Generator Estimation of Markov Jump Processes based on incomplete observations non-equidistant in time[*]

Philipp Metzner, Illia Horenko, Christof Schütte

Institute of Mathematics II, Free University Berlin,

Arnimallee 2-6, D-14195 Berlin, Germany

(Dated: October 29, 2007)

Abstract

Markov jump processes can be used to model the effective dynamics of observables in applications ranging from molecular dynamics to finance. In this paper we present a novel method which allows the inverse modeling of Markov jump processes based on incomplete observations in time: We consider the case of a given time series of the discretely observed jump process. We show how to compute efficiently the maximum likelihood estimator of its infinitesimal generator and demonstrate in detail that the method allows to handle observations non-equidistant in time. The new method is based on the work of Bladt and Sørensen (J. R. Statist. Soc. B, 39, 2005) but scales much more favorably than it with the length of the time series and the dimension/size of the state space of the jump process. We illustrate its performance on a toy problem as well as on data arising from simulations of biochemical kinetics of a genetic toggle switch.

Keywords: Markov jump process, generator estimation, inverse problem, maximum likelihood estimation, genetic toggle switch

[*] Supported by the DFG Research Center MATHEON "Mathematics for Key Technologies" (FZT86) in Berlin.

I. INTRODUCTION

In many application areas it is of interest to derive a reduced model of the effective dynamics of observables by a finite dimensional Markov process. In this paper we study the inverse problem of fitting Markov jump processes to discrete time series in situations where the time lags between consecutive observations are not necessarily equidistant. Such a situation arises naturally in a number of application, e.g. in finance or (bio-)chemical kinetics. In the case of equidistant time lags, several approaches can be found in the literature, e.g. [1–5]. In [6], we summarize, compare and discuss these approaches in detail. Furthermore, we therein present an enhanced version of the approach presented in [1, 5]: the enhanced maximum likelihood estimation method (enhanced MLE-method, see [6–8] for different variants) which drastically increase the efficiency of the method. However, the discussion in [6] is limited to time series with equidistant steps in time. In this paper, we discuss the generalization of the enhanced MLE-method for observations non-equidistant in time.

The likelihood approach is designed for the analysis of time series generated by first order Markov processes. But many physical processes exhibit long-term memory effects and, hence, it is not clear in advance if the time series under consideration is Markovian [9–13]. One option to handle the non-Markovian case within the likelihood framework is to consider a new process $Z(t) = (X(t), X(t+\tau), ..., X(t+(m-1)\tau))$, $\tau > 0$ where the order $m > 1$ respresenting the memory depth $m\tau$ of the original process $X(t)$ is known or has to be estimated from the time series. Then the new process is first order Markovian and the associated time series can be investigated with the MLE-method. However, this option is restricted to the case of equidistant observation times. There are some recent developments indicating that this restriction can be overcome, but their application is limited to time series with extremely short observation time lags [14].

The paper is organized as follows. After introducing some notation in section II, we revisit the likelihood approach in section II A. The main result of this paper – the derivation of the enhanced MLE-method for observations non-equidistant in time and the resulting algorithm – is presented in section III. Finally, in section IV we illustrate the performance of the new method in application to a small toy example and to data arising from simulations of the biochemical kinetics of a genetic toggle switch. The paper ends with concluding remarks in section V.

II. CONCEPTUAL FRAMEWORK AND NOTATION

Let us first consider a simple introductory example. Imagine an organism with two states: healthy (state 1) and sick (state 2), i.e., we have $d = 2$ states. Healthy individuals may stay healthy or get sick; sick ones may stay sick or get healthy. The rate of getting sick if healthy is α, say, while the rate of getting healthy if sick is β. In a Markov model this system is characterized by a 2×2 rate matrix

$$L = \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix},$$

where the entry L_{ij} is the transition rate from state i into state j, and where the sum of the transition rates in each row is 0. If we wait for a period t starting in 0, then the transition probabilities $p_{ij}(t)$ from state i to state j during this period are given by the entries of the transition matrix $P(t) = \exp(tL)$. Asymptotically, that is for $t \to \infty$, the process converges to a stationary distribution π (corresponding to the left eigenvector of L associated to eigenvalue 0) with a proportion $\pi_2 = \alpha/(\alpha + \beta)$ sick and $\pi_1 = \beta/(\alpha + \beta)$ healthy.

In order to formalize this, let $\{X(t), t \geq 0\}$ be a continuous-time homogeneous Markov jump process on a finite state space $S \cong \{1, \ldots, d\}$. The transition matrix of $\{X(t), t \geq 0\}$ is the time-dependent matrix

$$P(t) = \big(p_{ij}(t)\big)_{i,j} \in \mathbb{R}^{d \times d}, \qquad p_{ij}(t) = \mathbb{P}(X(t) = j \mid X(0) = i)$$

containing the transition probabilities $p_{ij}(t)$. If the limit

$$L = \lim_{t \to 0} \frac{P(t) - Id}{t}$$

exists, then the transition matrix can be expressed as the matrix exponential

$$P(t) = \exp(tL)$$

and L is called the *infinitesimal generator or rate matrix* of the Markov process $\{X(t), t \geq 0\}$. A matrix $L \in \mathbb{R}^{d \times d}$ generates a continuous-time Markov process if and only if all off-diagonal entries are nonnegative and the sum over each row equals zero, and the set of all generators will be denoted by

$$\mathfrak{G} = \left\{ L = (l_{ij})_{i,j} \in \mathbb{R}^{d\times d} : l_{ij} \geq 0 \quad \text{for all } i \neq j, \quad l_{ii} = -\sum_{j\neq i} l_{ij} \right\}. \tag{II.1}$$

A stationary probability distribution $\pi = (\pi_1, \dots, \pi_d)^T$ of a Markov process $X(t)$ satisfies the global balance equation ([15], Sect. 8.3.2)

$$0 = \pi^T L = L^T \pi, \tag{II.2}$$

or written in expanded form,

$$-\pi_i l_{ii} = \sum_{k\neq i} \pi_k l_{ki}, \quad i = 1, \dots, d.$$

In the following, we denote an incomplete observation of a Markov jump process $X(t)$ by

$$Y = \{y_0 = X(t_0), \dots, y_N = X(t_N)\}, \; t_0 < t_1 < \dots < t_N$$

and the observation time lags by $\tau_k = t_{k+1} - t_k$.

