UNIVERSITY OF BATH

Citation for published version:
Lencastre, P, Raischel, F, Rogers, T & Lind, PG 2016, 'From empirical data to time-inhomogeneous continuous Markov processes', *Physical Review E*, vol. 93, no. 3, 032135. https://doi.org/10.1103/PhysRevE.93.032135

DOI:
10.1103/PhysRevE.93.032135

Publication date:
2016

Document Version
Peer reviewed version

Link to publication

University of Bath

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 18. Sep. 2019

From empirical data to continuous Markov processes: a systematic approach

Pedro Lencastre,[1] Frank Raischel,[2] Tim Rogers,[3] and Pedro G. Lind[4,5]

[1]*Mathematical Department, FCUL, University of Lisbon, 1749-016 Lisbon, Portugal*
[2]*Instituto Dom Luiz, University of Lisbon, 1749-016 Lisbon, Portugal*
[3]*Centre for Networks and Collective Behaviour, Department of Mathematical Sciences, University of Bath, Claverton Down, BA2 7AY, Bath, UK*
[4]*ForWind - Center for Wind Energy Research, Institute of Physics, Carl-von-Ossietzky University of Oldenburg, DE-26111 Oldenburg, Germany*
[5]*Institut für Physik, Universität Osnabrück, Barbarastrasse 7, 49076 Osnabrück, Germany*
(Dated: October 27, 2015)

We present an approach for testing for the existence of continuous generators of discrete stochastic transition matrices. Typically, the known approaches to ascertain the existence of continuous Markov processes are based in the assumption that only time-homogeneous generators exist. Here, a systematic extension to time-inhomogeneity is presented, based in new mathematical propositions incorporating necessary and sufficient conditions, which are then implemented computationally and applied to numerical data. A discussion concerning the bridging between rigorous mathematical results on the existence of generators to its computational implementation. Our detection algorithm shows to be effective in more than 80% of tested matrices, typically 90% to 95%, and for those an estimate of the (non-homogeneous) generator matrix follows. We also solve the embedding problem analytically for the particular case of three-dimensional circulant matrices. Finally, a discussion of possible applications of our framework to problems in different fields is briefly addressed.

PACS numbers: 02.50.Ga, 05.10.Gg, 02.10.Yn, 89.65.Gh
Keywords: Continuous Markov Processes, Embedding Problem, Inhomogeneous Generators, Master Equation

arXiv:1510.07282v1 [physics.data-an] 25 Oct 2015

I. MOTIVATION

While models describing the evolution of a set of variables are typically continuous, observations and experiments retrieve discrete sets of values. Therefore, to bridge between models and reality one has to know if it is reasonable to assume a continuous "reality" underlying the discrete set of measurements. When the evolution has a non negligible stochastic contribution, one typically extracts from the set of measurements the distribution $\vec{P}(X, t-\tau)$ of the observed values $X(t-\tau)$, from which the probability density function (PDF) can be inferred. By knowing the distribution $\vec{P}(X,t)$ at a future time t, one is then able to define a transition matrix $\mathbf{T}(t,\tau)$ that satisfies:

$$\vec{P}(X,t) = \mathbf{T}(t,\tau)\vec{P}(X,t-\tau), \tag{1}$$

if we know the fraction of transitions T_{kj} from each observed value $X_j(t-\tau)$ at time $t-\tau$ to a value $X_k(t)$ at time t. The transition matrix $\mathbf{T}(t,\tau)$ has all its elements T_{kj} in the interval $[0,1]$, has row-sums one, $\sum_k T_{kj} = 1$, and has non-negative entries, $T_{kj} \geq 0$.

In this paper we address the problem of determining whether or not the evolution of a system is governed by a time-continuous Markov master equation. This problem is usually called the embedding problem [1]. Time-continuous Markov processes, have particular mathematical properties, namely they memoryless stochastic processes: the probability of transition between states $X(t)$ and $X(t+\tau)$ does not depend on the states of the system for times previous to t, for any $\tau > 0$. If the stochastic process is time-continuous and Markovian, than the transition matrix can be defined for infinitely small τ, obeying an equation of the form

$$\frac{d\mathbf{T}(t,\tau)}{d\tau} = \mathbf{Q}(t)\mathbf{T}(t,\tau)\,, \tag{2}$$

where $\mathbf{Q}(t)$ is called the generator matrix of the process, having zero row-sums and non-negative off-diagonal entries. Notice that, $\mathbf{T}(t,\tau)$ is a transition matrix for all t and τ, i.e. with non-negative real elements and unity row-sums, if and only if it obeys Eq. (2) for some $\mathbf{Q}(t)$[2].

Both equations above allow us to write the continuous-time evolution of a PDF. In other words, the time evolution of such PDF can be described by a master equation in continuous time. The Master Equation Approach is a fundamental tool in Statistical Physics used to derive important results in Thermodynamics[3] and in several interdisciplinary applications[4].

The transition matrix $T(t,\tau)$, solution of Eq. (2), defines the evolution equation, Eq. (1), of the PDF. Thus, the entries Q_{kj} of the generator matrix represent the transition rate between states j and k at time t. Time-continuity is a property that results from the fact that all entries of $\mathbf{Q}(t) \equiv \mathbf{Q}$, i.e. all transition rates, are finite. The general solution of Eq. (2) yields the relation between the empirical transition matrix and the "continuous" generator which, in the particular case of a time-homogeneous transition matrix, has the form

$$\mathbf{T}(t,\tau) \equiv \mathbf{T}(\tau) = \exp(\mathbf{Q}\tau), \tag{3}$$

for all times t. In general, the embedding problem reduces to the problem of being able to write the transition matrix $\mathbf{T}(t,\tau)$ as solution of Eq. (2) and typically one considers the particular case of a time-homogeneous solution, Eq. (3).

While time-homogeneity is a useful common assumption, it is in several cases too restrictive. Assuming time-homogeneity has the advantage of knowing all future evolution of a time-homogeneous Markov process from the law of the change of system's configuration in two distinct instants (see Eq. (3)), but simultaneously one is not able to address more realistic cases of non-stationary systems. Previous progress in this topic has been made recently. Shintani and Shinomoto have examined an optimized Bayesian rate estimator in cases where the probability density function is not constant in time[5].

In this scope, there are three main reasons for considering an empirical transition matrix to not be time-homogeneous embeddable. The first one is when the underlying process is not Markovian. Such scenario was previously addressed by us[6, 7]. One second reason is the statistical error any empirical data set is subjected to. Typically, one defines for these cases an interval of confidence (a distance) beyond which embeddability is rejected. The third reason is, of course, that the underlying process is itself not time-homogeneous. In this case, there is no time-homogeneous generator, but there is sill the chance that an inhomogeneous generator exists.

In this paper, we address analytically and numerically the case of time-inhomogeneous generators and test their implementation in one framework to address synthetic numerical data, dealing with statistical error of transition matrices. We will also review the time-homogeneous embedding problem, introduced in 1937 by Elfving[1], providing an analytical example in three-dimensions.

We start in Sec. II by describing the standard time-homogeneous problem with the main mathematical theorems that give the necessary and sufficient conditions for a generator matrix to exist. In Sec. III we illustrate this standard time-homogeneous embedding problem by applying the results to the specific case of a circulant transition matrix. Sections IV and V are the heart of this paper, the former establishes the main mathematical theorems that are still valid for the general case of inhomogeneous generators and the latter describes their implementation in a framework that is then tested with synthetic data. Finally, discussions and conclusions are given in Sec. VI.

II. THE HOMOGENEOUS EMBEDDING PROBLEM

The question of knowing if a time-homogeneous generator $\mathbf{Q}$ (see Eq. (3)) exists is known as homogeneous embedding problem[1] and, from the mathematical point of view is currently an open problem for matrices with dimension $n \geq 3$. The problem in dimension two was solved in 1962 by Kingman[2], who proved that, for $n = 2$, a matrix is embeddable if and only if its determinant is positive. More recently, developments in three dimensions were done with the study of matrices with repeated negative eigenvalues[8].

Part of the difficulty when addressing the embedding problem arises from the fact that the logarithm of a matrix is, in general, not unique. This is crucial when deriving a generator $\mathbf{Q}$, by inverting Eq. (3). Indeed, the logarithm of a matrix has counter-intuitive properties, namely:

(i) The product of two embeddable transition matrices $\mathbf{T}_1$ and $\mathbf{T}_2$ is also a transition matrix not necessarily embeddable.

