DATA-BASED INFERENCE OF GENERATORS FOR MARKOV JUMP PROCESSES USING CONVEX OPTIMIZATION *

DAAN CROMMELIN[†] AND ERIC VANDEN-EIJNDEN[‡]

Abstract. A variational approach to the estimation of generators for Markov jump processes from discretely sampled data is discussed and generalized. In this approach, one first calculates the spectrum of the discrete maximum likelihood estimator for the transition matrix consistent with the discrete data. Then the generator that best matches the spectrum is determined by solving a convex quadratic minimization problem with linear constraints (quadratic program). Here, we discuss the method in detail and position it in the context of maximum likelihood inference of generators from discretely sampled data. Furthermore, we show how the approach can be generalized to estimation from data sampled at non-constant time intervals. Finally, we discuss numerical aspects of the algorithm for estimation of processes with high-dimensional state spaces. Numerical examples are presented throughout the paper.

Key words. Markov jump process, generator, estimation, quadratic program

AMS subject classifications. 62G05, 60J75, 60J27, 60J35, 65C40, 62M10

1. Introduction.

Markov jump processes with finite state-space are used in many scientific disciplines: physics, chemistry, biology, finance, sociology, etc. The properties of these models are specified by their generator, here denoted by Q, also known as rate matrix or intensity matrix ([1, 21]). An important practical issue for modeling is the inference of the generator from time-series data. This issue is the topic of the present paper.

When the time-series is sampled continuously, the natural framework to infer the generator of the chain is to maximize the likelihood function associated with these continuous time data. In this case, there is an analytic expression for the maximum likelihood estimator (MLE) of Q which involves quantities easily computable from the data (e.g. [5]). Continuously sampled time-series, however, are rarely available in real applications. Most often, one is given a time series of the process sampled at discrete points in time. Maximum likelihood estimation can be generalized to discretely sampled data, but in this case, inference is not as straightforward as it is in the continuously sampled case. There is no analytical expression for the MLE of Q in this case, and its calculation is a nontrivial computational task in general because the likelihood function associated with the discrete time data may have several local maxima. Even worse, the MLE may not exist (i.e. the likelihood function may be non-coercive).

These difficulties were discussed in details in [6] and their origin is quite simple: Since the discrete time data only carry incomplete information about the continuous-time process, it is possible that several continuous-time Markov jump processes oscillating at different rates between the states are consistent with the data. This leads to multiplicity of the (local) MLE. It may also be that by considering generators Q

[*] Revised version, March 20, 2009. Accepted for publication in SIAM Multiscale Model. Simul.

[†] Centrum Wiskunde & Informatica (CWI), Amsterdam, The Netherlands (Daan.Crommelin@cwi.nl). This author's work was supported by the Netherlands Organization for Scientific Research (NWO) through the research cluster Nonlinear Dynamics of Natural Systems.

[‡] Courant Institute of Mathematical Sciences, New York, USA (eve2@cims.nyu.edu). This author's work was supported by NSF grants DMS02-39625 and DMS07-08140, and by ONR grant N00014-04-1-0565.

consistent with faster and faster oscillations between the states (i.e. by making the amplitude of the entries of Q larger and larger), one keeps on increasing the likelihood of Q given the data. In this case the MLE will not exist. These problems typically tend to disappear when the sampling rate of the discrete time data is increased (i.e. the time lag between the observations is shortened). However, this may not be a viable option in practice, for example because there is some practical limit imposed on the sampling rate of the data. Furthermore, the underlying continuous-time process may in fact not be exactly Markov at very short time scales. The latter situation is often encountered in practical applications, and in such cases one would like to infer a generator consistent with the data sampled at time intervals which are long enough to have filtered out the non-Markov properties.

For all of these reasons, it may be preferable to use another framework than maximum likelihood estimation to infer generators from discrete time data in situations when the sampling lag is not small enough. An alternative procedure was proposed in [8]. The basic idea behind this procedure is simple, especially when the time lag between the observations is constant (an assumption that we will relax below). In this case, the data can be viewed as a sample of a discrete-time Markov chain. The MLE for the transition matrix can be easily computed from the discretely sampled data. The spectrum of the MLE is calculated, and, consistent with this spectrum carrying the relevant information about the process, it is used to identify the generator Q whose spectrum is the closest to the spectrum of the MLE. Here, closeness is measured in terms of a convex quadratic objective function which needs to be minimized subject to a set a linear constraints that guarantee that the minimizer is a generator.

Our main objective here is to analyze in more detail the mathematical and computational aspects of the procedure proposed in [8]. In doing so we will simplify the procedure, and generalize it to inference from time series sampled at non-constant lags. Overall, the approach has several advantages. First, it bypasses completely the issue related to non-uniqueness or nonexistence of the MLE of Q because the constrained minimization problem, being convex, always has a unique generator Q as solution. This is especially appealing when studying data that is not exactly consistent with a Markov jump process, a problem which, as mentioned above, is rather common in practical applications. Second, this solution can be computed efficiently, even when the number of states in the chain is large, using the numerical tools of quadratic programming established after years of improvements by the numerical optimization community. Third, the method is versatile and can be adapted to Markov jump processes of special type (e.g. birth-death process) or to impose additional constraints in order that the generator matches exactly rather than approximately one or more elements from the spectrum (e.g. the invariant distribution).

The paper is organized as follows. In section 2 we review some basic properties of generators and discuss convergence of the estimates of the spectrum. We present a simplified and improved version of the estimation procedure from [8] and position it with respect to the maximum likelihood estimation procedure (as in [6]). The generalization to estimation from timeseries with non-constant sampling interval is the topic of section 3. We present a procedure to obtain estimates for the generator spectrum from such timeseries and we apply it to infer a generator from data with random sampling intervals (drawn from a gamma distribution). Section 4 deals with numerical aspects of large-scale quadratic programs, relevant for estimation of jump processes in high-dimensional state spaces. For such processes, an adequate formulation of the optimization problem and suitable choice of solution method are needed. In section 5 we conclude and we briefly discuss the relation with estimation of diffusion processes from discretely sampled data.

2. Generator inference using the eigenspectrum.

2.1. Basic properties of generators. We will consider the continuous-time Markov jump process $\{X_t : t \geq 0\}$ on the finite state space $S = \{1, \dots, d\}$. Assuming time-homogeneity, we will denote the transition matrices of this process by $P(\tau) = (p_{i,j}(\tau))_{i,j \in S}$, i.e.

$$p_{ij}(\tau) = \mathbb{P}(X_{t+\tau} = j | X_t = i), \qquad t, \tau \geq 0\,. \tag{2.1}$$

We will assume that the process is ergodic, implying that there exists a unique vector $\mu = (\mu_1, \dots, \mu_d)^T$ with positive entries $\mu_i > 0$ such that

$$\mu^T P(\tau) = \mu^T, \qquad \sum_{i \in S} \mu_i = 1\,. \tag{2.2}$$

This vector is referred to as the stationary distribution of the Markov jump process. If $P(\tau)$ is right-differentiable at zero, i.e. if the following limit exists

$$\lim_{\tau \to 0+} \frac{P(\tau) - I}{\tau} = Q \tag{2.3}$$

where I denotes the identity matrix, the matrix $Q = (q_{ij})_{i,j \in S}$ is called the (infinitesimal) generator of the process $\{X_t : t \geq 0\}$. The entries of the generator Q satisfy the two conditions:

$$q_{ij} \geq 0 \qquad \text{for all} \quad i, j \in S,\ i \neq j \tag{2.4a}$$

$$\sum_{j \in S} q_{ij} = 0 \qquad \text{for all} \quad i \in S \tag{2.4b}$$

Conversely, any matrix satisfying these two conditions is the generator of some continuous-time Markov jump process. Note that (2.3) implies that $P(\tau)$ is related to Q via the matrix exponential:

$$P(\tau) = \exp(Q\tau). \tag{2.5}$$

Also, in terms of Q (2.2) reads

$$\mu^T Q = 0. \tag{2.6}$$

2.2. Inference from discretely sampled data: maximum likelihood estimation and its difficulties. Let $X_{t_1}, \dots, X_{t_{N+1}}$ be a series of observations of $\{X_t : t \geq 0\}$ at the discrete points in time $0 < t_1 < \cdots < t_{N+1}$. In this section, we assume that the process is observed at a constant sampling interval, i.e. $t_{n+1} - t_n = \Delta t$ is constant for all $n = 1, \dots, N$. In section 3 we will relax this condition. When the sampling interval is constant, the data $X_{t_1}, \dots, X_{t_1 + N\Delta t}$ can be viewed as the sample path of a discrete-time Markov jump process. Consistently, we can calculate the discrete-time based maximum likelihood estimator (MLE) $\hat{P}$ of the transition probability matrix $P(\tau = \Delta t)$ from these data, i.e. the maximizer of

$$L_D(P) = \prod_{i,j \in S} p_{ij}^{K_{ij}^{(N)}} \tag{2.7}$$

where $K_{ij}^{(N)}$ denotes the number of transition from state i to state j observed in the data X_{t_1}, $X_{t_1+\Delta t}$, ..., $X_{t_1+N\Delta t}$. The likelihood function (2.7) can be maximized explicitly (e.g. [5]) and the entries $\hat{p}_{ij}$ of $\hat{P}$ are given by

$$\hat{p}_{ij} = \begin{cases} \dfrac{K_{ij}^{(N)}}{\sum_{j\in S} K_{ij}^{(N)}} & \text{if} \quad \sum_{j\in S} K_{ij}^{(N)} \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.8}$$

The MLE $\hat{P}$ is related to the discrete-time based maximum likelihood estimator $\hat{Q}$ of the generator Q, i.e. the maximizer of

$$L_C(Q) = \prod_{i,j\in S} \left(e^{Q\Delta t}\right)_{ij}^{K_{ij}^{(N)}} \tag{2.9}$$

where $(e^{Q\Delta t})_{ij}$ denote the entries of the matrix $\exp(Q\Delta t)$. Indeed, if the equation

$$\exp(Q\Delta t) = \hat{P} \tag{2.10}$$

admits a solution Q which is a generator (i.e. such that (2.4) is satisfied), then this solution is a MLE $\hat{Q}$ since we then have $L_C(\hat{Q}) = L_D(\exp(\hat{Q}\Delta t))$. There are, however, two difficulties, as discussed in detail in [6]. The first is that (2.10) may admit several solutions which are generators. If this is the case, each of these solutions is a MLE $\hat{Q}$, i.e. the MLE of the generator is non-unique. In statistical terms, this means that the parameterization of the distribution of the data $X_{t_1}, ..., X_{t_1+N\Delta t}$ by Q may not be identifiable. The second, more serious difficulty is that (2.10) may have no solution which is a generator. This difficulty is related to the so-called imbedding problem which states that the law of a discrete-time Markov chain is not always the law of a continuous-time Markov chain sampled at discrete times. If (2.10) has no solution which is a generator, the MLE $\hat{Q}$ may or may not exist. Even when it does exist, it is nontrivial to identify this MLE numerically because the likelihood function (2.9) is nonconvex and its gradient with respect to Q is complicated. For recent algorithmic advances in this direction, see [6, 16, 17].

2.3. Sampling error and spectral decomposition. To understand better the origin of the difficulties discussed in the last section, let us assume that the generator Q admits the spectral decomposition

$$Q = UD_\lambda U^{-1} \tag{2.11}$$

where U denotes the matrix whose columns are eigenvectors of Q, and D_λ is the diagonal matrix with the eigenvalues of Q on the diagonal, $D_\lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$. We order the eigenvalues λ_i by increasing amplitude of their absolute value. Then by ergodicity we have $\lambda_1 = 0$ and $\text{Re}\,\lambda_i < 0$ for $i = 2, \ldots d$. Using the decomposition (2.11), the matrix $P(\Delta t) = \exp(Q\Delta t)$ can be computed explicitly and is given by

$$P(\Delta t) = UD_\Lambda U^{-1} \tag{2.12}$$

where $D_\Lambda = \text{diag}(\Lambda_1, \Lambda_2, \ldots, \Lambda_d)$ with

$$\Lambda_i = e^{\lambda_i \Delta t}, \qquad i = 1, \ldots, d. \tag{2.13}$$

This implies that we can infer U, U^{-1} and D_λ in (2.11) from data by constructing the MLE $\hat{P}$, calculating its spectral decomposition and using the relation (2.13). In this section, we investigate how sampling errors on $\hat{P}$ affect the estimates of D_λ and of U and U^{-1}.

First of all, (2.13) implies that for all eigenvalues Λ_i such that $\operatorname{Re}\lambda_i\Delta t$ is very large negative, the modulus of Λ_i will be very small. Therefore, if one uses the MLE $\hat{P}$ to get estimates of the Λ_i , one expects that the estimates of the large Λ_i will be better than those of the small Λ_i. This effect can be quantified as follows:

THEOREM 1. *Assume that Q admits the spectral decomposition (2.11) and let $\hat{D}_\Lambda = diag(\hat{\Lambda}_1, \hat{\Lambda}_2, \ldots, \hat{\Lambda}_d)$ be the eigenvalues of the MLE $\hat{P}$ given in (2.8). Assume also that the multiplicity of each eigenvalue is one. Then as $N \to \infty$ for fixed Δt, we have*

$$\sqrt{N}\left(\hat{D}_\Lambda - D_\Lambda\right) \to diag(r_1, \ldots, r_d) \tag{2.14}$$

in probability. Here $r = (r_1, \ldots, r_d)$ is a Gaussian vector with mean zero and covariance

$$\mathbb{E}(r_i r_j) = \sum_{k,l\in S} \frac{\bar{u}_{ik}\bar{u}_{jk}(p_{kl} - \Lambda_i\Lambda_j\delta_{kl})u_{li}u_{lj}}{\mu_k}, \qquad i,j = 1,\ldots,d \tag{2.15}$$

where $\mathbb{E}$ denotes expectation over the law of the process $\{X(t) : t \geq 0\}$, u_{ij} denote the entries of U, and $\bar{u}_{ij}$ those of U^{-1}.

