5. Generator Estimation of Markov Jump Processes

We have seen that discrete TPT is a powerful tool for the investigation of the ensemble of reaction trajectories of Markov jump processes. The central object of discrete TPT is the committor function which satisfies a system of linear equation involving the infinitesimal generator of the process. In applications, however, the generator is not given but only an *incomplete* observation of the process is available. This chapter is devoted to the problem of finding a generator of a Markov jump process based on an *incomplete* observation of the process. The results of this chapter are published in [63, 64].

We will focus on two methods for the estimation of a generator. The maximum likelihood method (MLE-method) introduced by Asmussen 1996 in [5] and reinvented by Bladt and Sørensen 2005 in [9] finds a generator via an EM-algorithm which maximizes the likelihood of the given incomplete observation. Furthermore, we will discuss a significant algorithmic improvement of the MLE-method, we call it *enhanced* MLE-method, which was independently derived by Holmes and Rubin 2002 in [49]. Moreover, we introduce an adaption of the enhanced MLE-method to the case of reversible Markov jump processes. The quadratic programming approach introduced by Crommelin and Vanden-Eijnden in [19], determines a generator via the approximation of the eigenstructure of the empirical transition matrix.

After a comparison of both methods via their numerical performance on small test examples, we will apply the enhanced MLE-method to data from a molecular dynamics simulation of glycine in water. The resulting estimated generator is the basis for the investigation of the conformational dynamics of glycine via discrete TPT (cf. Chap. 4). Finally, we will demonstrate the performance of the enhanced MLE-method on an example with non-constant observation time steps.

5.1. The Embedding Problem

Let $\{X(t), t \geq 0\}$ be a Markov jump process on a finite state space $S \cong \{1, \ldots, d\}$ and let $L \in \mathbb{R}^{d \times d}$ be its generator. Then the time-dependent transition matrix $P(t)$ of the process can be expressed as the matrix exponential (cf. Sect. 2.2)

$$P(t) = \exp(tL) = \sum_{k=0}^{\infty} \frac{t^k}{k!} L^k.$$

In the following, the set of all generators with respect to a fixed dimension d will be denoted by

$$\mathfrak{G} = \left\{ L = (l_{ij})_{i,j} \in \mathbb{R}^{d \times d} : l_{ij} \geq 0 \quad \text{for all } i \neq j, \quad l_{ii} = -\sum_{j \neq i} l_{ij} \right\}. \tag{5.1}$$

Now suppose that a process is only partially observed, i.e. the process is only given by a finite sampling $Y = \{y_0 = X(t_0), \ldots, y_N = X(t_N)\}$ at discrete times $t_0 < t_1 < \ldots < t_N$. In this chapter we consider the problem of how to determine the generator if only an incomplete observation Y is available.

Several difficulties must be taken into account. First, from a finite number of samples it is impossible to tell if the underlying process is actually Markovian. Second, it is not clear if the observed data originates indeed from discrete samples of a continuous-time Markov chain with some generator L, or rather from a discrete-time Markov chain which cannot be embedded into a time-continuous counterpart. In the latter case, a generator does not exist because the transition matrix of the discrete chain does not belong to the set

$$\mathfrak{P} = \left\{ P \in \mathbb{R}^{d\times d} : \text{ there is a } L \in \mathfrak{G} \text{ such that } P = \exp(L) \right\}.$$

It is well-known that $\mathfrak{P}$ is a subset of all stochastic matrices, but the so-called *embedding problem*, i.e. the question what characterizes the elements of $\mathfrak{P}$, is widely open for $d > 3$ (cf. [9, 19] and references therein). A third difficulty is the fact that the matrix exponential function *is not injective* if the eigenvalues of the generator are complex. Hence, some matrices $P \in \mathfrak{P}$ can be represented as $P = \exp(L) = \exp(\bar{L})$ with two different generators $L \neq \bar{L}$. And finally, the question whether the time points t_n of the observations are equidistant plays an important role. In case of a constant time lag $\tau = t_{n+1} - t_n$ an estimate of the transition matrix $P(\tau)$ is available by counting the number of transitions between each pair of states, but in case of variable time lags the sampled data is typically not sufficient for reasonable approximations of the transition matrix.

Due to these problems the above question has to be modified: how can we find the generator that "agrees best" with a finite observation $Y = \{y_0 = X(t_0), \ldots, y_N = X(t_N)\}$ of a process?

5.2. The Maximum Likelihood Method

In this section we explain in detail the maximum likelihood method introduced in [5] and elaborated further in [9]. Furthermore, we present in detail the enhanced MLE-methods which is based on results in [49]. The idea behind the MLE-method is to find a generator $\tilde{L}$ such that it maximizes the *discrete likelihood* of the given time series.

5.2.1. Continuous and Discrete Likelihood Functions

The basis objects in the MLE-method is the continuous and discrete likelihood function. 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

$$R_i(T) = \int_0^T \mathbf{1}_{\{i\}} X(s) \mathrm{d}s$$

and 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 [9]

$$\mathcal{L}_c(L) = \prod_{i=1}^{d} \prod_{j \neq i} l_{ij}^{N_{ij}(T)} \exp(-l_{ij} R_i(T)), \qquad L = (l_{ij}). \tag{5.2}$$

By definition, the *maximum likelihood estimator* (MLE) $\tilde{L}$ maximizes the likelihood function (5.2). Exploiting the monotonicity of the *log*-function, $\tilde{L}$ is also the maximizer of

$$\log \mathcal{L}_c(L) = \sum_{i=1}^{d} \sum_{j \neq i} \left[N_{ij}(T) \log(l_{ij}) - l_{ij} R_i(T) \right], \tag{5.3}$$

i.e. $\tilde{L}$ is the null of the partial derivatives of $\log \mathcal{L}_c(L)$ with respect to l_{ij} and the Hessian matrix of $\log \mathcal{L}_c(L)$ evaluated at $\tilde{L}$ is negative definite. A short calculation shows

$$\frac{\partial \log \mathcal{L}_c(\tilde{L})}{\partial l_{ij}} = 0 \Longleftrightarrow \tilde{l}_{ij} = \frac{N_{ij}(T)}{R_i(T)} \tag{5.4}$$

and

$$\frac{\partial \log \mathcal{L}_c(\tilde{L})}{\partial l_{ij} \partial l_{kl}} = -\frac{N_{ij}(T)}{\tilde{l}_{ij}^2} \mathbf{1}_{\{k\}}(i) \mathbf{1}_{\{l\}}(j) \leq 0.$$

In the case where the process has only been observed at discrete time points $0 = t_0 < t_1 < \ldots < t_N = T$ 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 $P(t) = \exp(tL)$,

$$\mathcal{L}_d(L) = \prod_{k=0}^{N-1} p_{y_k, y_{k+1}}(\Delta t_k) = \prod_{s=1}^{r} \prod_{i,j \in S} \left[p_{ij}(\tau_s) \right]^{c_{ij}(\tau_s)}, \tag{5.5}$$

where $p_{y_k, y_{k+1}}(\Delta t_k)$ is the probability that the process makes a transition from state y_k to the state y_{k+1} in time Δt_k, $\tau_s \in \{\tau_1, \ldots, \tau_r\} = \cup_{k=1}^{N-1} \{\Delta t_k\}$ is an observed time lag and the entry $c_{ij}(\tau_s)$ in the *frequency matrix* $C(\tau_s) = (c_{ij}(\tau_s)),\ i, j \in S$, defined according to

$$c_{ij}(\tau_s) \stackrel{def}{=} \sum_{n=1}^{N-1} \mathbf{1}_{\{i\}}(X(t_n)) \mathbf{1}_{\{j\}}(X(t_{n+1})) \mathbf{1}_{\{\tau_s\}}(\Delta t_n), \tag{5.6}$$

provides the number of consecutively observed transitions in Y from state i to state j in time τ_s.

Unfortunately, even in the simplified case of a constant time lag, i.e. $\tau = const$, the derivative of the discrete log-likelihood function,

$$\log \mathcal{L}_d(L) = \sum_{k=0}^{N-1} \log p_{y_k, y_{k+1}}(\tau) = \sum_{i,j \in S} \left[\log p_{ij}(\tau) \right]^{c_{ij}(\tau)}, \tag{5.7}$$

with respect to the entries of L, that is

$$\begin{gathered} \tfrac{\partial}{\partial L} \log \mathcal{L}_d(L) = \textstyle\sum_{n=1}^{\infty} \sum_{k=1}^{n} \frac{\tau^n}{n!} (L^T)^{k-1} Z (L^T)^{n-k}, \\ \text{with } Z = (z_{ij})_{i,j \in S},\ z_{ij} = c_{ij} / \exp(\tau L)_{ij}, \end{gathered}$$

has such a complicated form that the null cannot be found analytically. Hence no analytical expression for the MLE with respect to L is available. The derivative of $\log \mathcal{L}_d$ with respect to the transition matrix $P(\tau)$ can analytical be obtained and the maximizer is simply given by

$$\hat{P} = (\hat{p}_{ij})_{i,j} \quad \text{with entries} \quad \hat{p}_{ij} = \frac{c_{ij}(tau)}{\sum_{j=1}^{d} c_{ij}(tau)}, \tag{5.8}$$

where $c_{ij} = c_{ij}(\tau)$. Notice that reversibility of the transition matrix $\hat{P}$ can easily be achieved by considering the symmetric frequency matrix $C^{REV}(\tau) = (c_{ij}^{REV})$ with entries given by

$$c_{ij}^{REV}(\tau) = c_{ij}(\tau) + c_{ji}(\tau).$$

Then the transition matrix which results via 5.8 on the basis of $C^{REV}(\tau)$ is reversible with respect to the probability distribution

$$\pi = Z^{-1} \left(\sum_{k=1}^{d} c_{1k}^{REV}(\tau), \ldots, \sum_{k=1}^{d} c_{dk}^{REV}(\tau) \right),$$

where $Z = \sum_{i,j=1}^{d} c_{ij}^{REV}$ is the normalization constant.

