tau \exp(s\hat{F}^{(i)})(\Sigma\Sigma^\top)^{(i)} \exp(s\hat{F}^{(i)})^\top ds$

$$\frac{\partial \mathcal{L}}{\partial R^{(i)}} = 0 \quad \Longrightarrow \quad R^{(i)}(\tau) = \frac{1}{\sum_k^{T-1} \nu_{k+1}(i)} \sum_{k=1}^{T-1} \nu_{k+1}(i) d_k d_k^\top \tag{30}$$

$$d_k = \left(Z_{k+1} - \mu^{(i)} - B^{(i)} \left(Z_k - \mu^{(i)} \right) \right).$$

For a fixed sequence (ν_k) the equations (28)-(29) have a unique solution that we can write down analytically, which gives us *explicit formulas* for the optimal estimators $\hat{\mu}_i$, $\hat{\Sigma}\hat{\Sigma}^\top$ and $exp(\tau\hat{F}^{(i)})$ as given in the following theorem which is a direct consequence of Lemmas 6.3 and 6.4 (see Appendix C).

Theorem 3.1 *Let $i \in \{1, \ldots, L\}$ and let the parameters $\nu_k(i)$ be the probabilities to observe the hidden process in state i in the time k. The running average, the covariance matrix and normalized autocorrelation of the time series for state i are defined by*

$$\begin{aligned}
\bar{Z}^{(i)} &= \frac{1}{\sum_{l=1}^{T-1} \nu_{k+1}(i)} \sum_{k=1}^{T-1} \nu_{k+1}(i) Z_k, \\
Cov^{(i)}(Z) &= \frac{1}{\sum_{l=1}^{T-1} \nu_{k+1}(i)} \sum_{k=1}^{T-1} \nu_{k+1}(i) (Z_k - \bar{Z}^{(i)})(Z_k - \bar{Z}^{(i)})^\top, \\
Cor^{(i)}(Z) &= \frac{1}{\sum_{l=1}^{T-1} \nu_{k+1}(i)} \sum_{k=1}^{T-1} \nu_{k+1}(i) (Z_{k+1} - \bar{Z}^{(i)})(Z_k - \bar{Z}^{(i)})^\top \cdot Cov^{(i)}(Z)^{-1}.
\end{aligned}$$

Suppose that $Cov^{(i)}(Z)$ is positive definite. Then, the optimal estimators $\exp(\tau\hat{F}^{(i)})$ and $\hat{\mu}^{(i)}$ are given by

$$\begin{aligned}
\exp(\tau\hat{F}^{(i)}) &= Cor^{(i)}(Z), \\
\hat{\mu}^{(i)} &= \bar{Z}^{(i)} - (\mathrm{Id} - Cor^{(i)}(Z))^{-1} \delta^{(i)}.
\end{aligned}$$

where

$$\delta^{(i)} = \frac{1}{\sum_{l=1}^{T-1} \nu_{k+1}(i)} \sum_{k=1}^{T-1} \nu_{k+1}(i)(Z_{k+1} - Z_k).$$

The estimator $\hat{F}^{(i)}$ inherits the property $\mathrm{spec}(\hat{F}^{(i)}) \subset \mathbf{C}^-$ iff $\|Cor^{(i)}(Z)\|_2 < 1$; as outlined above this is the generic situation for the cases considered as long as $\tau > 0$.

In addition, the optimal noise intensity matrix estimator $(\hat{\Sigma}\hat{\Sigma}^\top)^{(i)}$ is then determined by

$$(\hat{\Sigma}\hat{\Sigma}^\top)^{(i)} = -\Big((Cov^{(i)}(Z) + E^{(i)})(\hat{F}^{(i)})^\top + \hat{F}^{(i)}(Cov^{(i)}(Z) + E^{(i)}) \Big), \tag{31}$$

where $E^{(i)}$ is a symmetric matrix that satisfies the Sylvester equation

$$Cor^{(i)}(Z) E^{(i)} Cor^{(i)}(Z)^\top - E^{(i)} = f^{(i)}, \tag{32}$$

where $f^{(i)}$ is just the following symmetric-matrix-valued, quadratic function that generalizes the one of Theorem 2.1 to a single hidden state and still vanishes if computed for an ergodic, infinitely long time series. Whenever $\mathrm{spec}(\hat{F}^{(i)}) \subset \mathbf{C}^-$ (32) has a unique symmetric solution $E^{(i)}$.

3.2 Resulting HMMSDE algorithm

The resulting overall algorithm then has the following iterative form:

1. Initialization of the EM-algorithm (set the number of hidden states L, zeroth iterates for the local SDE parameters $\lambda^{(i)} = (\exp(\tau\hat{F}^{(i)}), (\hat{\Sigma}\hat{\Sigma}^\top)^{(i)}, \hat{\mu}^{(i)})$, the transition matrix $T = \exp(\tau\mathcal{R})$, and further HMM parameters).

2. E-step: compute new iterates for occupation probabilities $\nu_k(i)$, $i = 1, \ldots, L$, and transition matrix T, based on the present iterates of the SDE parameters $\lambda^{(i)}$, $i = 1, \ldots, L$ (according to standard HMM schemes).

3. M-step: compute new iterates for the SDE parameters $\lambda^{(i)}$, $i = 1, \ldots, L$, based on the present iterates for the occupation probabilities $\nu_k(i)$, according to Thm. 3.1.

4. Repeat steps 2 and 3 until the increase in the likelihood $\mathcal{L}(\lambda|Z, i)$ in the last iteration is smaller than the predefined threshhold value.

5. Compute the Viteri path (according to standard Viterbi schemes).

6. Determine the optimal number of metastable states $\hat{L} \leq L$ (via PCCA as pointed out above), and aggregate the parameter sets accodingly.

The proposed HMMSDE scheme has several nice properties: (i) the M-step (and therefore the EM-algorithm as whole) scales linearly wrt. the length of the observation sequence Z, (ii) in order to compute the optimal HMM parameters and the optimal estimators $(\hat{F}^{(i)}, (\hat{\Sigma}\hat{\Sigma}^\top)^{(i)}, \hat{\mu}^{(i)})$ we do *not* need to compute the estimator $\hat{F}^{(i)}$ but only the estimator of its exponential $\exp(\tau\hat{F}^{(i)})$.

4 Non-Uniqueness Problems

We will now discuss and solve the non-uniqueness problems indicated on page 5.

4.1 Uniqueness of $\hat{F}$

Since the matrix logarithm in general is not injective (there are $F_1 \neq F_2$ such that $\exp(F_1) = \exp(F_2)$), how can the optimal estimator $\hat{F}$ be identified from the optimal estimator $\exp(\tau\hat{F})$ in a unique way? This problem affects only non-symmetric F, i.e., we do not need to consider it whenever $\exp(\tau\hat{F})$ is symmetric.

In the following we assume that a sequence of correlation matrices $(\exp(\tau_k\hat{F}))_{k=1,\ldots,L}$ is explicitly available for $\tau_1 < \tau_2 < \ldots < \tau_L$ instead of just a single matrix for a single value of the lag time τ.

The resolvent. We can approximate $\hat{F}$ via its resolvent:

$$R(s, \tau_1) = \int_{\tau_1}^{\infty} \exp(-s(\tau - \tau_1)) \exp(\tau\hat{F}) d\tau. \tag{33}$$

For $s > 0$, we have $R(s, \tau_1) = (s - \hat{F})^{-1} e^{\tau_1\hat{F}}$. Thus, whenever the sequence of nodes $(\tau_k)_{k=1,\ldots,N}$ is appropriate, we can approximate $R(s, \tau_1)$ by approximating the integral by numerical quadrature for an appropriate choice of s, and then compute

$$\hat{F} = s\mathrm{Id} - e^{\tau_1\hat{F}} R(s, \tau_1)^{-1}.$$

This method is known to have limited numerical accuracy (depending on whether the nodes $(\tau_k)_{k=1,\ldots,N}$ are appropriate quadrature nodes for the unknown integrand). However, it in principle offers a way to identify $\hat{F}$ based on sufficient information on $\exp(\tau\hat{F})$ as functions in τ. For further details see Appendix D.

Spectral approach. Let us assume that $\hat{F}$ is diagonalizable (non-defective), or, equivalently, that the matrices of the sequence $(\exp(\tau_k \hat{F}))_{k=1,\dots,L}$ are diagonalizable (non-defective). Then, there exists a diagonal matrix $\Lambda \in \mathrm{Mat}_n(\mathbf{C})$ and some regular $U \in \mathrm{Mat}_n(\mathbf{C})$ such that $U \exp(\tau_k \hat{F}) U^{-1} = \exp(\tau_k \Lambda)$. Thus, our problem can be transformed into the problem of identifying $\lambda = a + ib \in \mathbf{C}$ from the sequence $(e_k)_{k=1,\dots,L}$, $e_k = \exp(\tau_k \lambda)$. While the real part a is easily uniquely identified from $a = (\log \|e_1\|)/\tau_1$, identification of b requires the solution of

$$\text{find } b \quad \text{s.t.} \quad \forall k = 1, \dots, L: \quad \cos(b\tau_k) = \Re(e_k)/|e_k| \text{ and } \sin(b\tau_k) = \Im(e_k)/|e_k|.$$

For the typical case $\tau_k = k\tau$, the solution of this problem is not unique. Instead, the complete solution family $(\hat{b}_p)_{p \in \mathbf{Z}}$ is given by $\hat{b}_p = b + 2\pi p/\tau$. However, if

$$\{\tau_1, \dots, \tau_L\} = \{k\tau | k = 1, \dots, L_1\} \cup \{ls | l = 1, \dots, L_2\}$$

with $\tau > 0$ and $s > 0$ such that s/τ is irrational, then the problem has just the single solution b since there are no $p, q \in \mathbf{Z}$ such that $ps = q\tau$. In Appendix D we will sketch an algorithm that allows to identify $\hat{F}$ along the lines of this argument.

4.2 Identifiability Problem

4.2.1 Overdamped Langevin Models

The identifiability problems are related to the following observation, here first presented for the overdamped Langevin equation but likewise valid for the full/hypoelliptic one: Let a quadrupel $(\gamma, D, \sigma, \mu_{eq}) \in \mathbf{OL}(n)$ be given. The associated solution is exactly the same for all equations of the family

$$\mathcal{A}\gamma\dot{q} = -\mathcal{A}D(q - \mu_{eq}) + \mathcal{A}\sigma\dot{W}, \tag{34}$$

with $\mathcal{A} \in \mathrm{Mat}_n(\mathbf{R})$ being an *arbitrary* regular matrix. That is, there is an *equivalence relation*, to be denoted by $\sim$, on $\mathbf{OL}(n)$, defined by

$$(\gamma, D, \sigma, \mu_{eq}) \sim (\gamma', D', \sigma', \mu'_{eq}) \quad \Leftrightarrow \quad \exists \mathcal{A} \in \mathrm{Reg}_n(\mathbf{R}) \text{ s.t. } (\mathcal{A}\gamma, \mathcal{A}D, \mathcal{A}\sigma, \mu_{eq}) = (\gamma', D', \sigma', \mu'_{eq}),$$

where $\mathrm{Reg}_n(\mathbf{R})$ denotes the set of regular $n \times n$ matrices with real-valued entries. The associated equivalence classes,

$$[(\gamma, D, \sigma, \mu_{eq})]_\sim = \{(\gamma', D', \sigma', \mu'_{eq}) \in \mathbf{OL}(n) | (\gamma, D, \sigma, \mu_{eq}) \sim (\gamma', D', \sigma', \mu'_{eq})\},$$

for $(\gamma, D, \sigma, \mu_{eq}) \in \mathbf{OL}(n)$ are those subsets of $\mathbf{OL}(n)$ that contain only models that share the same solution process.

Thus, identification of the parameters just from observation of its solution *cannot* be unique but can at most identify the associated equivalence class from which the solution originated, i.e., it has to leave at least one full regular matrix un-identified.

We suggest to circumvent this identifiability problem by means of the following lemma (for a proof see Appendix E):

Lemma 4.1 *Let $(\gamma, D, \sigma, \mu_{eq}) \in \mathbf{OL}(n)$ be given. Then for each inverse temperature $\beta > 0$ there exists a unique element, denoted $(\hat{\gamma}, \hat{D}, \hat{\sigma}, \hat{\mu}_{eq})_\beta$, in $[(\gamma, D, \sigma, \mu_{eq})]_\sim$ that satisfies the fluctuation-dissipation relation combined with the requirement that $\hat{D}$ be symmetric.*

We will call the parameter tupel $(\hat{\gamma}, \hat{D}, \hat{\sigma}, \hat{\mu}_{eq})_\beta$ *physically generic* iff $\hat{D} > 0$. Since always $\hat{\sigma}\hat{\sigma}^\top \geq 0$, validity of the fluctuation-dissipation relation also guarantees $\gamma \geq 0$ (such that there can be no gain of energy through friction). We will suppress the index β in $(\hat{\gamma}, \hat{D}, \hat{\sigma}, \hat{\mu}_{eq})_\beta$ when the choice of the inverse temperature is clear from the context.

This means that when interested in fitting to the overdamped Langevin dynamics to a given time series we can *always* assume that the fluctuation-dissipation equation is valid since the equivalence class of indistinguishable models from $\mathbf{OL}(n)$ *always* contains an element for which this is the case. Thus,

having computed $\hat{F}$, $\hat{\mu}$ and $\hat{\Sigma}\hat{\Sigma}^\top$ based on Theorem 2.1 and *setting* a certain value for β, we can identify $(\hat{\gamma}, \hat{D}, \hat{\sigma}, \hat{\mu}_{eq})_\beta$ from

$$-\hat{\gamma}^{-1}\hat{D} = \hat{F}, \quad \hat{\gamma}^{-1}\hat{\sigma}\hat{\sigma}^\top(\hat{\gamma}^{-1}) = \hat{\Sigma}\hat{\Sigma}^\top, \quad \hat{\gamma} + \hat{\gamma}^\top = \beta\hat{\sigma}\hat{\sigma}^\top, \quad \hat{\mu}_{eq} = \hat{\mu},$$

with the constraint that $\hat{D}$ has to be symmetric. We will discuss in detail what it means to *set* a certain value for β. Including the constraint the above equations have the (unique) solution

$$\begin{aligned}
\hat{\gamma}^{-1} &= -\beta\,\hat{F}\,(\mathrm{Cov}(Z) + E), && (35)\\
\hat{D}^{-1} &= \beta\,(\mathrm{Cov}(Z) + E), && (36)\\
\hat{\sigma}\hat{\sigma}^\top &= \frac{1}{\beta}(\hat{\gamma} + \hat{\gamma}^\top) && (37)\\
\hat{\mu}_{eq} &= \hat{\mu}.
\end{aligned}$$

Whether this parameter tupel is physically generic or not depends on whether $\mathrm{Cov}(Z) + E > 0$ or not. Whenever we assume that the time series $Z = \{Z_1, \ldots, Z_T\}$ under consideration comes from an ergodic, infinitely long time series $Z_\infty = \{Z_1, \ldots, Z_T, \ldots\}$ whose associated covariance matrix $\mathrm{Cov}(Z_\infty)$ is positive definite, then $\mathrm{Cov}(Z) \to \mathrm{Cov}(Z_\infty) > 0$ and $E \to 0$ for $T \to \infty$ so that it is guaranteed that $\mathrm{Cov}(Z)+E > 0$ for large enough T and $\hat{D} = (\hat{D}^{-1})^{-1}$ does exist. Then, the parameter tupel is physically generic, and $\mathrm{spec}(\hat{F}) \subset \mathbf{C}^-$ guarantees in addition that the required matrix inverse $\hat{\gamma} = (\hat{\gamma}^{-1})^{-1}$ does exist. (Whenever they do not exist we can take appropriate pseudo-inverses which then has an interesting physical interpretation on its own).