(ii) Having two transition embeddable matrices with generators $\mathbf{Q}_1$ and $\mathbf{Q}_2$, if their product is embeddable then its generator is not necessarily $\mathbf{Q}_1 + \mathbf{Q}_2$, unless the transition matrices commute.

(iii) It is possible that the product of two matrices, $\mathbf{T}_1\mathbf{T}_2$, is embeddable, but the product $\mathbf{T}_2\mathbf{T}_1$ is not.

Since the logarithm of a matrix is not unique, one defines the so-called principal logarithm of one matrix $\mathbf{T}$ as[9]

$$\log \mathbf{T} = \frac{1}{2\pi i} \int_\gamma \log z \, (z\mathbf{I} - \mathbf{T})^{-1} \, dz \,, \tag{4}$$

where γ is a path in the complex plane which does not intersect the negative real semi-axis and encloses all eigenvalues of $\mathbf{T}$. Computationally, one uses the Taylor expansion of the logarithmic function, yielding

$$\log \mathbf{T} = \sum_{n=1}^{\infty} (-1)^{n+1} \frac{(\mathbf{T} - \mathbf{I})^n}{n} \,, \tag{5}$$

which is the the principal branch of the complex logarithm in Eq. (4) or other numerical methods, such as Schur decomposition.

To ascertain if the principal logarithm is computable one has the following proposition[10]:

Proposition II.1. *Let $S = \max\left((a-1)^2 + b^2\right)$ where a and b are real coefficients of an eigenvalue $\lambda = a + ib$ of the transition matrix $\mathbf{T}$. If $S < 1$ then the polynomial series of the $\log \mathbf{T}$, Eq. (5), converges to a matrix with zero row-sums.*

While the existence of the logarithm of a transition matrix is necessary for our purposes, it does not solve the full embedding problem. One must assure further that a valid generator exists, i.e. a matrix with non-negative off-diagonal entries and zero row-sums. Moreover, it is also true that, if $S > 1$, one cannot claim that $\mathbf{T}$ has no generator: another generator may exist in a different branch.

We are interested in the general case of knowing if there is a valid generator, and if there is, to find it. For that, we need to solve the full embedding problem. The full embedding problem comprises a set of propositions which are separated in four different categories:

(A) Conditions for the convergence of the principal logarithm, as presented above in Proposition II.1, that determine if the matrix defined in Eq. (5) has finite entries Q_{kj}.

(B) Necessary conditions for the existence of a generator.

(C) Sufficient conditions for the existence of a generator.

(D) Uniqueness conditions of the generator for properly defining the underlying continuous process.

The conditions for the convergence of the principal logarithm are mainly included in Proposition II.1. Most of the other known results, comprehending categories (B), (C) and (D) are enumerated in the papers by Israel and co-workers[10] and Davies[9]. In the following we present an overview of the most relevant propositions.

Regarding the necessary conditions, important for establishing that a generator cannot exist, there are three highly used propositions easy to implement[10]. The first one is:

Proposition II.2. *If a transition matrix $\mathbf{T}$ obeys one of the following conditions*

a) $\mathrm{Det}\,(\mathbf{T}) \leq 0$,

b) $\mathrm{Det}\,(\mathbf{T}) > \prod_i T_{ii}$,

c) $T_{ij} = 0$ *and there is an integer n such that* $(\mathbf{T}^n)_{ij} \neq 0$,

then no valid generator exists.

For $\mathbf{Q} = \log(\mathbf{T})$, the equality

$$\mathrm{Tr}\,(\mathbf{Q}) = \log(\mathrm{Det}\,(\mathbf{T})) \tag{6}$$

gives the right insight to the property a) in Prop. II.2 since the logarithm of a real number is only defined for positive values. Property b) is related with the definition of determinant. As for property c), suppose that a minimum of t transitions are needed to go from i to j. If the processes is not time-continuous and transitions do not occur more than once in a time period Δt, then an entity can only go from i to j in a number of transitions larger than $(t-1)/\Delta t$. This naturally is not true for time-continuous processes, since there is always a non-zero probability of making t transitions between different states over any time-window. For a complete proof of Prop. II.2 see Ref. [10].

The second proposition is:

Proposition II.3. *For a transition matrix $\mathbf{T}$ with distinct eigenvalues, a generator $\mathbf{Q}$ exists only if, given any eigenvalue of $\mathbf{Q}$ in the form $\lambda = a + ib$, it satisfies the condition* $|b| \leq |\log(\mathrm{Det}\,\mathbf{T})|$.

Proposition II.3 is related to the previous one. Consider $\mathbf{T}$ embeddable and define $k \equiv \mathrm{Tr}\,(\mathbf{Q}) = \log(\mathrm{Det}\,(\mathbf{T}))$ (see Eq. (6)). All entries of matrix $\mathbf{Q}' = \mathbf{Q} - \mathbf{1}k$ are non-negative and its row-sums are equal to $-k$. From Perron-Frobenius Theorem we know that all eigenvalues of matrix $\mathbf{Q}'$ have an absolute value not smaller than $-k$. Since $\lambda = a + ib$ is an eigenvalue of $\mathbf{Q}$, then $\lambda' = (a-k) + ib$ is an eigenvalue of $\mathbf{Q}'$, yielding $k > |\lambda'| > |b|$.

A third necessary condition defines the region of the complex plane that contains the eigenvalues of $\mathbf{T}$, if a generator exists:

Proposition II.4. *If $\mathbf{T}$ is a $n \times n$ matrix and has a generator, then its eigenvalue spectrum is given by $\lambda_k = r_k \exp(i\theta_k)$, where $-\pi \leq \theta \leq \pi$ and*

$$r \leq \exp\left(-\theta \tan\left(\tfrac{\pi}{n}\right)\right). \tag{7}$$

The proof of this proposition, and a general description of the inverse eigenvalue problem can be found in Ref. [11, 12]. It is related with the inverse eigenvalue problem, and can also be used when studying the existence of stochastic roots of matrices.

One additional necessary condition for time-homogeneous generators that will be usefull below when comparing with time-inhomogeneous generators is the following one:

Proposition II.5. *If $\mathbf{T}$ is embeddable, then every negative eigenvalue of $\mathbf{T}$ has even algebraic multiplicity.*

In general, Prop. II.5 is usefull for the cases when $\mathbf{T}$ has negative real eigenvalues.

Sufficient conditions for the existence of one generator, usually deal with considering different branches of the logarithm of the transition matrix and check if they are valid generators, i.e., if their off-diagonal entries are real and positive, and their row-sums are one. In the particular case when it is known that the only possible generator is the principal logarithm, then computing Eq. (5) gives a complete answer to whether or not a valid generator exists. In case all necessary conditions hold, it is legitimate to raise the hypothesis a generator may exist, but there is still the question if the generator is unique.

The following two propositions are sufficient conditions for the uniqueness of one homogeneous generator[10]. The first one reads:

Proposition II.6. *Let $\mathbf{T} \in \mathbb{R}^{n\times n}$ be a transition matrix.*

a) If $\mathrm{Det}\,(\mathbf{T}) > \frac{1}{2}$, *then $\mathbf{T}$ has at most one generator.*

b) If $\mathrm{Det}\,(\mathbf{T}) > \frac{1}{2}$ *and* $||\mathbf{T} - \mathbf{I}|| < \frac{1}{2}$ *using any operator norm, then* $\log(\mathbf{T})$ *is the only possible generator of* $\mathbf{T}$.

c) If $\mathbf{T}$ has distinct eigenvalues, and $\mathrm{Det}\,(\mathbf{T}) > \exp(-\pi)$, *then* $\log(\mathbf{T})$ *is the only possible generator of* $\mathbf{T}$.

The second property b) guarantees that, when there are no repeated eigenvalues, only a finite number of generators exist. Such property is particularly relevant, since in this case it is often possible to find all generators[10].

The second proposition for the uniqueness of one generator is:

Proposition II.7. *If $\mathbf{T}$ is a Markov matrix with distinct eigenvalues $\lambda_1, \dots, \lambda_n$. We have that*

a) Only a finite number of solution $e^{\mathbf{Q}} = \mathbf{T}$ can be Markov Generators.

b) If $|\lambda_r| > \exp\left(-\pi \tan\left(\frac{\pi}{n}\right)\right)$ *for all r, then the principal logarithm is the only $\mathbf{Q}$ such that* $\exp(\mathbf{Q}) = \mathbf{T}$.

The proof of both Props. II.6 and II.7 can be found in Ref. [10].