In the following neither we assume a constant observation time lag nor we assume reversibility.

5.2.2. Likelihood Approach Revisited

In the likelihood approach, introduced by Bladt and Sørensen in [9], a generator $\tilde{L}$ for a given time series is determined such that $\tilde{L}$ maximizes the discrete likelihood function (5.5) for the time series. As pointed out in the previous section the discrete likelihood function $\mathcal{L}_d$ does not permit an analytical maximum likelihood estimator. On the other hand, the MLE (5.4) for a continuous time observation can be obtained analytically but for an incomplete observation the information between two consecutive observations is *hidden* and, hence, the observables $R_i(T)$ and $N_{ij}(T)$ are unknown.

Nevertheless, the discrete likelihood $\mathcal{L}_d$ can iteratively be maximized by means of an *Expectation-Maximization algorithm* (EM-algorithm). 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. This step is called expectation step (E-Step) and formally consists of the computation of the *conditional log-likelihood* function

$$\mathcal{G} : L \mapsto \mathbb{E}_{\tilde{L}_0} \left[\log \mathcal{L}_c(L) | Y \right], \tag{5.9}$$

where $L \in \mathfrak{G}$ and for reasons of 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 [23]

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

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 $\mathcal{L}_d$. The resulting algorithm is stated in Algorithm 3.

Algorithm 3 General EM-algorithm

Input: Time series $Y = \{y_0 = X(t_0), \dots, y_N = X(t_N)\}$, initial guess of generator $\tilde{L}_0$.

Output: MLE $\tilde{L}$.

(1) Expectation step (E-step):
Compute the function $L \mapsto \mathcal{G}(L; \tilde{L}_k)$.

(2) Maximization step (M-Step):
$\tilde{L}_{k+1} = \arg\max_{L \in \mathfrak{G}} \mathcal{G}(L; \tilde{L}_k)$

(3) Go to Step (1), unless a certain convergence criterion is satisfied.

For our particular likelihood function in (5.2) the conditional log-likelihood function $\mathcal{G}$ in the E-Step reduces to

$$\mathcal{G}(L; \tilde{L}_0) = \sum_{i=1}^{d} \sum_{j \neq i} \Big[\log(l_{ij}) \mathbb{E}_{\tilde{L}_0}[N_{ij}(T)|Y] - l_{ij} \mathbb{E}_{\tilde{L}_0}[R_i(T)|Y] \Big] \tag{5.10}$$

and the maximizer $\tilde{L} = (\tilde{l}_{ij}),\ i, j \in S$ of (5.10) takes the form (cf. (5.4))

$$\tilde{l}_{ij} = \begin{cases} \dfrac{\mathbb{E}_{\tilde{L}_0}[N_{ij}(T)|Y]}{\mathbb{E}_{\tilde{L}_0}[R_i(T)|Y]}, & i \neq j \\ -\sum_{k \neq i} \tilde{l}_{ik}, & \text{otherwise.} \end{cases} \tag{5.11}$$

The non-trivial task which remains is to evaluate the conditional expectations $\mathbb{E}_{\tilde{L}_0}[N_{ij}(T)|Y]$ and $\mathbb{E}_{\tilde{L}_0}[R_i(T)|Y]$, respectively. The first step towards their computation is the observation that by the Markov property and the homogeneity of the Markov jump process the conditional expectations in (5.10) can be expressed as sums [9]

$$\begin{aligned} \mathbb{E}_{\tilde{L}_0}[R_i(T)|Y] &= \sum_{s=1}^{r} \sum_{k,l=1}^{d} c_{kl}(\tau_s) \mathbb{E}_{\tilde{L}_0}[R_i(\tau_s)|X(\tau_s) = l, X(0) = k], \\ \mathbb{E}_{\tilde{L}_0}[N_{ij}(T)|Y] &= \sum_{s=1}^{r} \sum_{k,l=1}^{d} c_{kl}(\tau_s) \mathbb{E}_{\tilde{L}_0}[N_{ij}(\tau_s)|X(\tau_s) = l, X(0) = k]. \end{aligned} \tag{5.12}$$

Next, the conditional expectations in the right hand sides in (5.12) can be decomposed further by using the identities

$$\begin{aligned} \mathbb{E}_L[R_i(t)|X(t) = l, X(0) = k] &= \frac{\mathbb{E}_L\left[R_i(t)\mathbf{1}_{\{l\}}(X(t))|X(0) = k\right]}{p_{kl}(t)}, \\ \mathbb{E}_L[N_{ij}(t)|X(t) = l, X(0) = k] &= \frac{\mathbb{E}_L\left[N_{ij}(t)\mathbf{1}_{\{l\}}(X(t))|X(0) = k\right]}{p_{kl}(t)}. \end{aligned} \tag{5.13}$$

Finally, the authors in [5, 9] realized that the auxiliary functions defined by

$$\begin{aligned} M_{kl}^{i}(t) &\stackrel{def}{=} \mathbb{E}_L\left[R_i(t)\mathbf{1}_{\{l\}}(X(t))|X(0) = k\right], \\ F_{kl}^{ij}(t) &\stackrel{def}{=} \mathbb{E}_L\left[N_{ij}(t)\mathbf{1}_{\{l\}}(X(t))|X(0) = k\right] \end{aligned} \tag{5.14}$$

satisfy systems of ordinary differential equations. For example, let $i, j \in S$ be fixed. Then the vectors $M_k^i(t) = (M_{k1}^i(t), \ldots, M_{kd}^i(t))$ and $F_k^{ij}(t) = (F_{k1}^{ij}(t), \ldots, F_{kd}^{ij}(t))$ satisfy the two systems of ODEs

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} M_k^i(t) &= M_k^i(t)L + A_k^i(t), \quad M_k^i(0) = 0 \\ &\text{with } A_k^i(t) = p_{ki}(t)e_i, \\ \frac{\mathrm{d}}{\mathrm{d}t} F_k^{ij}(t) &= F_k^{ij}(t)L + A_k^{ij}(t), \quad F_k^{ij}(0) = 0 \\ &\text{with } A_k^{ij}(t) = l_{ij}p_{ki}(t)e_j, \end{aligned} \tag{5.15}$$

where e_i and e_j are the i^{th} and j^{th} unit vectors. To summarize, the computation of the function $\mathcal{G}(L; \tilde{L})$ in the E-step reduces to solving the systems of ODEs given in (5.15). Solving these ODEs numerically, however, causes prohibitive computational costs when the number of states of the system is large. Another option is to approximate the matrix-exponentials which are involved in the analytic solutions of (5.15)

$$\begin{aligned} M_k^i(t) &= \int_0^t A_k^i(s) \exp((t-s)L)\mathrm{d}s, \\ F_k^{ij}(t) &= \int_0^t A_k^{ij}(s) \exp((t-s)L)\mathrm{d}s \end{aligned} \tag{5.16}$$

via the so-called uniformization method [67]. Choose $\alpha = \max_{i=1,\ldots,d}\{-l_{ii}\}$, and define $B = I + \alpha^{-1}L$. Then, e.g., $M^i(t) = (M_{kl}^i(t))_{k,l \in S}$ is given by

$$M^i(t) = \exp(-\alpha t)\alpha^{-1} \sum_{n=0}^{\infty} \frac{(\alpha t)^{n+1}}{(n+1)!} \sum_{j=0}^{n} B^j (e_i e_i^T) B^{n-j}.$$

with e_i^T denoting the transpose of the unit vector e_i. However this expansion is fairly time consuming and for high dimensional matrices intractable. Moreover the infinite sum has to be cut off at a finite n which entails inaccuracies.

We will choose an alternative way to compute the left hand sides in (5.13) which avoids the treatment of the ODEs. We will explain the approach in detail in the next subsection. In Algorithm 4, we state the resulting EM-algorithm due to [9].

5.2.3. Enhanced Computation of the Maximum Likelihood Estimator

In [48], the authors showed that the conditional expectations $\mathbb{E}_L[N_{ij}(t)|X(t) = l, X(0) = k]$ and $\mathbb{E}_L[R_i(t)|X(t) = l, X(0) = k]$ can analytically be expressed in terms of the generator L. Recalling the notation of the transition matrix $P(s) = \exp(sL)$, they proved the identities

$$\begin{aligned} \mathbb{E}_L[R_i(t)|X(t) = l, X(0) = k] &= \frac{1}{p_{kl}(t)} \int_0^t p_{ki}(s)p_{il}(t-s)\mathrm{d}s, \\ \mathbb{E}_L[N_{ij}(t)|X(t) = l, X(0) = k] &= \frac{l_{ij}}{p_{kl}(t)} \int_0^t p_{ki}(s)p_{jl}(t-s)\mathrm{d}s. \end{aligned} \tag{5.17}$$

The crucial observation is now that an eigendecomposition of the generator L leads to closed form expressions of the integrals in (5.17). To be more precise, consider

Algorithm 4 MLE-method (Bladt,Sørensen,[5, 9])

Input: Time series $Y = \{y_0 = X(t_0), \dots, y_N = X(t_N)\}$, initial guess of generator $\tilde{L}_0$.

Output: MLE $\tilde{L}$.