Remark 1. Continuous-time Markov jump processes are quite prevalently used in physics, chemistry, or biology. Examples are spin system or lattice gas dynamics, Master equations in systems biology (like the system discussed in Sec. IV B) or polymerization modelling [16–18], or birth-death models on networks, e.g., in bioinformatics [19, 20].

A. Likelihood approach revisited

In the likelihood approach, introduced in [1] and re-invented by Bladt and Sørensen in [5], a generator $\tilde{L}$ for a given time series is determined such that $\tilde{L}$ maximizes the discrete likelihood function (II.5) for the time series. For the convenience of the reader, we recall the likelihood function associated with the case of a complete and incomplete observation in time of a Markov jump process, respectively.

Suppose that the Markov jump process $X(t)$ has been observed continuously in a certain time interval $[0, T]$. Let the random variable $R_i(T)$ be the time the process spent in state i before time T,

$$R_i(T) = \int_0^T \chi_{\{i\}}(X(s))\mathrm{d}s,$$

where the characteristic function $\chi_{\{i\}}(X(s))$ is equal to one if $X(s) = i$ and zero otherwise. Moreover, denote by $N_{ij}(T)$ the number of transitions from state i to state j in the time interval $[0, T]$. The *continuous time likelihood function* $\mathcal{L}_c$ of an observed trajectory $\{X_t : 0 \leq t \leq T\}$ is given by [5]

$$\mathcal{L}_c(L) = \prod_{i=1}^{d} \prod_{j \neq i} l_{ij}^{N_{ij}(T)} \exp(-l_{ij} R_i(T)) \tag{II.3}$$

and the *maximum likelihood estimator* (MLE) $\tilde{L} = (\tilde{l}_{ij}), i, j \in S$, i.e. the generator which maximizes the likelihood function (II.3), takes the form

$$\tilde{l}_{ij} = \begin{cases} \dfrac{N_{ij}(T)}{R_i(T)}, & i \neq j \\ -\sum_{k \neq i} \tilde{l}_{ik}, & i = j. \end{cases} \tag{II.4}$$

In the case where the process has only been observed at discrete time points $0 = t_0 < t_1 < \ldots < t_N = T$ the process between two consecutive observations is *hidden* and, hence, the observables $R_i(T)$ and $N_{ij}(T)$ are unknown. The *discrete likelihood function* $\mathcal{L}_d$ of a time series $Y = \{y_0 = X(t_0), \ldots, y_N = X(t_N)\}$ is given in terms of the transition matrix $P(t) = \exp(tL)$

$$\mathcal{L}_d(L) = \prod_{k=0}^{N-1} p_{y_k, y_{k+1}}(\tau_k) = \prod_{s=1}^{r} \prod_{i,j \in S} \left[p_{ij}(\tau_s)\right]^{c_{ij}(\tau_s)}, \tag{II.5}$$

where $\tau_s \in \{\tau_1, \ldots, \tau_r\}$ is an observed time lag and the entry $c_{ij}(\tau_s)$ in the *frequency matrices* $C(\tau_s) = (c_{ij}(\tau_s)),\ i, j \in S$, defined according to

$$c_{ij}(\tau_s) = \sum_{n=1}^{N-1} \chi_{\{i\}}(X(t_n)) \chi_{\{j\}}(X(t_{n+1})) \chi_{\{\tau_s\}}(\Delta t_n)$$

provides the number of consecutively observed transitions in Y from state i to state j in time τ_s. Unfortunately, no analytical maximizer of the discrete likelihood function (II.5) with respect to the generator is available.

Nevertheless, the discrete likelihood $\mathcal{L}_d$ can iteratively be maximized by means of an *Expectation-Maximization algorithm* (EM-algorithm). The idea is to assume that the hidden (not observed) process behind the incomplete observations in Y is given by a initial guess, say $\tilde{L}_0$. Then averaging over all possible realization of $\tilde{L}_0$ *conditional* on the observation Y allows to compute the *conditional expected* values of $R_i(T)$ and the $N_{ij}(T)$. This step is called expectation step (E-Step). Formally, the E-Step consists of the computation of the *conditional log-likelihood* function

$$\mathcal{G} : L \mapsto \mathbb{E}_{\tilde{L}_0}\left[\log \mathcal{L}_c(L)|Y\right], \tag{II.6}$$

where $L \in \mathfrak{G}$. Notice that for algebraical simplicity and without loss of generality the *log-likelihood* function, $\log \mathcal{L}_c$, is considered. The crucial observation is now that the maximizer (M-step)

$$\tilde{L}_1 = \arg\max_{L\in\mathfrak{G}} \mathcal{G}(L;\tilde{L}_0)$$

of the conditional log-likelihood function $\mathfrak{G}(L;\tilde{L}_0)$ satisfies [21]

$$\mathcal{L}_d(\tilde{L}_1) \geq \mathcal{L}_d(\tilde{L}_0).$$

Hence, taking the maximizer as a new guess of the hidden process, the iteration of the two described steps allows to approximate a (local) maximum of the discrete likelihood function $\mathcal{L}_d$.