4.2.2 Full/hypoelliptic Langevin Models

Above we considered models from $\mathbf{FL}(n)$ and how to describe time series $(Z_1, \ldots, Z_T)$ where the states $Z_k(q_k, p_k)$ include information on position q_k and *momenta* $p_k = M\dot{q}_k$ for each time k. This allowed to construct an optimal estimator $\hat{F}$ for the matrix in the associated first-order system:

$$F = \begin{pmatrix} 0 & M^{-1} \\ -D & -\gamma M^{-1} \end{pmatrix}.$$

Unfortunately, in most applications the available time series of the system under consideration give us positions q and *velocities* $\dot{q}$ instead of momenta p. In this case we have to consider the following form of the linear full Langevin equation:

$$\begin{aligned}
\dot{q}(t) &= v(t)\\
\dot{v}(t) &= -M^{-1}D(q(t) - \mu_{eq}) - M^{-1}\gamma v(t) + M^{-1}\sigma\dot{W}(t).
\end{aligned}$$

This parameter estimation will have to leave one of the matrices M, D, γ undetermined since the estimator $\hat{F}$ now estimates the matrix

$$F = \begin{pmatrix} 0 & \mathrm{Id} \\ -M^{-1}D & -M^{-1}\gamma \end{pmatrix},$$

instead of the above one.

Let a 5-tupel $(M, \gamma, D, \sigma, \mu_{eq}) \in \mathbf{FL}(n)$ be given. The associated solution is exactly the same for all equations of the family $(\mathcal{A}M, \mathcal{A}\gamma, \mathcal{A}D, \mathcal{A}\sigma, \mu_{eq})$ with $\mathcal{A} \in \mathrm{Mat}_n(\mathbf{R})$ being an *arbitrary* regular matrix. That is, the equivalence relation $\sim$ on $\mathbf{FL}(n)$ here is

$$\begin{aligned}
(M, \gamma, D, \sigma, \mu_{eq}) &\sim (M', \gamma', D', \sigma', \mu'_{eq})\\
&\Leftrightarrow \quad \exists \mathcal{A} \in \mathrm{Reg}_n(\mathbf{R}) \text{ s.t. } (\mathcal{A}M, \mathcal{A}\gamma, \mathcal{A}D, \mathcal{A}\sigma, \mu_{eq}) = (M', \gamma', D', \sigma', \mu'_{eq}),
\end{aligned}$$

where $\mathrm{Reg}_n(\mathbf{R})$ denotes the set of regular $n \times n$ matrices with real-valued entries. The associated equivalence classes $[(M, \gamma, D, \sigma, \mu_{eq})]_\sim \subset \mathbf{FL}(n)$ contain only models that share the same solution process.

We can again circumvent this identifiability problem:

Lemma 4.2 *Let $(M, \gamma, D, \sigma, \mu_{eq}) \in \mathbf{FL}(n)$ be given. Then for each inverse temperature $\beta > 0$ there exists a unique element, denoted $(\hat{M}, \hat{\gamma}, \hat{D}, \hat{\sigma}, \hat{\mu}_{eq})_\beta$, in $[(M, \gamma, D, \sigma, \mu_{eq})]_\sim$ that satisfies the fluctuation-dissipation relation combined with the requirement that $\hat{D}$, and $\hat{M}$ be symmetric.*

The proof of this lemma can again be found in Appendix E. We will again call the parameter tupel $(\hat{M}, \hat{\gamma}, \hat{D}, \hat{\sigma}, \hat{\mu}_{eq})_\beta$ *physically generic* iff $\hat{D} > 0$, and $\hat{M} > 0$.

Thus, having computed $\hat{F}$, $\mathrm{Cov}(Z) + E$, $\hat{\mu}$, and $\hat{\Sigma}\hat{\Sigma}^\top$ based on Theorem 2.1, we will first have to decide whether the form of $\hat{F}$ and $\hat{\Sigma}\hat{\Sigma}^\top$ justify the use of the full/overdamped Langevin model: We will have to *decide whether* they approximately satisfy the following block forms

$$\hat{F} = \begin{pmatrix} 0 & \mathrm{Id} \\ \hat{F}_{21} & \hat{F}_{22} \end{pmatrix}, \quad \hat{\Sigma}\hat{\Sigma}^\top = \begin{pmatrix} 0 & 0 \\ 0 & S_{22} \end{pmatrix}, \quad \mathrm{Cov}(Z) + E = \begin{pmatrix} C_{11} & C_{12} \\ C_{12}^\top & C_{22} \end{pmatrix}.$$

We will subsequently assume that this is the case; however, a general strategy of how to make this decision is not a topic of this article (see Sec. 4.3.1 below). The identity $\hat{F}(\mathrm{Cov}(Z) + E) + (\mathrm{Cov}(Z) + E)\hat{F}^\top = -\hat{\Sigma}\hat{\Sigma}^\top$ then yields

$$C_{12} = 0, \quad \hat{F}_{21}C_{11} + C_{22} = 0, \quad \hat{F}_{22}C_{22} + C_{22}\hat{F}_{22}^\top = -S_{22}.$$

This supposed and by setting a certain value for β, we can identify $(\hat{M}, \hat{\gamma}, \hat{D}, \hat{\sigma}, \hat{\mu}_{eq})_\beta$ from

$$-M^{-1}\hat{D} = \hat{F}_{21}, \quad -M^{-1}\hat{\gamma} = \hat{F}_{22}, \quad \hat{M}^{-1}\hat{\sigma}\hat{\sigma}^\top\hat{M}^{-1} = S_{22}, \quad \hat{\gamma} + \hat{\gamma}^\top = \beta\hat{\sigma}\hat{\sigma}^\top, \quad \hat{\mu}_{eq} = \hat{\mu},$$

with the constraint that $\hat{D}, \hat{M}$ have to be symmetric. Including these constraints the above equations have the (unique) solution

$$\begin{aligned} \hat{\gamma} &= -\hat{M}\hat{F}_{22}, && (38) \\ \hat{M}^{-1} &= \beta\, C_{22}, && (39) \\ \hat{D}^{-1} &= \beta\, C_{11} = -\hat{M}\hat{F}_{21} && (40) \\ \hat{\sigma}\hat{\sigma}^\top &= \frac{1}{\beta}(\hat{\gamma} + \hat{\gamma}^\top) && (41) \\ \hat{\mu}_{eq} &= \hat{\mu}, \end{aligned}$$

with E as in Thm. 2.1. Whether this parameter tupel is physically generic or not depends on whether $\mathrm{Cov}(Z) + E > 0$ or not. As above, whenever we assume that the time series $Z = \{Z_1, \ldots, Z_T\}$ under consideration comes from an ergodic, infinitely long time series whose associated covariance matrix $\mathrm{Cov}(Z_\infty)$ is positive definite, then $\mathrm{Cov}(Z) + E > 0$ for large enough T.

4.2.3 Selection of β

Whenever the time series originates from simulations or experiments with a heat bath or a constant thermostatic setting, like, e.g., it is the case in molecular dynamics simulations using the canonical ensemble, the value of the temperature is given by this setting and should be used for selecting β. However, even then we might simply set $\beta = 1$ which then means that we select specific *physical units* in which this is the case.

Whenever the setting does not imply which value has to be taken then estimation of β from data will be required. Several approaches seem possible depending on the nature of the problem: For example, if the setting seems to allow the assumption of the equi-partition of energy as in statistical mechanics, one may estimate β from the average energy of a small part of the system, e.g., an atom, or a single bond. However, even purely statistical approaches have been discussed, e.g., see [23].

4.3 Further Remarks and Generalizations

4.3.1 Langevin or Not?

After identifying the optimal parameters for the (F, μ, Σ)-model, we still have to decide whether the optimal model has the form of either a hypo-elliptic or an overdamped Langevin model, or whether if does

not belong to this class of processes. We propose an approach along the following lines of reasoning: Let the states in the time series Z_k contain positions and momenta (or positions and velocities, respectively), i.e., $Z_k = (q_k, p_k) \in \mathbf{R}^n \times \mathbf{R}^n$. Then compute $(\hat{\mu}, \exp(\tau\hat{F}), \hat{\Sigma}\hat{\Sigma}^T)$ due to Thm. 2.1. If the resulting $\hat{F}$ and $\hat{\Sigma}$ are close to the structure given in (8) then the time series is accepted to allow description by a (linear) hypo-elliptic Langevin model. If, in contrast, $\exp(\tau\hat{F})$ and $\hat{\Sigma}$ are close enough to being block-diagonal, then the time series of the positions (q_k) can be approximated by an n-dimensional overdamped Langevin model (cf. Appendix C, paragraph "Autocorrelation for large friction"). If neither of these conditions is satisfied to an acceptable degree then the time series belongs to some $2n$-dimensional Ornstein-Uhlenbeck process that has no Langevin but a more general form. We will illustrate this procedure in detail in the subsequent numerical experiments in Section 5.2. However, it should be obvious, that the decisions of whether the optimal parameter matrices are *close* in structure to a given form crucial depend on the uncertainty of the matrix entries. Thus, a more elaborated scheme for these decisions should be based on advanced statistical procedures, e.g., via confidence intervals of the parameters or in the framework of a full Bayesian approach.

4.3.2 Partially Observed States

We will now shortly address the case in which velocities as well as momenta are *un*-observed, i.e., the given time series contains information on the positions $q = (q_k)$ only but one wants to find the optimal full Langevin dynamics of form (8). Then, one can calculate the derivatives of the log–likelihood function (19) (where still $Z_k = (q_k, p_k)$) with respect to the momenta p_k and set them to zero. Note that in contrast to the approach presented in [23], we are dealing with the exact (i.e., without discretization in time) form of the log-likelihood function for the full Langevin dynamics of type (8)). Therefore, the resulting estimation scheme is more general than that presented in [23].

It is however easy to verify that, as for the approach of [23], the resulting equations for the optimal estimators consist of $T \times n$ linear equations with $T \times n$ unknowns with a banded matrix on the right-hand side. Consequently, the applicability of the resulting scheme is limited for large dimensions and long time series.

4.3.3 Memory and Unresolved Scales

Our reduced model satisfies the Markov property. The original system (which yields the given time series via observations with constant time lag τ), however, may not be Markovian. If the time scale of its memory is τ_*, this will *not* result in problems whenever $\tau > \tau_*$. Thus, whenever we can choose the time lag τ of the time series (as it is the case if it comes from simulations), then we should estimate τ_*, i.e., by computing (partial) autocorrelations, and choose larger time lags. Whenever this is not possible the use of the methods developed herein is not advisable and one should consider using other approaches, e.g., via ARIMA models or reduced models with build-in memory like generalized Langevin models [19, 36].

Similar problems may result if the time lag τ is much larger than the fastest modes in the original system. Then, as is intuitively clear, the optimal estimators based on time lag τ may fail to reproduce the dynamics on and mechanics of the fastest scales. Whether this can be tolerated or not depends on the system under investigation. Compare [37] for similar considerations.

5 Numerical Examples

In this section we aim at testing the numerical performance of the suggested method on two multidimensional systems. (i) The first system is constructed out of two Gaussian wells in two dimensions and of N harmonic potentials coupled to it. This system exhibits metastable behavior in two dimensions originating from a double–well potential for the first two of the $N + 2$ degrees of freedom. In order to make the task of finding these metastable subsets more demanding, we additionally rotate the derived potential in $(N+2)$–dimensional space with a randomly chosen rotation matrix. (ii) The second numerical example deals with identification of the conformations in a realistic molecular system (12-Alanine).

We use the numerical strategy presented above to identify the conformations in the time series of torsion angles. We will compare the resulting local Langevin models and interpret the physical meaning.

5.1 Multidimensional Langevin Dynamics in a Skew Double-Well potential

As a first example we consider realizations of the Langevin equation

$$\ddot{q}(t) \quad = \quad -\mathrm{grad}\, U(q(t)) - \gamma\dot{q}(t) + \sigma\dot{W}(t), \quad (M = \mathrm{Id}), \tag{42}$$

with $q = (x, y) \in \mathbf{R}^2 \times \mathbf{R}^N$ and the perturbed two–hole potential

$$U(x,y) \quad = \quad \sum_{l=1}^{2} a_l \exp\left(-(x-\mu_l^{sys})^{\mathcal{T}} D_l^{sys}(x-\mu_l^{sys})\right) + \frac{1}{2} y^\top D_{bath} y \tag{43}$$

$$+\delta_0\Big(\cos(2\pi k(x_1+x_2)) + \cos(2\pi k(x_1-x_2))\Big), \tag{44}$$

where $\delta_0 \ll 1$ is a small perturbation parameter. The N harmonic bath variables are denoted by y, whereas x labels the two "metastable" dimensions that live in the plane of the double well potential. We have chosen the following parameter values

$$\begin{array}{ll} \mu_1^{sys} = (1.8, 2.2)^\top, & \mu_2^{sys} = (1.8, 0.8)^\top, \\ D_1^{sys} = \begin{pmatrix} 1 & -1 \\ -1 & 3 \end{pmatrix}, & D_2^{sys} = \begin{pmatrix} 1 & 1 \\ 1 & 3 \end{pmatrix}, \\ a_1 = -6 & a_2 = -6 \end{array}$$

such that we get two contiguously placed skew wells (see Fig. 3) and make identification of the metastable sets more challenging compared to a well–separated situation. The parameter matrices D_{bath} and γ have been chosen to be symmetric, positive definite, and tri-diagonal, with 10.0 on the main diagonal and 5.0 on secondary diagonals for D_{bath} (5.0 and 2.5 respectively for γ). The noise parameter σ was taken as a diagonal matrix with 4.0 on the diagonal. Thus the degrees of freedom x and y are coupled through the friction only. These parameter settings imply two important properties: (1) The system is metastable because the barrier is sufficiently larger than the average kinetic energy in the system. (2) The fluctuation dissipation relation is *not* valid here, but since the input of the algorithm consists of positions and *momenta* we do not have an identifiability problem of the sort discussed above.