(1) E-step: **FOR ALL** $\tau_s \in \{\tau_1, \dots, \tau_r\}$ **DO**

i) Compute for $i, j, l, k = 1, \dots, d$ the conditional expectations
$\mathbb{E}_{\tilde{L}_k}[R_i(\tau_s)|X(\tau_s) = l, X(0) = k]$,
$\mathbb{E}_{\tilde{L}_k}[N_{ij}(\tau_s)|X(\tau_s) = l, X(0) = k], i \neq j$ via (5.15),(5.13).
END FOR

ii) Compute $\mathbb{E}_{\tilde{L}_k}[R_i(T)|Y]$ and $\mathbb{E}_{\tilde{L}_k}[N_{ij}(T)|Y]$ via (5.12).

(2) M-Step: Setup the next guess $\tilde{L}_{k+1}$ of the generator by

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

(3) Go to Step (1), unless a certain convergence criterion is satisfied.

the eigendecomposition of a generator L, that is

$$L = U D_\lambda U^{-1}, \tag{5.18}$$

where the columns of the matrix U consist of all eigenvectors to the corresponding eigenvalues of L in the diagonal matrix $D_\lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$. 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 (5.17), that is [49]

$$\int_0^t p_{ab}(s) p_{cd}(t-s) \mathrm{d}s = \sum_{p=1}^{d} u_{ap} u_{pb}^{-1} \sum_{q=1}^{d} u_{cq} u_{qd}^{-1} \Psi_{pq}(t), \tag{5.19}$$

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

$$\Psi_{pq}(t) = \begin{cases} t e^{t\lambda_p} & \text{if } \lambda_p = \lambda_q \\ \frac{e^{t\lambda_p} - e^{t\lambda_q}}{\lambda_p - \lambda_q} & \text{if } \lambda_p \neq \lambda_q. \end{cases} \tag{5.20}$$

Remark 5.2.1. *We are aware of the fact that in general an eigenvalue decomposition does not have to exist. For example, consider the matrix*

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

The characteristic polynomial of L is $-x(x+5)^2$, hence -5 is an eigenvalue with multiplicity two but the dimension of the corresponding eigenspace is one. However, in all of our numerical experiments this non-decomposable case did not show up.

For the convenience of the reader we state the resulting *enhanced MLE-method* in Algorithm 5. In a single iteration step for each single observation time lag, d^2 conditional expectations have to be computed where each one is decomposed into d^2 conditional expectations. Hence, the computational cost of a single iteration step in Algorithm 4 and in Algorithm 5 is $\mathcal{O}(r \cdot d^4 \cdot T_{\mathbb{E}})$ where r is the number of the different observed time lags and $T_{\mathbb{E}}$ denotes the computational cost to compute a single conditional expectation in the E-Step. The numerical considerations in [9] lead to a total computational cost per iteration in Algorithm 4 of at least $\mathcal{O}(r \cdot d^6)$. According to the closed form expressions for the expectations (5.19), the computational cost of a single iteration in the enhanced MLE-method (Algorithm 5) is $\mathcal{O}(r \cdot d^5)$ which is achieved by a simultaneously computation of the unknowns via matrix multiplication. For example, define for a fixed $i \in S$ the matrix $M^i_{kl} = \mathbb{E}_L\left[R_i(\tau)|X(\tau) = l, X(0) = k\right]$. Let U_i^{-1} denote the i^{th} row of the matrix U^{-1} and U_i the i^{th} column of U. Then M^i can be computed by

$$M^i = U\left[(U_i^{-1}U_i) * \Psi\right]U^{-1},$$

where $A * B$ is the Hadamard (entrywise) product of two matrices A and B. We want to emphasize that the algorithm works in principal 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.

Algorithm 5 Enhanced MLE-method

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 (5.18) of $\tilde{L}_k$.

(2) E-step: **FOR ALL** $\tau_s \in \{\tau_1, \ldots, \tau_r\}$ **DO**

i) Compute the auxiliary matrix $\Psi(\tau_s)$ (5.20).

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 (5.19),(5.17).
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 (5.12).

(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 $\|\tilde{L}_{k+1} - \tilde{L}_k\| < TOL$.

5.2.4. Reversible Case

In the reversible case the homogeneous Markov jump process, given by its generator L, admits a unique stationary distribution $\pi = (\pi_i)_{i \in S}$ and, moreover, *detailed balance* holds:

$$l_{ji} = \frac{\pi_i}{\pi_j} l_{ij}.$$

This has two important consequences for the EM-algorithm. The first one is that detailed balance guarantees a special representation of L which improves the stability

and accuracy of the EM-algorithm. Furthermore, one has to take into account that the M-step in general does not preserve the reversibility. To understand the first issue, notice that L can be written as

$$L = D_\pi^{-1/2} \mathbf{S} D_\pi^{1/2} \tag{5.21}$$

with a symmetric matrix $\mathbf{S}$ which can be decomposed as

$$\mathbf{S} = V D_\lambda V^T,$$

where $\lambda_1, \dots, \lambda_d \in \mathbb{R}$ are the eigenvalues of $\mathbf{S}$ and V is an orthogonal matrix, i.e. $VV^T = I$. Combining things, we end up with [48]

$$P(t) = D_\pi^{-1/2} V \exp(D_\lambda) V^T D_\pi^{1/2},$$

where $D_\pi^{1/2} = \operatorname{diag}(\sqrt{\pi_1}, \dots, \sqrt{\pi_d})$. Consequently, the integrals in (5.17) reduce to

$$\int_0^t p_{ab}(s) p_{cd}(t-s) \mathrm{d}s = \left(\frac{\pi_b \pi_d}{\pi_a \pi_c} \right)^{1/2} \sum_{p=1}^{d} v_{ap} v_{bp} \sum_{q=1}^{d} v_{cq} v_{dq} \Psi_{pq}(t), \tag{5.22}$$

where Ψ is defined in (5.20).

Next, we turn our attention to the problem of the non-preservation of the reversibility in the M-Step. The first idea could be to exploit the fact that detailed balance implies the bisection of the unknowns because l_{ji} is determined by π_i, π_j and l_{ij}. Then one could proceed as follows: Firstly, compute the MLE $\tilde{L}$ via the EM-algorithm as usual and then define a reversible generator $\tilde{L}^{REV} = (\tilde{l}_{ij}^{REV})_{i,j \in S}$ by

$$\tilde{l}_{ij}^{REV} = \begin{cases} \tilde{l}_{ij} & \text{if } i \leq j \\ \dfrac{\pi_j}{\pi_i} \tilde{l}_{ji} & \text{otherwise.} \end{cases}$$

This would work in principle but it does not guarantee that the resulting generator $\tilde{L}^{REV}$ is the MLE *subject* to the space of reversible generators. As a remedy, we include the restriction to that space explicitly in the log-likelihood function (5.10) via Lagrange multiplier:

$$\mathcal{G}^{REV}(L; \tilde{L}_0) = \mathcal{G}(L; \tilde{L}_0) + \sum_{i=1}^{d} \sum_{j>i}^{d} \mu_{ij} \left(\pi_i l_{ij} - \pi_j l_{ji} \right).$$

Performing the usual steps, we end up with the MLE $\tilde{L}^{REV}$, given by

$$\tilde{l}_{ij}^{REV} = \begin{cases} \dfrac{\mathbb{E}_{\tilde{L}_0}[N_{ij}(T)|Y]}{-\mu_{ij}\pi_i + \mathbb{E}_{\tilde{L}_0}[R_i(T)|Y]}, & i < j \\ \dfrac{\pi_i}{\pi_j} \tilde{l}_{ij}^{REV}, & \text{otherwise} \end{cases} \tag{5.23}$$

where the Lagrange multiplier can be determined by

$$\begin{aligned} \mu_{ij} = &\left[\frac{\mathbb{E}_{\tilde{L}_0}[R_j(T)|Y]}{\pi_j \mathbb{E}_{\tilde{L}_0}[N_{ji}(T)|Y]} - \frac{\mathbb{E}_{\tilde{L}_0}[R_i(T)|Y]}{\pi_i \mathbb{E}_{\tilde{L}_0}[N_{ij}(T)|Y]} \right] \\ &\times \left[- \frac{\mathbb{E}_{\tilde{L}_0}[N_{ij}(T)|Y] \cdot \mathbb{E}_{\tilde{L}_0}[N_{ji}(T)|Y]}{\mathbb{E}_{\tilde{L}_0}[N_{ij}(T)|Y] + \mathbb{E}_{\tilde{L}_0}[N_{ji}(T)|Y]} \right]. \end{aligned} \tag{5.24}$$

Combining both issues leads to Algorithm 6.

Algorithm 6 Enhanced MLE-method for the reversible case

Input: Time series $Y = \{y_0 = X(t_0), \dots, y_N = X(t_N)\}$, the set of observed time lags $\{\tau_1, \dots, \tau_r\}$, the tolerance TOL, initial guess of reversible generator $\tilde{L}_0^{REV}$.

Output: Reversible MLE $\tilde{L}^{REV}$.

(1) Compute eigendecomposition (5.21) of $\tilde{L}_k^{REV}$.

(2) E-step: **FOR ALL** $\tau_s \in \{\tau_1, \dots, \tau_r\}$ **DO**

i) Compute the auxiliary matrix $\Psi(\tau_s)$ (5.20).

ii) Compute for $i, j, l, k = 1, \dots, d$ the conditional expectations
$\mathbb{E}_{\tilde{L}_k^{REV}}[R_i(\tau_s)|X(\tau_s) = l, X(0) = k]$,
$\mathbb{E}_{\tilde{L}_k^{REV}}[N_{ij}(\tau_s)|X(\tau_s) = l, X(0) = k], i \neq j$ via (5.22),(5.17).
END FOR

iii) Compute $\mathbb{E}_{\tilde{L}_k^{REV}}[R_i(T)|Y]$ and $\mathbb{E}_{\tilde{L}_k^{REV}}[N_{ij}(T)|Y]$ via (5.12).