Theorem 1 follows from a central limit theorem for Markov chains [2]; the proof is given in Appendix A. From the theorem it can be seen that the sampling error does not propagate among the eigenvalues of $\hat{P}$. More specifically, if the estimate $\hat{\Lambda}_i$ is such that $\sqrt{N}|\hat{\Lambda}_i| \gg 1$, then this eigenvalue is accurately estimated, even if there are some other estimates $\hat{\Lambda}_j$ that are inaccurate because $\sqrt{N}|\hat{\Lambda}_j| \lesssim 1$. Based on this we can obtain a precise classification of which eigenvalues are reliable. We introduce a constant $0 < \sigma \ll 1$, whose magnitude can be interpreted as a tolerance for the relative error on $\hat{\Lambda}_i$. For large N we have $\hat{\Lambda}_i - \Lambda_i \approx \frac{1}{\sqrt{N}} r_i$, see Theorem 1. Thus, the expectation of the squared error is (approximately) $\frac{1}{N}\,\mathbb{E}(r_i^2)$. The estimate $\hat{\Lambda}_i$ is classified as reliable if this expectation is smaller than $\sigma^2\,|\hat{\Lambda}_i|^2$, i.e. if $\frac{1}{N}\,\mathbb{E}(r_i^2) < \sigma^2\,|\hat{\Lambda}_i|^2$. We do not know $\mathbb{E}(r_i^2)$ but if we replace the elements of U, U^{-1}, P and μ in (2.15) by their estimates $\hat{U}, \hat{U}^{-1}, \hat{P}$ and $\hat{\mu}$ we obtain an a posteriori error estimate $\mathbb{E}(\hat{r}_i^2)$. Using this gives the following reliability criterion:

$$\text{The estimate } \hat{\Lambda}_i \text{ is reliable if } \sigma|\hat{\Lambda}_i| > \frac{1}{\sqrt{N}}\sqrt{\mathbb{E}(\hat{r}_i^2)}\ . \tag{2.16}$$

The results above have important consequences for estimating the spectrum of the underlying generator Q via the spectrum of the MLE $\hat{P}$. If $\hat{\Lambda}_i \neq 0$ (which is increasingly likely as $N \to \infty$ since $\Lambda_i \neq 0$), we can compute

$$\check{\lambda}_i = \frac{1}{\Delta t}\log\hat{\Lambda}_i, \tag{2.17}$$

and this quantity gives an estimate of an eigenvalue λ_i of Q. We may wonder whether the condition in (2.16) will guarantee reliability of the $\check{\lambda}_i$ calculated from (2.17).

Unfortunately, $\check{\lambda}_i$ may be unreliable even if $\hat{\Lambda}_i$ is reliable. This can happen if the sampling interval is small, so that $\hat{\Lambda}_i$ is close to 1. The reverse situation is also possible: for $\hat{\Lambda}_i$ close to 0, $\check{\lambda}_i$ may be reliable while $\hat{\Lambda}_i$ is not. This typically occurs if the sampling interval is large.

Provided one picks the right branch of the logarithm when $\hat{\Lambda}_i$ is complex, Theorem 1 implies

$$\sqrt{N}(\check{\lambda}_i - \lambda_i) \to \frac{1}{\Delta t} e^{-\lambda_i \Delta t} r_i \quad \text{in probability as } N \to \infty \tag{2.18}$$

This leads to a reliability criterion similar to (2.16):

$$\text{The estimate } \check{\lambda}_i \text{ is reliable if } \sigma|\check{\lambda}_i| > \frac{1}{\Delta t\,\hat{\Lambda}_i}\frac{1}{\sqrt{N}}\sqrt{\mathbb{E}(\hat{r}_i^2)} \tag{2.19}$$

where σ is again a tolerance on the relative error. Comparing (2.16) with (2.19) it can be seen that if $\Delta t\,|\check{\lambda}_i| < 1$, it is possible that the criterion in (2.16) is satisfied but the one in (2.19) is not (i.e., $\hat{\Lambda}_i$ is reliable while $\check{\lambda}_i$ is not). If $\Delta t|\check{\lambda}_i| > 1$, the reverse situation can occur.

We note that Theorem 1 concerns the limit $N \to \infty$ with Δt fixed. It must be kept in mind that p_{kl} and Λ_i in (2.15) both depend on Δt. In Appendix B the small Δt limit of the errors on the estimated eigenvalues is discussed.

Regarding the left and right eigenvectors of P associated with an eigenvalue of multiplicity one, we have the following result.

Theorem 2. *Assume that P and $\hat{P}$ admit the spectral decompositions $P = UD_\Lambda U^{-1}$ and $\hat{P} = \hat{U}\hat{D}_\Lambda\hat{U}^{-1}$, where $D_\Lambda = diag(\Lambda_1, \Lambda_2, ..., \Lambda_d)$ and $\hat{D}_\Lambda = diag(\hat{\Lambda}_1, \hat{\Lambda}_2, ..., \hat{\Lambda}_d)$, both ordered by decreasing $|\Lambda_j|$. Assume also that the eigenvalue Λ_i has multiplicity one. We denote its associated left and right eigenvectors by ψ_i and ϕ_i. Then as $N \to \infty$ for fixed Δt, we have*

$$\sqrt{N}(\hat{\phi}_i - \phi_i) \to \sum_{j \neq i} \frac{\psi_j^T\, S\, \phi_i}{\Lambda_j - \Lambda_i}\, \phi_j \tag{2.20a}$$

$$\sqrt{N}(\hat{\psi}_i - \psi_i) \to \sum_{j \neq i} \frac{\psi_i^T\, S\, \phi_j}{\Lambda_j - \Lambda_i}\, \psi_j \tag{2.20b}$$

in probability. As before, S is a Gaussian matrix whose elements have mean zero and covariance given by (A.2).

The theorem is proven in Appendix C. From (2.20a) it follows that in the limit $N \to \infty$,

$$\begin{aligned}\mathbb{E}\,\|\hat{\phi}_i - \phi_i\| \to& \mathbb{E}\left\| \sum_{j\neq i} \frac{1}{\sqrt{N}(\Lambda_j - \Lambda_i)} \left(\psi_j^T S \phi_i\right) \phi_j \right\| \quad \text{in probability} \\ \leq& \frac{1}{\sqrt{N}\, d\Lambda_i} \sum_{j \neq i} \|\phi_j\|\, \mathbb{E}\, |\psi_j^T S \phi_i|\,,\end{aligned} \tag{2.21}$$

where

$$d\Lambda_i = \min_{j\,(j\neq i)} |\Lambda_j - \Lambda_i|\,. \tag{2.22}$$

Using, as before, the tolerance σ, the requirement $\mathbb{E}\,\|\hat{\phi}_i - \phi_i\| < \sigma\|\hat{\phi}_i\|$ is satisfied if $\sigma\sqrt{N}\,d\Lambda_i\,\|\hat{\phi}_i\| > \sum_{j\neq i}\|\phi_j\|\,\mathbb{E}\,|\psi_j^T S\phi_i|$, see (2.21). If we replace all ϕ_i, ψ_i and Λ_i by their estimates and assume the eigenvectors are normalized such that $\|\hat{\phi}_i\| = \|\hat{\phi}_j\|$ for all i, j, the latter condition simplifies to $\sigma\sqrt{N}\,d\hat{\Lambda}_i > \sum_{j\neq i}\mathbb{E}\,|\hat{\psi}_j^T\hat{S}\hat{\phi}_i|$. For $\hat{\psi}_i - \psi_i$ a similar expression holds. Therefore we formulate the following reliability criteria:

$$\hat{\phi}_i \text{ is reliable if } \;\sigma\, d\hat{\Lambda}_i > \frac{1}{\sqrt{N}}\sum_{j\neq i}\mathbb{E}\,|\hat{\psi}_j^T\hat{S}\hat{\phi}_i| \tag{2.23a}$$

$$\hat{\psi}_i \text{ is reliable if } \;\sigma\, d\hat{\Lambda}_i > \frac{1}{\sqrt{N}}\sum_{j\neq i}\mathbb{E}\,|\hat{\psi}_i^T\hat{S}\hat{\phi}_j| \tag{2.23b}$$

As can be seen, the reliability of the estimated eigenvectors $\hat{\phi}_i$ and $\hat{\psi}_i$ depends crucially on $d\hat{\Lambda}_i$, the separation of the i-th eigenvalue $\hat{\Lambda}_i$ from all other eigenvalues.

If we denote by $\hat{U}$ the matrix of eigenvectors of $\hat{P}$ and assume that this matrix has full rank, we have the following estimate for the solution of (2.10):

$$\check{Q} = \hat{U}\check{D}_\lambda\hat{U}^{-1} \tag{2.24}$$

where $\check{D}_\lambda = \operatorname{diag}(\check{\lambda}_1, \ldots, \check{\lambda}_d)$ and the $\check{\lambda}_i$ are given by (2.17). For the errors on $\hat{P}$ we have the central limit theorem (A.1); by contrast, the errors introduced by the unreliable part of the spectrum will propagate through the matrix $\check{Q}$ given by (2.24) in a way that is difficult to control. In many instances, $\check{Q}$ will not even be a generator: it will violate the constraints (2.4) and/or have complex matrix elements.

Summing up: In situations where one or more of the eigenvalues or eigenvectors of $\hat{P}$ are unreliable in the sense of (2.16) and (2.23), $\check{Q}$ is unsuitable as an estimate for the solution of (2.10) because it contains many unreliable elements. If we observe a jump process at discrete points in time, we expect that the MLE $\hat{P}$ may very well be non-embeddable because of sampling error, even though the true stochastic matrix $P(\Delta t)$ is embeddable. In many practical cases, the sampling interval Δt is too large to infer the fast timescale features of the process $\{X_t : t \geq 0\}$, resulting in a $\hat{P}$ that has an unreliable part in its spectrum. In these situations, we expect that the approach based on maximum likelihood estimation will encounter the difficulties discussed in section 2.2. However, the results above suggest that even if part of the spectrum of $\hat{P}$ is unreliable, there are important features of the process which may be inferred from the reliable part of the spectrum (we stress "may" because of the problem of the choice of the branch of the logarithm, to be discussed below). How to use this reliable part of the spectrum to infer Q is the core of the approach we propose, as explained next.

2.4. Inference as a convex optimization problem.

We now explain our inference procedure assuming that the MLE $\hat{P}$ inferred from the data admits the spectral decomposition

$$\hat{P} = \hat{U}\hat{D}_\Lambda\hat{U}^{-1} \tag{2.25}$$

If the decomposition (2.25) does not exist, i.e. if the geometric multiplicity of some eigenvalues Λ_i is lower than their algebraic multiplicity (in which case $\hat{U}$ does not have full rank), the procedure described below can be generalized using the Jordan decomposition of $\hat{P}$.

From $\hat{D}_\Lambda$ we compute $\check{D}_\lambda$ according to (2.17) using the principal branch of the logarithm. We classify the eigenvalues in $\check{D}_\lambda$ as (un)reliable according to (2.19).

Since the eigenvalues $\hat{\Lambda}_i$ are estimates of $e^{\lambda_i \Delta t}$, any real negative $\hat{\Lambda}_i$ should have even multiplicity if it is estimated accurately. Therefore we also classify as unreliable any $\check{\lambda}_i$ whose corresponding $\hat{\Lambda}_i$ is real negative with odd multiplicity, by adjusting σ so that this $\hat{\Lambda}_i$ falls in $N\sigma^2|\hat{\Lambda}_i|^2 \leq \mathbb{E}(\hat{r}_i^2)$. All other eigenvalues are classified as reliable. We then construct the diagonal matrix $\hat{D}_\lambda = \mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_d)$, where the $\hat{\lambda}_i$ are

$$\hat{\lambda}_i = \check{\lambda}_i \qquad \text{(reliable } \check{\lambda}_i\text{)} \tag{2.26a}$$

$$\hat{\lambda}_i = \check{\lambda}_i \qquad \text{(unreliable } \check{\lambda}_i \text{ with } \hat{\Lambda}_i \notin (-\infty, 0]\text{)} \tag{2.26b}$$

$$\hat{\lambda}_i = \frac{1}{\Delta t} \log \max(|\hat{\Lambda}_i|, \delta) \qquad \text{(unreliable } \check{\lambda}_i \text{ with } \hat{\Lambda}_i \in (-\infty, 0]\text{)} \tag{2.26c}$$

where $\delta > 0$ is a threshold parameter such that $|\hat{\Lambda}_i| \gg \delta$ for all reliable $\hat{\Lambda}_i$. This parameter is introduced in case some of the $\hat{\Lambda}_i$ are identically 0. The eigenvalues (2.26c) will turn out to have little impact on the reliable part of the spectrum of our inferred Q, so the precise recipe in (2.26c) is not important except for ensuring that the magnitude of unreliable eigenvalues computed from (2.26c) is larger than the one of the reliable eigenvalues computed from (2.17). We remark that is not necessary for the eigenvalues to be simple. Although for the error analysis in the previous section we assumed multiplicity one for each $\hat{\Lambda}_i$, the procedure described here remains applicable if there are repeated eigenvalues.