Simulation of the model has been realized with the Euler–Maruyama integrator (discretization time step $\Delta t_{Euler} = 0.0002$) and total time length 500. Each hundredth instance of the resulting time series has been taken for a subsequent parameter estimation (resulting in observation time step $\tau = 0.02$) such that $T = 25.000$.

Furthermore, in order to make our model system more realistic and mimic the features inherent in biological systems, we rotate the resulting time series in the (N+2) dimensional space. We do it in such a way, that the metastability of the system becomes *hidden* in all the dimensions of the system, see Fig. 2.

Application of the HMM–Langevin method (local SDE models from $\mathbf{FL}(n)$) to the time series results in identification of two metastable states. In order to interpret the quality of the identification, we rotate the time series back, color the elements according to the corresponding metastable state and plot them atop of the original potential surface in (x_1, x_2). As we can see in Fig. 3, the local Langevin models are correctly situated at the wells of the double–well potential in the metastable dimensions and the elements of the time series are assigned in a proper way. Fig. 4 shows the identified parameter matrices of the local Langevin models (after the backwards rotation to bring them in the form comparable to (42)). We can see that all of the parameters are estimated with satisfactory accuracy and the difference between the local Langevin models as expected only lies in first two dimensions of D (corresponding to the metastable degrees of freedom (x_1, x_2)).

Finally, we test the performance of the method with respect to the spatial dimension of the problem (see Fig. 5). We compare the relative errors for a time series of length 15.000 ($\tau = 0.02$) generated for different dimensions N of the oscillator bath. For all of the model parameters we have a linear growth of the estimation error with the dimension.

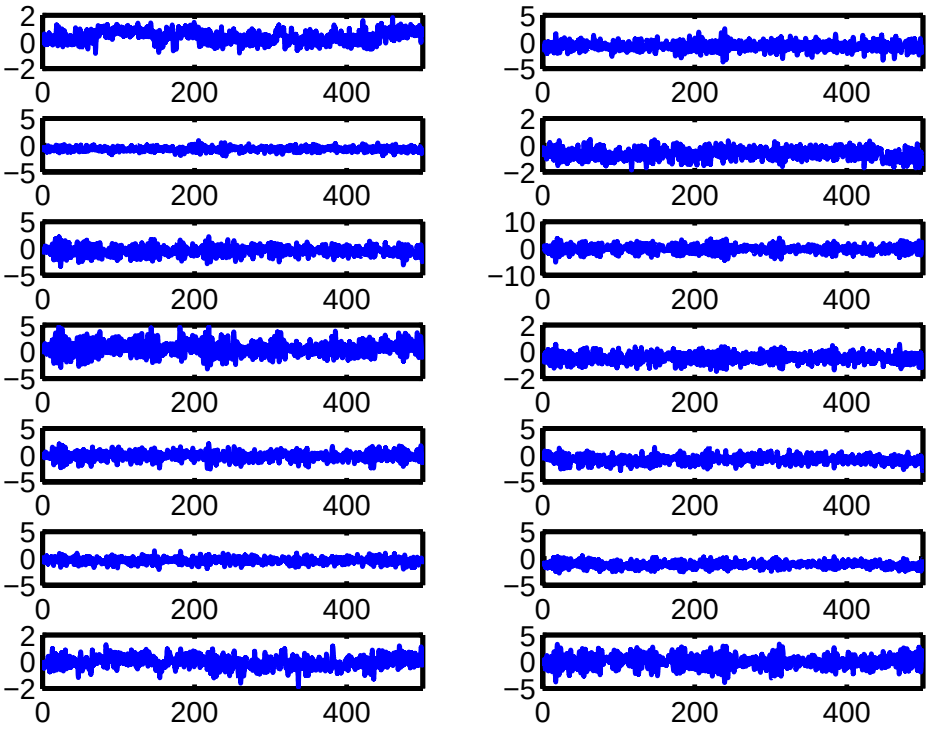

Fig. 2 Time series of the skew double–well model in $N + 2 = 14$ dimensions (observation step $\tau = 0.02$, $T = 25.000$, generated as described in the text, illustration after rotation).

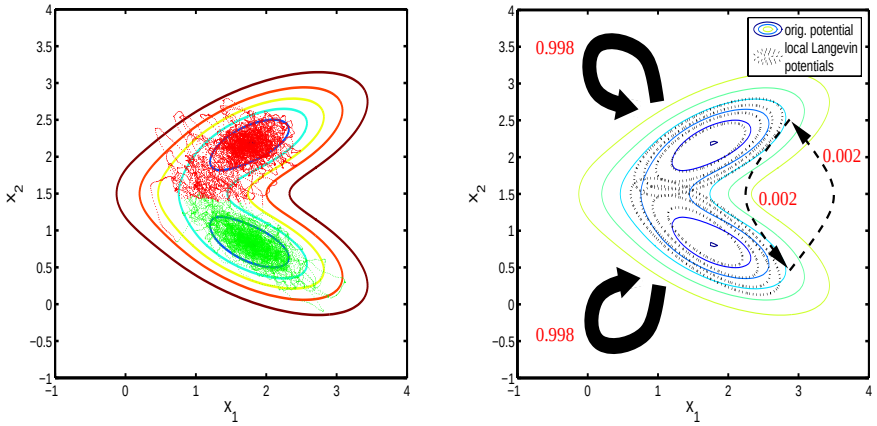

Fig. 3 Left: Projection of the time series shown in Fig. 2 (after back-rotation) onto the two metastable dimensions; coloring according to assignment to the two metastable sets as resulting from the HMM–Langevin Viterbi path. Right: Visualization of the reduced HMM–Langevin model in two metastable dimensions (also after back-rotation).

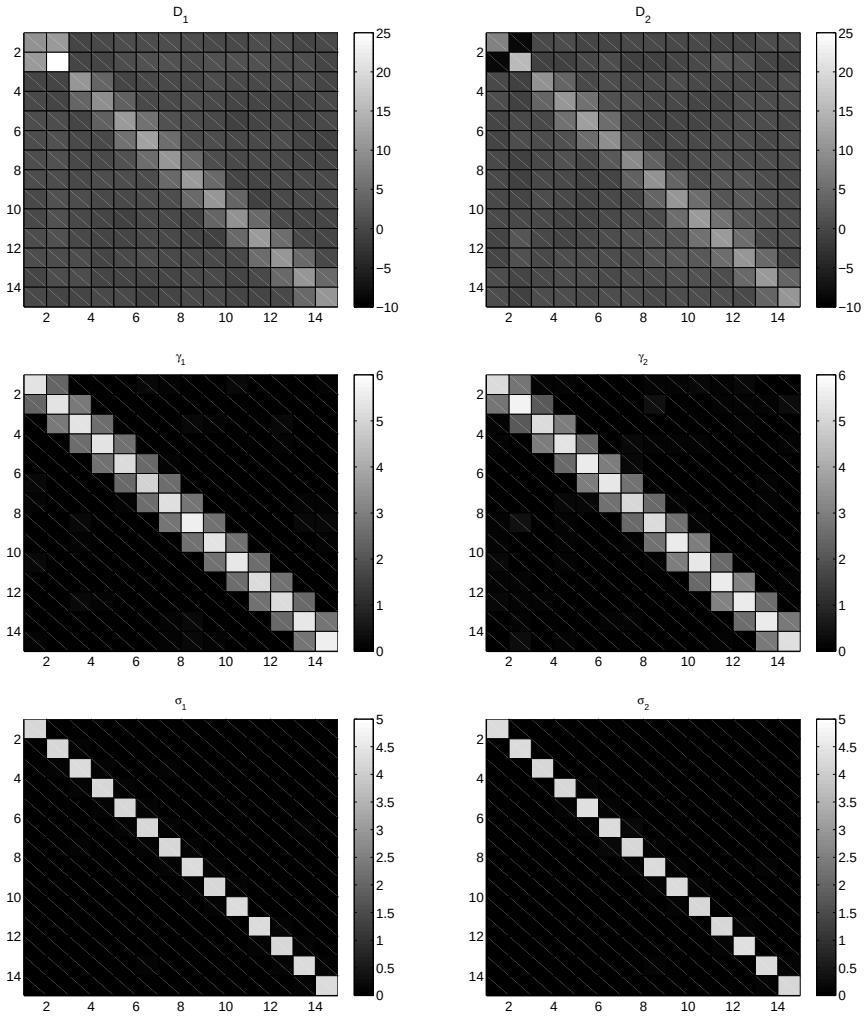

Fig. 4 Estimated parameters of two local Langevin models (left vs. right) resulting from the time series shown in Fig. 2; from top to bottom: stiffness matrices D_i, friction matrices γ_i, and noise intensity matrices σ_i. The relative error in all cases is less then 0.05 in maximum norm. Here, we deliberately used the knowledge about $M = \mathrm{Id}$ such that D_i, γ_i, σ_i can be directly taken from the respective blocks of $\hat{F}_i$ and $\hat{\Sigma}_i$.

5.2 Dynamics of the 12–Alanine Peptide

In order to illustrate the approach on a realistic molecular system, we choose a simulation of 12-Alanine in water at 300K. The molecular dynamics simulation was performed with the GROMOS force–field and implicit water box with a 2 fs time step and total length in time of 1.3 milliseconds. The time series has been provided by Frank Noe in a form of a 22–dimensional time series describing the dynamics of the internal (ϕ, ψ)–peptide angles of 12-Alanine (torsion angles corresponding to freely rotating end-groups were neglected). The associated velocities have been computed by numerical differentiation. The resulting time series of angles and associated velocities have been analyzed with the observation time step τ ranging from $\tau = 100$ fs to $\tau = 1$ ps (which means T ranging from $T = 13.000.000$ to $T = 1.300.000$, $n = 22$).

In the following we present results on the estimation of parameter matrices of processes of the form

$$\begin{aligned} \dot{z}(t) &= F^{(i(t))}(z - \mu^{(i(t))}) + \Sigma^{(i(t))}\,\dot{W}(t) \\ i(t) &= \text{Markov jump process with states } 1, \ldots, L. \end{aligned} \tag{45}$$

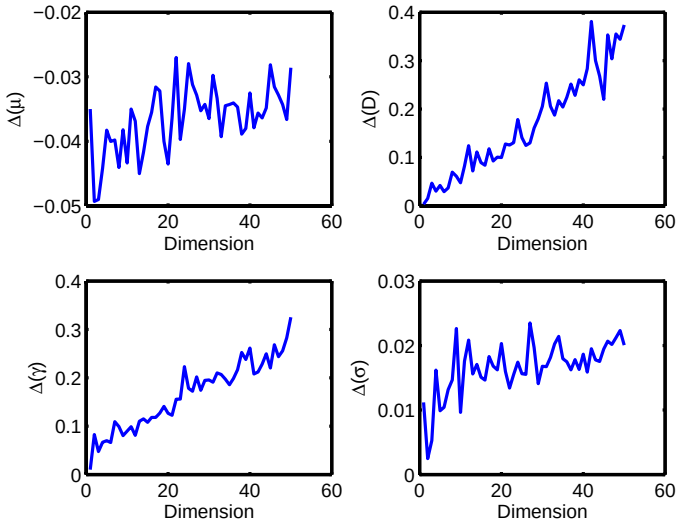

Fig. 5 Relative error of the parameter estimation in maximum–norm for a time series with 15.000 elements and $\tau = 0.05$.

We will call the SDEs (45) *second order* if $z = (q, \dot{q}) \in \mathbf{R}^{2n}$ (positions/angles and corresponding velocities), *first order* if $z = q \in \mathbf{R}^n$ (positions/angles only). From the optimal parameters $(\hat{F}^{(i)}, (\hat{\Sigma}\hat{\Sigma}^\top)^{(i)}, \hat{\mu}^{(i)})$ we will try to extract the optimal parameters for $\mathbf{FL}(2n)$-models (second order case; if the matrices $\hat{F}^{(i)}$ and the associated covariance matrices have the appropriate form) or for $\mathbf{OL}(n)$-models (first order case).

The transition matrix of the Hidden Markov chain with $G = 7$ hidden states calculated in the context of HMM–Langevin approach with positions/angles and their respective velocities (second order) indicates the presence of two most pronounced metastable states (see Fig. 6). The corresponding aggregated Viterbi–path is shown in the right panel of Fig. 6; it proves robust to changes of the initial number of G hidden states and to changes in the initial settings of the HMM procedure. The optimal estimator $\hat{\mu}$ for the two metastable states has a direct interpretation: it gives the mean configuration of the corresponding conformation in configuration space. These configurations are visualized in Fig. 7. From this figure it becomes visible that the second state corresponds to an α-helical structure, whereas the first has a mean configuration that lies between α–helix and β–sheet (it is a mixture between some misfolded and partially folded less metastable conformations that are separated from each other when more than the two leading eigenvalues of the transition matrix are taken into account).

When checking the estimators for the optimal estimators $(\hat{F}^{(i)}, (\hat{\Sigma}\hat{\Sigma}^\top)^{(i)})$ for the two resulting local second order models, however, we find that the estimator $\hat{F}^{(1)}$ for the helical conformation has the form illustrated in Fig. 8. We observe that $\hat{F}^{(1)}$ is far from the form typical for the hypo-elliptic Langevin model. Instead $\hat{F}^{(1)}$ (and $(\hat{\Sigma}\hat{\Sigma}^\top)^{(1)}$, too) is almost block-diagonal which seems to indicate that the dynamics of positions/angles q are essentially decoupled from the corresponding velocities. This, however, means that it is justified to use first order dynamics (i.e., $\mathbf{OL}(n)$-models) as local models of the dynamics, see Appendix C ("Autocorrelations for Large Friction") for further details.

When restarting the procedure with local $\mathbf{OL}(n)$-models, we get precisely the same Viterbi-path, almost the same transition matrix, and approximately the same mean positions μ as already computed with second order models and illustrated in Figs. 6 and 7.

The estimators of the parameter matrices of the first order model $\dot{q} = F(q - \mu) + \Sigma\dot{W}$ for the *helical* conformation are shown in Figs. 9-10.

Fig. 9 shows the estimator $\hat{F}^{(1)}$; from comparison with Fig. 8 we observe that it is almost identical with the upper left block of the estimator $\hat{F}^{(1)}$ computed with second order models.