For our particular likelihood function (II.3) the conditional log-likelihood function $\mathcal{G}$ in the E-Step reduces to

$$\mathcal{G}(L;\tilde{L}_0) = \sum_{i=1}^{d}\sum_{j\neq i}\Big[\log{(l_{ij})}\mathbb{E}_{\tilde{L}_0}\left[N_{ij}(T)|Y\right] - l_{ij}\mathbb{E}_{\tilde{L}_0}\left[R_i(T)|Y\right]\Big] \tag{II.7}$$

and the maximizer $\tilde{L} = (\tilde{l}_{ij}),\ i,j \in S$ of (II.7) takes the form (cf. (II.4))

$$\tilde{l}_{ij} = \begin{cases} \dfrac{\mathbb{E}_{\tilde{L}_0}\left[N_{ij}(T)|Y\right]}{\mathbb{E}_{\tilde{L}_0}[R_i(T)|Y]}, & i \neq j \\ -\sum_{k\neq i}\tilde{l}_{ik}, & \text{otherwise.} \end{cases} \tag{II.8}$$

The non-trivial task which remains is to evaluate the conditional expectations $\mathbb{E}_{\tilde{L}_0}\left[N_{ij}(T)|Y\right]$ and $\mathbb{E}_{\tilde{L}_0}\left[R_i(T)|Y\right]$, respectively. Exploiting the Markov property and the homogeneity of the Markov jump process the conditional expectations in (II.7) can be expressed as sums [5]

$$\begin{aligned} \mathbb{E}_{\tilde{L}_0}\left[R_i(T)|Y\right] &= \sum_{s=1}^{r}\sum_{k,l=1}^{d} c_{kl}(\tau_s)\mathbb{E}_{\tilde{L}_0}\left[R_i(\tau_s)|X(\tau_s)=l, X(0)=k\right], \\ \mathbb{E}_{\tilde{L}_0}\left[N_{ij}(T)|Y\right] &= \sum_{s=1}^{r}\sum_{k,l=1}^{d} c_{kl}(\tau_s)\mathbb{E}_{\tilde{L}_0}\left[N_{ij}(\tau_s)|X(\tau_s)=l, X(0)=k\right]. \end{aligned} \tag{II.9}$$

Thus, the computation of $\mathbb{E}_{\tilde{L}_0}\left[N_{ij}(T)|Y\right]$ and $\mathbb{E}_{\tilde{L}_0}\left[R_i(T)|Y\right]$ is reduced to the computation of $\mathbb{E}_{\tilde{L}_0}\left[R_i(\tau_s)|X(\tau_s)=l,X(0)=k\right]$ and $\mathbb{E}_{\tilde{L}_0}\left[N_{ij}(\tau_s)|X(\tau_s)=l,X(0)=k\right]$ which is explained in the next section.

The idea is to approximate the hidden (not observed) information between the incomplete observations in Y by the *expected* (averaged) information *conditional* on the data and on a given guess of the hidden process.

III. ENHANCED COMPUTATION OF THE MAXIMUM LIKELIHOOD ESTIMATOR

In [8], it has been realized that the conditional expectations $\mathbb{E}_L\left[N_{ij}(t)|X(t)=l,X(0)=k\right]$ and $\mathbb{E}_L\left[R_i(t)|X(t)=l,X(0)=k\right]$ can analytically be expressed in terms of the generator transition matrix $P(t)=\exp(tL)$. The following identities are proved

$$\begin{aligned}\mathbb{E}_L\left[R_i(t)|X(t)=l,X(0)=k\right] &= \frac{1}{p_{kl}(t)}\int_0^t p_{ki}(s)p_{il}(t-s)\mathrm{d}s,\\ \mathbb{E}_L\left[N_{ij}(t)|X(t)=l,X(0)=k\right] &= \frac{l_{ij}}{p_{kl}(t)}\int_0^t p_{ki}(s)p_{jl}(t-s)\mathrm{d}s.\end{aligned} \tag{III.1}$$

The key observation now is that an eigendecomposition of the generator L leads to closed form expressions of the integrals in (III.1). To be more precise, consider the eigendecomposition of a generator L, that is

$$L = UD_\lambda U^{-1}, \tag{III.2}$$

where the columns of the matrix U consist of all eigenvectors to the corresponding eigenvalues of L in the diagonal matrix $D_\lambda=\operatorname{diag}(\lambda_1,\ldots,\lambda_d)$. Consequently, the expression of the transition matrix $P(t)$ simplifies to

$$P(t)=\exp(tL)=U\exp(tD_\lambda)U^{-1}$$

and we finally end up with a closed form expression of the integrals in (III.1), that is

$$\int_0^t p_{ab}(s)p_{cd}(t-s)\mathrm{d}s = \sum_{p=1}^{d} u_{ap}u_{pb}^{-1}\sum_{q=1}^{d} u_{cq}u_{qd}^{-1}\Psi_{pq}(t), \tag{III.3}$$

where the symmetric matrix $\Psi(t)=(\Psi_{pq}(t))_{p,q\in S}$ is defined as

$$\Psi_{pq}(t)=\begin{cases} te^{t\lambda_p} & \text{if } \lambda_p=\lambda_q\\ \frac{e^{t\lambda_p}-e^{t\lambda_q}}{\lambda_p-\lambda_q} & \text{if } \lambda_p\neq\lambda_q.\end{cases} \tag{III.4}$$

Combining all issues, we finally end up with the *enhanced MLE-method for non-equidistant time lags* as stated in Algorithm 1.