We wish to construct a generator Q such that, if there are n reliable eigenvalues in $\hat{D}_\lambda$ and n associated eigenvectors in $\hat{U}$, Q has n eigenvalues and associated eigenvectors which closely match them. We also want the other eigenvalues of Q to have a larger magnitude than these n smallest eigenvalues. To get such a Q, we propose to minimize, subject to the constraints (2.4), the following objective function

$$E(Q) = \|\hat{U}^{-1} Q \hat{U} - \hat{D}_\lambda\|_c^2 \tag{2.27}$$

Here $\|\cdot\|_c^2$ denotes a weighted Frobenius norm: given any square matrix A with entries a_{ij},

$$\|A\|_c^2 = \sum_{i,j \in S} c_i c_j |a_{ij}|^2 \tag{2.28}$$

where c is a vector with positive entries $c_i > 0$ which is introduced to put more weight on the reliable eigenvalues and less on the unreliable ones. This can be achieved by making the c_i dependent on the typical magnitude of the errors of $\hat{\lambda}_i$. By setting $c_i = \tilde{c}_i |\hat{\lambda}_i|^{-1}$ $(i > 1)$, relative errors of equal magnitude are weighted equally if all $\tilde{c}_i = 1$. As the typical relative error of $\hat{\lambda}_i$ is proportional to $(|\hat{\lambda}_i||\hat{\Lambda}_i|)^{-1}$, see (2.18), a straightforward choice is $\tilde{c}_i = |\hat{\lambda}_i||\hat{\Lambda}_i|$ and thus $c_i = |\hat{\Lambda}_i|$.

The objective function (2.27) is a simplification of the one originally proposed in [8], and it has desirable features both from a theoretical and a computational viewpoint. Because $\hat{U}$ has full rank by assumption, the quadratic $E(Q)$ is strictly convex. Therefore, since the admissible region for the entries q_{ij} in $\mathbb{R}^{d^2}$ imposed by (2.4) is a convex region, there is a unique minimizer Q^*. If $Q^{**} = \hat{U} \hat{D}_\lambda \hat{U}$ satisfies the constraints (2.4), then $Q^* = Q^{**}$ and $E(Q^*) = 0$. If Q^{**} does not satisfy the constraints (2.4), then $Q^* \neq Q^{**}$ and $E(Q^*) > 0$; in this case the minimizer Q^* can be identified using well-established quadratic programming techniques, as discussed in section 4.2.

As was shown in the previous section, $\hat{U}$, $\hat{U}^{-1}$ and $\hat{D}_\Lambda$ all converge to their true values: as $N \to \infty$ and Δt fixed, $\hat{U} \to U$, $\hat{U}^{-1} \to U^{-1}$ and $\hat{D}_\Lambda \to D_\Lambda$. For reversible jump processes, all Λ_i are real positive so that the logarithm in (2.17) determines the λ_i uniquely. In that case, $\hat{D}_\lambda \to D_\lambda$ and thus the minimizer Q^{**} is a consistent estimator of $Q = UD_\lambda U^{-1}$: $Q^{**} \to Q$ as $N \to \infty$ with Δt fixed. Moreover, for large enough N, $\hat{P}$ must be embeddable (because the underlying process is a continuous-time jump process) and thus Q^{**} must satisfy (2.4), so that $Q^* = Q^{**}$. For jump processes with complex Λ_i, the convergence of $\hat{D}_\lambda$ hinges on the selection of the correct branch of the logarithm for complex $\hat{\Lambda}_i$. This issue is discussed in detail below.

If $\hat{P}$ is embeddable, the MLE $\hat{Q}$ exists and satisfies $\exp(\hat{Q}\,\Delta t) = \hat{P}$. Then $\hat{Q}$ must admit the spectral decomposition $\hat{Q} = \hat{U}\tilde{D}_\lambda\hat{U}^{-1}$ with $\tilde{D}_\lambda = \mathrm{diag}(\tilde{\lambda}_1, ...\tilde{\lambda}_d)$ such that $\hat{D}_\Lambda = \exp(\tilde{D}_\lambda\,\Delta t)$. If there are several $\tilde{D}_\lambda$ such that $\hat{D}_\Lambda = \exp(\tilde{D}_\lambda\,\Delta t)$ and $\hat{U}\tilde{D}_\lambda\hat{U}^{-1}$ satisfies (2.4), $\hat{Q}$ is non-unique. If $\hat{Q}$ is unique, it coincides with the minimizer Q^* if for the latter we picked the correct branch of the logarithm when determining the $\hat{\lambda}_i$ via (2.17). In case all eigenvalues of $\hat{P}$ are real positive because the process is reversible, non-uniqueness of the logarithm is not an issue and $\hat{Q}$ must coincide with Q^*.

By adjusting the weights in c, we can ensure that the part of the spectrum of Q^* associated with its n eigenvalues of lower magnitude matches the reliable part of the spectrum of Q^{**}; and that the other $d-n$ eigenvalues of Q^* have larger magnitude than the n first. The MLE $\hat{P}$ may be non-embeddable but the reliable part of its spectrum must be close to the spectrum of P, which in turn is embeddable if X_t is a jump process. Thus, there must exist a generator with a spectrum that closely matches the reliable part of the spectrum of $\hat{P}$.

We note that the choice of the weights in c does not affect the absolute, unconstrained minimizer Q^{**}. Provided all c_i remain positive, Q^{**} is invariant under changes in c. Thus, if Q^{**} satisfies (2.4), minimization of (2.27) always results in the same $Q^* = Q^{**}$. However, if Q^{**} does not satisfy (2.4), Q^* lies on the boundary of the feasible set defined by (2.4) and its precise location depends on c.

The non-uniqueness of the logarithm can complicate the determination of complex eigenvalues through (2.26a). Assume for simplicity that the estimate $\hat{\Lambda}_j$ is without error, i.e. $\hat{\Lambda}_j = \exp(\Delta t \lambda_j)$. If we use the principal branch of the logarithm in (2.26a), we obtain an estimate $\hat{\lambda}_j$ that can differ from the true eigenvalue λ_j by $i2\pi m/\Delta t$:

$$\hat{\lambda}_j = \lambda_j + i\frac{2\pi m}{\Delta t} \qquad \text{with } m \in \mathbb{Z}. \tag{2.29}$$

In principle, the value of m is not available to us. This complication is closely related to the fact that X_t is only observed at a single fixed sampling interval Δt. As will be discussed below, the problem disappears if X_t is observed at several different sampling intervals satisfying rather mild conditions.

Before discussing the situation with non-constant sampling intervals, we have two remarks about the situation with a single Δt. First, if desired one can carry out a search among different branches of the logarithms of the complex $\hat{\Lambda}_j$. This search is finite (see e.g. [14] and references therein) but can be very expensive if d is large. Our second remark is of a heuristic nature: Increasing $|m|$ for some complex pair $(\hat{\lambda}_j, \hat{\lambda}_j^*)$ implies increasing the rate of rotation associated with the decay of this eigenmode. Thus, by increasing $|m|$ one introduces ever faster oscillations in the corresponding two-dimensional eigenspace, although the sampling interval is too long to actually observe these fast oscillations. We consider this to be undesirable. By choosing the principal branch of the logarithm, we effectively pick the slowest oscillation consistent with $\hat{\Lambda}_j$.

As already mentioned, the non-uniqueness of the logarithm no longer poses a problem if the jump process is observed at at least two sampling intervals Δt_1 and Δt_2 whose ratio is irrational (i.e. $\Delta t_1/\Delta t_2 \in \mathbb{R}\backslash\mathbb{Q}$). In that case, if we have $\hat{\Lambda}_{j,1} = \exp(\Delta t_1 \lambda_j)$ and $\hat{\Lambda}_{j,2} = \exp(\Delta t_2 \lambda_j)$, the correct λ_j can be identified uniquely. If we cannot calculate $\hat{\Lambda}_{j,1}$ and $\hat{\Lambda}_{j,1}$ explicitly because X_t is observed at random sampling intervals, the correct λ_j can still be recovered if the sampling intervals are drawn from a non-atomic distribution. This will be explained in detail in section 3 where we consider inference from data with random sampling intervals.

2.5. Numerical examples. We conclude this section with two examples to illustrate the advantages of our procedure.

Example 1. The first example demonstrates that our approach performs well even when the spectrum of the MLE $\hat{P}$ has an unreliable part. We consider a Markov jump process with $d = 24$ states on a periodic ring and a generator with elements

$$q_{ij} = \begin{cases} e^{\beta(V_i - V_j)} & \text{if} \quad j = i \pm 1 (\text{mod } d) \\ -e^{\beta(V_i - V_{i+1})} - e^{\beta(V_i - V_{i-1})} & \text{if} \quad j = i \\ 0 & \text{otherwise} \end{cases} \tag{2.30}$$

where $\beta > 0$ is a parameter and

$$V = (0, 4, 8, 12, 16, 13, 10, 7, 4, 7, 10, 13, 16, 12, 8, 4, 0, 4, 8, 12, 16, 12, 8, 4)^T \tag{2.31}$$

This chain models the motion of a particle in the sawtooth triple-well potential V, which has minima at states 1, 9 and 17 where $V_1 = V_{17} = 0$ and $V_9 = 4$, separated by maxima at states 5, 13 and 21 where $V_5 = V_{13} = V_{21} = 16$. The parameter β plays the role of an inverse temperature. Taking $\beta = 1/8$, the chain is metastable over the states 1, 9 and 17 in the sense that it remains for long periods of times in or near these states, and moves very quickly through the intermediate states during its (infrequent) transitions from one metastable state to the other. This metastable behavior is apparent from the eigenvalues of Q, with the magnitudes of λ_2 and λ_3 being much smaller than those of $\lambda_4, \ldots, \lambda_{20}$, see table 1 and figure 1. The first three leading eigenvalues and their associated eigenvectors explain the long time dynamics of the chain, hence they are the important quantities to capture in this example.

To construct $\hat{P}$, we first compute

$$P = \exp(Q\Delta t) \qquad \text{with } \Delta t = 20 \tag{2.32}$$

This value of the sampling lag Δt captures the slow transitions between the metastable states $1, 9$ and 17, but it is too long to capture the phenomena arising on the shorter time scales associated with λ_4, etc.(for instance, $\Lambda_4 = e^{\lambda_4 \Delta t} = 2.7 \times 10^{-5}$ whereas $\Lambda_2 = e^{\lambda_2 \Delta t} = 0.758$ and $\Lambda_3 = e^{\lambda_3 \Delta t} = 0.624$). As a result, the sampling errors on $\hat{\Lambda}_i$ and $\hat{\lambda}_i$ will be large for all $i > 3$, unless the time series is exceptionally long.

We generate a time series of $N = 10^6$ data points by simulating the Markov chain with stochastic matrix (2.32). From the time series we construct the MLE $\hat{P}$ (2.8), calculate its spectral decomposition (2.25) and the associated estimates for the generator eigenvalues (2.26). As expected, the estimates $\hat{\lambda}_2 = -0.0139$ and $\hat{\lambda}_3 = -0.0236$ are accurate, whereas all the other $\hat{\lambda}_i$ are not (see figure 1). Correspondingly, any value of σ between 0.005 and 4 will result in a classification of $\hat{\lambda}_2$, $\hat{\lambda}_3$ as reliable and $\hat{\lambda}_i$ with $i \geq 4$ as unreliable. Also, the matrix $\check{Q}$ in (2.24) is not a generator.

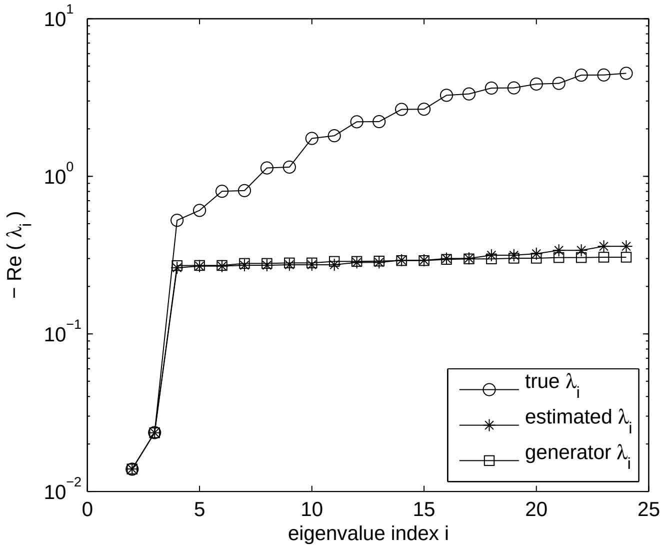

Fig. 1. *Eigenvalue spectra for Example 1. The circles indicate the original eigenvalues λ_i of Q; the stars the estimates $\hat{\lambda}_i$, inferred from a time series of length $N = 10^6$; and the squares the eigenvalues of the generator Q^* inferred by our procedure. The first two non-zero eigenvalues, λ_2 and λ_3 are well-captured despite the fact that the others are not.*

Table 1

Example 1: The first few eigenvalues λ_i computed from Q, their estimates $\hat{\lambda}_i$ and the corresponding eigenvalues λ_i^* of the inferred generator Q^*.

index i	2	3	4	5
λ_i	-0.0139	-0.0236	-0.5254	-0.6067
$\hat{\lambda}_i$	-0.0139	-0.0236	-0.2627	-0.2690 + i 0.0426
λ_i^*	-0.0139	-0.0236	-0.2708	-0.2716 + i 0.0160

We minimize (2.27) subject to (2.4) with

$$c_1 = 100, \quad c_2 = 100|\hat{\lambda}_2|^{-2}, \quad c_3 = 100|\hat{\lambda}_3|^{-2}, \quad \text{and} \quad c_i = |\hat{\lambda}_i|^{-2} \quad \text{for } i = 4, \dots, 24 \tag{2.33}$$

to put most weight on the reliable eigenvalues. The minimization is carried out using the Matlab internal QP solver `quadprog`, requiring about 30 seconds computation time on a modern PC. The eigenvalues λ_i^* of the minimizer Q^* are displayed in figure 1, and the value of the first few are listed in table 1. The first two nontrivial eigenvalues λ_2^* and λ_3^* are in excellent agreement with the exact λ_2 and λ_3. The first three eigenvectors of Q^* are also in very good agrement with those of Q (results not shown). Thus, Q^* accurately captures the long timescale features of Q. The fast (short timescale) features of Q are not accurately captured, due to the large sampling interval and the finite sample size. Correspondingly, the eigenvalues λ_i^* with $i \geq 4$ are inaccurate (but note that most of them are quite close to $\hat{\lambda}_i$). Because of sampling error, the minimizer Q^* is different from the original Q entry-wise; however, those errors only significantly affect the fast features of Q^* and not its slow features.