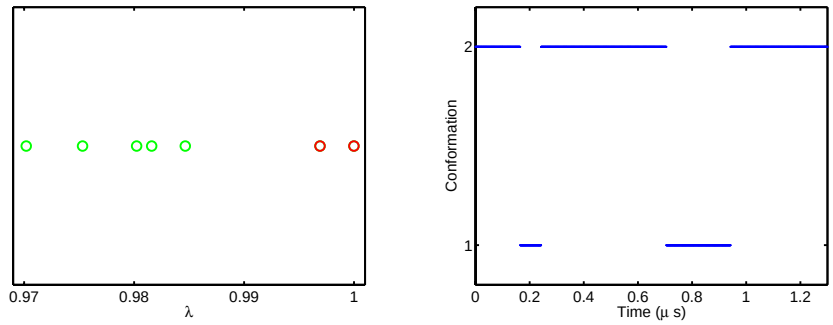

Fig. 6 12-Alanine: The spectrum of the transition matrix of the hidden Markov chain ($L = 7$) indicates two metastable states (just two eigenvalues close or identical to 1, left panel); the corresponding aggregated Viterbi–path is shown right.

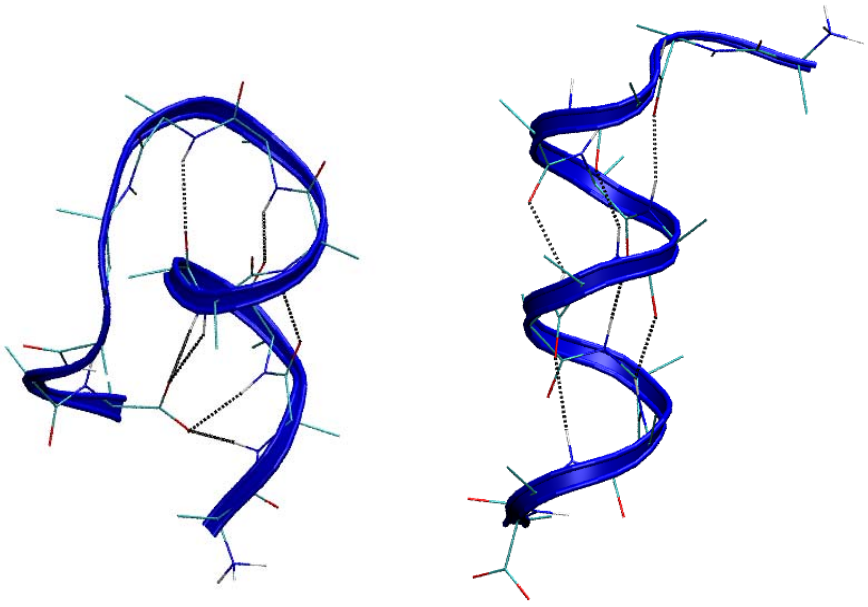

Fig. 7 12-Alanine: Mean configurations of the two metastable states as computed from parameters $\hat{\mu}^{(i)}$

Fig. 10 shows the estimators $\hat{D}$, $\hat{\gamma}$, and $\hat{\sigma}$ for the local **OL**(n)-model

$$\gamma\dot{q} = -D(q-\mu) + \sigma\dot{W},$$

of the helical conformation as computed from $\hat{F}^{(1)}$, and $(\hat{\Sigma}\hat{\Sigma}^{\top})^{(1)}$ and the corresponding covariance matrix as explained in Section 1 (setting $\beta = 1$ herein). As a measure for the quality of the resulting approximation we use the matrix E that we understand as an indicator for how well the covariances have been sampled. This matrix is illustrated in the lower right panel of Fig. 10; we observe that it is satisfactorily small.

This section is about demonstrating the performance and applicability of the proposed estimation techniques such that we cannot go into details of the physical validation of the results. However, let us add the following comments on possible interpretation of the results: The existence of a helical conformation of 12-Alanine is known and expected from other investigations concerning the alanine family; the existence of partially unfolded and mis-folded β-sheet-like conformations also. The resulting

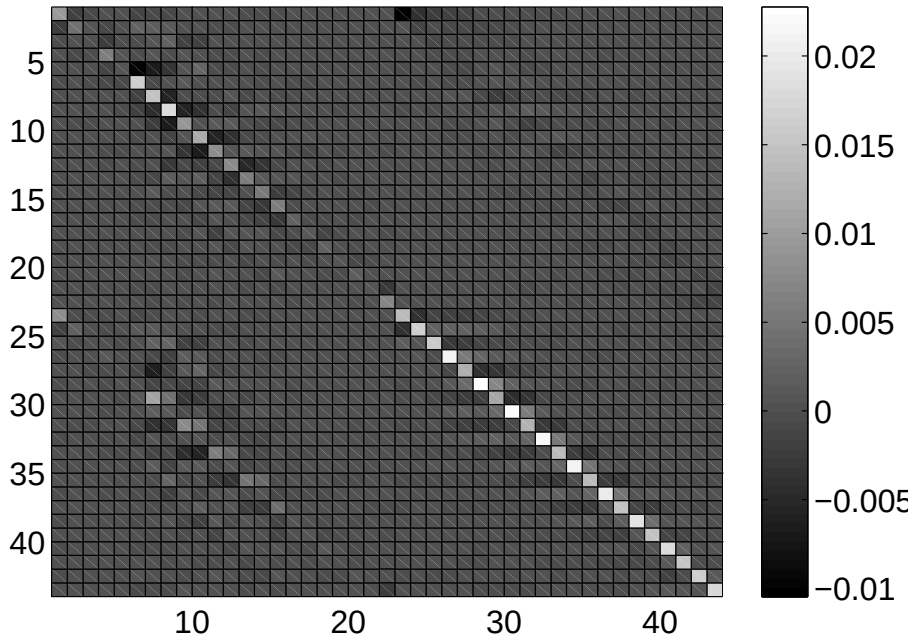

Fig. 8 12-Alanine: Estimator of the 44×44 matrix $\hat{F}^{(1)}$ for the helical conformation using a second order diffusion model $\dot{x} = F(x - \mu) + \Sigma\dot{W}$ with $x = (q, \dot{q})$ computed via the positions/angles and their respective velocities. $\hat{F}^{(1)}$ has been computed due to the algorithm introduced in the appendix (case II) for a series of values of the lag time τ ranging from 100 fs to 2000 fs.

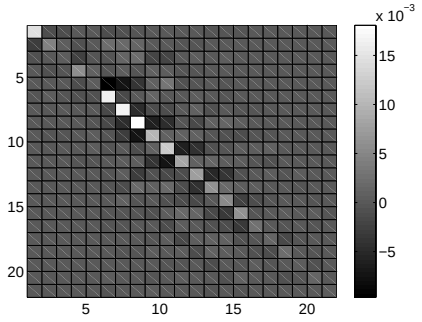

Fig. 9 12-Alanine: Estimator of the 22×22 matrix $\hat{F}^{(1)}$ for the helical conformation for the first order model $\dot{q} = F(q - \mu) + \Sigma\dot{W}$ computed via the positions/angles only. $\hat{F}^{(1)}$ has been computed computed using the algorithm introduced in the appendix (case II) for a series of values of the lag time τ ranging from 100 fs to 2000 fs.

stiffness matrix $\hat{D}$ of the helical conformation shows that the internal alanine peptides are rather stiffly packed while the alanine "end-groups" can move more flexibly; its band-like structure means that the interaction is dominated by the groups within one helical loop length. All this also is no surprise. The structure of the friction and noise intensity matrices, $\hat{\gamma}$ and $\hat{\sigma}$, would have also been expected: friction/dissipation should be largest for the internal groups while the noise intensity matrix should be

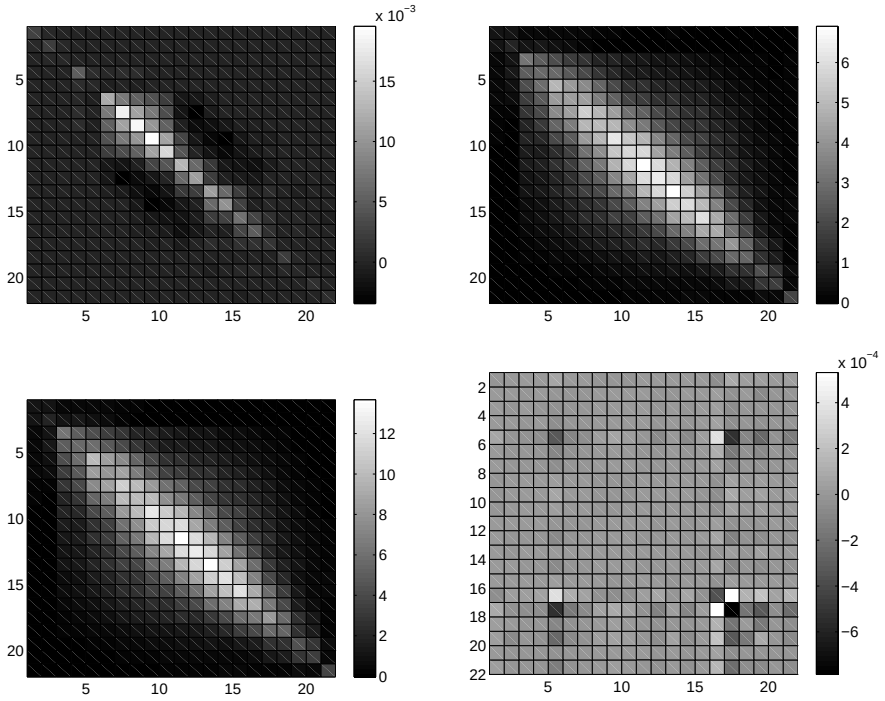

Fig. 10 12-Alanine: estimators $\hat{D}$ (left, top), $\hat{\gamma}$ (right, top), $\hat{\sigma}\hat{\sigma}^T$ (left, bottom) for the parameters of the linear overdamped Langevin model $\gamma\dot{q} = -D(q-\mu) + \sigma\dot{W}$ parameters for the helical conformation (setting $\beta = 1$); the fourth sub-figure (right, bottom) shows the error indicator matrix E (for details, see text).

diagonally-dominated since the stochastic excitation comes from the surrounding solvent molecules and should be essentially local, and only neighboring groups should add internal random-like excitations.

In order to conclude our discussion of 12-Alanine, we present in Fig. 11 the comparison of the correlation matrices Cor_T (for definition see Appendix C) for different values of τ or T, respectively, with the correlation matrices $\exp(\tau\hat{F})$ computed from the overdamped linear Langevin dynamics for the helical conformation. When computed from the HMM-Langevin estimators with second order local models the results look identical (deviations between second order and first order smaller than 0.1%).

6 Concluding Remarks

In this article we presented the HMM-Langevin approach to model reduction and parameter estimation of *metastable systems* based on time-series with constant lag time between observation resulting from long simulation runs that contain information on "positions" q as well as corresponding "velocities" $\dot{q}$. We in detail considered the appearing identifiability problems, shed light on the role of the fluctuation-dissipation relation, and illustrated possible reduction to overdamped Langevin / diffusion models that typically occur in application from biology, biophysics, or biochemistry. We demonstrated the applicability to realistic time series from molecular dynamics simulations and discussed the resulting performance of the novel approach.

In general, the novel approach has no restriction on the dimensionality of the system (scaling d^3 with dimension), on the length T of the time series (linear in T), or on the lag time (need not be "small"). However, we already discussed the possible pitfalls in computing the Jacobian $\hat{F}$ from the estimator $\exp(\tau\hat{F})$ (see last section of the appendix). Furthermore, we ignored the fact that the given time series might not be Markovian (in the sense that its partial autocorrelation time might be larger than the lag

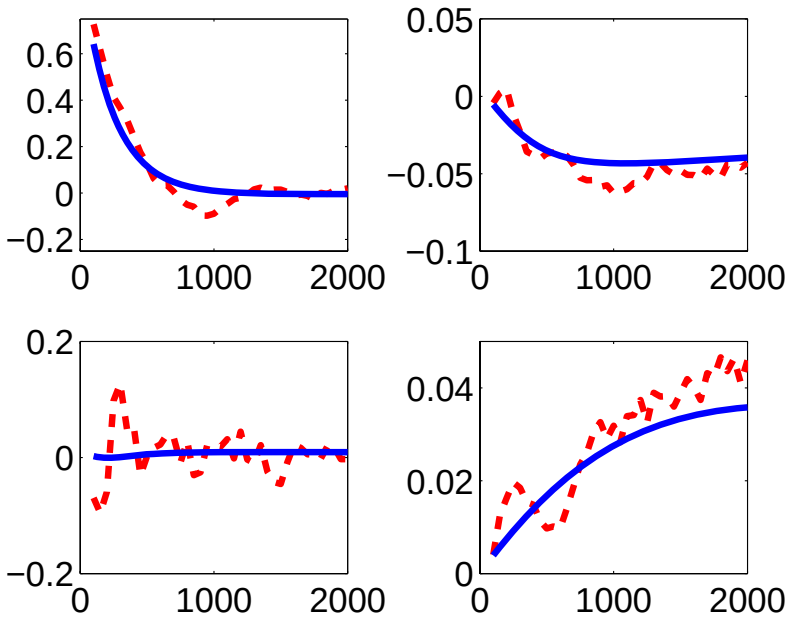

Fig. 11 12-Alanine: Dependence of the matrix Cor_T (blue solid lines) on the lag time τ and comparison with $\exp(\tau\hat{F})$ (red dashed lines) for four randomly chosen entries of the respective matrices. Top panel: $\mathrm{Cor}_{T,2,2}$ and $\exp(\tau\hat{F})_{2,2}$ (left) and $\mathrm{Cor}_{T,2,5}$ / $\exp(\tau\hat{F})_{2,5}$ (right). Lower panel: $\mathrm{Cor}_{T,6,8}$ and $\exp(\tau\hat{F})_{6,8}$ (left) and $\mathrm{Cor}_{T,21,19}$ and $\exp(\tau\hat{F})_{21,19}$ (right).

time). This case definitely needs careful consideration but is postponed to future investigations (here we can "always" avoid it by sub-sampling the given time series with some "large enough" lag time).