(4) Compute Lagrange multipliers μ_{ij} via (5.24).

(5) M-Step: Setup the next guess $\tilde{L}_{k+1}^{REV}$ of the generator via (5.23).

(4) Go to Step (1) unless $\|\tilde{L}_{k+1}^{REV} - \tilde{L}_k^{REV}\| < TOL$.

5.2.5. Scaling

We prove that the maximizer (5.11) in the (enhanced) MLE-method respects the time invariance of the semigroup $P(t) = \exp(tL)$. Consequently, in the case of a constant observation time step τ we can estimate a generator $\tilde{L}(\tau')$ with respect to $\tau' = 1$ and regain the generator with respect to τ by $\tilde{L}(\tau) = \tilde{L}(1)/\tau$.

Lemma 5.2.2. *Let $\tilde{L}(\tau)$ be the MLE with respect to the time lag τ and $\tilde{L}(1)$ with respect to $\tau' = 1$. Then for both cases the general and the reversible case the following relation holds:*

$$\tilde{L}(\tau) = \frac{1}{\tau}\tilde{L}(1). \tag{5.25}$$

Proof:
A short calculation shows that

$$\int_0^\tau p_{ab}(s)p_{cd}(\tau - s)\mathrm{d}s = \tau \int_0^1 [\exp(s\bar{L})_{ab}(\exp((1-s)\bar{L})_{cd}]\mathrm{d}s,$$

where $\bar{L} = \tau L$. But this immediately implies

$$\mathbb{E}_L[R_i(\tau)|X(\tau) = l, X(0) = k] = \tau\mathbb{E}_{\bar{L}}[R_i(1)|X(1) = l, X(0) = k]$$

and, by noting that $l_{ij} = \frac{1}{\tau}\bar{l}_{ij}$,

$$\mathbb{E}_L[N_{ij}(\tau)|X(\tau) = l, X(0) = k] = \mathbb{E}_{\bar{L}}[N_{ij}(1)|X(1) = l, X(0) = k]$$

which proves (5.25). In the reversible case, the same reasoning shows that the Lagrange multipliers scale linearly with τ and therefore (5.25) also holds.

5.2.6. Enhanced MLE-Method vs. MLE-Method

The eigendecomposition approach has several advantages compared to the numerical considerations proposed in [9]. Let d be the dimension of the discrete state space. As

explained in Section 5.2.2, the computational cost is reduced to $\mathcal{O}(r \cdot d^5)$ thanks to the closed form expression (5.19). Moreover, there is no longer an explicit dependency on the length of the time series. The second advantage is the exact computation of the conditional expectations involved in the E-step of the EM-algorithm. The steps which introduce numerical errors are the eigendecomposition and the computation of U^{-1}. As before, the explicit inversion of U can be avoided by considering the left eigenvectors of $\tilde{L}$. We are aware that the eigendecomposition of non-symmetric matrices can be *ill-conditioned*, but any reliable numerical solver should indicate this. Nevertheless, the computational cost of both steps ($\mathcal{O}(d^3)$) and their numerical stability are superior compared to any numerical approximation scheme for solving the ODEs in (5.15).

5.3. An Alternative Approach: The Quadratic Optimization Method

The approach introduced by Crommelin and Vanden-Eijnden [19] yields an estimate $\tilde{L}$ such that the spectral properties of the empirical transition matrix $\hat{P}$, i.e. eigenvalues and eigenvectors, are well approximated by the spectral properties of $\exp(\tau\tilde{L})$.

Let $\hat{P} \approx P(\tau)$ be the approximative transition matrix computed by Equation (5.8). Now suppose an eigendecomposition

$$\hat{P} = U\Lambda U^{-1} \tag{5.26}$$

with a diagonal matrix $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ containing the eigenvalues exists, and that $\lambda_k \neq 0$ for all k. (Note that U^{-1} can be obtained without explicit matrix inversion since its rows are the left eigenvectors of $\hat{P}$.) Then, the matrix

$$\tilde{L} \quad = \quad UZU^{-1} \quad \text{with} \quad Z = \mathrm{diag}(z_1, \ldots, z_d), \quad z_k = \frac{\log(\lambda_k)}{\tau} \tag{5.27}$$

can be defined, and the approximative transition matrix can be expressed in terms of the matrix exponential

$$\exp(\tau\tilde{L}) = \exp\left(U \log(\Lambda) U^{-1}\right) = U\Lambda U^{-1} = \hat{P}.$$

In spite of this relation, $\tilde{L}$ cannot be considered as a reasonable estimate for the generator because $\tilde{L} \notin \mathfrak{G}$ in many cases. In order to find an estimate with the correct structural properties, Crommelin and Vanden-Eijnden propose to compute the generator $\tilde{L} \in \mathfrak{G}$ which agrees best with the eigendecomposition (5.27). This is motivated by the fact that many properties of a continuous-time Markov chain (such as, e.g., its stationary distribution) depend strongly on the eigenvalues and eigenvectors of its generator. Therefore, in [19] the generator is estimated by solving the quadratic minimization problem

$$\begin{aligned} \tilde{L}_{QP} = \arg\min_{L \in \mathfrak{G}} \sum_{k=1}^{d} \Big(\alpha_k |U_k^{-1} L - z_k U_k^{-1}|^2 \quad &+ \quad \beta_k |LU_k - z_k U_k|^2 \\ &+ \quad \gamma_k |U_k^{-1} L U_k - z_k|^2 \Big), \end{aligned} \tag{5.28}$$

where U_k denotes the k^{th} column of U, U_k^{-1} is the k^{th} row of U^{-1}, and

$$\alpha_k = a_k |z_k U_k^{-1}|^{-2}, \quad \beta_k = b_k |z_k U_k|^{-2} \quad \text{and} \quad \gamma_k = c_k |z_k|^{-2}$$

are weights with suitably chosen coefficients a_k, b_k, c_k. The problem (5.28) can be solved with a standard quadratic optimizer such as the MATLAB `quadprog` command after reformulating (5.28) as

$$\tilde{L}_{QP} = \arg\min_{L \in \mathfrak{G}} \frac{1}{2} \langle L, HL \rangle + \langle F, L \rangle + E_0$$

with a tensor $H \in \mathbb{R}^{d \times d \times d \times d}$ and a matrix $F \in \mathbb{R}^{d \times d}$; see [19] for details. If d is so large that the tensor H cannot be stored, the problem (5.28) can still be solved with `quadprog`, but this requires a function for the evaluation of Hv for arbitrary v without composing H explicitly.

5.4. Numerical Examples for Equidistant Observation Times

5.4.1. Preparatory Considerations

In order to compare the performance of the quadratic programming approach (QP) and the maximum likelihood method (MLE), we will first restrict ourselves to the case of equidistant observation times and we will apply the approaches to a series of model problems. In Section 5.5, we will focus on the case of non-equidistant observation times.

A rather straightforward test would proceed as follows:

1. Choose an arbitrary generator $L \in \mathfrak{G}$ and a time lag τ.
2. Compute the corresponding transition matrix $P(\tau) = \exp(\tau L)$.
3. Produce a time series $Y = \{y_0 = X(t_0), \ldots, y_N = X(t_N)\}$ by sampling from $P(\tau)$.
4. Pass this data to each of the two methods and compute an estimate $\tilde{L} \approx L$.
5. Compare the errors of the two approaches.

Although such a test seems to be somewhat reasonable, we will *not* use this procedure. The reason for our refusal is the fact that the time series produced in step 3 is just a single realization. Hence, the result of this test is random, too, and applying the test several times to the methods yields different results even though the input L remains unchanged. In fact, both methods are affected by the sampling error

$$\|P(\tau) - \hat{P}\| \qquad \text{with} \quad \hat{P} = (\hat{p}_{ij})_{i,j} \quad \text{and} \quad \hat{p}_{ij} = \frac{c_{ij}}{\sum_{j=1}^{d} c_{ij}}. \tag{5.29}$$

(Here and below, $\|\cdot\|$ denotes the matrix 2-norm.) Roughly speaking, the sampling error indicates how well the frequency matrix of a time series "represents" the underlying transition matrix. In the limit $N \to \infty$ one may expect the sampling error to vanish, but for a finite number of observations the deviation can be considerable. Since the outcome of a numerical method cannot be better than the input data, the error of both methods are bounded from below by the sampling error.

Therefore, our numerical experiments are designed in a different way:

1. a) Choose a generator $L \in \mathfrak{G}$ and a time lag τ and compute the corresponding transition matrix $P(\tau) = \exp(\tau L)$,

   or

   b) choose a transition matrix P. This allows to test the performance of the methods in situations where no underlying generator exists. In this case, the time lag does not matter, and we can set $\tau = 1$.

2. Define a virtual frequency matrix by multiplying each row of the transition matrix $P(\tau)$ with the corresponding entry of the stationary distribution $\pi = (\pi_i), i \in S$ and the length N of the (virtual) time series:

$$c_{ij} = \text{round}(N\pi_i p_{ij}). \tag{5.30}$$

   This is the frequency matrix which, up to rounding errors, reflects the underlying transition matrix in an optimal way.