Similar results were obtained with other values of the weights in c, indicating that the procedure is robust against the choice of these parameters.

EXAMPLE 2. Our approach is tailored to capture the dynamics arising on the time scale of the lag Δt at which the data is observed, regardless of the details of the dynamics arising on much shorter time scales. As a result, it does not matter much whether the process $\{X_t : t \geq 0\}$ is actually Markov on these short time scales. To illustrate this point, we present the following example.

Consider the process $(X_t, Y_t) \in S = \{1, 2, .., d\} \times \{1, 2\}$ whose generator consists of two parts: $Q = Q^X \otimes Q^Y$. The elements of Q^X are given by

$$q_{ij}^X = \begin{cases} 3s & \text{if} \quad j = i + 1 (\text{mod } d) \\ 4s & \text{if} \quad j = i - 1 (\text{mod } d) \\ -7s & \text{if} \quad j = i \\ 0 & \text{otherwise} \end{cases} \tag{2.34}$$

and Q^Y reads

$$Q^Y = \frac{1}{\epsilon} \begin{pmatrix} -r_1 & r_1 \\ r_2 & -r_2 \end{pmatrix} \tag{2.35}$$

The value of s depends on Y_t: $s(Y_t = 1) = s_1$, $s(Y_t = 2) = s_2$. The (positive) parameters ϵ, r_1, r_2, s_1 and s_2 will be determined later on. As is clear, the combined process (X_t, Y_t) is Markov; the process X_t by itself is not. If s were fixed, X_t could be regarded as an asymmetric random walk on a discrete periodic domain with d states. However, s is not constant but switches between the two values s_1 and s_2 with switching rates $\epsilon^{-1} r_1, \epsilon^{-1} r_2$.

Although X_t is non-Markov, in the limit $\epsilon \to 0+$, when Y_t switches much faster than X_t, the law of the discrete time process $X_0, X_{\Delta t}, X_{2\Delta t}, ...$ with $\Delta t \gg \epsilon$ approaches that of the discretely observed Markov jump process with generator Q^{a} which is identical to Q^X with s fixed at $s^{\mathrm{a}} = s_1 r_2 (r_1 + r_2)^{-1} + s_2 r_1 (r_1 + r_2)^{-1}$, cf. (2.34). Thus, at timescales much longer than the switching timescale of Y_t, the effective jump rate of X_t is determined by the average of s_1 and s_2 weighted by the invariant distribution of Y_t. We refer to [15, 22] for more details on the asymptotic analysis of this type of two timescale stochastic processes.

We take parameters $s_1 = 1$, $s_2 = 3$, $r_1 = 1$, $r_2 = 2$, $\epsilon = 10^{-3}$, $d = 10$ and generate a sample path of the process (X_t, Y_t) with total time length T. We record X_t every $\delta t = 10^{-2}$ time units; from this time series we construct the MLE $\hat{P}$ for X_t only. We use every data point so that the the number of data points is $N = 10^2 T$. As explained before, the generator estimated from these observations should approach Q^{a} if $\Delta t \gg \epsilon$, where Q^{a} equals Q^X with $s = 5/3$.

The experiment is repeated for several values of T. In table 2 we show, for $T = 10^2, 10^3, 10^4, 10^5$, the eigenvalues $\hat{\lambda}_2, \hat{\lambda}_4$ inferred from the data, the eigenvalues $\lambda_2^{\mathrm{a}}, \lambda_4^{\mathrm{a}}$ of Q^{a} and λ_2^*, λ_4^* of Q^* (the minimizer of (2.27) with $c_i = 1$ for all i). For all three sets of eigenvalues, $\lambda_1 = 0$ and λ_3, λ_5 are the complex conjugates of λ_2, λ_4; those are therefore not included. We also show the difference $\|Q^* - Q^{\mathrm{a}}\| / \|Q^{\mathrm{a}}\|$ (using Frobenius norm).

3. Inference from data with non-constant sampling intervals.

In this section we show how the inference procedure described in section 2.4 can be generalized so that it can handle estimation from timeseries with non-constant sampling intervals. In case of a constant sampling interval Δt, estimates of the eigenvalues and eigenvectors of Q can be readily obtained through the spectral decomposition

Table 2

Example 2: The eigenvalues λ_2, λ_4 inferred from timeseries of total length T, and those of the inferred generator Q^* and the predicted generator Q^a (see text). Also shown are the differences $\|Q^* - Q^a\|/\|Q^a\|$ in Frobenius norm.

T	N		inferred from data	generator Q^*	generator Q^{a}	$\frac{\|Q^*-Q^{\mathrm{a}}\|}{\|Q^{\mathrm{a}}\|}$
10^2	10^4	λ_2	-2.17+0.91i	-2.23+0.91i	-2.23+0.98i	0.099
		λ_4	-7.81+1.45i	-7.92+1.44i	-8.06+1.59i	
10^3	10^5	λ_2	-2.25+1.01i	-2.27+1.01i	-2.23+0.98i	0.033
		λ_4	-8.13+1.62i	-8.16+1.63i	-8.06+1.59i	
10^4	10^6	λ_2	-2.23+1.01i	-2.23+1.01i	-2.23+0.98i	0.011
		λ_4	-8.07+1.64i	-8.07+1.64i	-8.06+1.59i	
10^5	10^7	λ_2	-2.23+0.97i	-2.23+0.97i	-2.23+0.98i	0.0039
		λ_4	-8.05+1.57i	-8.05+1.57i	-8.06+1.59i	

of the MLE $\hat{P}$, which in turn is easily calculated from the frequency matrix $K^{(N)}$, cf. (2.8). The meaning of $\hat{P}$ is clear in this case: its elements are estimates of the transition probabilities (2.1) with $\tau = \Delta t$. However, if the sampling interval is not constant throughout the timeseries, it is less obvious what $\hat{P}$ given by (2.8) means. Because there is no constant Δt, the relation (2.13) can no longer be used to infer the eigenvalues of Q through the eigenvalues of $\hat{P}$. Nevertheless, it is still possible to infer the spectrum of Q from the timeseries, as will be shown in this section. Once the spectrum of Q is estimated, the inference of Q proceeds in the same way as described in section 2.4.

We note that in [17] an algorithm is presented for estimation from inhomogeneously sampled data using the MLE approach. The computational cost of this MLE based algorithm scales linearly with the number of different observed sampling intervals, limiting its use in practice to data where the sampling interval takes on only a few different values. For the procedure described in this paper, there is no such limitation.

3.1. Inferring the spectrum. Suppose we observe a jump process X_t at discrete points in time, resulting in a timeseries that consists of observations at times t_n, $n = 1, ..., N+1$. We want to infer the spectrum of eigenvalues and eigenvectors of the generator Q from the timeseries. As mentioned, the sampling intervals $\tau_n := t_{n+1} - t_n$ are not constant, but depend on n. We will focus on the case where the τ_n are random.

Let us denote the probability distribution of the sampling intervals by π and the expectation with respect to the law of τ_n by $\mathbb{E}_\tau$. As before, we denote by $P(\tau_n)$ the transition probability matrix with associated time step $\tau_n (\geq 0)$, cf. (2.1) and (2.5). We define P^e as the expectation of $P(\tau_n)$:

$$P^e := \mathbb{E}_\tau \, P(\tau_n) = \int_0^\infty P(\tau) d\pi(\tau) \tag{3.1}$$

We propose the estimator $\hat{P}^e$ for P^e, with its elements defined by

$$\hat{p}^e_{ij} := \frac{\sum_{n=1}^N \mathbf{1}(X_{t_n} = i)\, \mathbf{1}(X_{t_{n+1}} = j)}{\sum_{n=1}^N \mathbf{1}(X_{t_n} = i)}\,. \tag{3.2}$$

Both P^e and $\hat{P}^e$ are stochastic matrices by construction. We will prove strong consistency of the estimator $\hat{P}^e$ under the following condition:

All sampling intervals τ_n are i.i.d. random variables with distribution π.
π may have atoms but not at zero. (3.3)

Theorem 3. *Let $X_{t_1}, ..., X_{t_{N+1}}$ be the discrete sampling of an ergodic jump process with unique invariant distribution μ. Under condition (3.3), $\hat{P}^e \to P^e$ almost surely as $N \to \infty$.*

The theorem is proven in appendix D. Note that there is no loss of generality by requiring that π has no atom at zero. After all, $\tau_n = 0$ would imply that the observation X_{t_n} is simply repeated in the data ($t_{n+1} = t_n$), making the observation $X_{t_{n+1}}$ redundant (it is the same as X_{t_n}). Furthermore, the case where the process X_t is sampled at a constant sampling interval, $\tau_n = \Delta t$ for all n, corresponds to a distribution π that is a single atom at Δt.

Assuming that Q admits the spectral decomposition $Q = UD_\lambda U^{-1}$ and thus $P(\tau) = UD_{\Lambda(\tau)}U^{-1}$, P^e can be decomposed as

$$P^e = UD^e_\Lambda U^{-1} \tag{3.4}$$

where $D^e_\Lambda = \mathrm{diag}(\Lambda^e_1, ..., \Lambda^e_d)$, with

$$\Lambda^e_i = \mathbb{E}_\tau \mathrm{e}^{\lambda_i \tau_n} \tag{3.5}$$

Because we have assumed that the process generated by Q has a unique invariant distribution μ, P^e has the same unique invariant distribution and its eigenvalues satisfy $\Lambda^e_1 = 1$ and $|\Lambda^e_i| < 1$ for all $i > 1$.

We estimate U, U^{-1} and D^e_Λ in (3.4) by computing the spectral decomposition of the estimator $\hat{P}^e$:

$$\hat{P}_e = \hat{U}\hat{D}^e_\Lambda \hat{U}^{-1} \tag{3.6}$$

with $\hat{D}^e_\Lambda = \mathrm{diag}(\hat{\Lambda}^e_1, ..., \hat{\Lambda}^e_d)$. Given (estimates of) Λ^e_i, the generator eigenvalues λ_i follow from (3.5). Hence, we construct estimates $\hat{\lambda}_i$ by solving

$$f(\lambda_i; \hat{\Lambda}^e_i) := \hat{\Lambda}^e_i - \frac{1}{N}\sum_{n=1}^{N} \mathrm{e}^{\lambda_i \tau_n} = 0 \tag{3.7}$$

This equation can easily be solved with Newton's method, using the exact gradient

$$\frac{d}{d\lambda_i} f = -\frac{1}{N}\sum_{n=1}^{N} \tau_n\, \mathrm{e}^{\lambda_i \tau_n} \tag{3.8}$$

If $\hat{\Lambda}^e_i$ is real, the solution $\hat{\lambda}_i$ must satisfy $\sum_n \exp(\tau_n \operatorname{Re} \hat{\lambda}_i)\, \sin(\tau_n \operatorname{Im} \hat{\lambda}_i) = 0$. This implies $\operatorname{Im} \hat{\lambda}_i = 0$ except for degenerate distributions of τ_n. If $\hat{\Lambda}^e_i$ and λ_i are both real and $\hat{\Lambda}^e_i > 0$, (3.7) has a unique solution (< 0) because $f(0; \hat{\Lambda}^e_i) = \hat{\Lambda}^e_i - 1 < 0$, $f(-\infty; \hat{\Lambda}^e_i) = \hat{\Lambda}^e_i > 0$ and $df/d\lambda_i < 0$. There is no solution if $\hat{\Lambda}^e_i, \lambda_i \in \mathbb{R}$ and $\hat{\Lambda}^e_i \leq 0$.