Appendix A: Positive Definite Matrices

A matrix $A \in \mathrm{Mat}_n(\mathbf{R})$ is called positive definite iff $\psi^\top A\psi > 0$ for all $0 \neq \psi \in \mathbf{R}^n$. We then write $A > 0$. Positive definiteness is often only defined for symmetric matrices. We here have that for some (possibly non-symmetric) $A \in \mathrm{Mat}_n(\mathbf{R})$

$$A > 0 \quad \Leftrightarrow \quad A_{sym} = \frac{1}{2}(A + A^\top) > 0 \quad \Leftrightarrow \quad \min \mathrm{spec}(A_{sym}) > 0.$$

It is important to note that in general $\Re\mathrm{spec}(A) > 0$ does *not* imply $A > 0$. The 2×2-matrix

$$A = \begin{pmatrix} 0 & -1 \\ 1 & 1 \end{pmatrix}$$

may be an example: its eigenvalues all have positive real part but for $e = (1, 0)^\top$ we have $e^\top Ae = 0$ since A_{sym} has 0 as an eigenvalue.

The following lemma collects some results on positive (semi-)definite matrices:

Lemma 6.1 *Let $A \in \mathrm{Mat}_n(\mathbf{R})$. Then $A > 0$ implies that*

1. *A is regular and $A^{-1} > 0$,*
2. *there is exactly one matrix $B \in \mathrm{Mat}_n(\mathbf{R})$ with $B > 0$ such that $A = B^2$,*
3. *if also $B > 0$ then $A + B > 0$,*

For symmetric $A \in \mathrm{Mat}_n(\mathbf{R})$ we have $A > 0$ if and only if there exists a regular $B \in \mathrm{Mat}_n(\mathbf{R})$ such that $A = BB^\top$. If $A \in \mathrm{Mat}_n(\mathbf{R})$ is just symmetric, positive semi-*definite (i.e. $\psi^\top A\psi \geq 0$ for all $0 \neq \psi \in \mathbf{R}^n$) of rank r, then a $B \in \mathrm{Mat}_n(\mathbf{R})$ of rank r such that $A = BB^\top$.*

Appendix A: Solution to the Sylvester Equation

Linear matrix equations of the form

$$AX + XB = C, \tag{46}$$

where A, B, C are (in our case square) given matrices and X is the sought-after solution matrix, are called *Sylvester equation*. It is well-known [38] that (46) has a unique solution if and only if

$$\alpha + \beta \neq 0 \quad \forall \alpha \in \mathrm{spec}(A), \beta \in \mathrm{spec}(B), \tag{47}$$

where spec$(\cdot)$ denotes the spectrum of a matrix. This general result yields the following two consequences that are important herein:

Lemma 6.2 *Let F be a square matrix whose spectrum is contained in the left half plane of the complex plane, i.e.,*

$$\mathrm{spec}(F) \subset \mathbf{C}^- = \{z \in \mathbf{C} : \Re(z) < 0\}.$$

Then the two Sylvester equations

$$XF^T + FX = C \tag{48}$$

$$Xe^{\tau F^T} - e^{-\tau F}X = E \tag{49}$$

have unique solutions for arbitrary square matrices C and E, and $\tau > 0$. Moreover, whenever F has the form

$$F = \begin{pmatrix} 0 & M^{-1} \\ -D & -\gamma M^{-1} \end{pmatrix},$$

with some positive definite square matrix γ then spec$(F) \subset \mathbf{C}^-$.

Proof: The last statement directly follows from [39]. The uniqueness for (48) is guaranteed since for arbitrary $\alpha \in \mathrm{spec}(F) \subset \mathbf{C}^-$ and $\beta \in \mathrm{spec}(F^T) = \mathrm{spec}(F) \subset \mathbf{C}^-$ we have

$$\Re(\alpha + \beta) < 0 \text{ such that } \alpha + \beta \neq 0.$$

In analogy the uniqueness for (49) results since for arbitrary $\alpha \in \mathrm{spec}(-e^{-\tau F})$ and $\beta \in \mathrm{spec}(e^{\tau F^T})$ we find $\lambda, \nu \in \mathrm{spec}(F)$ such that $\alpha = -e^{-\tau\lambda}$ and $\beta = e^{\tau\nu}$ such that

$$\begin{aligned} \alpha + \beta \neq 0 \qquad & \text{if and only if} \qquad \exp(\tau(\lambda + \nu)) \neq 1 \\ & \text{if and only if} \qquad \lambda + \mu \neq 0, \end{aligned}$$

which is again valid since $\mathrm{spec}(F) \subset \mathbf{C}^-$.

Under certain conditions, the solution of (49) is easily available: Let us denote (49) in the form

$$E = XA^\top - A^{-1}X, \quad A = e^{\tau F}, \tag{50}$$

and assume that A allows complex diagonalization of the form:

$$A = J\Lambda J^{-1} \tag{51}$$

where $\Lambda \in \mathbf{C}^{n\times n}$ is a complex diagonal matrix with non-vanishing entries, and $J \in \mathbf{C}^{n\times n}$ invertible. Inserting (51) into (50) and multiplying both sides of the equation with J^{-1} from the left and $J^{-\top}$ from the right we get:

$$J^{-1}EJ^{-\top} = Y\Lambda - \Lambda^{-1}Y \tag{52}$$

$$Y = J^{-1}XJ^{-\top} \tag{53}$$

It is easy to see that a componentwise solution of (52) can be written as

$$Y_{ij} = \frac{\left(J^{-1}EJ^{-\mathcal{T}}\right)_{ij}}{\Lambda_j^{\top} - \Lambda_i^{-1}} \tag{54}$$

Then the solution of (50) is

$$X = JYJ^{\top} \tag{55}$$

Appendix C: Explicit Expressions for Optimal Estimators

Autocorrelations

We again consider

$$\dot{z} = F\,(z-\mu) + \Sigma\dot{W}. \tag{56}$$

We subsequently assume that the spectrum of F lies in the left half plane $\{z \in \mathbf{C} : \Re(z) < 0\}$ of the complex plane. This assumption is satisfied, e.g., in the case of

$$F = \begin{pmatrix} 0 & M^{-1} \\ -D & -\gamma M^{-1} \end{pmatrix}, \quad \text{or} \quad F = -\gamma^{-1}D \tag{57}$$

with γ and D being positive definite.

In order to compute the autocorrelation of the process z (as they are needed to determine the optimal estimators) we first transfer to new coordinates

$$\xi = z - \mu$$

such that the equation of motion reads

$$\dot{\xi} = F\xi + \Sigma\dot{W}.$$

Then by standard techniques we get the following expression for the un-normalized autocorrelation

$$\langle \xi(t)\xi(0)^T \rangle = \frac{1}{2\pi}\int_{-\infty}^{\infty} (-F + i\omega\mathrm{Id})^{-1}\Sigma\Sigma^T((-F + i\omega\mathrm{Id})^H)^{-1}e^{i\omega t}d\omega.$$

We now evaluate this integral by considering the following sequence of integrals in the complex plane:

$$C_r(t) = \frac{1}{2\pi i}\int_{\mathcal{C}(r)} (-F + z\mathrm{Id})^{-1}\Sigma\Sigma^T(-F^T - z\mathrm{Id})^{-1}e^{zt}dz,$$

where $\mathcal{C}(r)$ denotes the closed curve that is composed of the path from $-ir$ to ir along the imaginary axis and the half circle $\{z \in \mathbf{C} : z = r\exp(i\phi), \phi = \pi/2, \ldots, 3\pi/2\}$. For large enough r the spectrum of F is contained in the interior of $\mathcal{C}(r)$. Moreover, for $t > 0$ and large enough r, the integrand $I(z)$ decays like $|I(r\exp(i\phi))| < Mr^{-k}$ for some constants $M > 0$ and $k > 1$ such the the contribution of the half circle to the integral $C_r(t)$ vanishes for $r \to \infty$. This means that

$$\lim_{r\to\infty} C_r(t) = \langle \xi(t)\xi(0)^T \rangle.$$

In order to evaluate the integral $C_r(t)$ for large r, we first observe that

$$(z\mathrm{Id} - F)^{-1}\Sigma\Sigma^T(-F^T - z\mathrm{Id})^{-1} = (z\mathrm{Id} - F)^{-1}A + A(-F^T - z\mathrm{Id})^{-1}, \tag{58}$$

where the square matrix A is the unique solution of

$$AF^T + FA = -\Sigma\Sigma^T. \tag{59}$$

Uniqueness and specific forms of the solution will be discussed below. This decomposition of the integral yields

$$C_r(t) = \frac{1}{2\pi i}\left(\int_{\mathcal{C}(r)} (z\mathrm{Id} - F)^{-1} e^{zt} dz\, A - A \int_{\mathcal{C}(r)} (z\mathrm{Id} + F^T)^{-1} e^{zt} dz\right).$$

For large enough r, the spectrum of F is contained in the interior of $\mathcal{C}(r)$, while the spectrum of $-F^T$ entirely lies in its exterior. Thus, typical residue theorems yield

$$\lim_{r\to\infty} C_r(t) = \exp(tF)A.$$

Putting everything together we find that the autocorrelation satisfies

$$\langle (z(t)-\mu)(z(0)-\mu)^T\rangle \cdot \langle (z(0)-\mu)(z(0)-\mu)^T\rangle^{-1} = \exp(tF). \tag{60}$$

Let us return to the uniqueness and specific form of A. The uniqueness obviously follows from Appendix A, since the spectrum of F and F^T are contained in the left half plane of $\mathbf{C}$. In addition, we observe that the equation for A is identical with the equation (17) for $R(\infty)$, i.e., $A = R(\infty)$, the asymptotic limit of the covariance of the centered process $\xi = z - \mu$.

Let us furthermore consider the typical case that γ and D, both, are positive definite matrices, that D is symmetric in addition, that

$$\Sigma \;=\; \begin{pmatrix} 0 & 0 \\ 0 & \sigma \end{pmatrix}, \tag{61}$$

and that the fluctuation dissipation result holds, i.e., that

$$\gamma + \gamma^\top = \beta\sigma\sigma^T,$$

for some positive inverse temperature β. Then we uniquely have

$$A = \frac{1}{\beta}\begin{pmatrix} D^{-1} & 0 \\ 0 & M \end{pmatrix}.$$

For the case of overdamped linear Langevin dynamics ($F = -\gamma^{-1}D, \Sigma = \gamma^{-1}\sigma$) we find $A = D^{-1}/\beta$.

Autocorrelations for Large Friction γ

As we have seen the time dependence of the autocorrelation is given by the exponential of F. The following result is shown in [40] based on higher order linear perturbation theory:

$$\exp(t\begin{pmatrix} 0 & M^{-1} \\ -D & -\gamma M^{-1} \end{pmatrix}) = \begin{pmatrix} \exp(-t\gamma^{-1}D) & 0 \\ 0 & \exp(-t\gamma M^{-1}) \end{pmatrix} + \mathcal{O}(\frac{1}{\|\gamma\|}).$$

This means, that for large enough friction γ the autocorrelation matrix of the full Langevin dynamics is approximately block-diagonal with blocks given by the autocorrelations of

$$\gamma\dot{q} = -D(q-\mu) + \sigma\dot{W},$$

which is overdamped Langevin dynamics, and

$$\dot{p} = -\gamma M^{-1}p + \sigma\dot{W},$$

which is the analogue for the momentum part of full Langevin dynamics.

Properties of the Estimator

Suppose that an observation sequence $Z = (Z_t)_{t=1,2,\ldots}$ and the occupation probabilities $\nu_k(i), i = 1,\ldots,L$ of the hidden Markov chain are given. Then, denote the mean and covariance of the subsequences $(Z_1,\ldots,Z_T)$ (first T steps of the given time series) in the state i by

$$\begin{aligned}\bar{Z}_T^{(i)} &= \frac{1}{\sum_{k=1}^{T-1}\nu_{k+1}(i)}\sum_{k=1}^{T-1}\nu_{k+1}(i)Z_k \\ \mathrm{Cov}_T^{(i)}(Z) &= \frac{1}{\sum_{k=1}^{T-1}\nu_{k+1}(i)}\sum_{k=1}^{T-1}\nu_{k+1}(i)(Z_k-\bar{Z}_T^{(i)})(Z_k-\bar{Z}_T^{(i)})^\top.\end{aligned}$$

In all of the following, we assume that for large enough T the covariance $\mathrm{Cov}_T^{(i)}$ is a positive definite matrix.

Furthermore, let the one-step correlation be defined as

$$\mathrm{Cor}_T^{(i)}(Z) = \frac{1}{\sum_{k=1}^{T-1}\nu_{k+1}(i)}\sum_{k=1}^{T-1}\nu_{k+1}(i)(Z_{k+1}-\bar{Z}_T^{(i)})(Z_k-\bar{Z}_T^{(i)})^\top\cdot\mathrm{Cov}_T^{(i)}(Z)^{-1},$$

and, for given $\mu^{(i)}$ also

$$\begin{aligned}A_1^{(i)}(T,\mu^{(i)}) &= \frac{1}{\sum_{k=1}^{T-1}\nu_{k+1}(i)}\sum_{k=1}^{T-1}\nu_{k+1}(i)(Z_{k+1}-\mu)(Z_k-\mu^{(i)})^\top. \\ A_2^{(i)}(T,\mu^{(i)}) &= \frac{1}{\sum_{k=1}^{T-1}\nu_{k+1}(i)}\sum_{k=1}^{T-1}\nu_{k+1}(i)(Z_k-\mu^{(i)})(Z_k-\mu^{(i)})^\top,\end{aligned}$$

Then, the equations (28) and (29) that determine the optimal estimators $B^{(i)}, \hat{\mu}^{(i)}$ based on subsequence $(Z_1,\ldots,Z_T)$ read

$$B^{(i)} = A_1^{(i)}(T,\hat{\mu}^{(i)})\cdot A_2^{(i)}(T,\hat{\mu}^{(i)})^{-1} \quad (62)$$

$$\hat{\mu}^{(i)} = \bar{Z}_T^{(i)} + (\mathrm{Id}-B^{(i)})^{-1}\delta_T^{(i)}, \quad (63)$$

where $\delta_T^{(i)}$ for the sake of convenience denotes

$$\delta_T^{(i)} = \frac{1}{\sum_{k=1}^{T-1}\nu_{k+1}(i)}\sum_{k=1}^{T-1}\nu_{k+1}(i)(Z_{k+1}-Z_k).$$

note that in the case of a single hidden state (i. e. for the case when $\nu_1=\nu_2=\cdots=\nu_T=1$) the expression for δ can be further simplified resulting in

$$\delta_T^{(i)} = \frac{Z_T-Z_1}{T-1}.$$

We easily observe now that

$$\begin{aligned}A_1^{(i)}(T,\hat{\mu}^{(i)}) &= \mathrm{Cov}_T^{(i)}(Z) + (\bar{Z}_T^{(i)}-\hat{\mu}^{(i)})\cdot(\bar{Z}_T^{(i)}-\hat{\mu}^{(i)})^\top \\ A_2^{(i)}(T,\hat{\mu}^{(i)}) &= \mathrm{Cor}_T^{(i)}(Z)\mathrm{Cov}_T^{(i)}(Z) \\ &\quad +(\bar{Z}_T^{(i)}-\hat{\mu}^{(i)})\cdot(\bar{Z}_T^{(i)}-\hat{\mu}^{(i)})^\top + \delta_T^{(i)}\cdot(\bar{Z}_T^{(i)}-\hat{\mu}^{(i)})^\top.\end{aligned}$$