Algorithm 1 Enhanced MLE-method for non-equidistant time lags

```
Input: Time series $Y = \{y_0 = X(t_0), \ldots, y_N = X(t_N)\}$, the set of observed time lags $\{\tau_1, \ldots, \tau_r\}$,
    the tolerance $TOL$, initial guess of generator $\tilde{L}_0$.
Output: MLE $\tilde{L}$.
  (1) Compute eigendecomposition (III.2) of $\tilde{L}_k$.
  (2) E-step: FOR ALL $\tau_s \in \{\tau_1, \ldots, \tau_r\}$ DO
          i)   Compute the auxiliary matrix $\Psi(\tau_s)$ (III.4).
          ii)  Compute for $i, j, l, k = 1, \ldots, d$ the conditional expectations
                 $\mathbb{E}_{\tilde{L}_k}\left[R_i(\tau_s)|X(\tau_s) = l, X(0) = k\right]$,
                 $\mathbb{E}_{\tilde{L}_k}\left[N_{ij}(\tau_s)|X(\tau_s) = l, X(0) = k\right], i \neq j$ via (III.3),(III.1).
            END FOR
          iii) Compute $\mathbb{E}_{\tilde{L}_k}\left[R_i(T)|Y\right]$ and $\mathbb{E}_{\tilde{L}_k}\left[N_{ij}(T)|Y\right]$ via (II.9).
  (3) M-Step: Setup the next guess $\tilde{L}_{k+1}$ of the generator by
          $\tilde{l}_{ij} = \begin{cases} \mathbb{E}_{\tilde{L}_k}\left[N_{ij}(T)|Y\right]/\mathbb{E}_{\tilde{L}_k}\left[R_i(T)|Y\right], & i \neq j \\ -\sum_{k \neq i} \tilde{l}_{ik}, & \text{otherwise.} \end{cases}$
  (4) Go to Step (1) unless $\|\tilde{L}_{k+1} - \tilde{L}_k\| < TOL$.
```

The computational cost of a single iteration step in Algorithm 1 is $\mathcal{O}(r \cdot d^5)$ where r is the number of the different observed time lags and d is the dimension of the finite state space S. We want to emphasize that the algorithm in principal works even in the case of pairwise different time lags, i.e. $r = N - 1$ where N is the number of observations, but in practise this would lead to unacceptable computational costs.

IV. NUMERICAL EXAMPLES

In this section we demonstrate the performance of the enhanced MLE-method for non-equidistant observation times on a test example and for a process arising in the approximation of a genetic toggle switch. In both examples, we re-identify a generator L of a Markov jump process from an associated artificially generated incomplete observation. To be more precise, we drew from a generator L a continuous time realization

$\{X(t), 0 \leq t \leq T\}$ for a prescribed $T > 0$ and extracted out of it an incomplete observation $Y = \{y_0 = X(t_0), \ldots, y_N = X(t_N)\}$ with respect to a prescribed set of time lags $\{\tau_1, \ldots, \tau_r\}$, $r > 1$, as follows: Suppose $t_k < T$ is the observation time last considered then the next observation time t_{k+1} is given by $t_{k+1} = t_k + \tau$ where τ is uniformly drawn from the set of time lags $\{\tau_1, \ldots, \tau_r\}$. We terminate that procedure if $t_{k+1} > T$.

A. Test example

In the first example we demonstrate the performance of the enhanced MLE-method on a small toy example. To this end we consider a five-state Markov jump process given by its generator

$$L = \begin{pmatrix} -6 & 2 & 2 & 1 & 1 \\ 1 & -4 & 0 & 1 & 2 \\ 1 & 0 & -4 & 2 & 1 \\ 2 & 1 & 0 & -3 & 0 \\ 1 & 1 & 1 & 1 & -4 \end{pmatrix} \in \mathfrak{G}. \tag{IV.1}$$

For the reconstruction of L, we extracted from a realization of total time $T = 3.7 \cdot 10^6$ a time series of $N = 10^7$ observations with respect to the set of time lags $\{\tau_1 = 0.01, \tau_2 = 0.1, \tau_3 = 1\}$. In (IV.2) we state the estimated generator resulting from Algorithm 1 with the prescribed tolerance $TOL = 10^{-6}$. One clearly can see that $\tilde{L}$ approximates the original one very well.

$$\tilde{L} = \begin{pmatrix} -5.9803 & 2.0054 & 1.9863 & 0.9911 & 0.9975 \\ 1.0002 & -4.0018 & 0.0010 & 0.9938 & 2.0068 \\ 0.9921 & 0.0001 & -3.9768 & 1.9938 & 0.9909 \\ 1.9909 & 0.9951 & 0.0004 & -2.9871 & 0.0006 \\ 0.9982 & 1.0051 & 0.9993 & 1.0050 & -4.0075 \end{pmatrix} \in \mathfrak{G}. \tag{IV.2}$$

Next, we address the question of how the length of the respective time series and the number of different time lags do affect the outcome of the estimation procedure. To make things comparable, we generated three different time series of length $N = 10^8$ with respect to the time lags sets $\{0.01\}, \{0.01, 0.1\}$ and $\{0.01, 0.1, 1\}$, all subsampled from the same

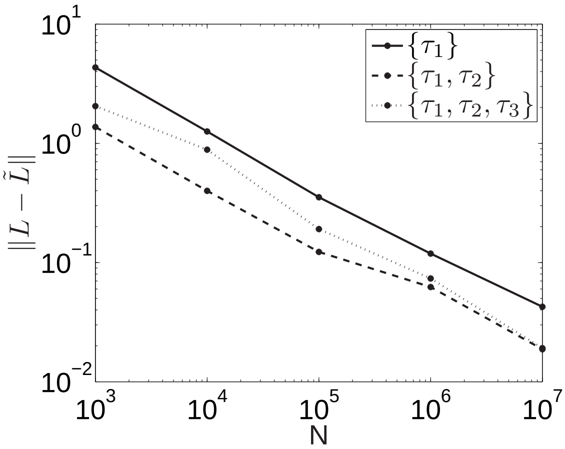

FIG. 1: Error of the estimated generator $\tilde{L}$ with respect to the original generator (IV.1), measured in the 2-norm $\|\tilde{L}-L\|$, as a function of the length N of the respective time series. Results for the three different sets of time lags $\{0.01\}, \{0.01, 0.1\}$ and $\{0.01, 0.1, 1\}$ and the tolerance $TOL = 10^{-6}$.

underlying continues time realization, respectively, and estimated for each time series a generator on the basis of the first $N = 10^3, N = 10^4 \ldots, N = 10^8$ observed states, respectively. Furthermore, we used for all estimations the same initial guess $\tilde{L}_0$. In Figure 1 we illustrate the dependence of the approximation error $\|\tilde{L}-L\|$ (measured in the 2-norm) with respect to the length N of the respective time series and the number of different time lags. The graphs reveal that the error $\|\tilde{L}-L\|$ decays exponentially with the length of the underlying time series approximately as $N^{\frac{1}{2}}$. The second observation is that the estimations based on multiple observation time lags give better results than the estimation on a single time lag. The authors are not aware of how to explain this observation.