If $\hat{\Lambda}^e_i$ is complex, establishing uniqueness is more complicated. It comes down to the question under what conditions on π the mapping $\lambda_i \mapsto \Lambda^e_i$ ($\lambda_i, \Lambda^e_i \in \mathbb{C}$) is one-to-one. We will not explore this question in full detail here, but only make the following observation. Suppose $\hat{\lambda}_i$ and $\hat{\lambda}'_i$ are both solutions to (3.7), so that $\sum_n (\mathrm{e}^{\hat{\lambda}_i})^{\tau_n} = \sum_n (\mathrm{e}^{\hat{\lambda}_i}\, \mathrm{e}^{\hat{\lambda}'_i - \hat{\lambda}_i})^{\tau_n}$. Clearly, this equality holds if $(\mathrm{e}^{\hat{\lambda}'_i - \hat{\lambda}_i})^{\tau_n} = 1$ for all n, which implies $\operatorname{Re} \hat{\lambda}_i = \operatorname{Re} \hat{\lambda}'_i$ and $\tau_n \operatorname{Im} (\lambda'_i - \lambda_i) = 2\pi m_n$ for all n ($m_n \in \mathbb{Z}$). The latter requirement cannot be met if some of the sampling intervals have irrational ratios, unless $\operatorname{Im}\hat{\lambda}_i = \operatorname{Im}\hat{\lambda}'_i$ (see also the discussion at the end of section 2.4). We will not consider the possibility that $\sum_n (\mathrm{e}^{\hat{\lambda}_i})^{\tau_n} = \sum_n (\mathrm{e}^{\hat{\lambda}_i}\,\mathrm{e}^{\hat{\lambda}'_i - \hat{\lambda}_i})^{\tau_n}$ and yet $(\mathrm{e}^{\hat{\lambda}'_i - \hat{\lambda}_i})^{\tau_n} \neq 1$ for some n.

Finding all $\hat{\lambda}_i$ with $i = 2, ..., d$ amounts to solving $d-1$ uncoupled one-dimensional numerical problems ($\hat{\Lambda}^e_1 = 1$ and $\hat{\lambda}_1 = 0$ by construction). Once we have found the $\hat{\lambda}_i$, we have inferred the spectrum of Q from the timeseries and we can start to infer Q itself using the convex optimization procedure described in section 2.4.

As a final remark, we expect (3.2) to be a consistent estimator of (3.1) also in many cases where the sampling intervals are not i.i.d. (as was required in condition (3.3)). Examples are situations where τ_n is itself a stochastic process or where τ_n is a periodic function of n. We will not investigate this generalization here.

3.2. Numerical example with random sampling intervals.

Example 3. We take the jump process whose generator Q has elements

$$q_{ij} = \frac{2d+i}{2d(i-j)^2} \qquad \text{if} \;\; i \neq j \tag{3.9}$$

where i and j run from 1 to $d = 10$. The diagonal elements of Q follow from the property (2.4b). The nonzero eigenvalues of Q range from $\lambda_2 = -0.678$ to $\lambda_{10} = -5.88$. We generate timeseries with $N+1$ points by Monte Carlo simulation, drawing the sampling intervals from a gamma distribution with density

$$\gamma(\tau; a, b) = \frac{1}{b^a\,\Gamma(a)}\,\tau^{a-1}\mathrm{e}^{-\tau/b} \tag{3.10}$$

and parameters

$$a = 4, \qquad b = \frac{1}{8|\lambda_2|}\,. \tag{3.11}$$

The gamma distribution has mean $ab = 0.5/|\lambda_2| = 0.737$ with these parameters. The variance is $ab^2 = 0.136$. Thus, the sampling intervals are on the order of the timescale of the slowest decaying eigenmode, but have significant variance.

For comparison, we also generate timeseries with constant sampling interval $\Delta t = 0.5/|\lambda_2|$, with the same length ($N+1$ points). We estimate the spectrum of eigenvalues of Q from both timeseries, either by using the procedure described in this section (for non-constant sampling intervals), or by calculating the spectrum $\{\hat{\Lambda}_i\}$ of the estimator (2.8) and taking the usual $\hat{\lambda}_i = (\Delta t)^{-1}\log\hat{\Lambda}_i$ (for constant sampling intervals). After reconstruction of the spectrum, we calculate Q^* following the procedure from section 2.4.

The test is carried out using various lengths of the timeseries ($N = 10^4$ and $N = 10^5$). We repeat the experiment (generating data, reconstructing the spectrum, calculating Q^*) 100 times for each N. In tables 3 and 4 the mean and standard deviation of the leading estimated eigenvalues are shown, together with the true eigenvalues of Q. In table 5 we show the mean and standard deviation of the average element-wise error $\sum_{i,j}|q^*_{ij} - q_{ij}|/d^2$. As can be seen in the tables, the mean errors and variance of the eigenvalues inferred from non-homogeneously sampled data are very similar to those of the eigenvalues obtained from data with constant sampling interval. The mean errors on the inferred generators themselves (table 5) are even smaller with the random sampling intervals than with the constant intervals.

Table 3

Estimation of generator eigenvalues from timeseries with random and constant sampling intervals. Shown are the leading true eigenvalues λ_i of Q and the mean and standard deviation of the estimated eigenvalues $\hat{\lambda}_i$ from 100 different timeseries of the same length ($N = 10^4$).

		random sampling intervals		constant sampling intervals	
i	λ_i	mean($\hat{\lambda}_i$)	std($\hat{\lambda}_i$)	mean($\hat{\lambda}_i$)	std($\hat{\lambda}_i$)
2	-0.678	-0.677	0.019	-0.678	0.017
3	-1.539	-1.535	0.044	-1.542	0.042
4	-2.369	-2.377	0.085	-2.369	0.069

Table 4

Same as table 3 but with $N = 10^5$.

		random sampling intervals		constant sampling intervals	
i	λ_i	mean($\hat{\lambda}_i$)	std($\hat{\lambda}_i$)	mean($\hat{\lambda}_i$)	std($\hat{\lambda}_i$)
2	-0.678	-0.679	0.0055	-0.678	0.0049
3	-1.539	-1.539	0.015	-1.540	0.012
4	-2.369	-2.372	0.028	-2.371	0.023

Table 5

Generator inference from timeseries with random and constant sampling intervals. Shown are the mean and standard deviation of the average element-wise error $\sum_{i,j} |q^*_{ij} - q_{ij}|/d^2$ from 100 different timeseries of the same length ($N + 1$).

	random sampling intervals		constant sampling intervals	
N	mean(error(Q^*))	std(error(Q^*))	mean(error(Q^*))	std(error(Q^*))
10^4	0.13	0.021	0.15	0.025
10^5	0.046	0.0064	0.070	0.017

4. Numerical aspects of high-dimensional problems.

In section 2, we showed that the estimation of Q can be cast as a minimization problem that falls in the class of convex quadratic programs (QP) with linear equality and inequality constraints. This means that the problem (2.27) has the form

$$Q^* = \operatorname{argmin} E(Q) \tag{4.1}$$

where the objective function E can be written compactly as

$$E = \tfrac{1}{2} V^T H V + V^T F + E_0\,. \tag{4.2}$$

The vector V contains the elements of Q. E must be minimized under the linear constraints (2.4).

The objective function E is strictly convex because $\hat{U}$ has full rank by assumption; this implies that the Hessian matrix H is positive definite. Since the constraints (2.4) define a convex domain, the constrained QP has a unique solution. It is straightforward to absorb the equality constraints (2.4b) into the objective function by eliminating the diagonal elements of Q from the QP, so that the problem can be reformulated as a strictly convex QP of lower dimension with only inequality constraints (nonnegative variables). Without restrictions on Q other than (2.4), the QP has d^2 variables with d equality constraints; thus, it can be reduced to a QP with $d^2 - d$ variables, without equality constraints. If the jump process is restricted to be of certain type, the problem reduces further: for example, if only jumps from i to $i \pm 1 (\mathrm{mod}\ d)$ are allowed (birth-death process on a periodic domain), the QP has only $2d$ degrees of freedom.

If the dimension of the QP becomes too large, numerical solution methods that require explicit storage of the Hessian matrix H (for example, the internal Matlab solver `quadprog`) become impractical. Instead, one has to use solution methods that do not ask for H itself, but only for matrix-vector products HV. As is shown in the next part of this section, these products can be calculated cheaply, without forming H explicitly. In the last part, we discuss numerical solution methods for large-scale QP, and present a large-scale numerical example ($d = 250$).

4.1. Efficient evaluation without explicit Hessian.

If we hold on to the matrix notation for Q (rather than convert it into the vector V), we can write the objective function (2.27) as

$$E(Q) = \|\hat{U}^{-1} Q \hat{U}\|_c^2 + \|\hat{D}_\lambda\|_c^2 - \mathrm{Trace}(QF) \tag{4.3}$$

where $F = \hat{U}\hat{D}^*_{c^2\lambda}\hat{U}^{-1}$ + complex conjugate, and $\hat{D}^*_{c^2\lambda} = \mathrm{diag}(c_1^2\hat{\lambda}_1^*, c_2^2\hat{\lambda}_2^*, ...)$. We define the matrices

$$\Phi = \hat{U}^* D_c \hat{U}^T \tag{4.4a}$$

$$\Psi = (\hat{U}^{-1})^T D_c (\hat{U}^{-1})^* \tag{4.4b}$$

in which $D_c = \mathrm{diag}(c_1, c_2, ...)$ and * denotes (element-wise) complex conjugation. Φ and Ψ are both Hermitian matrices. With these definitions, the quadratic term can be written as

$$\|\hat{U}^{-1} Q \hat{U}\|_c^2 = \mathrm{Trace}(Q^T \Psi\, Q\, \Phi). \tag{4.5}$$

The gradient of the quadratic term with respect to the elements of Q (i.e., the equivalent of the matrix-vector product HV mentioned earlier) is $\Psi\, Q\, \Phi + \Psi^T\, Q\, \Phi^T$. Its evaluation is cheap, requiring only a few matrix multiplications. Ψ and Φ need to be constructed only once, at the beginning of the minimization procedure. It must be stressed that Ψ and Φ are matrices of the same dimensions as Q.

As a final remark, we note that dealing with the equality constraints (2.4b) is straightforward. We take the diagonal elements q_{ii} to be dependent variables and the non-diagonal elements of Q as the independent variables of the minimization problem. The gradient of E with respect to the independent variables is

$$\begin{aligned}\frac{d\,E}{d\,q_{ij}} &= \frac{\partial\,E}{\partial\,q_{ij}} - \frac{\partial\,E}{\partial\,q_{ii}} \\ &= (\Psi\,Q\,\Phi + \Psi^T\,Q\,\Phi^T - F^T)_{ij} - (\Psi\,Q\,\Phi + \Psi^T\,Q\,\Phi^T - F^T)_{ii}\,.\end{aligned} \tag{4.6}$$

4.2. Large-scale QP solution methods and numerical examples. Solution methods for quadratic programs are covered extensively in the literature, see for example [20] for an overview and references. The special case of QP problems with bounds on the variables as only constraints is sometimes referred to as "box constrained" QP problems. Because of the relative simplicity of the box constraints, these problems can be solved in high dimensions. Large-scale QP problems with box constraints are studied in for example [18], [12], [7] and [10]. Below, we show two examples using the solution method described in detail in [10], which is a variant of the gradient method known as the projected alternating Barzilai-Borwein (PABB) method (due to [4]). This method is easy to implement and does not require explicit construction of the Hessian matrix. It is a non-monotonic method, meaning that the decrease of the objective function need not be monotone. The following two examples show that the PABB method makes fast minimization of the objective function possible even if the number of variables of the minimization problem is high.

Example 4. We consider the jump process with generator (3.9) and state-space dimension $d = 250$. This implies that the QP problem has 62250 independent variables, making the need for a large-scale solution algorithm obvious. As mentioned, we use the PABB method (without line search) described in [10] for solving the constrained QP problem. To see the numerical convergence of the PABB method without the results being influenced by sampling errors, we use in the objective function for $\hat{U}$ and $\hat{D}_\lambda$ the spectral decomposition of the true generator (3.9). As initial guess, we take $q_{ij} = 1$ for all non-diagonal elements $(i \neq j)$, which is clearly far from the optimum. The weights c_i in the objective function are all set to one: $c_i = 1\ \forall i$.

In the upper panel of figure 2 the value of the objective function E is shown during 1000 iterations with the PABB algorithm without line search (requiring about 57 seconds computation time using Matlab on a modern PC). The lower panel shows the error $\|Q - Q^{\text{true}}\|/\|Q^{\text{true}}\|$ in the Frobenius norm, $\|A\| = \sqrt{\sum_{i,j} a_{ij}^2}$. As can be seen, the algorithm converges rapidly to the correct solution.