This immediately shows that $A_1^{(i)}(T,\hat{\mu}^{(i)})$ is positive definite for all $\hat{\mu}^{(i)}$ since $\mathrm{Cov}_T^{(i)}$ is. Therefore, the inverse in equation (62) is justified. In order to check whether (62) makes sense as a whole, let us introduce the further abbreviation

$$\Delta_T^{(i)} = \bar{Z}_T^{(i)} - \hat{\mu}^{(i)}$$

and observe that (62) and (63) can be written as

$$\begin{aligned} B^{(i)} &= (\mathrm{Cor}_T^{(i)}\mathrm{Cov}_T^{(i)} + \Delta_T^{(i)}(\Delta_T^{(i)})^\top \\ &\quad +\delta_T^{(i)}(\Delta_T^{(i)})^\top) \cdot (\mathrm{Cov}_T^{(i)} + \Delta_T^{(i)}(\Delta_T^{(i)})^\top)^{-1} \qquad (64) \\ \delta_T^{(i)} &= -(\mathrm{Id} - B^{(i)})\Delta_T^{(i)}. \qquad (65) \end{aligned}$$

With the assumption that $\mathrm{Cov}_T^{(i)}$ is positive definite we get from (64) that

$$(\mathrm{Id} - B^{(i)})(\mathrm{Cov}_T^{(i)} + \Delta_T^{(i)}(\Delta_T^{(i)})^\top) = (\mathrm{Id} - \mathrm{Cor}_T^{(i)})\mathrm{Cov}_T^{(i)} - \delta_T^{(i)}(\Delta_T^{(i)})^\top.$$

Inserting (65) this directly yields

$$B^{(i)} = \mathrm{Cor}_T^{(i)}.$$

Then, using (65) again, we find

$$\Delta_T^{(i)} = -(\mathrm{Id} - \mathrm{Cor}_T^{(i)})^{-1}\delta_T^{(i)}.$$

Thus, we have

Lemma 6.3 *Let $Cov_T^{(i)}$ be positive definite. Then the solution of (64) and (65) satisfies:*

$$\begin{aligned} B^{(i)} &= e^{\tau \hat{F}^{(i)}} = Cor_T^{(i)}, \qquad (66) \\ \Delta_T^{(i)} &= -(\mathrm{Id} - Cor_T^{(i)})^{-1}\delta_T^{(i)}. \qquad (67) \end{aligned}$$

Consequently, whenever $\|Cor_T^{(i)}\| < 1$ we have that $\hat{F}^{(i)}$ is well-defined and its spectrum satisfies $\mathrm{spec}(\hat{F}^{(i)}) \subset \mathbf{C}^-$.

In the case of a single hidden state there are the following direct consequences of this lemma: Whenever we can assume the subsequent convergence for $T \to \infty$ (compare [41, 42], but be aware that one has to expect extremely slow rates of convergence for large values of the lag-time τ):

$$\mathrm{Cov}_T \to \mathrm{Cov}_\infty, \quad \delta_T \to 0, \quad \mathrm{Cor}_T \to \mathrm{Cor}_\infty, \quad \bar{Z}_T \to \bar{Z}_\infty,$$

then (62) and (63) yield that for $T \to \infty$

$$\begin{aligned} \hat{\mu}_T &\to \bar{Z}_\infty \qquad (68) \\ e^{\tau \hat{F}_T} &\to \mathrm{Cor}_\infty. \qquad (69) \end{aligned}$$

If the SDE (56) satisfies appropriate ergodicity assumption and $(Z_t)_{t=1,2,\ldots}$ results from τ-sampling the corresponding solution process, then in probability $\mathrm{Cov}_T \to \mathrm{Cov}_\infty$, $\delta_T \to 0$, and especially

$$\begin{aligned} \bar{Z}_T &\to \mu, \\ \mathrm{Cor}_T &\to \langle (z(\tau) - \mu)(z(0) - \mu)^\top \rangle \cdot \langle (z(0) - \mu)(z(0) - \mu)^\top \rangle^{-1}, \end{aligned}$$

such that for $T \to \infty$

$$\begin{aligned} \hat{\mu}_T^{(i)} &\to \mu \qquad (70) \\ \hat{F}_T &\to F. \qquad (71) \end{aligned}$$

where the last convergence follows from

$$e^{\tau \hat{F}_T} \to \langle (z(\tau) - \mu)(z(0) - \mu)^\top \rangle \cdot \langle (z(0) - \mu)(z(0) - \mu)^\top \rangle^{-1} = e^{\tau F}.$$

Moreover, for several hidden states, if $\mathrm{spec}(\hat{F}^{(i)}) \subset \mathbf{C}^-$ and $\tau \to \infty$ then

$$\begin{aligned} \exp(F^{(i)}\tau) &\to 0, \\ \mu^{(i)} &\to \frac{1}{\sum_{k=1}^{T-1} \nu_{k+1}(i)} \sum_{k=1}^{T-1} \nu_{k+1}(i) Z_{k+1} \end{aligned}$$

and

$$R^{(i)}(\tau) \quad \rightarrow \quad \frac{1}{\sum_{k=1}^{T-1}\nu_{k+1}(i)}\sum_{k=1}^{T-1}\nu_{k+1}(i)(Z_{k+1}-\mu^{(i)})(Z_{k+1}-\mu^{(i)})^{\top}$$

which means that the likelihood function and estimators considered herein converge towards the likelihood and estimators of the HMM coupled to multivariate Gaussian distributions as derived by Liporace *et. al* [43] and corresponding limiting expressions for μ and R become respectively the optimal estimators of expectation value and covarince matrix of the multidimensional Gaussian distribution. This means that for $\tau \rightarrow \infty$ the SDE–dynamics reaches its equilibrium and the conditional probability distribution (15) approaches the multivariate Gaussian distribution unconditioned on the previous observations, i. e. system loses the memory about its previous positions. In this case the expressions above demonstrate that the HMM-SDE algorithm becomes equivalent to the HMM-Gaussian approach from [43].

Finally, we will consider the consequences of these observations concerning the estimator $\hat{\Sigma}\hat{\Sigma}^{\top}$ of the noise intensity matrix for which we have to solve equation (22). Combining (22) with (17) gives a linear matrix equation for the optimal noise intensity matrix estimator

$$e^{-\tau\hat{F}}W(\hat{F}) \quad = \quad \hat{\Sigma}\hat{\Sigma}^{\top}e^{\tau\hat{F}^{\top}} - e^{-\tau\hat{F}}\hat{\Sigma}\hat{\Sigma}^{\top}, \tag{72}$$

where

$$\begin{aligned} W(\hat{F}) &= \left(\frac{1}{T-1}\sum_{k=1}^{T-1}\hat{d}_k\hat{d}_k^{\top}\right)\hat{F}^{\top} + \hat{F}\left(\frac{1}{T-1}\sum_{k=1}^{T-1}\hat{d}_k\hat{d}_k^{\top}\right), \\ \hat{d}_k &= \left(Z_{k+1}-\hat{\mu}-e^{\tau\hat{F}}\left(Z_k-\hat{\mu}\right)\right). \end{aligned} \tag{73}$$

We present the derivation for a single hidden state; for several hidden states the definition of $W(\hat{F})^{(i)}$ for state i just includes the occupation probabilities in a way analogous to what we discussed above.

Under our main assumption $\mathrm{spec}(\hat{F}) \subset \mathbf{C}^{-}$, equation (72) for $\hat{\Sigma}$ has a unique solution (see Appendix B, Lemma 6.2).

With the help of

$$d_k \quad = \quad (Z_{k+1}-\bar{Z}_T) - \mathrm{Cor}_T(Z_k-\bar{Z}_T) + (\mathrm{Id}-\mathrm{Cor}_T)(\bar{Z}_T-\hat{\mu}_T).$$

and our above abbreviations we then get

$$\begin{aligned} \frac{1}{T-1}\sum_{k=1}^{T-1}d_kd_k^{\top} &= \mathrm{Cov}_T - \mathrm{Cor}_T\mathrm{Cov}_T\mathrm{Cor}_T^{\top} - \delta\delta^{\top} \\ &\quad + \frac{1}{T-1}\left((Z_T-\bar{Z}_T)(Z_T-\bar{Z}_T)^{\top} - (Z_1-\bar{Z}_T)(Z_1-\bar{Z}_T)^{\top}\right). \end{aligned}$$

Since $\mathrm{Cor}_T = e^{\tau\hat{F}_T}$ this can be decomposed into

$$\begin{aligned} \frac{1}{T-1}\sum_{k=1}^{T-1}d_kd_k^{\top} &= D_T + f(Z_1,Z_T,\bar{Z}) \\ D_T &= \mathrm{Cov}_T - e^{\tau\hat{F}_T}\mathrm{Cov}_T e^{\tau\hat{F}_T^{\top}}, \\ f(Z_1,Z_T,\bar{Z}) &= -\delta\delta^{\top} \\ &\quad + \frac{1}{T-1}\left((Z_T-\bar{Z}_T)(Z_T-\bar{Z}_T)^{\top} - (Z_1-\bar{Z}_T)(Z_1-\bar{Z}_T)^{\top}\right). \end{aligned} \tag{74}$$

Inserting these results into (73) lets us infer that

$$\begin{aligned} e^{\tau\hat{F}_T}\hat{\Sigma}_T\hat{\Sigma}_T^\top e^{\tau\hat{F}_T^\top} - \hat{\Sigma}_T\hat{\Sigma}_T^\top &= D_T\hat{F}_T^\top + \hat{F}_T D_T + \eta_T\hat{F}_T^\top + \hat{F}_T\eta_T, \\ &= \mathrm{Cov}_T\hat{F}_T^\top + \hat{F}_T\mathrm{Cov}_T \\ &\quad -e^{\tau\hat{F}_T}\left(\mathrm{Cov}_T\hat{F}_T^\top + \hat{F}_T\mathrm{Cov}_T\right)e^{\tau\hat{F}_T^\top} \\ &\quad +f(Z_1,Z_T,\bar{Z})\hat{F}_T^\top + \hat{F}_T f(Z_1,Z_T,\bar{Z}). \end{aligned}$$

This shows that the estimates satisfy the *approximate fluctuation-dissipation result*

$$-\hat{\Sigma}_T\hat{\Sigma}_T^\top = (\mathrm{Cov}_T + E_T)\hat{F}_T^\top + \hat{F}_T(\mathrm{Cov}_T + E_T), \tag{75}$$

with some symmetric matrix E_T that is the solution of

$$f(Z_1,Z_T,\bar{Z}) = E_T - e^{\tau\hat{F}_T}E_T e^{\tau\hat{F}_T^\top}.$$

Again this solution is unique if $\mathrm{spec}(\hat{F}) \subset \mathbf{C}^-$.

We summarize:

Lemma 6.4 *Suppose the correlation matrix $Cor_T = \exp(\tau\hat{F})$ exists and $\|Cor_T\| < 1$ be satisfied. Then the optimal estimator of the noise intensity matrix has the form:*

$$-\hat{\Sigma}_T\hat{\Sigma}_T^\top = (Cov_T + E_T)\hat{F}_T^\top + \hat{F}_T(Cov_T + E_T),$$

with a symmetric matrix E_T that is the unique solution of

$$f(Z_1,Z_T,\bar{Z}) = E_T - e^{\tau\hat{F}_T}E_T e^{\tau\hat{F}_T^\top},$$

with f being defined in (74).

The following is worth emphasizing: the fluctuation-dissipation theorem –due to the above results– can be written in the following general form

$$\mathrm{Cov}_\infty F^\top + F\mathrm{Cov}_\infty = -\Sigma\Sigma^\top. \tag{76}$$

where Cov_∞ is the analytical covariance matrix of the invariance measure of the model (F,Σ,μ). Eq. (76) indeed justifies to call (75) an *approximate fluctuation-dissipation result.*

In addition we observe that E_T will vanishes for $T \to \infty$ along an ergodic, infinitely long time series. Thus, the exact fluctuation-dissipation result is recovered in the limit. In this case, we finally get consistency for the noise intensity estimator

$$\hat{\Sigma}_T\hat{\Sigma}_T^\top \to \Sigma\Sigma^\top.$$

Error Estimation. Since the estimators result from a likelihood optimization, we can always estimate their statistical uncertainty by computing the corresponding variance of the likelihood understood as a probability distribution in parameter space. This can be done via Monte Carlo, Gibbs, or Langevin samplers that at most need the first derivative of the likelihood wrt. the parameters (which is what we already computed above). In the subsequent example we will compute the variance of the likelihood relative to a flat prior (i.e., prior distribution is uniform/Lebesgue measure). If non-uniform prior information is available (e.g., local stationarity, or the fuctuation-dissipation relation holds) then this should be considered by taking an appropriate non-uniform prior. Alternatively, one could also apply the Fisher information matrix, cf. [44]. These options are not further discussed herein; they are topics of further investigation.

In order to illustrate the dependence of the error/uncertainty of the optimal estimators, we will now give a scalar example for one hidden state (thus the parameter space is three-dimension consisting of (B,R,μ) or (F,Σ,μ)).

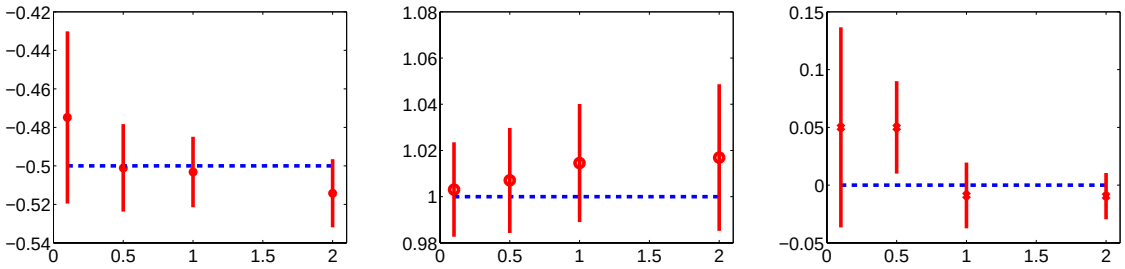

Fig. 12 Dependence of estimators (indicated by circles) and their standard variation (indicator by vertical bars) on τ for the scalar test case described in the text. For all τ the number of instances are fixed ($N = 5.000$). From left to right: $\hat{D}$ and $\Delta\hat{D}$, $\hat{\Sigma}$ and $\Delta\hat{\Sigma}$, and $\hat{\mu}/\Delta\hat{\mu}$. The blue dashed lines indicate the values for D, Σ, and μ in the original SDE.