B. Application to a genetic toggle switch model

In this example we apply the enhanced MLE-method to a *birth-death process* which arises as a stochastic model of a genetic toggle switch consisting of two genes that repress each others' expression [22]. Expression of the two different genes produces two different types of proteins; let us name them P_A and P_B. If we denote the number of molecules of type P_A by x and of type P_B by y, then the authors in [22] proposed the following birth-death process on the discrete state space $S = (\mathbb{Z} \times \mathbb{Z}) \cap ([0, d_1] \times [0, d_2]), d_1, d_2 > 0$, given by its generator acting on a function $f : S \mapsto \mathbb{R}$

$$\begin{aligned}(Lf)(x,y) =& c_1(x+1,y)(f(x+1,y)-f(x,y)) + \frac{x}{\tau_1}(f(x-1,y)-f(x,y)) \\ &+ c_2(x,y+1)(f(x,y+1)-f(x,y)) + \frac{y}{\tau_2}(f(x,y-1)-f(x,y)),\end{aligned} \tag{IV.3}$$

where

$$c_1(x+1,y) = \begin{cases} \frac{a_1}{1+(y/K_2)^n}, & \text{if } x \in [0,d_1) \\ 0, & \text{if } x = d_1, \end{cases}$$

$$c_2(x,y+1) = \begin{cases} \frac{a_2}{1+(x/K_1)^m}, & \text{if } y \in [0,d_2) \\ 0, & \text{if } y = d_2 \end{cases}$$

for describing the evolution of the numbers x and y of the respective proteins in the genetic toggle switch. For the biological interpretation of the involved parameters see [22]. Moreover, notice that the particular choice of the coefficients c_1 and c_2 on the right and upper boundary can be seen as reflecting boundary conditions.

A single realization of the jump process generated by L models the evolution of the numbers of proteins with respect to a specific initial value (x_0, y_0). The resulting evolution of the associated probability density function (PDF) in time is governed by the *Master-equation*: Let $p_0 \in \mathbb{R}^{|S|}$ be the initial PDF, then the PDF evolves in time according to

$$\frac{\partial p(t)}{\partial t} = L^T p(t), \quad p(0) = p_0, \; t > 0, \tag{IV.4}$$

where L^T denotes the transpose of the generator given in (IV.3). Its steady state $\pi = (\pi_i), \; i \in S$ of (IV.4) is called stationary distribution.

It is well-known that in the limit of large protein numbers the dynamics of the jump process or, more precisely, of the associated Master equation is given by a deterministic model of the biochemical kinetics in terms of the associated concentrations. The authors in [22] also consider this deterministic model in order to get a rough understanding of the switch dynamics. The model consists of two coupled ordinary differential equations,

$$\begin{aligned}\dot{x} &= \frac{a_1}{1+(y/K_2)^n} - \frac{x}{\tau_1}, \\ \dot{y} &= \frac{a_2}{1+(x/K_1)^m} - \frac{y}{\tau_2},\end{aligned} \tag{IV.5}$$

where the parameters are the same as in the stochastic model (IV.3). For the numerical experiments to be presented , we used the parameters $a_1 = 156, a_2 = 30, n = 3, m =$

$1, K_1 = K_2 = 1, \tau_1 = 1/7$ and $\tau_2 = 1/3$. For this particular choice the deterministic dynamics (IV.5) has two stable stationary points approximately at $(x, y) = (20, 0)$ and $(x, y) = (0, 8)$ and one unstable point approximately at $(x, y) = (6, 1)$. This insight in the deterministic approximation helps to understand the following analysis of the jump process:

For the sake of illustration, the left panel of Figure 2 shows $(-\log \pi_i), i \in S$ instead of the stationary distribution $\pi = (\pi_i), i \in S$ of the jump process itself. All states with almost vanishing stationary distribution are depicted by the white region. Moreover, in order to emphasize the states of interest, we chose a log-log representation. The color scheme is chosen such that the darker the color of a region the higher is the probability of finding the process there. One can clearly see that the process spends most of its time near the two stable stationary points approximately at $(x, y) = (20, 0)$ and $(x, y) = (0, 8)$.

In order to motivate the relevance of the following numerical experiment, suppose you measure the numbers of proteins of types P_A and P_B discretely in time; without knowing the generator, you are interested in fitting a Markov jump process. Assuming that the hidden process is Markovian, one can apply the enhanced MLE-method.

Before we describe our numerical example in detail, notice that the structure of a transition matrix P, i.e. the occupation of the entries in P, does not allow to infer on the structure of the underlying generator. For example, the generator of a dense transition matrix does not have to be dense too. This means that there is some freedom in the choice of the structure of the estimated generator $\tilde{L}$. In this example, we follow two options. One option – we call it option A – is to use the structure of the observed transition matrix as a blueprint for the structure of $\tilde{L}$. In option B we exploited knowledge about the hidden process. We know that the number of a gene's molecule can only increase or decrease by one in a single reaction while the number of the other one remains constant. Hence, it is natural to estimate the entries $\tilde{l}_{ij}$ if the states i and j (the numbers) have been observed and are adjacent in the sense of a single reaction.