III. A EXAMPLE: THE CIRCULANT TRANSITION MATRIX

As a mathematical problem, the embedding problem is still open for a general n-dimensional matrix, but it can be analytically solved for some subclasses of matrices. In this section we address in detail a simple example in three dimensions, namely the embedding of circulant transition matrices of the form:

$$\mathbf{T}_C = \begin{pmatrix} a & b & c \\ c & a & b \\ b & c & a \end{pmatrix}, \tag{8}$$

or simply $\mathbf{T}_C = \text{circ}(a, b, c)$, with $0 \geq a, b, c \geq 1$ and $a + b + c = 1$. Circulant transition matrices have two independent degrees of freedom: any pair of values (a, b) can represent a three-dimensional circulant transition matrices if $a + b < 1$, $a, b > 0$. See the triangular region in Fig. 1.

It is easy to check that all necessary conditions in Prop. II.2 for a generator to exist are fulfilled if $0 < a^3 + b^3 + c^3 - a^3b^3c^3 \leq a^3$. Further, according to Prop. II.3 a generator may exist if the argument of the eigenvalues of $\mathbf{T}_C$ are not larger than $\log(a^3 + b^3 + c^3 - a^3b^3c^3)$.

For the particular case of the circulant transition matrix, only Prop. II.4 matters, since in this case it turns out to be a necessary and sufficient condition as we next prove.

To that end, we write the transition matrix as $\mathbf{T}_C = \exp \mathbf{Q}_C$, since the exponential of a circulant matrix with real entries is itself a circulant matrix with real entries and consider $\mathbf{Q}_C$ in the form $\mathbf{Q}_C = \text{circ}(-\alpha, \beta, \gamma)$. For $\mathbf{Q}_C$ to be a generator we need to prove that $\alpha, \beta, \gamma > 0$.

The row-sums of $\mathbf{T}_C$ are equal to one by definition and this can only happen if the row-sums of $\mathbf{Q}_C$ are equal to zero. Thus, the equality $\alpha = \beta + \gamma$. Moreover, it can be shown that, computing the principal logarithm of $\mathbf{T}_C$, yields a matrix with negative diagonal elements. Thus we take $\alpha > 0$.

Since $\alpha > 0$ and all entries of the generator $\mathbf{Q}$ are real we need only to prove that β and γ are both non-negative. Since $\alpha = \beta + \gamma$, either β or γ must be positive. Therefore we only need to prove that $\beta\gamma > 0$.

Proposition II.4 gives a condition for the eigenvalues of the transition matrix $\mathbf{T}_C$ to have a generator matrix. It can be proven[13] that such condition hold if and only if an equivalent condition for $\mathbf{Q}_C$ holds, namely:

$$\left| \frac{\Im(\lambda_i)}{\Re(\lambda_i)} \right| < \tan(\frac{\pi}{3}), \tag{9}$$

where λ_i with $i = 1, 2, 3$ are the eigenvalues of $\mathbf{Q}_C$[14]:

$$\lambda_1 = 0, \tag{10a}$$
$$\lambda_2 = -\beta - \gamma + \beta k + \gamma k^*, \tag{10b}$$
$$\lambda_3 = -\beta - \gamma + \beta k^* + \gamma k = \lambda_2^*, \tag{10c}$$

with $k = e^{\frac{2\pi i}{3}}$ and k^* its complex conjugate.

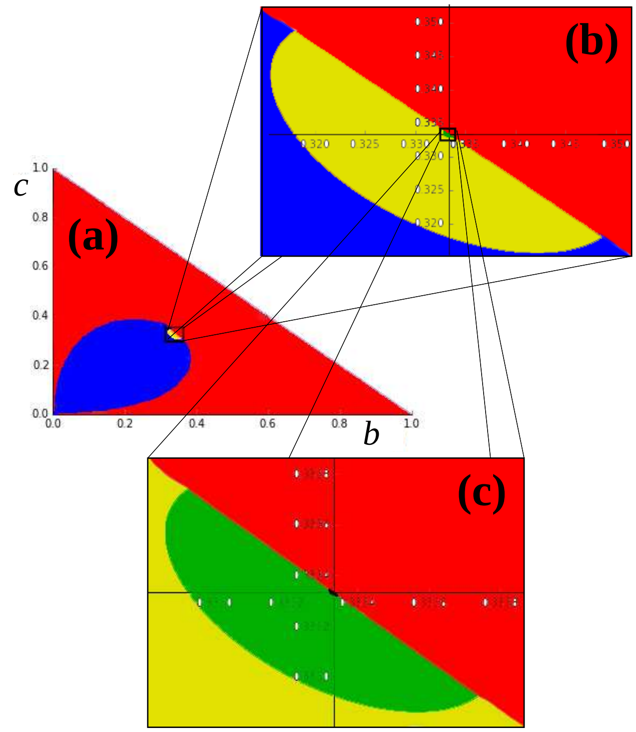

FIG. 1: (Color online) **(a)** Region in parameter-space of transition matrix $\mathbf{T}_C$, Eq. (8), for which a generator $\mathbf{Q}_C$ exists. The triangular (red) region is the one for which matrix $\mathbf{T}_C$ is a transition matrix, $a, b, c > 0$ and $a + b + c = 1$. The blue region indicates the region of parameter values for which only one generator exists, while the yellow (small) region indicates a region where two generators exist. **(b)** By zooming in this region shows a green region where three generators exist, and **(c)** continuing to zoom shows smaller and smaller regions, where a larger number of genertors exist (see text).

Using λ_2 in Eq. (10b) and substituting in Eq. (9), yields

$$\left| \frac{\lambda_2 - \lambda_2^*}{\lambda_2 + \lambda_2^*} \right| < \tan\left(\frac{\pi}{3}\right) \tag{11}$$

and through algebraic manipulation one arrives to

$$\left| \frac{\beta - \gamma}{\beta + \gamma} \right| < 1. \tag{12}$$

The last inequality implies necessarily that $\beta\gamma > 0$. A similar result is obtained by substituting in Eq. (9) one of the other eigenvalues λ_0 and λ_2.

Hence, in our particular case of a circulant matrix, Prop. II.4 is also a sufficient condition and one needs only to determine the inequality in Eq. (7) as a function of the degrees of freedom in matrix $\mathbf{T}_C$ for all its three eigenvalues

$$\lambda_1^{(T)} = 1, \tag{13a}$$
$$\lambda_2^{(T)} = \tfrac{1}{2}(2 - 3b - 3c) + \tfrac{\sqrt{3}}{2}(b - c)i, \tag{13b}$$
$$\lambda_3^{(T)} = \tfrac{1}{2}(2 - 3b - 3c) - \tfrac{\sqrt{3}}{2}(b - c)i, \tag{13c}$$

The first eigenvalue is independent of the parameters. The other two are complex conjugate, having the same norm r and symmetric arguments θ. Thus, we only need to consider one eigenvalue, say $\lambda_3^{(T)} = r\exp(i\theta)$, which according to Prop. II.4 for $\mathbf{T}_C$ to be embeddable must fullfil $r \le \exp(-\sqrt{3}\theta)$ with

$$r = \tfrac{1}{2}\left((2-3(b+c))^2 + 3(b-c)^2\right)^{1/2} \tag{14}$$

and

$$\theta = \begin{cases} \arctan\tilde\theta & \Leftarrow \; c < \frac{2}{3} - b, \\ \arctan\tilde\theta + \mathrm{sgn}\,(b-c)\pi & \Leftarrow \; c > \frac{2}{3} - b, \\ \frac{\pi}{2}\,\mathrm{sgn}\,(b-c) & \Leftarrow \; c = \frac{2}{3} - b, \end{cases} \tag{15}$$

where $\tilde\theta = \sqrt{3}(b-c)/(2-3b-3c)$.

Figure 1 shows the region within the triangle $1-b-c>0$, $b>0$ and $c>0$ where the circulant transition matrix $\mathbf{T}_C$ has a generator, i.e. the region where $a = 1-b-c$ and b and c obey the inequality in Eqs. (7), (14) and (15). The number of valid generators of $\mathbf{T}_C$, a three-dimensional circulant transition matrix, can also be determined from its eigenvalues, namely it is given by the largest integer smaller than $\left(\sqrt{3}\log\left(\Re^2(\lambda^{(T)}) + \Im^2(\lambda^{(T)})\right)\right)/(4\pi)$.

Figure 1a shows one blue region and one smaller yellow region. While the blue region indicates the set of parameter values for b and c for which only one generator exists, the yellow region comprehends the set of parameter values for which $\mathbf{T}_C$ has two or more generators. By zooming in this region, smaller and smaller regions appear, Figs. 1b and 1c, near the crossing point between the diagonal $c=b$ and the line $c = \frac{2}{3} - b$, where a larger number of generators exist.