3. Based on the virtual frequency matrix, define the virtual transition matrix

$$\hat{P}_{virt} = (\hat{p}_{ij})_{i,j} \quad \text{and} \quad \hat{p}_{ij} = \frac{c_{ij}}{\sum_{j=1}^{d} c_{ij}} \tag{5.31}$$

   and compute an estimate $\tilde{L} \approx L$ for the generator.

4. For both methods, compute and compare the errors:
   a) $\|\tilde{L} - L\|$ (only if L is available, i.e. if variant (a) of step 1 was used)
   b) $\|P(\tau) - \exp(\tau\tilde{L})\|$
   c) $\|\hat{P}_{virt} - \exp(\tau\tilde{L})\|$ with $\hat{P}_{virt}$ defined in (5.31).

The advantage of this approach to numerical experiments is illustrated by a simple example in the next section.

Of course, the choice of the initial value $\tilde{L}_0$ for the MLE-method is crucial for the convergence. If the matrix logarithm of $\hat{P}_{virt}$ exists, then a good initial value $\tilde{L}_0$ can easily be obtained by taking the absolute values of the off-diagonal entries of $\log(\hat{P}_{virt})/\tau$ and setting the diagonal entries to the corresponding negative row sums, respectively.

A Simple Example Illustrating the Effect of Sampling Errors

This example illustrates the influence of the sampling error on the optimal generator estimate. The transition matrix of the generator

$$L = \begin{pmatrix} -0.2 & 0.2 \\ 0.2 & -0.2 \end{pmatrix}$$

with respect to the time lag $\tau = 1$ is

$$P(\tau) = \begin{pmatrix} 0.8352 & 0.1648 \\ 0.1648 & 0.8352 \end{pmatrix}.$$

Suppose that sampling according to the transition matrix produces the time series

time t_n	0	1	2	3	4	5	6	7	8	9	10
state $X(t_n)$	2	1	1	1	1	2	2	1	1	2	1

such that the corresponding frequency matrix is

$$C = \begin{pmatrix} 4 & 2 \\ 3 & 1 \end{pmatrix}.$$

According to this data, the transition matrix seems to be

$$\hat{P} = \begin{pmatrix} 2/3 & 1/3 \\ 3/4 & 1/4 \end{pmatrix} \tag{5.32}$$

and since $\hat{P} = \exp(\hat{L})$ with

$$\hat{L} \approx \begin{pmatrix} -0.5003 & 0.5003 \\ 0.3752 & -0.3752 \end{pmatrix} \in \mathfrak{G} \tag{5.33}$$

the best result we can expect to obtain based on the time series is $\hat{L}$ instead of L. The errors $\|\hat{P} - P\| \approx 0.2670$ and $\|\hat{L} - L\| \approx 0.4916$ are caused by the time series and cannot be avoided by the two methods. However, these errors decrease if, according to the second test procedure, the frequency matrix is replaced by the virtual frequency matrix (5.30). Since in our example the stationary distribution is $\pi = (0.5, 0.5)$, one obtains

$$C = \begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix}.$$

The corresponding transition matrix

$$\hat{P}_{virt} = \begin{pmatrix} 0.8 & 0.2 \\ 0.2 & 0.8 \end{pmatrix}$$

is obviously a better approximation of the true transition matrix P than (5.32), and the generator estimate

$$\tilde{L} = \log(\hat{P}_{virt}) \approx \begin{pmatrix} -0.2554 & 0.2554 \\ 0.2554 & -0.2554 \end{pmatrix}$$

is evidently better than (5.33). In fact, the new errors are only $\|P - \hat{P}_{virt}\| \approx 0.0703$ and $\|L - \tilde{L}\| \approx 0.1108$.

5.4.2. Transition Matrix with Underlying Generator

In a first example we follow variant (a) of step 1 and consider the generator

$$L = \begin{pmatrix} -4.29 & 0.678 & 0.301 & 0.819 & 0.592 & 0.149 & 0.543 & 0.411 & 0.774 & 0.023 \\ 0.033 & -3.83 & 0.633 & 0.260 & 0.636 & 0.878 & 0.485 & 0.527 & 0.147 & 0.231 \\ 0.857 & 0.995 & -5.46 & 0.704 & 0.532 & 0.021 & 0.441 & 0.920 & 0.148 & 0.845 \\ 0.682 & 0.499 & 0.005 & -4.69 & 0.208 & 0.923 & 0.626 & 0.379 & 0.639 & 0.726 \\ 0.801 & 0.430 & 0.816 & 0.082 & -4.26 & 0.632 & 0.077 & 0.638 & 0.093 & 0.694 \\ 0.917 & 0.829 & 0.690 & 0.875 & 0.241 & -5.58 & 0.544 & 0.173 & 0.928 & 0.383 \\ 0.388 & 0.116 & 0.981 & 0.077 & 0.720 & 0.632 & -4.66 & 0.785 & 0.485 & 0.479 \\ 0.472 & 0.598 & 0.069 & 0.741 & 0.400 & 0.753 & 0.270 & -4.43 & 0.163 & 0.967 \\ 0.088 & 0.221 & 0.045 & 0.125 & 0.394 & 0.769 & 0.291 & 0.776 & -3.49 & 0.783 \\ 0.925 & 0.398 & 0.740 & 0.443 & 0.411 & 0.808 & 0.822 & 0.342 & 0.131 & -5.02 \end{pmatrix} \in \mathfrak{G}. \tag{5.34}$$

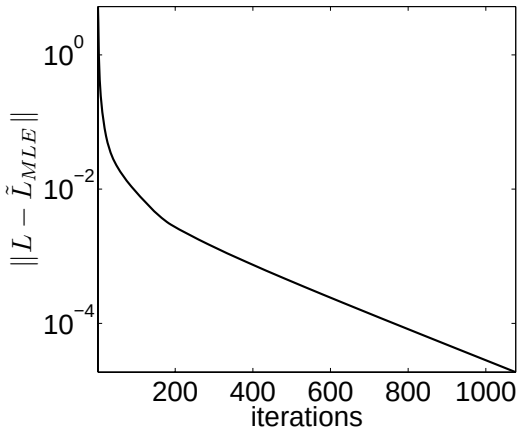

Figure 5.1.: Approximation error of $\tilde{L}_{MLE}$ with respect to the generator L in (5.34) as a function of the iteration steps.

	$\|L - \tilde{L}\|$	$\|\exp(\tau L) - \exp(\tau\tilde{L})\|$	$\|\hat{P}_{virt} - \exp(\tau\tilde{L})\|$
QP	$2.07 \cdot 10^{-8}$	$1.39 \cdot 10^{-9}$	$1.18 \cdot 10^{-14}$
MLE	$1.88 \cdot 10^{-5}$	$1.19 \cdot 10^{-6}$	$1.19 \cdot 10^{-6}$

Table 5.1.: Approximation errors of the estimated generators $\tilde{L}_{QP}$ and $\tilde{L}_{MLE}$ with respect to the given generator (5.34), the exact transition matrix $P(\tau)$ and the transition matrix $\hat{P}_{virt}$ constructed via (5.31). Results for the time lag $\tau = 0.2$ and the length of the virtual time series $N = 10^{10}$.

Based on the exact transition matrix $P(\tau)$ with $\tau = 0.2$, we computed the virtual transition matrix $\hat{P}_{virt}, N = 10^{10}$ via (5.31) and estimated the generator with both methods. The enhanced MLE-method (5) stopped after 1132 iteration steps because the increment-based stopping criterion $\|\tilde{L}_k - \tilde{L}_{k-1}\| \leq tol$ with $tol = 10^{-7}$ had been met. Figure 5.1 shows the error of $\tilde{L}_{MLE}$ with respect to L (5.34) as a function of the iteration steps.

Obviously, the convergence of the enhanced MLE-method is very slow. In contrast to the MLE-method, the QP-method converged after only one iteration step. In Table 5.1 the errors of both approaches are compared. The QP-approach approximates the original generator clearly better than the enhanced MLE-method. This is, however, not surprising because it has to be taken into account that the QP-approach approximates the eigendecomposition of $\hat{P}_{virt}$ and for the length $N = 10^{10}$ of a virtual time series the difference between the exact and the virtual transition matrix is only $\|P(0.2) - \hat{P}_{virt}\| = 1.39 \cdot 10^{-9}$.

Next, we investigate the influence of the sampling error on both estimation methods. Instead of considering realizations of the Markov jump process, we compute estimations of L for a number of virtual time series of increasing length N. Figure 5.2 shows the resulting errors of $\tilde{L}_{MLE}$ and $\tilde{L}_{QP}$ with respect to the generator L (5.34) as a function of the length N of the virtual time series. It reveals that for a time series of a realistic length ($N \leq 10^7$), the errors of $\tilde{L}_{QP}$ and $\tilde{L}_{MLE}$ are almost identical. The fact that the error of the MLE-method remains larger than 10^{-5} regardless of N is due to the chosen stopping criterion.

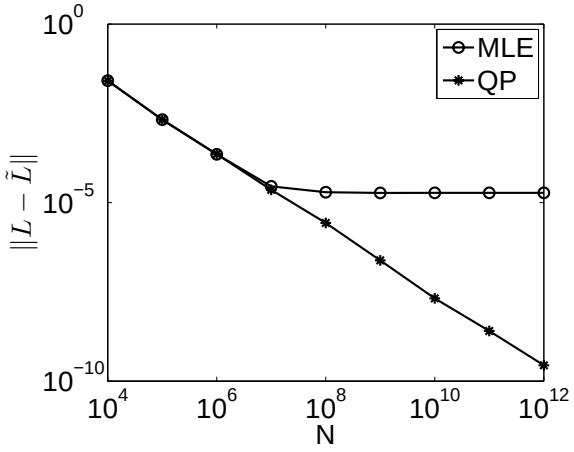

Figure 5.2.: Graphs of the errors of $\tilde{L}_{MLE}$ and $\tilde{L}_{QP}$ with respect to the generator L in (5.34), respectively, as a function of the length N of the virtual time series. The error of the MLE-method remains larger than 10^{-5} regardless of N due to the stopping criterion $\|\tilde{L}_k - \tilde{L}_{k-1}\| \leq 10^{-7}$.