For our numerical experiment, we generated a sufficiently long realization of the birth-death-process on the state space $\mathbb{Z}^2 \cap ([0, 30] \times [0, 30])$ and extracted out of it a time series of length $N = 10^8$ with respect to the set of time lags $\{\tau_1 = 0.0001, \tau_2 = 0.001, \tau = 0.01\}$. As one can see in Figure 2, the relative occupation of the states (right panel) is consistent with the exact stationary distribution (left panel).

The generated time series visits 225 states of 900 possible states, hence we had to estimate

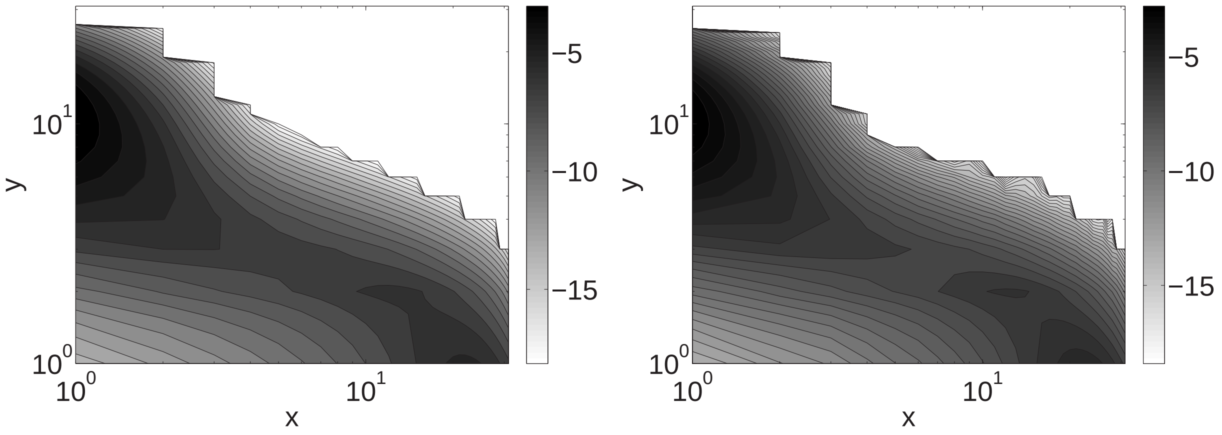

FIG. 2: Left: Log-log contour plot of $(-\log \pi_i), i \in S$, where $\pi = (\pi_i), i \in S$ is the stationary distribution of the process in (IV.3) computed via $\pi^T L = 0$ (cf. (II.2)). Right: Log-log contour plot of $(-\log \hat{\pi}_i), i \in S$ resulting from the observed distribution $\hat{\pi}$ of states in the time series. Result for $N = 10^8$.

a generator $\tilde{L} \in \mathfrak{G}$ on the state space $S \cong \{1, \ldots, 225\}$. In the following $\tilde{L}_A$ denotes the estimated generator resulting from the estimation option A and $\tilde{L}_B$ via option B. For both estimation options we used the tolerance $TOL = 10^{-2}$. Figure 3 shows $(-\log \pi_i), i \in S$ resulting from the stationary distribution associated with $\tilde{L}_A$ (left panel) and with $\tilde{L}_B$ (right panel). From the viewpoint of stationarity, one can see that both estimated generators are good approximations of the original one (compare left panel of Figure 2). In order to make things more precise, we compare in the following the estimated generators with the original generators of (IV.3) *restricted* on the set of observed states. Formally, we consider the restricted generator $\bar{L} \in \mathfrak{G}, S \cong \{1, \ldots, 225\}$ defined according to

$$\bar{l}_{ij} = \begin{cases} l_{ij}, & \text{if } i \neq j \text{ were visited by the time series,} \\ -\sum_k \bar{l}_{ik}, & \text{if } i = j \text{ was visited by the time series.} \end{cases} \tag{IV.6}$$

Now we compare the spectral properties of the estimated generators with those of the restricted generator from (IV.6) in more detail. In the left panel of Figure 4 we depict the real parts of the 30 largest eigenvalues of $\tilde{L}_A$ and $\tilde{L}_B$ with those of the restricted generator $\bar{L}$, respectively. Although the enhanced MLE-method is not designed to approximate spectral properties, notice that the real parts of considered eigenvalues of $\bar{L}$ are well reconstructed by both estimation options. Another important quantity in time series analysis is the auto-correlation function (ACF) of a process which reflects the speed of memory loss of

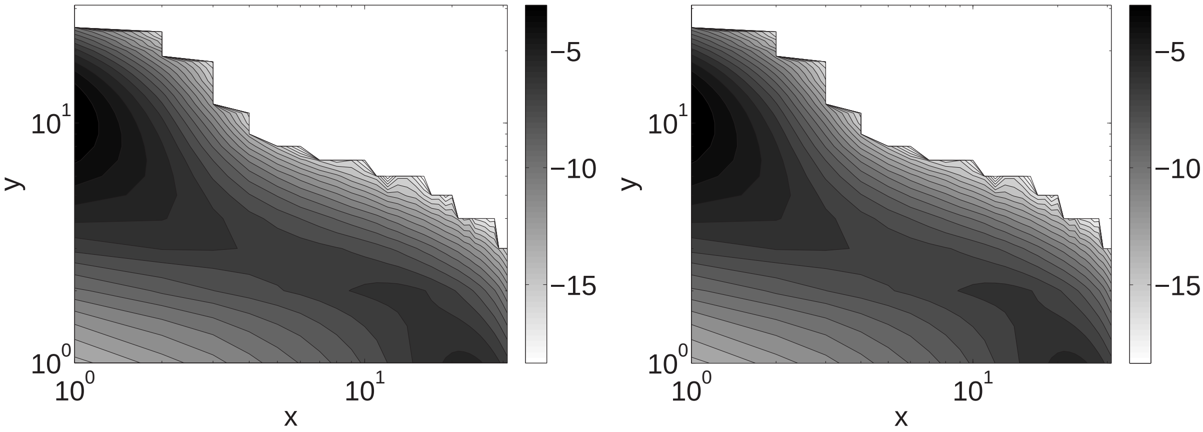

FIG. 3: Log-log contour plot of $(-\log \tilde{\pi}_i), i \in S$ associated with the estimated generators $\tilde{L}_A$ (left panel) and $\tilde{L}_B$ (right panel) where $\tilde{\pi}$ is the stationary distribution of the estimated generators computed via $\tilde{\pi}^T \tilde{L} = 0$, respectively.