Example 5. As our next example we consider a stochastic matrix P that is non-embeddable because it has real negative eigenvalues of multiplicity one:

$$p_{ij} = \begin{cases} \dfrac{d_i}{|i-j|} & \text{if } \ i \neq j \\ d_i & \text{if } \ i = j \end{cases} \tag{4.7}$$

where $d_i = (1 + \sum_{i\neq j} \frac{1}{|i-j|})^{-1}$. As in the previous example, we set the state space dimension to $d = 250$. The matrices $\hat{U}$ and $\hat{D}_\lambda$ needed in the objective function are

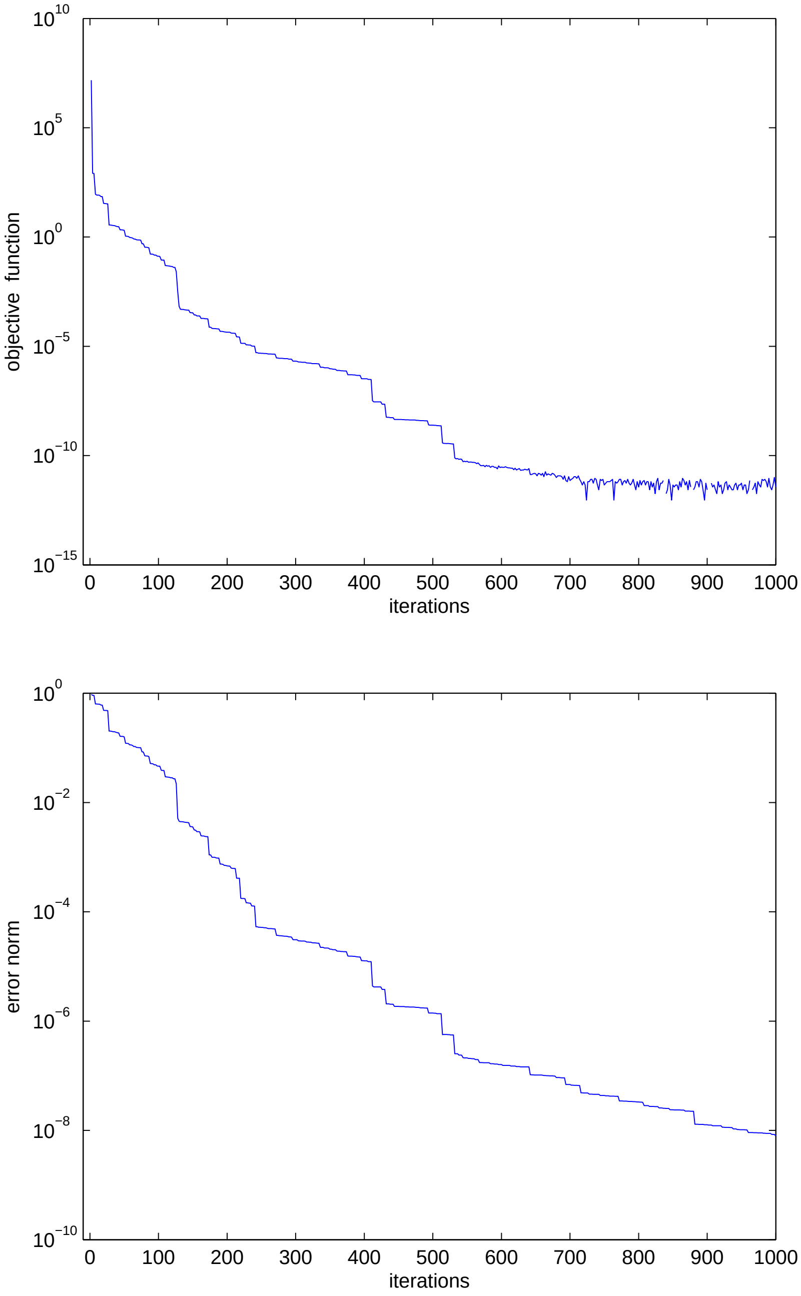

FIG. 2. *Generator reconstruction for large-scale system (generator (3.9) with $d = 250$ states). For numerical minimization the PABB method is used (see text), starting from initial guess $q_{ij} = 1$ for all i, j with $i \neq j$. Shown are the objective function E (upper panel) and the error norm $\|Q - Q^{true}\| / \|Q^{true}\|$ (lower panel) during 1000 iterations of the PABB minimization algorithm. All objective function weights were set to one: $c_i = 1\ \forall i$.*

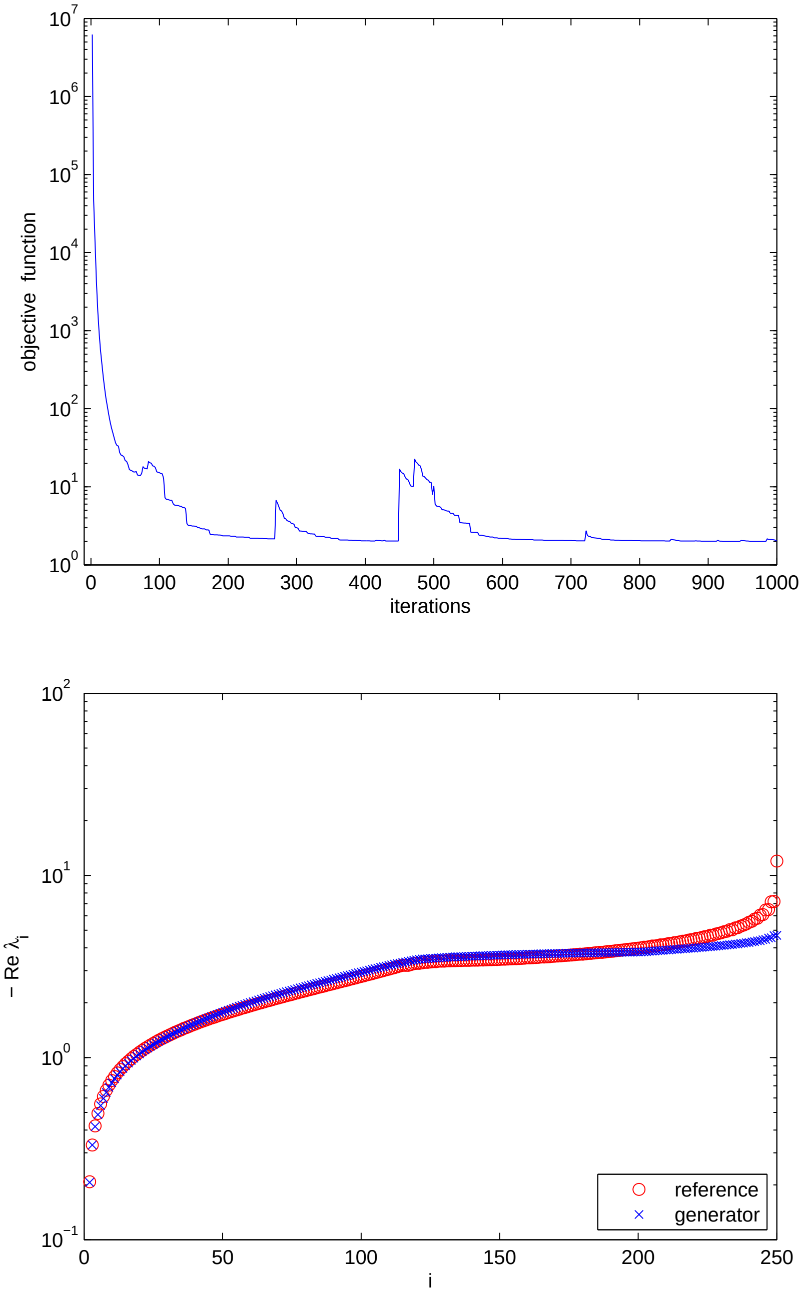

FIG. 3. *Generator reconstruction for large-scale system (stochastic matrix (4.7) with $d = 250$ states). For numerical minimization the PABB method is used (see text), starting from initial guess $q_{ij} = 1$ for all i, j with $i \neq j$. The upper panel shows the objective function E; the lower panel shows the reference eigenvalues derived from P given by (4.7) and the eigenvalues of the generator Q after 1000 iterations of the PABB minimization algorithm.*

taken from the spectral decomposition of P. We set the weights to $c_i = |\lambda_i|^{-1}$ for $i > 1$ and $c_1 = 10\, c_2$ in order to emphasize the leading eigenvalues and eigenvectors. In the upper panel of figure 3 the value of the objective function E is shown during 1000 iterations of the PABB algorithm (with initial guess, as before, $q_{ij} = 1$ for all i, j, with $i \neq j$). It can be seen that the PABB algorithm gives a non-monotonic decrease of the objective function. The lower panel of figure 3 shows the spectrum of eigenvalues of the generator after 1000 iterations as well as the reference eigenvalues $\hat{\lambda}_i$. Except for the trailing eigenvalues ($i \gtrsim 200$), the shape of the spectrum is well recovered; for the first few dozen eigenvalues, the match is particularly close.

5. Conclusion. The inference of generators for Markov jump processes from discretely sampled timeseries is a well-known problem, relevant in various fields of science. Maximum likelihood estimation, the preferred approach for many inference problems in statistics, faces several difficulties with this type of generator estimation, in particular if the sampling interval Δt of the data is not small, see sections 2.2 and 2.3. In this paper, a different approach to generator estimation is discussed that provides an alternative in cases where likelihood inference is not a viable option (for example, because of large Δt or high computational costs).

The approach was first introduced in [8]; in this paper we simplified the procedure (section 2.4) and generalized it to handle inference from data with non-constant sampling intervals (section 3). It consists of two steps: first the spectrum of the generator is estimated from the data, then the generator is fitted to the spectrum through a convex optimization procedure. The result is guaranteed to be a generator, since it satisfies (2.4). In the fitting step we can emphasize the part of the spectrum that is accurately estimated, thereby avoiding that the entire inferred generator is affected in an uncontrolled way by sampling error.

The minimization problem that must be solved to infer the generator is a quadratic program (QP): it has a strictly convex quadratic objective function and linear equality and inequality constraints, see sections 2.4 and 4. Numerical solution methods to find its unique minimum are well-studied and readily available; in section 4 we discussed numerical aspects of estimation for jump processes in high-dimensional state spaces.

The objective function measures the distance between the spectrum of the generator Q and the reference spectrum derived from the MLE $\hat{P}$. This distance is minimized by minimizing the objective function. We point out that adding linear equality constraints to a convex QP does not change its convex quadratic nature, therefore it is possible to construct generators that match elements from the reference spectrum exactly, rather than approximately. An example is the observed invariant distribution $\hat{\mu}$: adding the condition $\hat{\mu}^T Q = 0$ to the QP guarantees that the minimizer Q^* satisfies $\hat{\mu}^T Q^* = 0$ exactly.

Finally, we remark that the approach to generator estimation presented and discussed in this paper can also be used for the estimation of diffusion processes from discrete observations. A first step in this direction was made in [9]. In case of diffusions, the technical details are more complicated because of the presence of continuous instead of discrete state spaces. Nevertheless, the estimation can still be cast as a convex optimization problem. We expect that the generalization to inference from data with non-constant sampling intervals will carry over to the diffusion estimation as well. Further work on diffusion estimation using this approach will be presented elsewhere.

Appendix A. Proof of Theorem 1.

Theorem 1 is a consequence of a central limit theorem for Markov chains [2] which states that

$$\sqrt{N}\left(\hat{P} - P(\Delta t)\right) \to S \quad \text{in probability as } N \to \infty \tag{A.1}$$

where S is a Gaussian matrix whose entries s_{ij} have mean zero and covariance

$$\mathbb{E}(s_{ij}s_{i'j'}) = \delta_{ii'}p_{ij}(\Delta t)\left(\delta_{jj'} - p_{i'j'}(\Delta t)\right)/\mu_i, \qquad i,j,i',j' = 1,\dots,d \tag{A.2}$$

A standard result in matrix perturbation theory (e.g., [23]) states that if Λ_i is a simple eigenvalue of P, and P is perturbed to $P + \delta P$, then $P + \delta P$ has a unique eigenvalue $\Lambda_i + \delta\Lambda_i$ with

$$\delta\Lambda_i = \frac{\psi_i^T \delta P \phi_i}{\psi_i^T \phi_i} + O(\|\delta P\|^2) \tag{A.3}$$

where ψ_i and ϕ_i are the left and right eigenvectors of P associated with Λ_i. Without loss of generality we can assume that the eigenvectors are normalized so that $\psi_i^T\phi_i = 1$. Specializing to the perturbation $\hat{P}$ of $P(\Delta t)$ as in (A.1), this gives

$$\sqrt{N}\delta\Lambda_i = \psi_i^T S \phi_i + O(\frac{1}{\sqrt{N}}\|S\|^2). \tag{A.4}$$

Since the matrix elements of S have mean zero we find, as $N \to \infty$,

$$\mathbb{E}\,\sqrt{N}\delta\Lambda_i \to \mathbb{E}\,\psi_i^T S\phi_i = 0 \tag{A.5}$$

and

$$\mathbb{E}\,\sqrt{N}\delta\Lambda_i\,\sqrt{N}\delta\Lambda_j \to \mathbb{E}\,\psi_i^T S\phi_i\psi_j^T S\phi_j = \sum_{k,l,k',l'} \bar{u}_{ik}\bar{u}_{jk'}u_{li}u_{l'j}\mathbb{E}\,s_{kl}s_{k'l'} \tag{A.6}$$

Using (A.2) gives

$$\mathbb{E}\,\sqrt{N}\delta\Lambda_i\,\sqrt{N}\delta\Lambda_j \to \sum_{k,l} \frac{\bar{u}_{ik}\bar{u}_{jk}(p_{kl} - \Lambda_i\Lambda_j\delta_{kl})u_{li}u_{lj}}{\mu_k}. \tag{A.7}$$

This shows that $\sqrt{N}\delta\Lambda_i \to r_i$ in probability as $N \to \infty$, with $\mathbb{E}\,r_i = 0$ and $\mathbb{E}\,r_ir_j$ as in (2.15). □

Appendix B. Small Δt limit of eigenvalue estimates.

In this section we consider the Δt dependence of the sampling errors on the eigenvalues. In the limit $\Delta t \to \infty$, the real part of the factor $\Delta t^{-1}\exp(-\lambda_i\Delta t)$ blows up the amplitude of the random error (2.18). For convergence as $\Delta t \to \infty$, N must be exponentially large in $-(\mathrm{Re}\lambda_i)\Delta t$.