	$\|P - \exp(\tau\tilde{L})\|$	$\|\hat{P}_{virt} - \exp(\tau\tilde{L})\|$
QP	$1.74 \cdot 10^{-2}$	$1.74 \cdot 10^{-2}$
MLE	$2.86 \cdot 10^{-2}$	$2.86 \cdot 10^{-2}$

Table 5.2.: Approximation errors of $\exp(\tilde{L}_{QP})$ and $\exp(\tilde{L}_{MLE})$ with respect to the given transition matrix (5.35) and the transition matrix $\hat{P}_{virt}$ constructed via (5.31). Results of MLE-method for $tol = 10^{-7}$.

5.4.3. Transition Matrix without Underlying Generator

In contrast to the first case both estimation procedures are now applied to a transition matrix which does not possess a generator:

$$P = \begin{pmatrix} 0.645 & 0.037 & 0.033 & 0.039 & 0.046 & 0.062 & 0.040 & 0.003 & 0.031 & 0.059 \\ 0.014 & 0.792 & 0.054 & 0.06 & 0.010 & 0 & 0 & 0 & 0.016 & 0.051 \\ 0.049 & 0.065 & 0.751 & 0.069 & 0.000 & 0 & 0 & 0 & 0.046 & 0.015 \\ 0.020 & 0.056 & 0.057 & 0.723 & 0.061 & 0 & 0 & 0 & 0.022 & 0.057 \\ 0.037 & 0.044 & 0.039 & 0.061 & 0.707 & 0 & 0 & 0 & 0.066 & 0.043 \\ 0.010 & 0.057 & 0.025 & 0.012 & 0.020 & 0.727 & 0.032 & 0.053 & 0.050 & 0.009 \\ 0 & 0 & 0 & 0.069 & 0.047 & 0.016 & 0.753 & 0.069 & 0.029 & 0.014 \\ 0 & 0 & 0 & 0.019 & 0.019 & 0.040 & 0.055 & 0.770 & 0.052 & 0.042 \\ 0 & 0 & 0 & 0.019 & 0.035 & 0.057 & 0.004 & 0.059 & 0.776 & 0.047 \\ 0 & 0 & 0 & 0.065 & 0.004 & 0.039 & 0.045 & 0.032 & 0.033 & 0.778 \end{pmatrix} \notin \mathfrak{P} \quad (5.35)$$

One can immediately verify via Theorem A.6.8 cited in the Appendix that (5.35) cannot be generated since, e.g, the state 6 is accessible from state 2 via state 1 but $p_{2,6} = 0$. As Table 5.2 shows, the errors of the estimated transition matrices $\exp(\tau\tilde{L})$ are of the same order of magnitude and are larger than in the first example due to the additional difficulty that no generator exists. The error $\|P - \exp(\tilde{L}_{MLE})\|$ as a function of the first 10 iteration steps is shown in the left panel of Figure 5.3. Surprisingly, the best accuracy is obtained after only one iteration, but the following iterations increase the error again. The reason for this behavior is the fact that the MLE-method aims to maximizing the likelihood instead of minimizing the error, and the graph of the discrete log-likelihood, depicted in the right panel of Figure 5.3, clearly shows that the maximum likelihood was not attained after the first iteration.

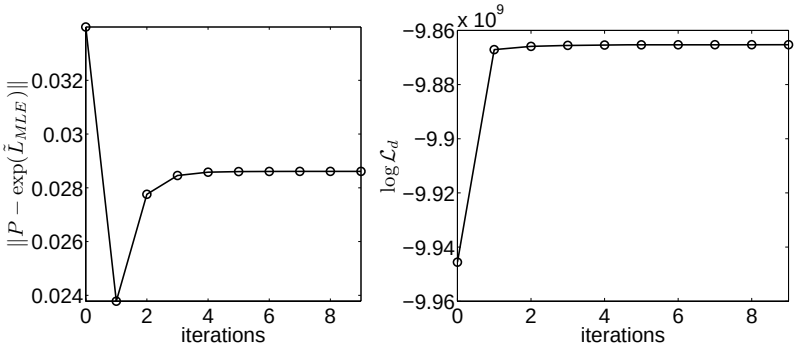

Figure 5.3.: Left: Error of $\exp(\tilde{L}_{MLE})$ with respect to the transition matrix P (5.35) as a function of the first 10 iteration steps. Right: The discrete log-likelihood $\mathcal{L}_d$ as a function of the 10 first iteration steps.

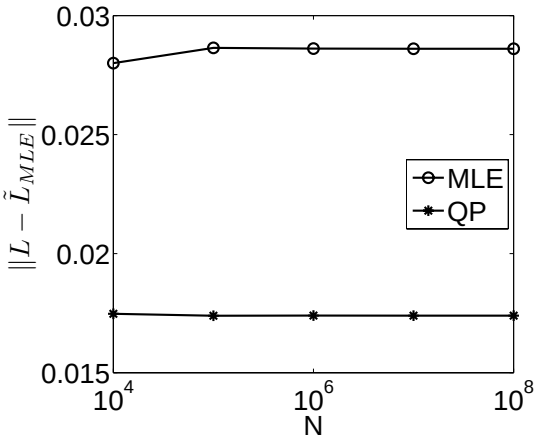

Figure 5.4.: The graphs of the errors of $\exp(\tilde{L}_{MLE})$ and $\exp(\tilde{L}_{QP})$ with respect to the transition matrix (5.35), respectively, as a function of the length N of the virtual time series.

In contrast to the first example, Figure 5.4 shows that here increasing the length of the virtual time series does not improve the estimation significantly in both methods.

5.4.4. Transition Matrix with Exact Generator under Perturbation

In the next example, we consider again the transition matrix $P(\tau)$ with $\tau = 0.2$ which is generated by the generator (5.34) given in the first example. In order to investigate the impact of perturbations due to, e.g., sampling from a time series, we estimate a generator based on a perturbed transition matrix

$$P_\epsilon(\tau) = \exp(\tau L) + k\epsilon, \quad k = 0, ..., 19,$$

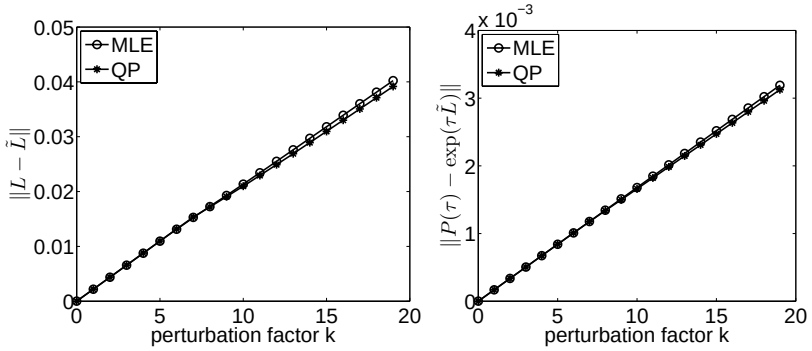

Figure 5.5.: Left: Approximation error of the generator estimates $\tilde{L}_{QP}$ and $\tilde{L}_{MLE}$ with respect to the unperturbed generator (5.34) as a function of the perturbation factor k. Right: Error of the estimated transition matrices $\exp(\tau\tilde{L}_{QP})$ and $\exp(\tau\tilde{L}_{MLE})$ with respect to the unperturbed transition matrix $\exp(\tau L)$ as a function of the perturbation factor k. Results for $\tau = 0.2$.

, where ϵ is the perturbation matrix

$$\epsilon = 10^{-5} \cdot \begin{pmatrix} 4.05 & -3.55 & 1.75 & 0.80 & -4.09 & -3.51 & 4.71 & 0.04 & 0.69 & -0.91 \\ 3.10 & -3.50 & -1.60 & 2.87 & 1.31 & -0.67 & 2.02 & 1.45 & 1.27 & -6.26 \\ -3.22 & -0.97 & -2.61 & 5.67 & -3.65 & 2.38 & 5.72 & -2.47 & 0.15 & -0.99 \\ 4.46 & -1.23 & -5.22 & 1.94 & -1.02 & -3.49 & 2.43 & -2.04 & 2.68 & 1.49 \\ 4.69 & -4.18 & -1.27 & 1.94 & -4.19 & -0.45 & -0.85 & 3.64 & -4.33 & 4.99 \\ 4.37 & -2.33 & -1.60 & 3.41 & 1.55 & 1.85 & -4.52 & -2.27 & 4.35 & -4.80 \\ 1.20 & -2.23 & 5.50 & -4.12 & -1.15 & -0.13 & -3.34 & -3.63 & 4.11 & 3.78 \\ 2.83 & -1.00 & 2.73 & -3.00 & -1.06 & -4.55 & 2.69 & 2.61 & 3.19 & -4.43 \\ -1.47 & 4.04 & -0.31 & -3.72 & -0.41 & 1.24 & 0.45 & -2.99 & -2.15 & 5.33 \\ -1.46 & -1.56 & 5.23 & -0.77 & -2.61 & 4.25 & -2.00 & -0.25 & 0.70 & -1.51 \end{pmatrix}.$$

The left panel of Figure 5.5 shows the deviation of the estimated generators from the unperturbed generator as a function of the perturbation factor k. The QP-method performs slightly better but both errors $\|L - \tilde{L}_{QP}\|$ and $\|L - \tilde{L}_{MLE}\|$ are of the same order of magnitude. Furthermore, the errors scale linearly with the perturbation factor k. This observation is plausible since for small perturbations the logarithm $\log(P + \epsilon)$ can be approximated by $\log(P) + \mathcal{O}(\epsilon)$. The right panel of Figure 5.5 illustrates the behavior of the errors of the estimated transition matrices $\exp(\tau\tilde{L}_{QP})$ and $\exp(\tau\tilde{L}_{MLE})$, respectively. A similar reasoning as above explains the linear scaling.