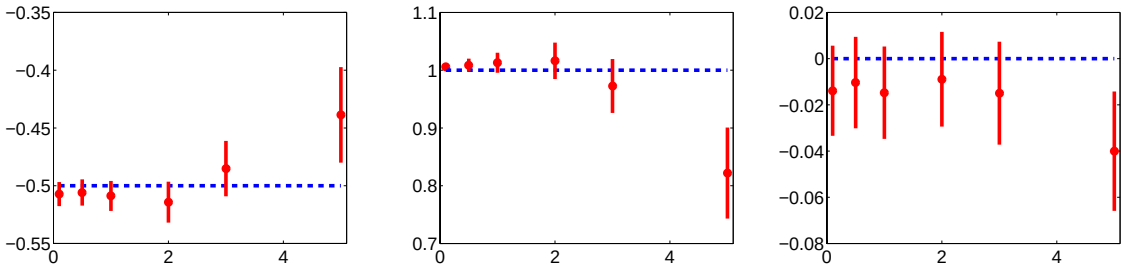

Fig. 13 Dependence of estimators (indicated by circles) and their standard variation (indicator by vertical bars) on τ for the scalar test case described in the text. For all τ all available instances are taken (from $N = 2.000$ for $\tau = 5$ up to $N = 100.000$ for $\tau = 0.1$). From left to right: $\hat{D}$ and $\Delta\hat{D}$, $\hat{\Sigma}$ and $\Delta\hat{\Sigma}$, and $\hat{\mu}/\Delta\hat{\mu}$. The blue dashed lines indicate the values for D, Σ, and μ in the original SDE.

To this end, we first computed a realization of the dynamics given by $\dot{q} = -Dq + \Sigma\dot{W}$ with one-dimensional q, $D = -0.5$ and $\Sigma = 1$. The realization has been computed by means of the Euler–Maruyama discretization with timestep $dt = 0.001$ and total length 10.000. By subsampling the resulting time series with lag times $\tau = m \cdot 0.1$ for $m = 1, \ldots, 50$ we produced 50 different time series with respective total lengths ranging between 2.000 and 100.000 instances. When computing an estimator E and its standard deviation ΔE (the square root of its variance), we have to understand both quantities as variables of the lag time, τ, and the number of instances, N, that have been taken into account: $E = E(\tau, N)$, $\Delta E = \Delta E(\tau, N)$.

Figure 12 illustrates the dependence of the estimators on τ when the number of instances $N = 5000$ is *fixed* (that is, for $\tau = 0.1$ only the first 5.000 instances of the available time series of length 100.000 have been considered; for $\tau = 2$ all available 5.000 instances). Figure 12 allows to observe that the standard deviations of the estimators do not increase with τ (rather decrease) as long as the same number of observations is available. As observed in standard theory of SDE parameter fitting, the standard deviation of the drift parameter $\hat{D}$ increases for $\tau \to 0$ and N fixed but rather decreases for $\hat{\Sigma}$.

In real world cases, the situation will mostly be characterized by the availability of a time series of given length N_{tot} with lag time $\tau_{\min}$ any subsampling of which with some other lag time $\tau = j\tau_{\min}$, $j \in \mathbf{N}$, will produce a time series whose number of instances $N = N_{tot}/j = N_{tot} \cdot \tau_{min}/\tau$ depends on τ. This scenario and the resulting dependence of the standard deviation on τ is illustrated in Fig. 13 which illustrates the obvious consequence: the standard deviation increases with decreasing number of instances and increasing τ.

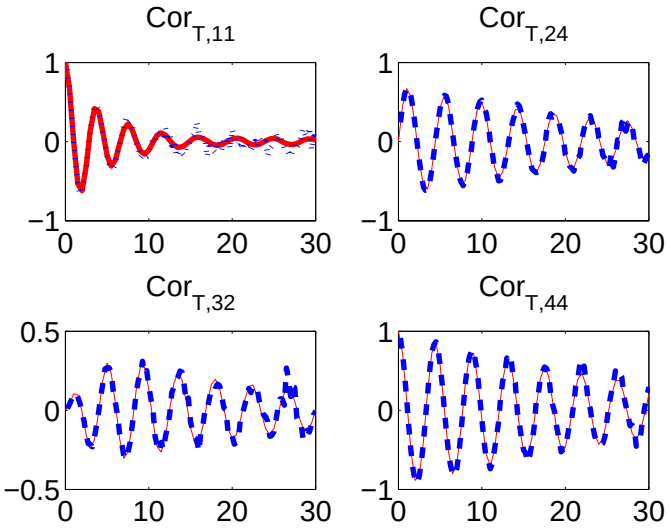

Fig. 14 Dependence of the matrix Cor_T (blue dashed lines) on the lag time τ as computed from sub-sampling a given time series of fixed length with lag time τ and comparison with $\exp(\tau F)$ (red lines). Top panel: $\mathrm{Cor}_{T,11}$ and $\exp(\tau F)_{11}$ (left) and $\mathrm{Cor}_{T,24}$ / $\exp(\tau F)_{24}$ (right). Lower panel: $\mathrm{Cor}_{T,32}$ and $\exp(\tau F)_{32}$ (left) and $\mathrm{Cor}_{T,44}$ and $\exp(\tau F)_{44}$ (right).

Appendix D: Determination of F

Let us now consider the linear, two-dimensional Langevin dynamics $M\ddot{q} = -Dq - \gamma\dot{q} + \sigma\dot{W}$ with

$$M = \mathrm{Id}, D = \begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}, \quad \gamma = \begin{pmatrix} 0.5 & 0.2 \\ 0.2 & 0.1 \end{pmatrix}, \quad \sigma = 0.3\mathrm{Id}, \quad \mu = (0,0,0,0)^\top.$$

We used an appropriate second order discretization in time [45] and produced a time series $\{Z(t_k)\}_{k=1,\ldots,N}$ consisting of positions and momenta by simulation of the dynamics with time step $\Delta t = 0.001$ from time $t = 0$ to $t = 5000$, i.e., with $N = 5 \cdot 10^6$ integration steps. By sub-sampling of this time series with different lag times τ, in this case for $\tau = m \cdot 0.25$, $m = 1, \ldots, 120$, we get 120 time series with different lag times τ and of different total lengths $T = 5000/\tau$. For each of these we computed $\bar{Z}_T$ and Cor_T as defined above. In Fig. 14 we compare the results for Cor_T with its limit value $\exp(\tau F)$; we observe that the agreement is remarkably good over the entire range of τ-values considered.

However, when computing the estimate $\hat{F} = \log(\mathrm{Cor}_T)/\tau$ for different lag times τ using the matrix logarithm as it is implemented in MATLAB, for example, we find the artifacts shown in Fig. 15: For very small τ the results are satisfactory. However, for larger τ we observe periodically repeated bursts of deviation between $\hat{F}$ and F with the effect that $\hat{F}$ is totally misleading even for medium size τ.

The reason for this lies in the properties of the principal matrix logarithm [46]: the eigenvalues of $\exp(\tau F)$ have the form $\exp(\tau(a + ib))$ with real $a < 0$ and b such that, as functions of τ, their graphs are spirals around the origin in the complex plane. However, the principal matrix logarithm seeks to compute a matrix $\hat{F}$ with eigenvalues $\tau a + i\xi$ where $\xi \in [0, 2\pi]$ such that $\xi + 2m\pi = \tau b$ for some integer m. Thus $\exp(\tau\hat{F})$ is a good approximation of $\exp(\tau F)$ but $\hat{F}$ may be far from F since the matrix function log cannot have information about the right Riemann plane as long as Cor_T is given for one value of T, or τ only, and thus chooses the principal one.

Let us address three options for dealing with this problem:

(I) We can compute $\hat{F}$ based on "small enough" values of τ; however, small enough τ may not be available and/or we can never know whether specific values of τ are small enough or not.

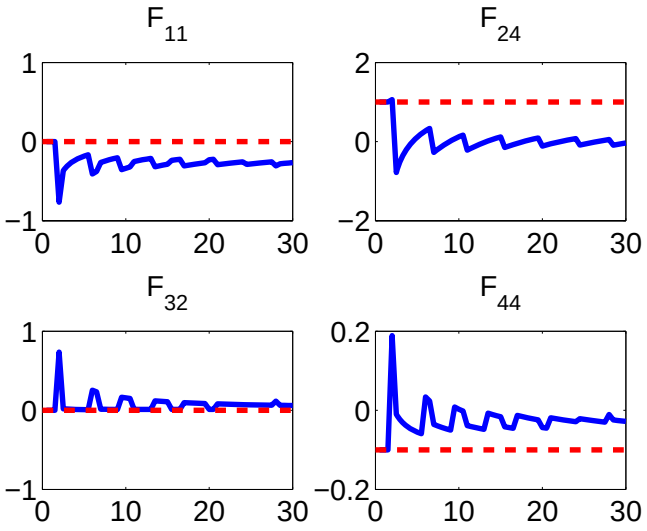

Fig. 15 Estimator $\hat{F}$ (blue solid lines) computed via the matrix logarithm from Cor_T for different as determined from sub-sampling a given time series of fixed length with lag time τ and comparison with the original matrix F (red dashed lines). Top panel: $\hat{F}_{11}$ and F_{11} (left) and $\hat{F}_{24}$ / F_{24} (right). Lower panel: $\hat{F}_{32}$ and F_{32} (left) and $\hat{F}_{44}$ and F_{44} (right).

(II) We can approximate F via its resolvent as already discussed in Sec. 4.1: Therefore the integral expression for the resolvent

$$R(s,\tau_*) = \int_{\tau_*}^{\infty} \exp(-s(\tau-\tau_*))\exp(\tau\hat{F})d\tau,$$

needs to be discretized by an appropriate quadrature rule. Whenever we explicitly have $B(\tau_k) = \exp(\tau_k\hat{F})$ for a sequence of τ's, e.g., $\tau_k = \tau_* + k\Delta\tau$, $k = 0,\ldots,L$, and some appropriate $s > 0$ has been selected then we can use the interpolating matrix function for $\tau \in [\tau_k, \tau_{k+1}]$:

$$B(\tau) = B(\tau_k) + \frac{\tau-\tau_k}{\Delta\tau}(B(\tau_{k+1}) - B(\tau_k)),$$

to approximate the resolvent by

$$\begin{aligned}\tilde{R}(s,\tau_*) &= \int_{\tau_*}^{\infty} \exp(-s(\tau-\tau_*))B(\tau)d\tau,\\ &= \sum_{k=0}^{L-1}\int_{\tau_k}^{\tau_{k+1}} \exp(-s(\tau-\tau_*))\Big(B(\tau_k) + \frac{\tau-\tau_k}{\Delta\tau}(B(\tau_{k+1})-B(\tau_k))\Big)d\tau\\ &\quad + \int_{\tau_L}^{\infty}\exp(-s(\tau-\tau_*))B(\tau_L)d\tau,\end{aligned}$$

where the integrals in the last expression can all be evaluated explicitly. Having computed $\tilde{R}$ we then set

$$\hat{F} = s\ -\ B(\tau_*)\tilde{R}(s,\tau_*)^{-1}.$$

For the case considered above we compute ($s = 1/5, \tau_* = 0$):

$$\hat{F} = \begin{pmatrix} -0.0373 & -0.0001 & 1.0052 & -0.0026\\ -0.0032 & 0.0220 & -0.0095 & 1.0207\\ -3.0351 & -0.0068 & -0.4897 & -0.1891\\ 0.0038 & -1.9992 & -0.2019 & -0.1307\end{pmatrix},$$

which is a reasonably good approximation of F with an error of $\|F - \hat{F}\| = 0.05$. For $s = 1/5, \tau_* = 4$ we get $\|F - \hat{F}\| = 0.15$. However, in many cases approximation via the resolvent is unreliable and can even be totally misleading. The reason for this lies in the effect of the noise and different scales on the quality of the numerical quadrature.

(III) Option (I) will not cure our problem if τ_0 is too large or the length of the observation sequence is too short. In the first case the numerical calculation of the resolvent integral in (33) gets more and more inaccurate when the discretization step $\Delta\tau$ gets longer and the total number of available discretization points for quadrature of the integral decreases. In the second case the variance of the error in the $\exp(\tau\hat{F})$ estimation increases when the length of the observation decreases which leads to the growing "noise" in the sub–integral function. Both of the problems can emerge in the analysis of realistic processes if the observation is available for a rather "short" total time with "long" lags between the single observations. In this case an alternative algorithm is needed where neither τ_0 nor $\Delta\tau$ are assumed to be particularly small.

In order to handle such scenarios we suggest to use an alternative method based on the spectral structure of the matrices $\exp(\tau\hat{F})$, as already discussed in Sec. 4.1. For increasing τ, the real-valued eigenvalues of this matrix are exponentially decaying whereas its complex eigenvalues exhibit a mixture of exponential decay with rotation in the complex plane. The rotation frequency is proportional to the imaginary part of the underlying matrix $\hat{F}$.