For the limit of small Δt the following lemma on the covariances of the random errors (2.15), (A.2) will be useful.

Lemma 1. *In the limit $\Delta t \to 0+$ we have*

$$\lim_{\Delta t \to 0+} \mathbb{E}\Big(\frac{r_ir_j}{\Delta t}\Big) = C_{ij} \tag{B.1a}$$

$$\lim_{\Delta t \to 0+} \mathbb{E}\Big(\frac{s_{ij}s_{i'j'}}{\Delta t}\Big) = C'_{iji'j'} \tag{B.1b}$$

where C_{ij} and $C'_{iji'j'}$ are defined as

$$C_{ij} = \sum_{k,l \in S} \frac{\bar{u}_{ik}\bar{u}_{jk}\big(q_{kl} - \delta_{kl}(\lambda_i + \lambda_j)\big)u_{li}u_{lj}}{\mu_k} \tag{B.2a}$$

$$C'_{iji'j'} = \frac{\delta_{ii'}(\delta_{jj'}q_{ij} - \delta_{ij}q_{i'j'} - \delta_{i'j'}q_{ij})}{\mu_i} \tag{B.2b}$$

Proof of Lemma 1. Substitution of the expansions $p_{kl} = \delta_{kl} + q_{kl}\Delta t + O(\Delta t^2)$ and $\Lambda_i = 1 + \lambda_i \Delta t + O(\Delta t^2)$ in (2.15) and (A.2) leads to $\mathbb{E}(r_i r_j) = C_{ij}\,\Delta t + O(\Delta t^2)$ and $\mathbb{E}(s_{ij}s_{i'j'}) = C'_{iji'j'}\,\Delta t + O(\Delta t^2)$. □

For the errors on $\hat{\Lambda}_i$ and $\check{\lambda}_i$ we have $N\,\mathbb{E}(\hat{\Lambda}_i - \Lambda_i)^2 \to \mathbb{E}(r_i^2)$ and $N\,\mathbb{E}(\check{\lambda}_i - \lambda_i)^2 \to (\Delta t^2\,\Lambda_i^2)^{-1}\,\mathbb{E}(r_i^2)$ as $N \to \infty$, cf. (2.14), (2.15) and (2.18). A time series with sampling interval Δt and total length T consists of $N = \lfloor T/\Delta t \rfloor$ data points. If we substitute $N = T/\Delta t$ we obtain

$$\lim_{T\to\infty} T\,\mathbb{E}(\hat{\Lambda}_i - \Lambda_i)^2 = \Delta t^2\,\mathbb{E}\Big(\frac{r_i^2}{\Delta t}\Big) \tag{B.3a}$$

$$\lim_{T\to\infty} T\,\mathbb{E}(\check{\lambda}_i - \lambda_i)^2 = \Lambda_i^{-2}\,\mathbb{E}\Big(\frac{r_i^2}{\Delta t}\Big) \tag{B.3b}$$

By (B.1a), the limit $\Delta t \to 0+$ gives

$$\lim_{\Delta t\to 0+}\ \lim_{T\to\infty} T\,\mathbb{E}(\hat{\Lambda}_i - \Lambda_i)^2 = 0 \tag{B.4a}$$

$$\lim_{\Delta t\to 0+}\ \lim_{T\to\infty} T\,\mathbb{E}(\check{\lambda}_i - \lambda_i)^2 = \Lambda_i^{-2}\,C_{ii} \tag{B.4b}$$

with C_{ii} defined as in Lemma 1. Thus, as already implied by (2.14) and (2.18), the errors on $\hat{\Lambda}_i$ and $\check{\lambda}_i$ disappear in the limit $T \to \infty$ with Δt fixed. However, (B.4) strongly suggests that the error on $\check{\lambda}_i$ does *not* vanish in the limit $\Delta t \to 0+$ with T fixed (even though $N \to \infty$ in this limit). By contrast, $T\,\mathbb{E}(\hat{\Lambda}_i - \Lambda_i)^2$ is of order Δt^2 for small Δt (and large T), suggesting that the error on $\hat{\Lambda}_i$ disappears in the limit $\Delta t \to 0+$ with T fixed (which should not be surprising, because $\Lambda_i \to 1$ for all i as $\Delta t \to 0+$).

Note that the MLE $\hat{Q}$ (assuming it exists) has a similar limiting behavior as $\Delta t \to 0+$. Because $\hat{P} - P(\Delta t) = \Delta t(\hat{Q} - Q) + O(\Delta t^2)$, we have

$$\lim_{T\to\infty} T\,\mathbb{E}(\hat{q}_{ij} - q_{ij})(\hat{q}_{i'j'} - q_{i'j'}) = \mathbb{E}\left(\frac{s_{ij}s_{i'j'}}{\Delta t}\right) + O(\Delta t) \tag{B.5}$$

Recalling (B.1b), we see that the error on $\hat{Q}$ does not vanish in the limit $\Delta t \to 0+$ with T fixed.

Appendix C. Proof of Theorem 2..

Assume that Λ_i is a simple eigenvalue of P with associated left and right eigenvectors ψ_i and ϕ_i, and P is perturbed to $P + \delta P$. From matrix perturbation theory it is known that if δP is sufficiently small, $P + \delta P$ has a simple eigenvalue $\Lambda_i + \delta\Lambda_i$ given by (A.3) and an associated right eigenvector $\phi_i + \delta\phi_i$ with

$$\delta\phi_i = \left(\sum_{j\neq i} \frac{\phi_j \psi_j^T}{\Lambda_j - \Lambda_i}\right)\delta P \phi_i + O(\|\delta P\|^2)\,. \tag{C.1}$$

If $\delta P = \hat{P} - P$ then $\delta P \to S/\sqrt{N}$ in probability as $N \to \infty$, see (A.1), and we find (2.20a); (2.20b) is found in a similar way. □

Appendix D. Proof of Theorem 3.

We will prove that

$$\text{(i)} \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}(X_{t_n} = i) \to \mu_i \quad \text{a.s.} \tag{D.1a}$$

$$\text{(ii)} \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}(X_{t_n} = i)\mathbf{1}(X_{t_{n+1}} = j) \to \mu_i p^e_{ij} \quad \text{a.s.} \tag{D.1b}$$

Convergence $\hat{p}^e_{ij} \to p^e_{ij}$ a.s. then follows from the almost sure version of Slutzky's theorem ([11]). We note that the observation times t_n are determined by the random sampling intervals τ_n:

$$t_n = t_1 + \sum_{m=1}^{n-1} \tau_m \tag{D.2}$$

First, we define

$$\Lambda_* := \mathbb{E}_\tau \mathrm{e}^{\lambda \tau_n} < 1 \tag{D.3}$$

where, as before, $\lambda = \mathrm{Re}\,\lambda_2 < 0$. It can be shown that $\Lambda^* < 1$ by considering the function $\Lambda(\lambda) = \mathbb{E}_\tau \mathrm{e}^{\lambda\tau_n}$ and noting that $\Lambda(\lambda = 0) = 1$ as well as $\frac{d\Lambda}{d\lambda} \geq 0$ for all $\lambda \in \mathbb{R}$. Furthermore, $\frac{d\Lambda}{d\lambda}(\lambda = 0) = \mathbb{E}_\tau \tau_n > 0$. The last inequality follows from the fact that π can not be atomic at zero, condition (3.3). Therefore, $\Lambda(\lambda)$ is strictly smaller than one if λ is strictly smaller than zero.

For notational convenience, we also define

$$S^i_N := \sum_{n=1}^{N} \mathbf{1}(X_{t_n} = i)\,, \quad S^{ij}_N := \sum_{n=1}^{N} \mathbf{1}(X_{t_n} = i)\mathbf{1}(X_{t_{n+1}} = j) \tag{D.4a}$$

$$U^i_n := \mathbf{1}(X_{t_n} = i) - \mu_i\,, \quad U^{ij}_n := \mathbf{1}(X_{t_n} = i)\mathbf{1}(X_{t_{n+1}} = j) - \mu_i p_{ij}(\tau_n) \tag{D.4b}$$

as well as

$$\tilde{p}_{ij}(t) := p_{ij}(t) - \mu_j \tag{D.5}$$

Furthermore, we write $\mathbb{E}_X$ for the expectation with respect to the law of X_t and $\mathbb{E}$ for the expectation with respect to the law of both X_t and τ_n, i.e. $\mathbb{E} = \mathbb{E}_\tau \mathbb{E}_X$. Finally, let ρ denote the probability distribution for X_{t_1}:

$$\rho_i := \mathbb{E}\,\mathbf{1}(X_{t_1} = i) \tag{D.6}$$

Because X_t is an ergodic process with unique invariant measure, there exists a positive constant C_1 such that for all $i, j \in S$ and all $t \geq 0$,

$$|\tilde{p}_{ij}(t)| \leq C_1\,\mathrm{e}^{\lambda t}\,. \tag{D.7}$$

Using (D.7) and the fact that the τ_n are i.i.d.,

$$\mathbb{E}_\tau |\tilde{p}_{ij}(t_{n'} - t_n)| \leq C_1 (\Lambda_*)^{n'-n}\,. \tag{D.8}$$

For any $\varepsilon > 0$,

$$\begin{aligned}\mathbb{P}\left(\left|\tfrac{1}{N}S_N^i - \mu_i\right| \geq \varepsilon\right) &= \mathbb{P}\left(\left(\tfrac{1}{N}S_N^i - \mu_i\right)^4 \geq \varepsilon^4\right)\\ &\leq \frac{1}{\varepsilon^4 N^4}\mathbb{E}\left(S_N^i - N\mu_i\right)^4 \end{aligned} \tag{D.9}$$

using Chebyshev's inequality. We prove that $\mathbb{E}\left(S_N^i - N\mu_i\right)^4 = O(N^2)$.

$$\begin{aligned}&|\mathbb{E}\left(S_N^i - N\mu_i\right)^4| = |\mathbb{E}(\sum_n U_n^i)^4|\\ &\leq \sum_n |\mathbb{E}(U_n^i)^4| + n_1 \sum_{n<n'} |\mathbb{E}U_n^i(U_{n'}^i)^3| + n_2 \sum_{n<n'<n''} |\mathbb{E}U_n^i U_{n'}^i (U_{n''}^i)^2|\\ &\qquad + n_3 \sum_{n<n'<n''<n'''} |\mathbb{E}U_n^i U_{n'}^i U_{n''}^i U_{n'''}^i| \end{aligned} \tag{D.10}$$

where n_1, n_2, n_3 are positive constants determined by permutations of n, n', n'', n'''. The summation $\sum\limits_{n<n'}$ is shorthand notation for $\sum\limits_{n'=1}^{N}\sum\limits_{n=1}^{n'-1}$, etcetera.

- $-1 \leq U_n^i \leq 1$ for all i, n, hence $(U_n^i)^4 \leq 1$ and $|\mathbb{E}\sum\limits_n (U_n^i)^4| \leq N$
- $\mathbb{E}_X U_n^i U_{n'}^i = \Big[\sum\limits_{i'} \rho_{i'} p_{i'i}(t_n - t_1)\tilde{p}_{ii}(t_{n'} - t_n) - \sum\limits_{i'} \rho_{i'}\mu_i \tilde{p}_{i'i}(t_{n'} - t_1)\Big]$. Using (D.8),
  we find $|\mathbb{E}\sum\limits_{n<n'} U_n^i (U_{n'}^i)^3| \leq |\mathbb{E}\sum\limits_{n<n'} U_n^i U_{n'}^i| \leq \sum\limits_{n<n'} \mathbb{E}_\tau |\mathbb{E}_X U_n^i U_{n'}^i|$
  $\leq C_1 \sum\limits_{n<n'} \left[(\Lambda_*)^{n'-n} + (\Lambda_*)^{n'-1}\right] \leq \dfrac{2C_1 N}{1-\Lambda_*}$.
- $\left|\mathbb{E}\sum\limits_{n<n'<n''} U_n^i U_{n'}^i (U_{n''}^i)^2\right| \leq \left|\mathbb{E}\sum\limits_{n<n'<n''} U_n^i U_{n'}^i\right| \leq N\left|\mathbb{E}\sum\limits_{n<n'} U_n^i U_{n'}^i\right| \leq \dfrac{2C_1 N^2}{1-\Lambda_*}$
- By arranging terms (and assuming $n < n' < n'' < n'''$) we can write

$$\begin{aligned}&\mathbb{E}_X U_n^i U_{n'}^i U_{n''}^i U_{n'''}^i =\\ &\quad \sum_{i'} \rho_{i'} \Bigg[\tilde{p}_{ii}(t_{n'''} - t_{n''})\tilde{p}_{ii}(t_{n''} - t_{n'})\tilde{p}_{ii}(t_{n'} - t_n)p_{i'i}(t_n - t_1)\\ &\quad -\mu_i\Big(\tilde{p}_{ii}(t_{n'''} - t_{n''})\tilde{p}_{ii}(t_{n''} - t_n)p_{i'i}(t_n - t_1) + \tilde{p}_{ii}(t_{n'''} - t_{n'})\tilde{p}_{ii}(t_{n'} - t_n)p_{i'i}(t_n - t_1)\\ &\quad -\tilde{p}_{ii}(t_{n'''} - t_{n''})\tilde{p}_{ii}(t_{n'} - t_n)p_{i'i}(t_n - t_1) - \tilde{p}_{ii}(t_{n'''} - t_{n''})\tilde{p}_{ii}(t_{n''} - t_{n'}) \times\\ &\quad (p_{i'i}(t_n - t_1) - p_{i'i}(t_{n'} - t_1))\Big) + \mu_i^2 \tilde{p}_{ii}(t_{n'''} - t_{n'})(p_{i'i}(t_{n'} - t_1) - p_{i'i}(t_n - t_1))\\ &\quad +(\mu_i)^2 \tilde{p}_{ii}(t_{n'''} - t_{n''})(p_{i'i}(t_{n''} - t_1) - p_{i'i}(t_{n'} - t_1)) - (\mu_i)^3 \tilde{p}_{i'i}(t_{n'''} - t_1)\Bigg]\end{aligned}$$