IV. THE TIME-INHOMOGENEOUS EMBEDDING PROBLEM

In this section we show which of the known theorems for time-homogeneous embedding problem hold when both transition matrix and its generator depend explicitly on time. In this scope, we provide three new conditions, two necessary and one sufficient, for the existence of a time-inhomogeneous generator. We also provide two additional necessary and sufficient conditions which enable the possibility for testing equivalent matrices.

The generator $\mathbf{Q}(t)$ is considered to explicitly depend on time t, as well as its corresponding transition matrix $\mathbf{T}(t,\tau)$. As stated in the introduction, a transition matrix is solution of Eq. (2), i.e. it has a generator if and only if it describes a time-continuous and Markov process, besides having the properties of a transition matrix (non-negative entries and unitary row-sums).

For time-inhomogeneity, the general solution of Eq. (2) is given by:

$$\mathbf{T}(t,\tau) = \sum_{k=0}^{\infty} X_k(t-\tau)\,, \tag{16}$$

with $X_0(t-\tau) \equiv X_0 = \mathbf{1}$ and

$$X_{k+1}(t-\tau) = \int_{t-\tau}^{t} X_k(s)\mathbf{Q}(s)ds\,. \tag{17}$$

Equation (17) is known as the Peano-Baker series[15]. In the particular case that $\mathbf{Q}(t)$ and $\mathbf{Q}(t')$ commute for all t and t' solution (16) reads

$$\mathbf{T}(t,\tau) = \exp\left(\int_{t-\tau}^{t} \mathbf{Q}(s)ds\right). \tag{18}$$

The first necessary proposition for time-inhomogeneous generators follows simply from the fact that $\mathbf{T}(t,\tau)$ is a transition matrix:

Proposition IV.1. *If a transition matrix $\mathbf{T}(t,\tau)$ has a negative determinant, then no generator $\mathbf{Q}(s)$ exists, for $t < s < t+\tau$.*

Proof. To prove the positiveness of the determinant we start by assuming that a generator $\mathbf{Q}(t)$ exists. Then, letting the arguments of $\mathbf{T}$ and $\mathbf{Q}$ drop for simplicity,

$$\frac{d}{dt}\,\mathrm{Det}\,\mathbf{T} = \mathrm{Det}\,\mathbf{T}\,\mathrm{Tr}\left(\mathbf{T}^{-1}\frac{d\mathbf{T}}{dt}\right) \tag{19a}$$

$$\frac{d\log\left(\mathrm{Det}\,\mathbf{T}\right)}{dt} = \mathrm{Tr}\left(\mathbf{T}^{-1}\mathbf{T}\mathbf{Q}\right) \tag{19b}$$

$$\mathrm{Det}\,\mathbf{T} = \exp\left(\int_{t-\tau}^{t}\mathrm{Tr}\,\mathbf{Q}ds\right) > 0. \tag{19c}$$

The final inequality stands true since the trace of $\mathbf{Q}(t)$ is always a real (negative) value. □

The second necessary proposition deals also with the fact that $\mathbf{T}$ is a transition matrix, namely that its entries are probabilities:

Proposition IV.2. *If a transition matrix $\mathbf{T}$ fulfills $\mathrm{Det}\,\mathbf{T} > \prod_i T_{ii}$, then no generator exists.*

Proof. If $\mathbf{T}$ has a generator, then,

$$\frac{dT_{kk}(t,\tau)}{dt} = \sum_j Q_{kj}(t+\tau)T_{jk}(t,\tau)\,, \tag{20}$$

and, since for $k \neq j$, $T_{kj} > 0$ and $Q_{kj} < 0$, one arrives to

$$\frac{dT_{kk}(t,\tau)}{dt} \ge Q_{kk}(t+\tau)T_{kk}(t,\tau). \tag{21}$$

Since $T_{kk}(t,0) = 1$, we can integrate the differential equation in Eq. (20) yielding

$$T_{kk}(t,\tau) \ge \exp\left(\int_{t-\tau}^{t} Q_{kk}(s)ds\right), \tag{22}$$

where Grönwall's inequality is used[16], and finally, from Eq. (19c), one arrives to

$$\begin{aligned}\prod_k T_{kk}(t,\tau) \ &\geq \ \prod_k \exp\left(\int_{t-\tau}^{t} Q_{kk}(s)ds\right)\\ &= \ \exp\left(\sum_k \int_{t-\tau}^{t} Q_{kk}(s)ds\right)\\ &= \ \exp\left(\int_{t-\tau}^{t} \mathrm{Tr}\left(Q_{kk}(s)\right)ds\right)\\ &= \ \mathrm{Det}\left(\mathbf{T}(t,\tau)\right). \qquad (23)\end{aligned}$$

□

The sufficient condition we will implement afterwards deals with the particular case of a LU decomposition:

Proposition IV.3. *If* $\mathbf{T}$ *has a LU decomposition with* $\mathbf{L}$ *and* $\mathbf{U}$ *having only non-negative elements, then* $\mathbf{T}$ *has an inhomogeneous generator* $\mathbf{Q}(t)$.

Proof. To prove this proposition it is important to know an auxiliary result, Prop. A.1 in Append. A, from which it follows that the property of having a time-dependent generator is preserved under multiplication. We use this results from proving that a matrix having an LU decomposition, with $\mathbf{L}$ and $\mathbf{U}$ with non-negative entries, can be modeled through a time-dependent generator. For that, it suffices to prove that the matrix $\mathbf{T}$ has an LU decomposition with $\mathbf{L}$ and $\mathbf{U}$ transition matrices.

Let us first define a diagonal matrix $\mathbf{D}$ with entries $D_{ii} = (\sum_j U_{ij})^{-1}$. Thus, $\mathbf{T}$, with dimension $n\times n$ can be written as $\mathbf{T} = \mathbf{LU} = \mathbf{LD}^{-1}\mathbf{DU} = \mathbf{L}'\mathbf{U}'$, with $\mathbf{L}' = \mathbf{LD}^{-1}$ and $\mathbf{U}' = \mathbf{DU}$ triangular matrices that have all non-negative elements since they are a multiplication of one diagonal matrix with one triangular matrix, all of them with non-negative elements. Furthermore their row-sums are one, since

$$\begin{aligned}\sum_j U'_{ij} \ &= \ \sum_j\sum_k D_{ik}U_{kj}\\ &= \ \sum_k D_{ik}(\sum_j U_{kj}) = D_{ii}\sum_j U_{ij})\\ &= \ (\sum_k U_{ik})^{-1}(\sum_j U_{kj}) = 1, \qquad (24)\end{aligned}$$

for all $i = 1,\ldots,n$. Analogously, since $\sum_j T_{ij} = 1$ for i, one has

$$\begin{aligned}\sum_j T_{ij} \ &= \ \sum_j\sum_k L'_{ik}U'_{kj}\\ &= \ \sum_k L'_{ik}(\sum_j U'_{kj})\\ &= \ \sum_k L'_{ik} = 1. \qquad (25)\end{aligned}$$

and therefore

$$\sum_k L'_{ik} \ = \ \underline{1}. \qquad (26)$$

□

Notice that in the LU factorization there are usually n^2+n variables and n^2 equations. By imposing the row-sums equal to one, we get n^2+n equations, and consequently the LU decomposition defined in this way is unique.

To end this Section we introduce two additional propositions, which are necessary and sufficient for both time-homogeneous and inhomogeneous cases. They are useful when implementing the computational framework for detecting generators, since they help to handle cases where the application of the above propositions do not provide satisfactory output for the embedding problem. With these equivalent matrices one aims to derive a class of matrices that are embeddable if and only if the "original" transition matrix $\mathbf{T}$ is embeddable, which expands the set of possible matrices one may properly test.

The first proposition uses the similarity of matrices through permutation matrices:

Proposition IV.4. *Let* $\mathbf{A} = \mathbf{P}^\top\mathbf{TP}$, *where* $\mathbf{P}$ *is a permutation matrix and* $\mathbf{T}$ *is a transition matrix.* $\mathbf{T}$ *is embeddable if and only if* $\mathbf{A}$ *is also embeddable.*

Proof. To prove this proposition, we will consider a reblabeling of the states i, j, etc. Notice that, under such relabeling, the properties of the transition matrix do not change. Therefore, since changing the transition matrix $\mathbf{T}$ by $\mathbf{P}^\top\mathbf{TP}$ one is, in fact, just relabeling the states, one intuitively concludes that if $\mathbf{T}$ is embeddable, then $\mathbf{P}^\top\mathbf{TP}$ should also be embeddable.