Finally, we consider the error of the estimated transition matrices $\exp(\tau\tilde{L}_{QP})$ and $\exp(\tau\tilde{L}_{MLE})$ with respect to the perturbed transition matrix $P_\epsilon(\tau) = \exp(\tau L) + k\epsilon$, depicted in Figure 5.6. Notice that the error $\|P_\epsilon(\tau) - \exp(\tau\tilde{L})\|$ is bounded from above, namely

$$\|P_\epsilon(\tau) - \exp(\tau\tilde{L})\| \leq \|\exp(\tau L) - \exp(\tau\tilde{L})\| + k\|\epsilon\|.$$

Indeed, Figure 5.6 shows that both errors obey that bound. For the perturbation factors up to $k = 8$, the matrix logarithm of P_ϵ is still a generator whereas for $k = 9, \ldots, 19$ the perturbation is apparently high enough to destroy the generator structure of the matrix logarithm of P_ϵ. However, the accuracy of both methods is again of the same order of magnitude.

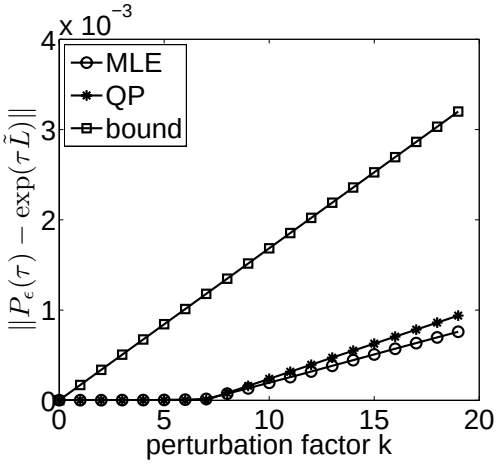

Figure 5.6.: Error of the estimated transition matrices $\exp(\tau\tilde{L}_{QP})$ and $\exp(\tau\tilde{L}_{MLE})$ with respect to the perturbed transition matrix $P_\epsilon(\tau) = \exp(\tau L) + k\epsilon$ as a function of the perturbation factor k. The upper bound was computed via $\tilde{L}_{MLE}$.

5.4.5. Application to a Time Series from Molecular Dynamics

In this example, we apply the enhanced MLE-method to a time series of two torsion angles extracted from a molecular simulation of glycine in water. The ball-and-stick representation of glycine together with the two considered torsion angles Φ and Ψ is shown in Figure 4.9(Sect. 4.3.2). The time series used herein was extracted out of a molecular simulation of the glycine-molecule embedded in a cubic box of edge length 3.51 nm with 1402 water molecules. The integration of the trajectory with total length $T = 5$ nanoseconds was realized with 2 femtoseconds time steps in the Leapfrog-integration scheme with GROMACS force field [8, 59] at room temperature of 300K. The left panel of Figure 5.7 shows the projection of the time series onto the torsion angles Φ and Ψ which reveals the metastable behavior. The Ramachandran plot of the time series, given in the right panel of Figure 5.7, illustrates the dependency among both torsion angles and indicates that the glycine-molecule attains four different main conformations.

As explained in Section 4.3.2, the identification of conformations amounts to identify metastable states in a coarse grained model of the dynamics. We considered a 20×20 box discretization of the torsion angle space which results in a state space of 374 visited boxes. In order to ensure the Markov property, we considered only every 100^{th} step of the original trajectory and estimated a reversible generator with respect to the reversible transition matrix $\hat{P}$ (cf. (5.8)) via the Algorithm 6. Moreover, instead of using the time lag $\tau = 2 \cdot 10^{-13}$, we performed the estimation with respect to $\tau = 1$ and re-scaled the resulting generator $\tilde{L}_{MLE}$ afterwards (cf. Sect. 5.2.5). As one can see in Figure 5.8, the estimation algorithm is already converged after 200 steps (here we used the log-likelihood function as an indicator for convergence).

In Table 5.3 we compare the dominant eigenvalues of the transition matrix $\hat{P}$ with those of $\tilde{P} = \exp(\tau\tilde{L}_{MLE})$. The spectral gap as well as the eigenvalues are more or less well reproduced. The panels of Figure 5.9 illustrate the decomposition

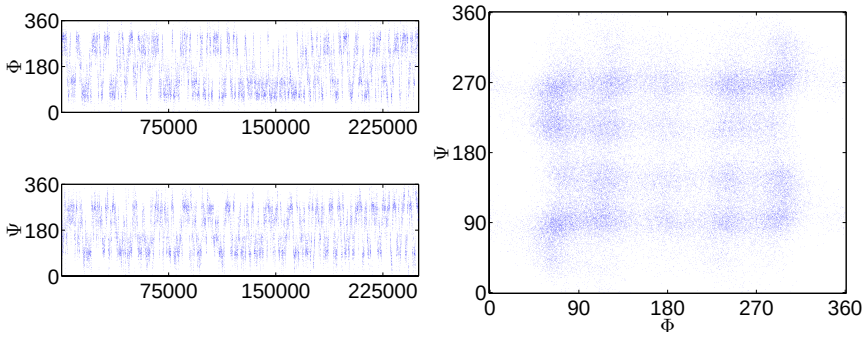

Figure 5.7.: Left: We show the projection of the time series (all atomic positions) on the torsion angles Φ and Ψ. Right: To illustrate the dependency of the torsion angles, we show the Ramachandran plot of the time series of the torsion angles Φ and Ψ. At first glance, the glycine-molecule attains four different main conformations indicated by the four clusters.

	λ_1	λ_2	λ_3	λ_4	λ_5
$\hat{P}$	1	0.9944	0.9942	0.9890	0.9718
$\exp(\tau\tilde{L}_{MLE})$	1	0.9988	0.9987	0.9977	0.9931

Table 5.3.: The five largest eigenvalues of the transition matrix $\hat{P}$ and the transition matrix computed from the estimated generator $\tilde{L}_{MLE}$. Results for an equidistant 20×20 box-discretization of the torsion angle space.

of the state space via PCCA which is based on the eigenvectors corresponding to the largest eigenvalues. The decomposition in the left panel results from $\hat{P}$ and the right panel corresponds to the transition matrix $\exp(\tau\tilde{L}_{MLE})$. The almost identical decompositions of the torsion angle space show that despite the slight deviations in the dominant eigenvalues, the estimated generator contains the essential information on the coarse grained dynamics. For a further analysis of the estimated Markov jump process via discrete TPT see Chapter 4.

5.5. Numerical Examples for Non-Equidistant Observation Times

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 $t_{k+1} > T$.

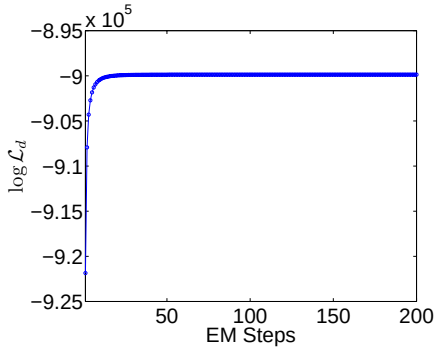

Figure 5.8.: The figure shows the log-likelihood function $\log \mathcal{L}_d$ as a function of the EM-steps. The constancy of the log-likelihood function indicates that the estimation procedure is converged already after 50 iteration steps.

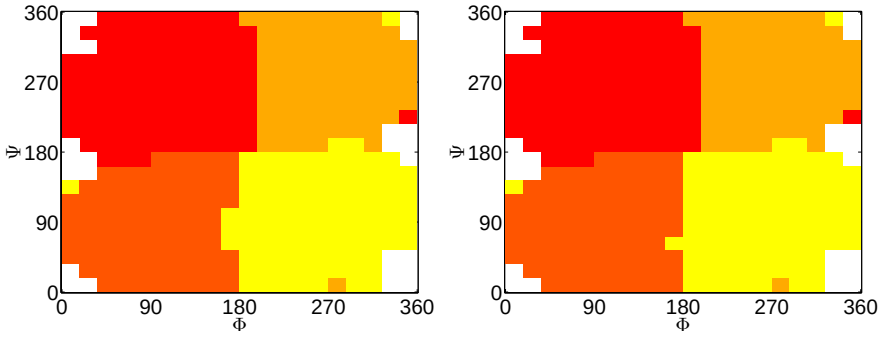

Figure 5.9.: Decomposition of the torsion angle state space into four metastable sets via PCCA. Left: Decomposition with respect to the observed transition matrix $\hat{P}$. Right: Decomposition with respect to the transition matrix $\tilde{P} = \exp(\tau \tilde{L}_{MLE})$. The decompositions are almost identical which indicates that the estimated process captures the essential dynamics in the coarse grained torsion angle space.