The basic idea of the alternative algorithm is as follows: Calculate the spectrum of $\exp(\tau\hat{F})$ for different values of τ, filter out the imaginary eigenvalues and determine the corresponding frequencies of rotation; then we can directly get access to the eigenvalues of $\hat{F}$. Due to the fact that the eigenvectors of $\exp(\tau\hat{F})$ are (approximately) identical with the eigenvectors of F, this will open the way to calculate matrix $\hat{F}$ itself. In order to realize this idea we need an algorithm that approximates the sequence of eigenvalues of the matrices $\exp(\tau\hat{F})$ with spirals around the origin of the complex plane (i.e., we need an optimal approximation algorithm). The main problem connected with such kind of approach is that if the matrix F has more then one pair of complex eigenvalues and it is not clear which eigenvalues corresponds to which spirals since the ordering of the spectrum of $\exp(\tau\hat{F}$ may be changing with τ. We suggest to solve this problem in two steps: (i) Divide the complex eigenvalues of $\exp(\tau\hat{F})$ with their absolute values in order to get rid of the exponential decay (we denote the resulting values with $\bar{\lambda}_i^t, t = \tau_1, \ldots, \tau_L$) and (ii) fit the resulting points in the three dimensional space (spanned by the real and imaginary parts of the normalized eigenvalues and the time τ) through J spirals (where J is the number of complex eigenvalues of F). The last step can be solved by minimization of the following functional:

$$\mathcal{L}_\omega \quad = \quad \sum_{k=1}^{J}\sum_{i=1}^{J}\sum_{l=1}^{L} \nu_{ki}(l)\left((\Im\bar{\lambda}_i^{\tau_l} - \sin(\tau_l\omega_k))^2 + (\Re\bar{\lambda}_i^{\tau_l} - \cos(\tau_l\omega_k))^2\right), \tag{77}$$

where $\nu_{ki}(t)$ is the probability for the complex eigenvalue i for $\tau = t\tau_0$ to be described by the spiral number k with rotation frequency ω_k, and $\Re$ and $\Im$ denote real and imaginary parts. Assuming for a moment the probabilities $\nu_{ki}(t)$ to be fixed and known, we can calculate the optimal spiral parameters by differentiating the functional (77) wrt. ω_k and setting the resulting derivative to zero. We get:

$$\sum_{i=1}^{J}\sum_{l=1}^{L} \nu_{ki}(l)\tau_l\left(\Re\bar{\lambda}_i^{\tau_l}\sin(\tau_l\omega_k) - \Im\bar{\lambda}_i^{\tau_l}\cos(\tau_l\omega_k)\right) \quad = \quad 0 \tag{78}$$

One can use a Newton–method to calculate ω_k for given $\nu_{ki}(l)$ and $\bar{\lambda}$. In order to define ω_k *and* the probabilities $\nu_{ki}(l)$ we can once again apply the standard form of the EM–algorithm as described before. As a starting point for optimization one can take the frequencies derived from the estimation of $\hat{F}$ with help of the resolvent as described in case (II).

To illustrate the proposed method we take the half of the time series used in the previous examples (i.e., we take the time series from $t = 0$ to $t = 2500$; this should increase the variance of the $\exp(\tau\hat{F})$ estimation error) and calculate the estimation of $\hat{F}$ based on 6 different values of $\exp(\tau\hat{F})$ with τ being

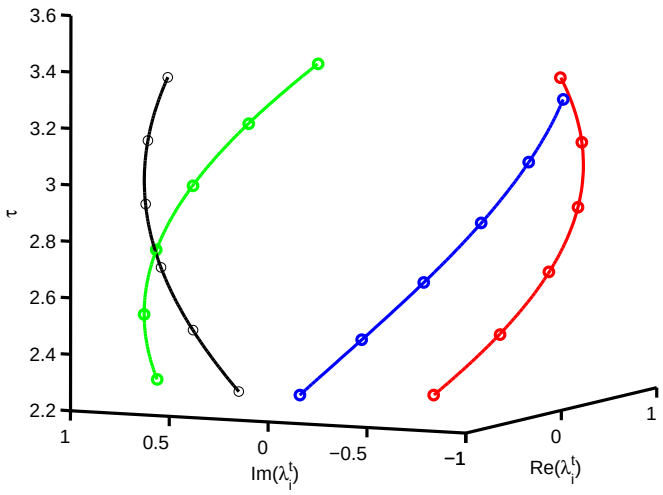

Fig. 16 Optimal fit of the normalized eigenvalues of the matrix $\mathrm{Cor}_{N/\tau}$ with four spirals as resulting from the application of EM algorithm for $J = 4$ ($\tau = [2.3, 2.5, 2.7, 2.9, 3.1, 3.3]$).

$[2.3, 2.5, 2.7, 2.9, 3.1, 3.3]$. The resulting matrix $\hat{F}$ as derived from the resolvent calculated from the third-order spline interpolated sub–integral function in (33) is

$$\hat{F} = \begin{pmatrix} -1.1229 & -0.0012 & 0.9292 & 0.0311 \\ -0.0250 & 1.2749 & 0.0092 & 0.8536 \\ -2.8488 & -0.0568 & -1.6451 & -0.2066 \\ 0.0135 & -1.7121 & -0.1444 & -1.3434 \end{pmatrix},$$

As it can be seen, the result is wrong. Applying the EM–based minimization of the functional (77) to the same data we get:

$$\hat{F} = \begin{pmatrix} 0.0037 & 0.0010 & 1.0009 & 0.0010 \\ -0.0113 & 0.0053 & -0.0128 & 0.9966 \\ -3.0656 & -0.0043 & -0.5540 & -0.2174 \\ 0.0245 & -1.9992 & -0.1733 & -0.0933 \end{pmatrix},$$

which is a reasonable approximation of F with an error of $\|F - \hat{F}\| = 0.08$. The identified spirals are shown in Fig. 16.

Appendix E: Proofs of Identifiability Lemmas

Proof of Lemma 4.1

Let $(\gamma, D, \sigma, \mu_{eq}) \in \mathbf{OL}(n)$ and an arbitrary inverse temperature $\beta > 0$ be given. Every other element of $[(\gamma, D, \sigma, \mu_{eq})]_\sim$ has the form

$$(\mathcal{A}\gamma, \mathcal{A}D, \mathcal{A}\sigma, \mu_{eq}) \in \mathbf{OL}(n)$$

with some $\mathcal{A} \in \mathrm{Reg}_n(\mathbf{R})$. The fluctuation-dissipation relation is valid for the respective model iff

$$\mathcal{A}\gamma + \gamma^\top \mathcal{A}^\top = \beta \mathcal{A}\sigma\sigma^\top \mathcal{A}^\top.$$

We have to show that there is exactly one such $\mathcal{A}$ under the additional condition that $\mathcal{A}D$ be symmetric. Multiplying the above equation by $\gamma^{-1}\mathcal{A}^{-1}$ from the left and $(\mathcal{A}^\top)^{-1}(\gamma^\top)^{-1}$ from the right gives

$$(\mathcal{A}^\top)^{-1}(\gamma^\top)^{-1} + \gamma^{-1}\mathcal{A}^{-1} = \beta\Sigma\Sigma^\top,$$

where $\Sigma = \gamma^{-1}\sigma$. By setting $B = \mathcal{A}D$ and $F = -\gamma^{-1}D$ this yields

$$(B^{-1})^\top F^\top + FB^{-1} = -\beta\Sigma\Sigma^\top,$$

which we can further simplify by exploiting the constraint ($B = B^\top$):

$$B^{-1}F^\top + FB^{-1} = -\beta\Sigma\Sigma^\top. \tag{79}$$

This is a Sylvester equation for B^{-1} with F satisfying $\mathrm{spec}(F) \subset \mathbf{C}^{-1}$. Lemma 6.2 (see Appendix B) states that equations of the form of (79) have unique symmetric solutions. In addition it is easy to see that B^{-1} is regular, such that we have just proved that there is a unique symmetric $B = \mathcal{A}D$ that satisfies equation (79) and thus there is a unique $\mathcal{A}$ such that the fluctuation-dissipation relation is valid and $\mathcal{A}D$ is symmetric.

Proof of Lemma 4.2

We will follow the same line of arguments as in the proof of Lemma 4.1. Let $(M, \gamma, D, \sigma, \mu_{eq}) \in \mathbf{FL}(n)$ and an arbitrary inverse temperature $\beta > 0$ be given. The fluctuation-dissipation relation is valid for the model $(\mathcal{A}M, \mathcal{A}\gamma, \mathcal{A}D, \mathcal{A}\sigma, \mu_{eq})$ from the respective equivalence class iff

$$\mathcal{A}\gamma + \gamma^\top\mathcal{A}^\top = \beta\mathcal{A}\sigma\sigma^\top\mathcal{A}^\top,$$

for $\mathcal{A} \in \mathrm{Reg}_n(\mathbf{R})$. With $A = \mathcal{A}M$ this yields

$$M^{-1}\gamma(A^{-1})^\top + A^{-1}\gamma^\top(M^{-1})^\top = \beta M^{-1}\sigma\sigma^\top(M^{-1})^\top. \tag{80}$$

By setting

$$B = \begin{pmatrix} \mathcal{A}M & 0 \\ 0 & \mathcal{A}D \end{pmatrix}.$$

together with the required symmetry of A and $\mathcal{A}D$, and equation (80) we find

$$(B^{-1})^\top F^\top + FB^{-1} = -\beta\Sigma\Sigma^\top,$$

which we can further simplify by exploiting the constraint ($B = B^\top$):

$$B^{-1}F^\top + FB^{-1} = -\beta\Sigma\Sigma^\top. \tag{81}$$

As above this is a Sylvester equation for B^{-1} with F satisfying $\mathrm{spec}(F) \subset \mathbf{C}^{-1}$ of which Lemma 6.2 (see Appendix B) states the existence of a unique symmetric solution.

Acknowledgement

We are indebted to John Maddocks for his comments on the role of the mass matrix and the fluctuation-dissipation relation, and to Andrew M. Stuart for insisting on the identifiability problem. We are also very grateful to Jeremy Smith and Frank Noe for the possibility to use the 12-Alanine MD-simulation data.

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] F. A. Bornemann and Ch. Schütte. Homogenization of Hamiltonian systems with a strong constraining potential. *Physica D*, 102(1-2):57–77, 1997.

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

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

[7] Christof Schütte, Alexander Fischer, Wilhelm Huisinga, and Peter Deuflhard. A direct approach to conformational dynamics based on hybrid Monte Carlo. *J. Comput. Phys.*, 151:146–168, 1999.

[8] Christof Schütte and Wilhelm Huisinga. Mathematical analysis and simulation of conformational dynamics of biomolecules. In P. G. Ciaret and J.-L. Lions, editors, *Handbook of Numerical Analysis*, volume Computational Chemistry. North–Holland, 2002. in preparation.

[9] Peter Deuflhard, Wilhelm Huisinga, Alexander Fischer, and Christof Schütte. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Lin. Alg. Appl.*, 315:39–59, 2000.

[10] I. Horenko, E. Dittmer, A. Fischer, and Ch. Schuette. Automated model reduction for complex systems exhibiting metastability. *Mult. Mod. Sim.*, 5:802–827, 2006.

[11] I. Horenko, E. Dittmer, and Ch. Schuette. Reduced stochastic models for complex molecular systems. *Comp. Vis. Sci.*, 9:89–102, 2005.

[12] H. Grubmueller and P. Tavan. Molecular dynamics of conformational substates for a simplified protein molecule. *J. Chem. Phys.*, 101:5047–5057, 1994.

[13] T. Schlick. *Molecular Modeling Simulation- An Interdisciplinary Guide*. Springer, 2000.

[14] G.F. Schroeder and H. Grubmüller. Maximum likelihood trajectories from single molecule fluorescence resonance energy transfer experiments. *J. Chem. Phys.*, 119:9920–9924, 2003.

[15] J.L. Barber. *Application of Optimal Prediction to Molecular Dynamics*. PhD Thesis, UC Berkeley, University of California, Berkeley, 2004.

[16] E. Cancès, F. Legoll, and G. Stoltz. Theoretical and numerical comparison of some sampling methods for molecular dynamics. *IMA Preprint Series*, 2040:1–35, 2005.

[17] A.J. Chorin, O.H. Hald, and R. Kupferman. Prediction from partial data, renormalization, and averaging. *J. Sci. Comp.*, 2005. `http://dx.doi.org/10.1007/s10915-006-9089-5`.

[18] O. Lange. *Collective Langevin Dynamics of Conformational Motions in Proteins*. PhD thesis, University of Goettingen, Abt. Helmut Grubmueller, 2005.

[19] O. Lange and H. Grubmueller. Collective Langevine dynamics of conformational motions in proteins. *J. Chem. Phys*, 124:2149, 2006.

[20] X. Mao and C. Yuan. *Stochastic Differential Equations with Markovian switching*. Imperial College Press, 2006.

[21] P. G. Blackwell. Bayesian inference for Markov processes with diffusion and discrete components. *Biometrica*, 90(3):613–627, 2003.

[22] A. Beskos, O. Papaspiliopoulos, G. Roberts, and P. Fearnhead. Exact and computational efficient likelihood-based estimation for discretely observed diffusion processes. *Journal of Royal Statistical Society B*, 68:1–29, 2006.

[23] I. Pokern, A. Stuart, and P. Wiberg. Parameter estimation for partially observed hypoelliptic diffusion. *preprint, University of Warwick*, 2006. (available via www.maths.warwick.ac.uk/ pokern).

[24] C. Penland. A stochastic model of indopacific sea surface temperature anomalies. *Physica D*, 98:534–558, 1996.

[25] L. E. Baum, T. Petrie, G. Soules, and N. Weiss. A maximization technique occuring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.*, 41(1):164–171, 1970.

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

[27] 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.

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

[29] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B*, 39(1):1–38, 1977.
[30] A.J. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Informat. Theory*, 13:260–269, 1967.
[31] L.R. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77(2):257–286, 1989.
[32] 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.
[33] 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.
[34] 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.
[35] P. Deuflhard and M. Weber. Robust Perron cluster analysis in conformation dynamics. *Lin. Alg. Appl.*, 398c:161–184, 2005.
[36] I. Horenko, C. Hartmann, Ch. Schuette, and F. Noe. Data-based parameter estimation of generalized multidimensional Langevin processes. *Phys. Rev. E*, 01:016706–, 2007.
[37] A. Stuart and G. Pavliotis. Parameter estimation for multiscale diffusions. *preprint, University of Warwick*, 2006. (available via www.maths.warwik.ac.uk/ stuart).
[38] P. Benner, E.S. Quintana-Orti, and G. Quintana-Orti. Solving stable Sylvester equations via rational iterative schemes. *to appear in J. Sci. Comp.*, 2006.
[39] J.H. Maddocks and M.L. Overton. Stability theory for dissipatively perturbed Hamiltonian systems. *Comm. Pure Appl. Math.*, 48:583–610, 1995.
[40] J. Maddocks, C. Hartmann, and C. Schütte. Perturbation results for the linear Langevin equation. *in preparation*, 2006. (to be available via www.biocomputing.mi.fu-berlin.de, Jan. 07).
[41] A. Le Breton and M. Musiela. Some parameter estimation problems for hypoelliptic homogeneous gaussian diffusion. *Seq. Meth. in Stat.*, 22:337–356, 1985.
[42] D. Florens-Zmirou. Approximate discrete-time schemes for statistics of diffusion processes. *Statistics*, 20:547–557, 1985.
[43] L.A. Liporace. Maximum likelihood estimation for multivariate observations of Markov sources, 1989.
[44] M. Schervish. *Theory of Statistics, Sec. 2.3.1.* Springer, 1995.
[45] B. Øksendal. *Stochastic differential equations: an introduction with applications.* Springer, Berlin, 2003.
[46] N.J. Higham. Evaluating Pade approximants of the matrix logarithm. *SIAM J. Matrix Anal. Appl.*, 22(4):1126–1135, 2001.