We start by assuming that $\mathbf{T}$ is embeddable,

$$\mathbf{T} = \exp\left(\mathbf{Q}\right) = \sum \frac{\mathbf{Q}^n}{n!}, \qquad (27)$$

where $\mathbf{Q}$ is the generator of $\mathbf{T}$. Since $e^{\mathbf{P}^\top\mathbf{QP}} = \mathbf{P}^\top e^{\mathbf{Q}}\mathbf{P} = \mathbf{P}^\top\mathbf{TP}$, we only need to prove that $\mathbf{Q}' = \mathbf{P}^\top\mathbf{QP}$ is a valid generator, i.e. it must have zero row-sums and non-negative off-diagonal entries.

Since $\mathbf{Q}$ is a valid generator one has

$$\begin{aligned}\sum_j Q'_{ij} \ &= \ \sum_j\sum_k\sum_l P^\top_{il}Q_{lk}P_{kj}\\ &= \ \sum_k\sum_l P^\top_{il}Q_{lk}(\sum_j Pkj)\\ &= \ \sum_k\sum_l P^\top_{il}Q_{lk}\\ &= \ \sum_l P^\top_{il}\sum_k Q_{lk}\\ &= \ \sum_l P^\top_{il}\times 0 = 0\,. \qquad (28)\end{aligned}$$

To prove that matrix $\mathbf{Q}'$ has non-negative off-diagonal entries we write for $k \neq l$ the off-diagonal entry $Q'_{kl} = \sum_{nm} P_{kn}Q_{nm}(P^\top)_{ml}$ and note that, since the matrix $\mathbf{P}$ has only one non-zero element per column and per row. Thus, being that column i and row j, one has $Q'_{kl} = P_{ki}Q_{ij}(P^\top)_{jl}$.

If $k\neq l$ and $i = j$, then $P_{ki} = 1$ and $(P^\top)_{il} = P_{li} = 1$ which contradicts the fact that $\mathbf{P}$ is a permutation matrix. Thus, if $k\neq l$ then $i\neq j$, and so there is a direct correspondence between off-diagonal elements of $\mathbf{Q}'$ and those of $\mathbf{Q}$: if all Q_{ij} are non-negative so are all Q'_{kl}.

Conversely, if $\mathbf{A}$ is embedabble, one just writes $\mathbf{T} = (\mathbf{P}^\top)^{-1}\mathbf{A}\mathbf{P}^{-1} = (\mathbf{P}')^\top \mathbf{A}(\mathbf{P}')^\top$ with $\mathbf{P}' = \mathbf{P}^{-1}$ and applies the same arguments as above. $\square$

The second proposition uses renormalization and transposition of the "original" transition matrix:

Proposition IV.5. *Let* $\mathbf{T}$ *be a transition matrix and consider* $\mathbf{B} = \mathbf{D}\mathbf{T}^\top$*, where* $\mathbf{D}$ *is the diagonal matrix* $D_{ii} = (\sum_j T_{ij}^\top)^{-1}$. $\mathbf{T}$ *is embeddable if and only if* $\mathbf{B}$ *is also embeddable.*

Proof. It is easy to see that if $\mathbf{T}$ is a transition matrix so is $\mathbf{B}$, since $\mathbf{B}$ is always normalized to have row-sums one, and if $\mathbf{T}$ has all its elements non-negative, so has $\mathbf{B}$. Notice that, while T yields the probabilities to which a present states transitates, B gives the probabilities *from* which a state has transitated. It was proven that for a fixed time t, a matrix $\mathbf{T}(t,\tau)$ has all its entries non-negative, $T_{ij}(t,\tau) > 0$, for all τ and is time-continuous, i.e. for any $\epsilon > 0$ there is one δ for which, if $|\tau_1 - \tau_2| < \delta$ then $||\mathbf{T}(t,\tau_1) - \mathbf{T}(t,\tau_2)|| < \epsilon$ if and only if there is a valid generator associated with $\mathbf{T}(t,\tau)$. Since $\mathbf{B}$ is the product of two matrices that are time continuous, $\mathbf{B}$ is also time-continuous. $\square$

V. COMPUTATIONAL IMPLEMENTATION: HOW "EMBEDDABLE" IS A MATRIX?

The mathematical conditions for the existence of a homogeneous generator from the embedding problem are useful more at a theoretical than at a computational level. They give a bivalent result that does not take into consideration neither noise generated from finite samples nor how distant an empirical process is from having a constant generator.

In this section we will describe how to adapt our mathematical results to be meaningful to empirical transition matrices in real situations. First, we evaluate how embeddable a transition matrix is, we define in Sec. V A a proper metric for each proposition above that measures how "close" the empirical transition matrix is from satisfying the corresponding proposition. Then, in Sec. V B, if one arrives to the conclusion that the transition matrix is indeed embeddable we describe proper ways to model its corresponding generator.

There are several differences between the time-homogeneous and the time-inhomogeneous problem:

- In the time-inhomogeneous problem there is no finite set of possible generators, as is usually the case in the time-homogeneous counterpart, namely when the transition matrix has no repeated eigenvalues[10]. If there is a non-homogeneous generator, then there is an infinite number of them.

- The product of two homogeneous embeddable matrices might not be time-homogeneous embeddable, whereas the product of two time-inhomogeneous matrices is always embeddable.

- In the inhomogeneous case, the existence of a real-valued logarithm is not a necessary condition for being embeddable.

- The necessary conditions of the time-homogeneous problem concerning the eigenvalues of the transition matrix, Props. II.4 and II.5, are not necessary conditions for the time-inhomogeneous problem.

As an illustrative example consider the matrix:

$$\mathbf{T}_E = \begin{bmatrix} 0.1179 & 0.0890 & 0.7931 \\ 0.0100 & 0.1000 & 0.8900 \\ 0.8901 & 0.0010 & 0.1089 \end{bmatrix}, \tag{29}$$

The matrix $\mathbf{T}_E$ is, according to Prop.IV.3, time-inhomogeneous embeddable, since it is a product of matrices that have a positive LU decomposition. However it is not time-homogeneous embeddable, since it has distinct negative eigenvalues, $\{1, -0,001490, -0,671710\}$, and thus it has no real-valued logarithm[17]. Moreover, the conditions in both Props. II.4 and II.5 are not fulfilled.

Regarding Prop. II.2, we have shown that conditions a) and b) are necessary conditions for the more general case of time-inhomogeneous generators. As for condition c), one can show that there is also a limit number of zero entries for the time-inhomogeneous case. See Prop. A.2 in Append. A.

Before proceeding, two important remarks. First, it is necessary to describe how to estimate the transition matrix directly from data series and then explain how to resample the transition matrix which will be necessary for evaluating if it is embeddable or not. Among several algorithms[18, 19], we concentrate in the so-called "Cohort Method", which counts the number N_{kj} of transitions from state k to state j in the desired time-interval $[t, t+\tau]$, defining the entries of the transition matrices as

$$T_{kj}(t) = \frac{N_{kj}(t)}{\sum_j N_{kj}(t)}, \tag{30}$$

with the associated error

$$\sigma_{T_{kj}} = \sqrt{\frac{T_{kj}(1-T_{kj})}{N_{kj}}}. \tag{31}$$

Second, in order to implement the set of propositions with an associated statistical error, we propose a method of resampling a given empirical transition matrix $\mathbf{T}(t,\tau)$. The set of resampling matrices obtained is then used to quantify the error associated to the estimates on the transition matrix: each metric that is applied to the empirical transition matrix retrieves a set of metric values when applied to the full set of resampling matrices, and the standard deviation of that value distribution is then taken as the error or uncertainty associated to the metric estimation.

More specifically, one generates number series from the distribution of states $P(X,t)$ at time t till the distribution $P(X,t+\tau)$ at $t+\tau$, and estimates the corresponding resampling matrix through the Cohort Method. See Eq. (30).

A. Embeddability metrics

The propositions of the embedding problem do not take in consideration the uncertainty in the estimation of $\mathbf{T}$, and thus we need to develop methods to determine, beyond statistical uncertainty, whether a generator exists or not. Notice that the embedding problem determines only if the process can be modeled as a time-continuous Markov process, but it cannot guarantee if the underlying process actually is a time-continuous Markov process. Thus we will use a proper null hypothesis, which if not rejected, one assumes that a suitable generator can be estimated. In the case of Props. IV.1 and IV.2 the null hypothesis states that a generator exists, while for Prop. IV.3 the null hypothesis states that such a generator does *not* exist. The null hypothesis is tested for each proposition separately.