the process. For a Markov jump process, it is easy to see that the ACF reduces to [4]

$$\mathbb{E}(X_{t+\tau} X_t) = \sum_{k=1}^{d} e^{\tau \lambda_k} \sum_{i,j \in S} i \cdot j \cdot \pi_i U_{ik} U_{kj}^{-1}, \tag{IV.7}$$

where $L = U \mathrm{diag}(\lambda_1, \ldots, \lambda_d) U^{-1}$ is the eigendecomposition of the generator L of the Markov jump process and $\pi = (\pi_i),\ i \in S$ its stationary distribution. The graphs of the normalized ACFs associated with $\tilde{L}_A$ and $\tilde{L}_B$ together with the ACF of the restricted generator $\bar{L}$ are given in Figure 5. As one can see, the ACFs associated with $\tilde{L}_A$ and $\tilde{L}_B$ are consistent with the ACF of the restricted process which shows that besides the eigenvalues even the eigenvectors of the restricted generator $\bar{L}$ are well reproduced by both estimated generators, respectively. The almost identical reproduction of the ACF of $\bar{L}$ by $\tilde{L}_B$ shows that the incorporation of theoretical knowledge of the hidden process leads to sightly better results.

V. SUMMARY AND DISCUSSION

A generalization of the enhanced MLE-method for the estimation of a generator based on a time series with non-equidistant observation time lags has been presented. Its performance has been validated numerically for a test case and an application to biochemical kinetics data has been given. In particular, the latter example has shown that the enhanced MLE-

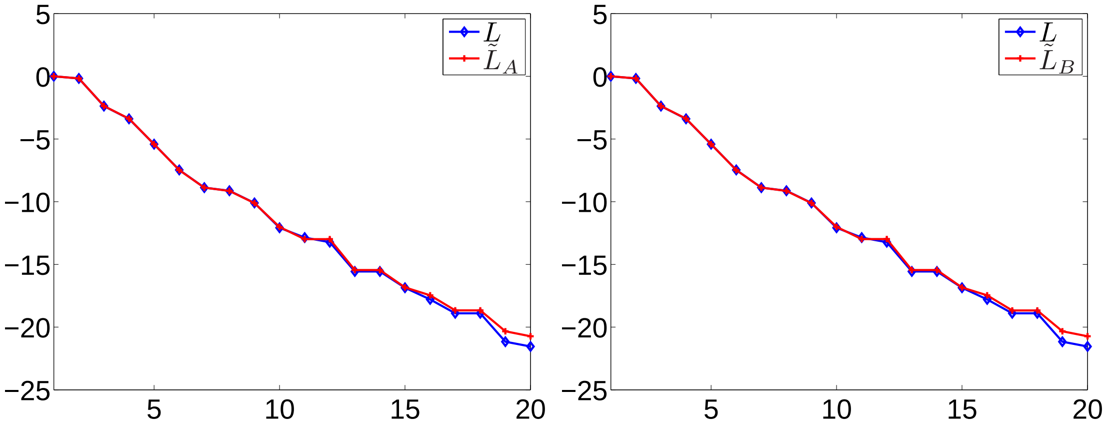

FIG. 4: The real parts of the first 30 largest eigenvalues of the estimated generators compared to the eigenvalues of the restricted generator $\bar{L}$ (IV.6). Left: Real parts of eigenvalues of $\tilde{L}_A$. Right: Real parts of eigenvalues of $\tilde{L}_B$.

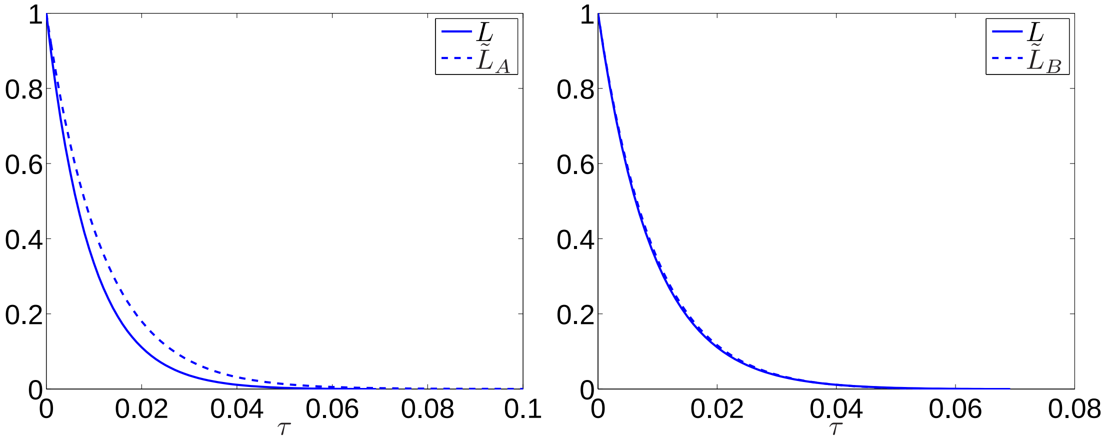

FIG. 5: The graphs of the ACFs associated with $\tilde{L}_A$ (left panel) and $\tilde{L}_B$ (right panel) compared to the ACF of the restricted generator $\bar{L}$, respectively.

method is applicable to processes in larger state spaces. As illustrated in Section IV B, the new method can be devised to respect specific sparsity patterns of the generators to be estimated; furthermore it can also be specified for the estimation of generators of reversible Markov jump processes (in analogy to the approach presented in [6]).