Using $p_{i'i}(t) - p_{i'i}(s) = \tilde{p}_{i'i}(t) - \tilde{p}_{i'i}(s)$ as well as (D.8), we find

$$
\begin{aligned}
&|\mathbb{E} \sum_{n<n'<n''<n'''} U^i_n U^i_{n'} U^i_{n''} U^i_{n'''}| \\
&\quad \le \sum_{n<n'<n''<n'''} \mathbb{E}_\tau |\mathbb{E}_X U^i_n U^i_{n'} U^i_{n''} U^i_{n'''}| \\
&\quad \le C_2 \sum_{n<n'<n''<n'''} \Big[\Lambda_*^{n'''-n''} \Lambda_*^{n''-n'} \Lambda_*^{n'-n} + \Lambda_*^{n'''-n''} \Lambda_*^{n''-n} + \Lambda_*^{n'''-n'} \Lambda_*^{n'-n} \\
&\qquad\quad + \Lambda_*^{n'''-n''} \Lambda_*^{n'-n} + \Lambda_*^{n'''-n''} \Lambda_*^{n''-n'} \Lambda_*^{n} + \Lambda_*^{n'''-n''} \Lambda_*^{n''-n'} \Lambda_*^{n'} \\
&\qquad\quad + \Lambda_*^{n'''-n'} \Lambda_*^{n'} + \Lambda_*^{n'''-n'} \Lambda_*^{n} + \Lambda_*^{n'''-n''} \Lambda_*^{n''} + \Lambda_*^{n'''-n''} \Lambda_*^{n'} + \Lambda_*^{n'''} \Big] \\
&\quad \le C_2' \sum_{n<n'<n''<n'''} \big[\Lambda_*^{n'''-n''} \Lambda_*^{n'-n} + \Lambda_*^{n'''-n''} \Lambda_*^{n} \big] \\
&\quad \le C_2' \Big(\sum_{n''<n'''} \Lambda_*^{n'''-n''} \Big) \Big(\sum_{n<n'} [\Lambda_*^{n'-n} + \Lambda_*^{n}] \Big) \\
&\quad \le \frac{2 C_2' N^2}{(1-\Lambda_*)^2}
\end{aligned}
$$

with positive constants C_2, C_2'.

Thus, we have established that $|\mathbb{E}\left(S^x_N - N\mu(x)\right)^4| \le C_3 N^2$ with some constant $C_3 > 0$. Therefore

$$
\mathbb{P}\left(\left|\tfrac{1}{N} S^i_N - \mu_i\right| \ge \varepsilon\right) \le \frac{C_3}{\varepsilon^4 N^2} \tag{D.11}
$$

By the Borel-Cantelli lemma, (D.1a) follows.

We prove (D.1b) in a similar way. In particular, we need to show that $\mathbb{E}(S^{ij}_N - N\mu_i p^e_{ij})^4 = O(N^2)$. We write

$$
\mathbb{E}(S^{ij}_N - N\mu_i p^e_{ij})^4 = \frac{1}{N^4} \mathbb{E}(\sum_n U^{ij}_n)^4 \tag{D.12}
$$

and decompose and bound $|\mathbb{E}(\sum_n U^{ij}_n)^4|$ in a similar manner as in (D.10):

- $-1 \le U^{ij}_n \le 1$ hence $(U^{ij}_n)^4 \le 1$ and $|\mathbb{E} \sum_n (U^{ij}_n)^4| \le N$
- $\mathbb{E}_X U^{ij}_n U^{ij}_{n'} = \sum_{i'} \rho_{i'} p_{ij}(\tau_n) p_{ij}(\tau_{n'}) \Big[p_{i'i}(t_n - t_1) \tilde{p}_{ji}(t_{n'} - t_{n+1}) - \mu_i \times$ $\tilde{p}_{i'i}(t_n - t_1) \Big]$, therefore $|\mathbb{E} \sum_{n<n'} U^{ij}_n (U^{ij}_{n'})^3| \le |\mathbb{E} \sum_{n<n'} U^{ij}_n U^{ij}_{n'}|$ $\le \sum_{n<n'} \mathbb{E}_\tau |\mathbb{E}_X U^{ij}_n U^{ij}_{n'}| \le \sum_{n<n'} \mathbb{E}_\tau \big(|\tilde{p}_{ji}(t_{n'} - t_{n+1})| + |\tilde{p}_{i'i}(t_n - t_1)| \big)$ $\le \sum_{n<n'} C_1 \big(\Lambda_*^{n'-n-1} + \Lambda_*^{n-1} \big) \le \dfrac{2 C_1 N}{\Lambda_* - \Lambda_*^2}$
- $|\mathbb{E} \sum_{n<n'<n''} U^{ij}_n U^{ij}_{n'} (U^{ij}_{n''})^2| \le |\mathbb{E} \sum_{n<n'<n''} U^{ij}_n U^{ij}_{n'}| \le N |\mathbb{E} \sum_{n<n'} U^{ij}_n U^{ij}_{n'}|$ $\le \dfrac{2 C_1 N^2}{\Lambda_* - \Lambda_*^2}$

- Assuming $n < n' < n'' < n'''$:

$$\begin{aligned}
\mathbb{E}_X U_n^{ij} U_{n'}^{ij} U_{n''}^{ij} U_{n'''}^{ij} = \sum_{i'} \rho_{i'} p_{ij}(\tau_{n'''}) p_{ij}(\tau_{n''}) p_{ij}(\tau_{n'}) p_{ij}(\tau_n) \times \\
\Big[\tilde{p}_{ji}(t_{n'''} - t_{n''+1}) \tilde{p}_{ji}(t_{n''} - t_{n'+1}) \tilde{p}_{ji}(t_{n'} - t_{n+1}) p_{i'i}(t_n - t_1) \\
-\mu_i \tilde{p}_{ji}(t_{n'''} - t_{n'+1}) \tilde{p}_{ji}(t_{n'} - t_{n+1}) p_{i'i}(t_n - t_1) \\
-\mu_i \tilde{p}_{ji}(t_{n'''} - t_{n''+1}) \tilde{p}_{ji}(t_{n''} - t_{n+1}) p_{i'i}(t_n - t_1) \\
+\mu_i \tilde{p}_{ji}(t_{n'''} - t_{n''+1}) \tilde{p}_{ji}(t_{n'} - t_{n+1}) p_{i'i}(t_n - t_1) \\
+\mu_i^2 \tilde{p}_{ji}(t_{n'''} - t_{n+1}) p_{i'i}(t_n - t_1) \\
+\mu_i \tilde{p}_{ji}(t_{n'''} - t_{n''+1}) \tilde{p}_{ji}(t_{n''} - t_{n'+1}) (p_{i'i}(t_n - t_1) - p_{i'i}(t_{n'} - t_1)) \\
+\mu_i^2 \tilde{p}_{ji}(t_{n'''} - t_{n'+1}) (p_{i'i}(t_{n'} - t_1) - p_{i'i}(t_n - t_1)) \\
+\mu_i^2 \tilde{p}_{ji}(t_{n'''} - t_{n''+1}) (p_{i'i}(t_{n''} - t_1) - p_{i'i}(t_{n'} - t_1)) - \mu_i^3 \tilde{p}_{i'i}(t_{n'''} - t_1) \Big]
\end{aligned}$$

We note that $t_{n'} - t_{n+1} \geq 0$ if $n < n'$, and find

$$\begin{aligned}
&|\mathbb{E} \sum_{n<n'<n''<n'''} U_n^{ij} U_{n'}^{ij} U_{n''}^{ij} U_{n'''}^{ij}| \\
&\leq \sum_{n<n'<n''<n'''} \mathbb{E}_\tau |\mathbb{E}_X U_n^{ij} U_{n'}^{ij} U_{n''}^{ij} U_{n'''}^{ij}| \\
&\leq C_4 \sum_{n<n'<n''<n'''} \Big[\Lambda_*^{n'''-n''-1} \Lambda_*^{n''-n'-1} \Lambda_*^{n'-n-1} + \Lambda_*^{n'''-n'-1} \Lambda_*^{n'-n-1} \\
&+\Lambda_*^{n'''-n''-1} \Lambda_*^{n''-n-1} + \Lambda_*^{n'''-n''-1} \Lambda_*^{n'-n-1} + \Lambda_*^{n'''-n-1} \\
&+\Lambda_*^{n'''-n''-1} \Lambda_*^{n''-n'-1} (\Lambda_*^{n} + \Lambda_*^{n'}) + \Lambda_*^{n'''-n'-1} (\Lambda_*^{n'} + \Lambda_*^{n}) \\
&+\Lambda_*^{n'''-n''-1} (\Lambda_*^{n''} + \Lambda_*^{n'}) + \Lambda_*^{n'''} \Big] \\
&\leq C_4' N^2
\end{aligned}$$

with some positive constants C_4, C_4'.

We conclude that for any $\varepsilon > 0$ there is a constant $C_5 > 0$ such that

$$\mathbb{P}\left(\left| \tfrac{1}{N} S_N^{ij} - \mu_i p_{ij}^e \right| \geq \varepsilon \right) \leq \frac{C_5}{\varepsilon^4 N^2} \tag{D.13}$$

Using the Borel-Cantelli lemma we find (D.1b). ▯

REFERENCES

[1] W. J. Anderson, *Continuous-time Markov chains*, Springer, New York, 1991.

[2] T. W. Anderson and L. A. Goodman, *Statistical inference about Markov chains*, Ann. Math. Statist., 28 (1957), pp. 89–110.

[3] S. Asmussen, O. Nerman and M. Olsson, *Fitting phase-type distributions via the EM algorithm*, Scan. J. Stat., 23 (1996), pp. 419–441.

[4] J. Barzilai and J. M. Borwein, *Two-point step size gradient methods*, IMA J. Numer. Anal., 8 (1988), pp. 141–148.

[5] P. Billingsley, *Statistical inference for Markov processes*, University of Chicago Press, Chicago, 1961.

[6] M. Bladt and M. Sørensen, *Statistical inference for discretely observed Markov jump processes*, J. R. Statist. Soc. B, 67 (2005), pp. 395–410.

[7] T. F. Coleman and Y. Li, *A reflective Newton method for minimizing a quadratic function subject to bounds on some of the variables*, SIAM J. Optim., 6 (1996), pp. 1040–1058.

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

[9] ———, *Reconstruction of diffusions using spectral data from timeseries*, Comm. Math. Sci., 4 (2006), pp. 651–668.

[10] Y. -H. and R. Fletcher, *Projected Barzilai-Borwein methods for large-scale box-constrained quadratic programming*, Numer. Math., 100 (2005), pp. 21–47.

[11] T. S. Ferguson, *A course in large sample theory*, Chapman & Hall, London, 1996.

[12] A. Friedlander and J. M. Martínez, *On the maximization of a concave quadratic function with box constraints*, SIAM J. Optim., 4 (1994), pp. 177–192.

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

[14] R. B. Israel, J. S. Rosenthal and J. Z. Wei, *Finding generators for Markov chains via empirical transition matrices, with applications to credit ratings*, Math. Finance, 11 (2001), pp. 245–265.

[15] T. G. Kurtz, *A limit theorem for pertrubed operator semigroups with applications to random evolution*, J. Functional Analysis, 12 (1973), pp. 55–67.

[16] P. Metzner, E. Dittmer, T. Jahnke and Ch. Schütte, *Generator estimation of Markov jump processes*, J. Comp. Phys., 227 (2007), pp. 353–375.

[17] P. Metzner,I. Horenko, I. and Ch. Schütte, *Generator estimation of Markov jump processes based on incomplete observations non-equidistant in time*, Phys. Rev. E, 76 (2007), pp. 066702

[18] J. J. Moré and G. Toraldo, *On the solution of large quadratic programming problems with bound constraints*, SIAM J. Optim., 1 (1991), pp. 93–113.

[19] T. Müller and M. Vingron, *Modeling amino acid replacement*, J. Comput. Biol., 7 (2000), pp. 761–776.

[20] J. Nocedal and S. J. Wright, *Numerical optimization* (2nd edition), Springer, New York, 2006.

[21] J. R. Norris, *Markov chains*, Cambridge University Press, Cambridge, 1997.

[22] G. C. Papanicolaou, *Introduction to the asymptotic analysis of stochastic equations*, in Modern Modeling of Continuum Phenomena, R. DiPrima, ed., Lectures in Applied Mathematics, Vol. 16 (1977), Amercian Mathematical Society, Providence RI, pp. 109–147.

[23] G. W. Stewart and J. Sun, *Matrix perturbation theory*, San Diego: Academic Press, San Diego, 1990.