To evaluate if the condition of Prop. IV.1 is fulfilled for a given transition matrix $\mathbf{T}$, we compute the following quantity,

$$d_{N_1} = -\frac{\det(\mathbf{T})}{\sigma_{det}} , \tag{32}$$

where σ_{det} is the standard deviation from the sample of the determinants calculated for each resampling matrix. If $d_{N_1} \geq 2$, we assume that the determinant of $\mathbf{T}$ is negative and the distribution of the resampled determinants are all negative within two standard deviations. In this case we reject the null hypothesis, i.e. no generator exists.

Regarding the condition in Prop. IV.2, we use the following metric,

$$d_{N_2} = -\frac{\prod_i T_{ii} - \det(\mathbf{T})}{\sigma_{prod} + \sigma_{det}} , \tag{33}$$

where σ_{prod} is the standard deviation associated to the variable $\prod_i T_{ii}$ according to the expression in Eq. (31). Again, if $d_{N_2} \geq 2$, then no generator exists.

Concerning the sufficient condition of the LU-decomposition with non-negative elements, Prop. IV.3, we can use the following distance:

$$d_{S_1} = \min\{m_L, m_U\} , \tag{34}$$

with

$$m_L = \min_{i,j} \left\{ \frac{L_{ij}}{\sigma_{L_{ij}}} \right\} , \tag{35a}$$

$$m_U = \min_{i,j} \left\{ \frac{U_{ij}}{\sigma_{U_{ij}}} \right\} , \tag{35b}$$

where L_{ij} and U_{ij} represent the entries of the matrices $\mathbf{L}$ and $\mathbf{U}$ respectively, and σ_{Lij} and σ_{Uij} their corresponding standard deviations. The quantities $\sigma_{L_{ij}}$ and $\sigma_{U_{ij}}$ are calculated solving the same system of equations of the LU decomposition, but using the uncertainties in the estimation of T_{ij} with the rules of error propagation. If $d_{S_1} > 2$, then we reject the null hypothesis, i.e. we assume that a generator exists.

Applying these three metrics to one transition matrix, if the null hypothesis cannot be rejected, we estimate a generator matrix as describe in the Sec. V B. To ascertain if the estimated generator matrix yields a transition matrix sufficiently close to the empirical transition matrix, we use it to generate auxiliary transition matrices $\tilde{\mathbf{T}}$. If the auxiliary matrices are typically close to the empirical transition matrix $\mathbf{T}$ we assume that the estimate is good. To that end, we introduce one additional metric to assert if the matrix $\mathbf{T}$ is close enough to a auxiliary matrix, $\tilde{\mathbf{T}}$, originated from a time-continuous Markov process with a generator $\mathbf{Q}(t)$, is to compute the quantity,

Metric	d_{N1} (Prop.IV.1)	d_{N2} (Prop. IV.2)	d_{S1} (Prop. IV.3)
> 2	166	199	178
< 2	34	1	22

TABLE I: Test results of the inhomogeneous framework detection for a set of 200 samples, each one with 10^4 points. When one of the metrics is larger than two, the null hypothesis cannot be rejected.

$$d_{est} = \frac{1}{R} \sum_{k=1}^{R} \Theta \left(||\mathbf{T} - \tilde{\mathbf{T}}||_F - ||\mathbf{T}' - \tilde{\mathbf{T}}||_F \right) , \tag{36}$$

where R is the number of auxiliary matrices, $\Theta(x)$ is the Heaviside function and $||\mathbf{X}||_F = \left(\sum_{i=1}^n \sum_{j=1}^n X_{ij}^2 \right)^{1/2}$ is the Frobenius norm of matrix $\mathbf{X}$. We assume that the empirical process, observed for the estimation $\mathbf{T}'$ is not close to the time-continuous Markov process with a transition matrix $\tilde{\mathbf{T}}$ if $d_{est} < 0.10$, i.e. if less than 10% of the auxiliary matrices are outside a confidence interval with significance value $p = d_{est}$.

If the distance d_{est} is too small a new matrix is generated within the conditions of Props. IV.4 and IV.5. In case that the new matrices pass the tests above, these propositions guarantee that the original matrix also passes.

To test all the metrics introduced above we generate a set of 200 samples of 10^4 points, each one from a different inhomogeneous transition matrix, as described below. We then compute numerically the transition matrix from each sample and apply all three metrics d_{N1}, d_{N2} and d_{S1}. The results, summarized in Tab. I, clearly show that in at least 80% of the cases the framework is able to correctly detect the inhomogeneity of an existing generator.

B. Modeling the generator matrix $\mathbf{Q}(t)$

In case the null hypothesis cannot be rejected (i.e. that a valid generator exists), we then derive an estimate $\mathbf{Q}(t)$ able to model the empirical process. Unlike the case of the time-homogeneous embedding problem, here we need to estimate a matrix which changes in time and therefore a different procedure is necessary. In general, for deriving an inhomogeneous generator, one solves the Peano-Baker series Eq. (16). Assume $\mathbf{Q}(t)$ can be modeled as a polynomial of degree N, i.e.

$$\mathbf{Q}(t) = \sum_{n}^{N} \mathbf{B}_n t^n , \tag{37}$$

where each matrix $\mathbf{B}_n$ is constant over time. Naturally, we need to make sure that no off-diagonal entry in $\mathbf{Q}(t)$ ever become negative in $t \in [0,1]$. Introducing Eq. (37) in Eq. (16) yields

$$\mathbf{T} = \sum_{k=0}^{\infty} \prod_{l=1}^{k} \sum_{n=0}^{N} \frac{\mathbf{B}_n}{l + \sum_{m=1}^{l} n_m}. \tag{38}$$

To invert Eq. (38) however is very cumbersome and computationally expensive. In this subsection, we propose an alternative for estimating inhomogeneous generators that is accurate and easily implementable.

Our procedure is based in the assumption that the original transition matrix is a product of a finite number of embeddable matrices, $\mathbf{T} = \mathbf{T}_1^* \dots \mathbf{T}_n^*$ with each $\mathbf{T}_i^*$ $(i = 1, \dots, n)$ having an homogeneous generator.

One starts with a decomposition of the form

$$\mathbf{T} = \mathbf{A}_1 \dots \mathbf{A}_n \mathbf{T}_0^* \mathbf{A}_{n+1} \dots \mathbf{A}_{2n}, \tag{39}$$

where $\mathbf{A}_i$ are embeddable matrices having one off-diagonal positive term. The objective here is to find an embeddable matrix $\mathbf{T}_0^*$ from the empirical matrix $\mathbf{T}$ through the multiplication by matrices $\mathbf{A}_i$. If $\mathbf{T} = e^{\mathbf{Q}}$ and $\mathbf{Q}$ has one negative off-diagonal entry, $Q_{ij} < 0$, we can try "correct" that entry by multiplying $\mathbf{T}$ by two matrices, $\mathbf{A}_l$ and $\mathbf{A}_{l+n}$, such that $(A_l)_{ik} > 0$ and $(A_{l+n})_{kj} > 0$ for a fixed index k. Intuitively, if there are transitions from a state k to a state j and only afterwards from another state i to state k, a time-inhomogeneous process might correspond to a logarithm with a negative off-diagonal entry if $Q_{ij} < 0$. Hence, one derives a first estimate $\mathbf{T}_0^*$ of the transition matrix $\mathbf{T}$. In case there is more than one negative off-diagonal element of $\mathbf{Q}$ one proceeds similarly for each element separately.

The algorithm proceeds then as follows:

1. Compute $\mathbf{Q}_0^* = \log \mathbf{T}_0^*$ and verify it is a valid generator. Note that, during the algorithm we must always use the same branch of the complex logarithm.

2. If the generator is not valid, i.e. it has at least one negative off-diagonal entry $(Q_0^*)_{ij}$, one finds a suitable integer k for which two matrices, $\mathbf{A}_1$ and $\mathbf{A}_{n+1}$, have entries $(A_1)_{kj} > 0$ and $(A_{n+1})_{ik} > 0$.

3. One considers the new estimate $\mathbf{T}_1^* = \mathbf{A}_1 \mathbf{T}_0^* \mathbf{A}_{n+1}$ and computes the generator estimate $\mathbf{Q}_1^* = \log \mathbf{T}_1^*$ and verifies if it is now a valid generator.

4. One proceeds recursively until for a certain recursive step i $\mathbf{Q}_i^* = \log \mathbf{T}_i^*$ has no negative off-diagonal entries.

5. The final estimate at step i is identified as the k-factor $\mathbf{T}_k^*$ in the assumed decomposition $\mathbf{T} = \mathbf{T}_1^* \dots \mathbf{T}_n^*$.