5.5.1. Test Example

In the first example we consider a five-state Markov jump process given by its generator

$$L = \begin{pmatrix} -6 & 2 & 2 & 1 & 1 \\ 1 & -4 & 0 & 1 & 2 \\ 1 & 0 & -4 & 2 & 1 \\ 2 & 1 & 0 & -3 & 0 \\ 1 & 1 & 1 & 1 & -4 \end{pmatrix} \in \mathfrak{G}. \tag{5.36}$$

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 (5.37) we state the estimated generator resulting from Algorithm 5 with the prescribed tolerance $TOL = 10^{-6}$. One clearly can see that $\tilde{L}$ approximates the original one very well.

$$\tilde{L} = \begin{pmatrix} -5.9803 & 2.0054 & 1.9863 & 0.9911 & 0.9975 \\ 1.0002 & -4.0018 & 0.0010 & 0.9938 & 2.0068 \\ 0.9921 & 0.0001 & -3.9768 & 1.9938 & 0.9909 \\ 1.9909 & 0.9951 & 0.0004 & -2.9871 & 0.0006 \\ 0.9982 & 1.0051 & 0.9993 & 1.0050 & -4.0075 \end{pmatrix} \in \mathfrak{G}. \tag{5.37}$$

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 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 5.10 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.

5.5.2. Application to a Genetic Toggle Switch Model

In the last 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 [77]. 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 generator in (4.47) describes their dynamics. For details on that process and its investigation via discrete TPT we refer the reader to Chapter 4, Section 4.3.3.

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

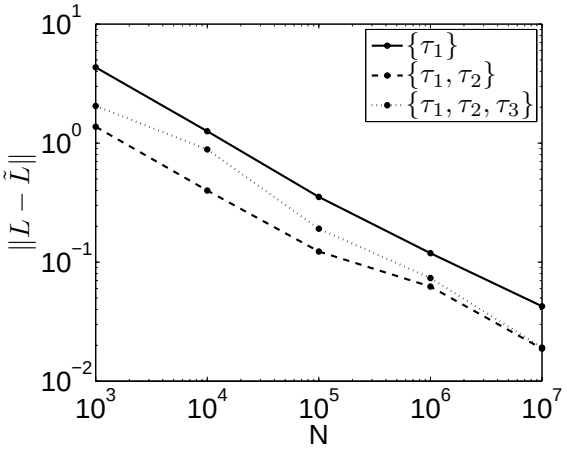

Figure 5.10.: Error of the estimated generator $\tilde{L}$ with respect to the original generator (5.36), measured in the 2-norm $\|\tilde{L} - L\|$, as a function of the length N of the respective time series. Results for the three different sets of time lags $\{0.01\}, \{0.01, 0.1\}$ and $\{0.01, 0.1, 1\}$ and the tolerance $TOL = 10^{-6}$.

choice the deterministic dynamics (4.48) 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 5.11 shows the discrete free energy, $-\log \pi$, of the jump process instead of its stationary distribution π itself. All states with almost vanishing stationary distribution are depicted by the white region and 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 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 $(x, y) = (0, 8)$.

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

$$\frac{\partial p(t)}{\partial t} = L^T p(t), \quad p(0) = p_0, \ t > 0, \tag{5.38}$$

where L^T denotes the transpose of the generator given in (4.47). 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

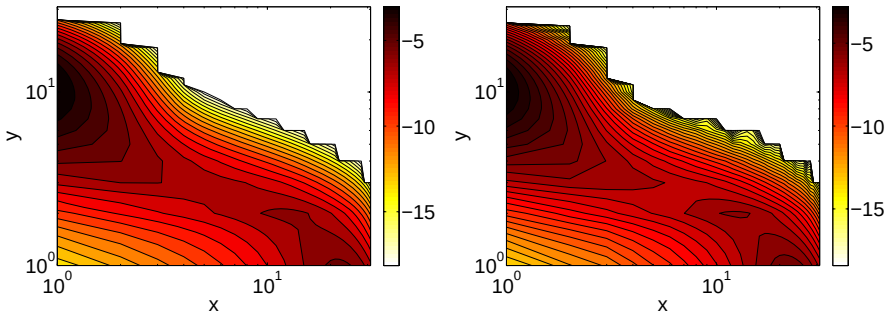

Figure 5.11.: Left: Log-log contour plot of the Gibbs energy, $-log\pi$, where $\pi = (\pi_i), i \in S$ is the stationary distribution computed via $\pi^T L = 0$. Right: Log-log contour plot of the Gibbs energy resulting from the observed distribution $\hat{\pi}$ of states in the time series. Result for $N = 10^8$.

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 $\{\tau_1 = 0.0001, \tau_2 = 0.001, \tau = 0.01\}$. The Gibbs energy resulting from the distribution of the observed states in the time series is shown in the right panel of Figure 5.11. As one can see, the relative occupation of the states is consistent with the exact stationary distribution depicted in the left panel.

The generated time series visits 225 states of 900 possible states, hence we had to estimate 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 $TOL = 10^{-2}$. The Figure 5.12 shows the free energies 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 (cf. left panel of Figure 5.11). In order to make things more precise, we compare in the following the estimated generators with the original generators of (4.47) *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

$$\bar{l}_{ij} = \begin{cases} l_{ij}, & \text{if } i \neq j \text{ were visited by the time series,} \\ -\sum_k \bar{l}_{ik}, & \text{if } i = j \text{ was visited by the time series.} \end{cases} \tag{5.39}$$

Now we compare the spectral properties of the estimated generators with those of the restricted generator from (5.39) in more detail. In the left panel of Figure 5.13

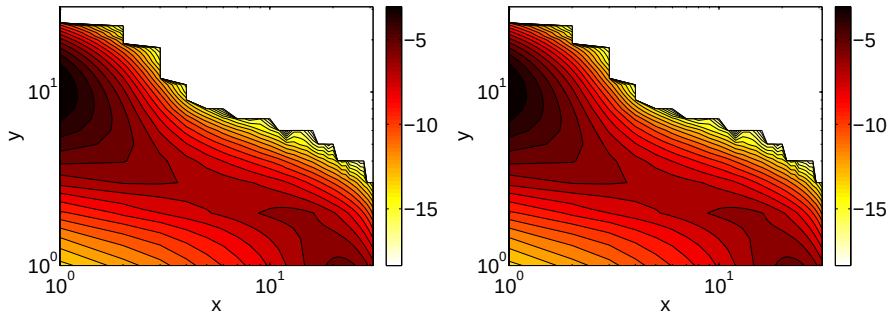

Figure 5.12.: Log-log contour plot of the free energies, $-\log\tilde{\pi}$, associated with the estimated generators $\tilde{L}_A$ (left panel) and $\tilde{L}_B$ (right panel) where $\tilde{\pi}$ is the stationary distribution of the estimated generators computed via $\tilde{\pi}^T\tilde{L} = 0$, respectively.

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 the process. For a Markov jump process, it is easy to see that the ACF reduces to [19]

$$\mathbb{E}(X_{t+\tau}X_t) = \sum_{k=1}^{d} e^{\tau\lambda_k} \sum_{i,j\in S} i\cdot j\cdot \pi_i U_{ik} U^{-1}_{kj}, \tag{5.40}$$

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 5.14. 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.

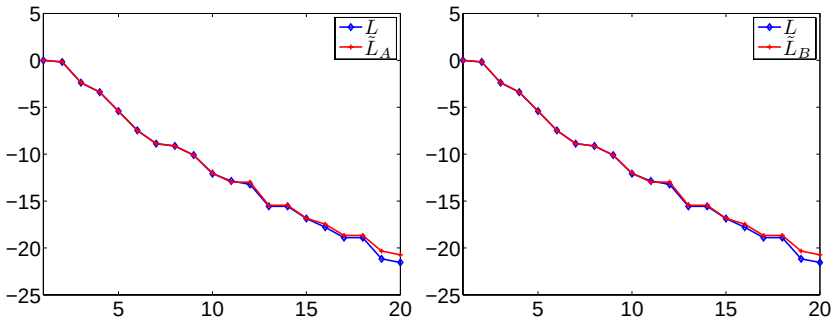

Figure 5.13.: The real parts of the first 30 largest eigenvalues of the estimated generators compared to the eigenvalues of the restricted generator $\bar{L}$ (5.39). Left: Real parts of eigenvalues of $\tilde{L}_A$. Right: Real parts of eigenvalues of $\tilde{L}_B$.

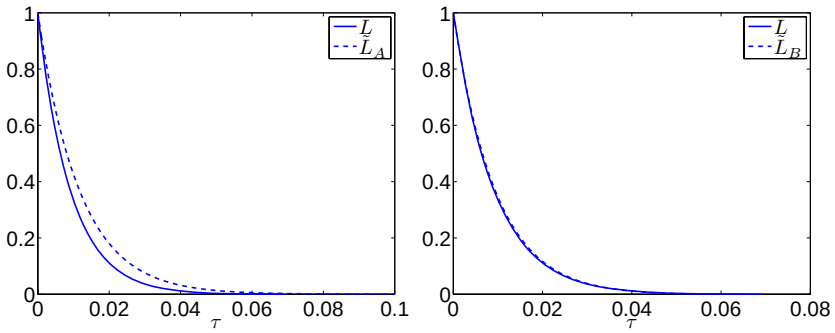

Figure 5.14.: The graphs of the ACFs associated with $\tilde{L}_A$ (left panel) and $\tilde{L}_B$ (right panel) compared to the ACF of the restricted generator $\bar{L}$, respectively.