Several remarks have to be made regarding possible pitfalls of the presented approach. First, the enhanced MLE-methods relies on the decomposability of the generator that appear in the course of the iteration. There is no reason to expect that each single one has a complex diagonalization; however in none of our quite extensive numerical experiments the situation of a non-decomposable generator appeared. If this would happen one would be able to fall back on the original "not-enhanced" algorithm (just for this step of the iteration). Second, the eigendecomposition for non-symmetrical matrices could be a numerical problem (it may even be ill-conditioned, see [23], for example). However, an appropriate numerical solver should indicate this by a warning message. This case also never appeared in our numerical experiments. Third, since the enhanced MLE-methods is an EM-algorithm, it can only be assured that it converges to a local maximum of the discrete likelihood function; global convergence is not guaranteed. The dependence of the result on the initial guess has not been addressed here and will be subject of further investigations. Fourth, the scaling of the computational effort with the dimension/size of the state space makes application to very large state spaces infeasible. One can circumvent this problem whenever one knows in advanced that the generator to be estimated has a certain sparsity pattern. If this is not the case the present authors are not aware of any cure to this "curse of dimension".

Acknowledgement

We are indebted to E. Dittmer and T. Jahnke for helpful discussions. We also thank the anonymous referee for useful comments.

[1] S. Asmussen, M. Olsson, and O. Nerman. Fitting phase-type distributions via the em algorithm. *Scand. J. Statist.*, 23(4):419–441, 1996.

[2] T. Müller. *Modellierung von Proteinevolution.* PhD thesis, Heidelberg, 2001.

[3] U. Nodelman, C.R. Shelton, and D. Koller. Expectation maximization and complex duration distributions for continuous time bayesian networks. In *Proceedings of the Twenty-first Conference on Uncertainty in AI (UAI)*, pages 421–430, Edinburgh, Scottland, UK, 2005.

[4] D. T. Crommelin and E. Vanden-Eijnden. Fitting timeseries by continuous-time Markov chains: A quadratic programming approach. *J. Comp. Phys.*, 217(2):782–805, 2006.

[5] M. Bladt and M. Sørensen. Statistical inference for discretely observed Markov jump processes. *J. R. Statist. Soc. B*, 39(3):395–410, 2005.

[6] P. Metzner, E. Dittmer, T. Jahnke, and Ch. Schütte. Generator estimation of Markov jump

processes. *J. Comp. Phys.*, 227(1):353–375, 2007.

[7] I. Holmes and G. M. Rubin. An expectation maximization algorithm for training hidden substitution models. *J. Mol. Biol.*, 317(5):753–764, 2002.

[8] A. Hobolth and J. L.Jensen. Statistical inference in evolutionary models of DNA sequences via the EM algorithm. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 2005.

[9] S. A. Adelman and J. D. Doll. Generalized Langevin equation approach for atom/solid surface scattering: General formulation for classical scattering of harmonic solids. *J. Chem. Phys.*, 64:2375–2388, 1976.

[10] J. B. Witkoskie, J. Wu, and J. Cao. Basis set study of classical rotor lattice dynamics. *J. Chem. Phys.*, 120:5695–5708, 2004.

[11] J. Peinke, A. Kittel, S. Barth, and M. Oberlack. Langevin models of turbulence. In B. Dubrulle, J-P. Lavale, and S. Nazarenko, editors, *Progress in Turbulence*, pages 77–86. Springer Berlin Heidelberg, 2005.

[12] T. Yamaguchi, T. Matsuoka, and S. Koda. Molecular dynamics simulation study of the transient response of solvation structure during the translational diffusion of solute. *J. Chem. Phys.*, 122:014512.1–014512.10, 2005.

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

[14] R. M. Yulmetyev et al. Non-Markov statistical effects of X-ray emission intensity of the microquasar Grs 1915+105. *Nonlin. Phenom. Compl. Systems*, 9(4):313–330, 2006.

[15] Pierre Brémaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues.* Springer, New York, 1999.

[16] P. Deuflhard and M. Wulkow. Computational treatment of polyreaction kinetics by orthogonal polynomials of a discrete variable. *IMPACT Comput Sci Eng*, 1(3):269–301, 1989.

[17] M. Wulkow. Adaptive treatment of polyreactions in weighted sequence spaces. *IMPACT Comput Sci Eng*, 4(2):153–193, 1992.

[18] M. Wulkow. The simulation of molecular weight distribution in polyreaction kinetics by discrete Galerkin methods. *Macromol Theory Simul*, 5(3):393–416, 1996.

[19] G. P. Karev, Y. I. Wolf, and E. V. Koonin. Simple stochastic birth and death models of genome evolution: was there enough time for us to evolve? *Bioinformatics*, 19(15):1889–1900, 2003.

[20] T. Müller, R. Spang, and M. Vingron. Estimating amino acid substitution models: A comparison of Dayhoff's estimator, the resolvent approach and a maximum likelihood method. *Mol Biol Evol*, 19(1):8–13, 2002.

[21] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc.*, 39(1):1–38, 1977.

[22] D. M. Roma, R. O'Flanagan, A. Ruckenstein, A. M. Sengupta, and R. Mukhopadhyay. Optimal path to epigenetic switching. *Phys. Review E*, 71(1):011902, 2005.

[23] P. Deuflhard and A. Hohmann. *Numerical Analysis. A First Course in Scientific Computation.* Springer, 2003.