6. One computes now $\mathbf{T}_{k+1}^* = (\mathbf{T}_1^* \dots \mathbf{T}_k^*)^{-1} \mathbf{T}$ and repeats the procedure.

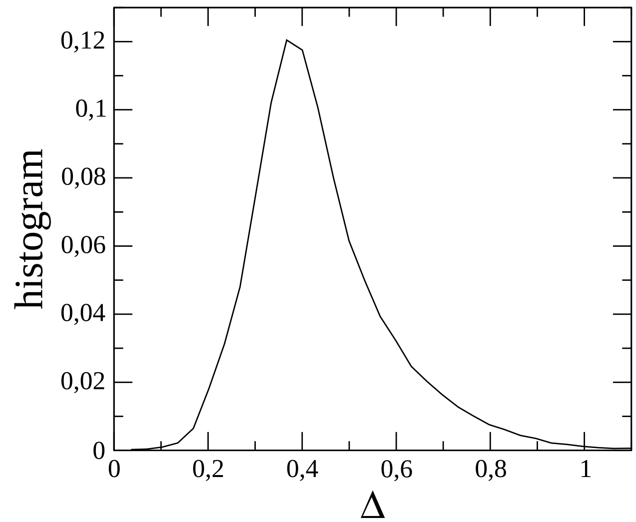

FIG. 2: Histogram of Δ values, Eq. (40), from a sample of 43222 matrices (see text).

7. The full algorithm ends when the last estimated matrix in the decomposition is either an embeddable matrix or a matrix sufficiently close to the identity. More specifically, when the matrix norm of the difference between the matrix and identity matrix is at least one order of magnitude smaller than the matrix norm of the estimated matrix. Alternatively, when the number of iterations exceeds a pre-fixed maximum number of iterations, typically a few thousand, the algorithm stops.

We tested 1000 matrices with principal logarithms having only one negative off-diagonal entry and a valid generator was found 945 times. If the number of negative entries is not too large at each step of the recursive procedure above ($< n^2$) similar results are obtain, which indicates an accuracy of around 90 and 95%.

In Sec. V A we generated matrices with an inhomogeneous generator $\mathbf{Q}(t) = \mathbf{Q}_0 + \mathbf{Q}_1(t)$, integrated them in order to compute a transition matrix $\mathbf{T}(t)$ and produce data series for testing our framework. Here, we use the subset of matrices that were correctly detected as time-inhomogeneous embeddable and estimate one generator as described above. A valid generator was found in around 90% of the generated samples.

To evaluate the accuracy of the estimates, we compare the modeled transition matrix $\mathbf{T}_{mod}(t, \tau)$ with the empirical one, $\mathbf{T}_{emp}(t, \tau)$. The comparison is based in a normalized distance given by the fraction of the matrix norm of the difference between both matrices and the matrix norm of the difference between the modeled matrix and the identity matrix (initial state):

$$\Delta = \frac{||\mathbf{T}_{mod}(t, \tau/2) - \mathbf{T}_{emp}(t, \tau/2)||_F}{||\mathbf{T}_{mod}(t, \tau/2) - \mathbf{Id}||_F}, \tag{40}$$

where $|| \cdot ||_F$ is the Frobenius norm. Figure 2 shows an histogram of computed values of the normalized distance Δ in Eq. (40) for all estimates. Typically the deviations are not larger than 40% of the deviations from the initial state, where no transition occur.

Such procedure closes the computational framework for uncovering Markov continuous processes from empirical data sets.

VI. DISCUSSION AND CONCLUSIONS

We have extend some theoretical results on the inhomogeneous embedding problem and established a framework which can evaluate empirical data for detecting the existence of continuous Markov processes. Eight new proposition were presented and demonstrated concerning the general case of processes having a time-inhomogeneous generator. While it was also recently proven that the problem of deriving a general algorithm capable of solving the embedding problem for any finite dimension n is NP-hard[20], our implimented algorithm presents acceptable results: when applied to synthetic data generated from pre-given generators our framework is able to detect at least 80% of them and, moreover, returns a good estimate of the generator underlying the data set. Thus, our algorithm enables one to find a time-inhomogeneous generator of any transition matrix with a real-valued logarithm.

Concerning the new proposition demonstrated above for inhomogeneous transition matrices, there are e.g. some extensions of the LU decomposition theorem, Prop. IV.3, that can be interesting for future work. Namely, the quasi LU decomposition[21], the ULU decomposition[22] and the LULU factorization[23].

This framework is now able to be straightforwardly applied to specific sets of data for evaluating hidden continuous Markov processes. Indeed, since the transition matrix defines a specific Markov chain, our framework can be taken as a possibility for accessing continuous hidden processes in (time-dependent) Markov chains found in e.g. models for polymer growth processes or enzyme activity.

For specific applications, our framework can be used for three types of stochastic data sets: (i) one where only the initial and final configuration of the system is given; (ii) one where all possible state transitions are defined through a probability value between the start and end of the observation period and (iii) the transition between the beginning of intermediate instants till the end of the observation. In this paper we dealt typically with type (ii) data sets, while in previous works[6, 7] we considered mainly type (iii). Type (i) is typically not well-defined and additional cautions must be taken.

One important interdisciplinary application is, of course, in economics and finance, when addressing rating matrices: if ratings do indeed reflect a natural (continuous) economic process, the extracted rating matrices must have a proper generator[24] . This problem as already addressed by us[6, 7] in the particular case of homogeneous transition matrices derived by rating agencies. Further, our methodology could be extended to other situations where correlation matrices are taken for describing the macroscopic state of a financial system[25]. With a proper normalization such correlation matrices can be taken, in an algebraic sense, as transition matrices and therefore the framework described above is applicable.

Acknowledgments

The authors thank *Deutscher Akademischer Austauschdienst* (DAAD) and *Fundação para a Ciência e a Tecnologia* (FCT) for support from bilateral collaboration DRI/DAAD/1208/2013. FR thanks FCT for the fellowship SFRH/BPD/65427/2009. TR acknowledges support from the Royal Society. PGL thanks German Environment Ministry for financial support.

Appendix A: Additional results on the time-inhomogeneous embeddable problem

Here we present additional results concerning the existence of inhomogeneous generators. These results serve for proving the theorems implemented above and provide theoretical consistency to our framework. The first result is a sufficient condition concerning a possible decomposition of transition matrices:

Proposition A.1. *If $\mathbf{T}$ is an n-dimensional triangular transition matrix, then it has an inhomogeneous generator, which can be defined from a decomposition of the transition matrix as $\mathbf{T} = e^{\mathbf{Q}_1} \cdots e^{\mathbf{Q}_{n-1}}$ where $\mathbf{Q}_i$ are time-homogeneous generators of some elementary transition matrix.*

Proof. The proof is given by induction. For $n = 2$, the triangular transition matrix $\mathbf{T}$ can be parameterised by one single parameter $p \in [0, 1]$:

$$\mathbf{T} = \begin{pmatrix} 1-p & p \\ 0 & 1 \end{pmatrix} . \tag{A1}$$

It is straightforward to see that $\mathbf{T} = e^{\mathbf{Q}}$ with

$$\mathbf{Q} = \begin{pmatrix} \log(1-p) & -\log(1-p) \\ 0 & 0 \end{pmatrix} . \tag{A2}$$

Since $\log(1-p) < 0$, $\mathbf{Q}$ is indeed a generator matrix.

We now consider an triangular transition matrix of arbitrary dimension n and treat the rightmost column separately, yielding

$$\mathbf{T} = \begin{pmatrix} \mathbf{A} & \underline{a} \\ \underline{0}^\top & 1 \end{pmatrix} , \tag{A3}$$

where $\mathbf{A}$ is an $(n-1) \times (n-1)$ triangular matrix, $\underline{a}$ is a column-vector with $n-1$ non-negative entries and $\underline{0}^\top$ is a row-vector of $n-1$ zeros. Since $\mathbf{T}$ is a transition matrix, for all $i = 1, \ldots, n-1$ one has

$$\sum_j A_{ij} = 1 - a_i . \tag{A4}$$

Introducing a $(n-1)$-dimensional triangular transition matrix $\mathbf{T}'$ with entries $T'_{ij} = \frac{A_{ij}}{1-a_i}$, one reads

$$\mathbf{T} = \begin{pmatrix} \mathbf{I} - \mathrm{diag}(\underline{a}) & \underline{a} \\ \underline{0}^\top & 1 \end{pmatrix} \begin{pmatrix} \mathbf{T}' & \underline{0} \\ \underline{0}^\top & 1 \end{pmatrix}, \tag{A5}$$

where diag$(\underline{a})$ is the $(n-1)$-dimensional diagonal matrix with entries taken from vector $\underline{a}$. The first matrix above is embeddable since

$$\begin{pmatrix} \mathbf{I} - \mathrm{diag}(\underline{a}) & \underline{a} \\ \underline{0}^\top & 1 \end{pmatrix} = \exp \begin{pmatrix} \mathrm{diag}(\log(\underline{1} - \underline{a})) & -\log(\underline{1} - \underline{a}) \\ \underline{0}^\top & 0 \end{pmatrix} \tag{A6}$$

and the second matrix can be further decomposed as

$$\mathbf{T} = \begin{pmatrix} \mathbf{I} - \mathrm{diag}(\underline{a}) & \underline{a} \\ \underline{0}^\top & 1 \end{pmatrix} \begin{pmatrix} \mathbf{I} - \mathrm{diag}(\underline{b}) & \underline{b} & 0 \\ \underline{0}^\top & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mathbf{T}' & \underline{0} & \underline{0} \\ 0 & 1 & 0 \\ \underline{0}^\top & 0 & 1 \end{pmatrix}. \tag{A7}$$

Therefore, we arrive to a decomposition of the form $\mathbf{T} = e^{\mathbf{Q}'_1} \cdots e^{\mathbf{Q}'_{n-1}}$ for generator matrices $\mathbf{Q}'_1, \dots, \mathbf{Q}'_{n-1}$ with

$$\mathbf{Q_1} = \begin{pmatrix} \mathrm{diag}(\log(\underline{1} - \underline{a})) & -\log(\underline{1} - \underline{a}) \\ \underline{0}^\top & 0 \end{pmatrix} \tag{A8}$$

and

$$\mathbf{Q_k} = \begin{pmatrix} \mathbf{Q}'_{k-1} & \underline{0} \\ \underline{0}^\top & 0 \end{pmatrix} \tag{A9}$$

for $k = 2, \dots, n-1$. □

One could implement Prop. A.1 by finding a product of n-dimensional square matrices $\prod_i \mathbf{A}^{(i)}$ where each matrix $\mathbf{A}^{(i)}$ has only one off-diagonal non-zero element and if for matrix $\mathbf{A}^{(k)}$ one has $A^{(k)}_{ij} \neq 0$, then for all other matrices $\mathbf{A}^{(l)}$ ($l \neq k$) one has $A^{(l)}_{ij} = 0$. If that product has $m = n(n-1)$ terms, we can solve $\prod_i \mathbf{A}^{(i)} = \mathbf{T}$ as a linear system of equations with n equations and n unknowns. Having this, we define the following distance for the $\mathbf{A}$-factorization:

$$d_{S_2} = \min_k \{ \min_{i,j} \{ \frac{A^k_{ij}}{\sigma_{A^k_{ij}}} \} \}, \tag{A10}$$

where $\sigma_{A^n_{ij}}$ is the dispersion associated with the entry A^n_{ij}. If $d_{S_2} > 2$, we statistically infer that a generator exists. Notice that, it is possible to prove that the LU decomposition is a particular case of the factorization in Eq. (A10).

One additional necessary condition that may be useful in some cases is the following one:

Proposition A.2. *An irreducible matrix* $\mathbf{T}$*, i.e. it cannot be placed into block upper-triangular form by simultaneous row or column permutations, is time-inhomogeneous embeddable only if, for at least in one row there is more than one non-zero off-diagonal entry.*

Proof. If $\mathbf{T}$ is time-inhomogeneous embeddable, then from Prop. A.1, $\mathbf{T}$ can be written as a product n of embeddable matrices $\mathbf{P}^{(k)} = \exp \mathbf{Q}^{(k)}$. Assume, without loss of generality that all matrices $\mathbf{P}^{(k)}$ are time-homogeneous embeddable.

Since no matrix $\mathbf{P}^{(i)}$ has no zeros in the diagonal entries, from Props. IV.1 and IV.2, the product of an irreducible matrix by an embeddable matrix is always irreducible. Notice that if any of the matrices $\mathbf{P}^{(k)}$ is time-homogeneous embeddable, then from Prop. II.2c), $\mathbf{T}$ will have no zero entries.

Let us consider $\mathbf{P}^{(k)}$ such that the product $\mathbf{P}^{(1)} \dots \mathbf{P}^{(k)}$ is irreducible but $\mathbf{P}^{(1)} \dots \mathbf{P}^{(k-1)}$ is not. Since we assume, without loss of generality that $\mathbf{P}^{(k)}$ is not the identity matrix, $P^{(k)}_{ij} > 0$ for at least one $j \neq i$. Then, for $m \leq k$ there is one l for which $P^{(m)}_{li} > 0$. Thus $T_{ij} > 0$ and $T_{lj} > 0$. □

Proposition A.2 is not a condition we can evaluate for empirical systems. Nonetheless it might be useful if one has some apriori knowledge about the dynamics of the system.

Another sufficient condition for time-inhomogeneous generators concerns situations when the matrices have non-negative entries:

Proposition A.3. *Totally non-negative transition matrices, i.e. matrices* $\mathbf{T}(t)$ *for which all submatrices have positive determinant, have an inhomogeneous generator* $\mathbf{Q}(t)$*.*

Proof. It was proved [26] that the LU factorization of any totally non-negative matrix is composed by a totally non-negative lower diagonal matrix $\mathbf{L}$ and a totally non-negative upper diagonal $\mathbf{U}$. If a matrix is totally non-negative, then it has only non-negative elements, thus in particular $\mathbf{L}$ and $\mathbf{U}$ are matrices with non-negative elements. □

[1] G. Elfving, *Zur theorie der Markoffschen ketten*, Acta Societatis scientiarum Fennicae, Nova series A, T.2, 8 (Harrassowitz, Leipzig, 1937).

[2] J. Kingman, Probability Theory and Related Fields **1**, 14 (1962).

[3] M. Esposito and C. Van den Broeck, Physical Review E **82**, 011143 (2010).

[4] V. Privman, *Nonequilibrium statistical mechanics in one dimension* (Cambridge University Press, 2005).

[5] T. Shintani and S. Shinomoto, Phys. Rev. E **85**, 041139 (2012).

[6] P. Lencastre, F. Raischel, P. G. Lind, and T. Rogers, in *3rd SMDTA Conference Proceedings* (2015) pp. 405–414.

[7] P. Lencastre, F. Raischel, and P. G. Lind, Journal of Physics: Conf. Ser. **574**, 012151 (2015).

[8] Y. Chen and J. Chen, Journal of Theoretical Probability **24**, 928 (2011).

[9] E. Davies, Electronic Journal of Probability **15**, 1474 (2010).

[10] R. Israel, J. Rosenthal, and J. Wei, Mathematical Finance **11**, 245 (2001).

[11] F. I. Karpelevich, Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya **15**, 361 (1951).

[12] N. J. Higham and L. Lin, L Algebra and its Applications **435**,

448 (2011).
[13] J. Runnenberg, Proceedings of the KNAW **65**, 536 (1962).
[14] R. M. Gray, Communications and Information Theory **2**, 155 (2005).
[15] M. Baake and U. Schlägel, Proceedings of the Steklov Institute of Mathematics **275**, 155 (2011).
[16] B. Pachpatte, *Inequalities for differential and integral equations* (San Diego, Academic Press, 1998).
[17] N. J. Higham, *Functions of matrices: theory and computation* (Siam, 2008).
[18] C. Sherlaw-Johnson, S. Gallivan, and J. Burridge, Journal of the Operational Research Society , 405 (1995).
[19] Y. Jafry and T. Schuermann, Journal of Banking & Finance **28**, 2603 (2004).
[20] T. S. Cubitt, J. Eisert, and M. M. Wolf, Physical review letters **108**, 120503 (2012).
[21] M. Bueno and C. Johnson, Linear Algebra and its Applications **427**, 99 (2007).
[22] T. Toffoli, Linear algebra and its applications **259**, 31 (1997).
[23] G. Strang, Linear algebra and its applications **265**, 165 (1997).
[24] A. Metz and R. Cantor, *Introducing Moody's Credit Transition Model* (Moody's Analytics, 2007).
[25] M. Münnix, T. Shimada, R. Schäfer, F. Leyvraz, T. Seligman, T. Guhr, and H. Stanley, Scientific Reports **2**, 644 (2012).
[26] C. W. Cryer, Linear Algebra and its Applications **7**, 83 (1973).