Journal of Physics A:
Mathematical and Theoretical

PAPER • OPEN ACCESS

State space reconstruction of Markov chains via autocorrelation structure

OPEN ACCESS
IOP Publishing

Journal of Physics A: Mathematical and Theoretical

J. Phys. A: Math. Theor. **57** (2024) 315701 (17pp)

https://doi.org/10.1088/1751-8121/ad6224

State space reconstruction of Markov chains via autocorrelation structure

Antal Jakovác[1,*], **Marcell T Kurbucz**[1,2] and **András Telcs**[1,3]

[1] Department of Computational Sciences, Institute for Particle and Nuclear Physics, HUN-REN Wigner Research Centre for Physics, 29-33 Konkoly Thege Miklós Street, Budapest H-1121, Hungary
[2] Department of Statistics, Institute of Data Analytics and Information Systems, Corvinus University of Budapest, 8 Fővám Square, Budapest H-1093, Hungary
[3] Department of Quantitative Methods, Faculty of Business and Economics, University of Pannonia, 10 Egyetem Street, Veszprém H-8200, Hungary

E-mail: jakovac.antal@wigner.hun-ren.hu, kurbucz.marcell@wigner.hun-ren.hu and telcs.andras@wigner.hun-ren.hu

Received 19 February 2024; revised 28 June 2024
Accepted for publication 10 July 2024
Published 19 July 2024

Abstract

Understanding the state space of observed Markov processes is essential for advancing causal inference in a wide range of scientific fields. This paper demonstrates how the previously unknown state space can be reconstructed by exploring the spectrum of the time-delay embedding matrix derived from the autocorrelation sequence of the observed series. It also highlights that the eigenvector associated with the smallest eigenvalue can provide valuable insights into the hidden data generation process itself. The presented results provide a deeper understanding of the complex dynamics of Markov chains and hold promise for enhancing various scientific applications.

Keywords: Markov chain, hidden Markov model, state space reconstruction

1. Introduction

Causal discovery is a fundamental challenge in scientific inquiry. In various physics models, this challenge is effectively addressed by formulating dynamical evolution equations,

[*] Author to whom any correspondence should be addressed.

which are governed by an equation of motion (EoM). The classical physics paradigm encompasses disciplines such as mechanics, hydrodynamics, electrodynamics, and quantum mechanics, where the respective EoMs, such as the *equation of mechanics* (Newton's equation) and the Navier–Stokes equation for hydrodynamics, serve as foundational principles. However, in more complex systems, the EoM is often unknown, necessitating concerted efforts for its reconstruction. In recent years, attempts have been made to uncover hidden dynamics through the integration of artificial intelligence (AI) techniques (see, e.g. [1–4]).

The mathematical foundation upon which the reconstruction of dynamical systems rests is provided by Takens' theorem [5]. This theorem asserts that the state space of a deterministic dynamic system can be reconstructed from observations of a single system variable, provided a specific embedding dimension is chosen (see also [6, 7]). If the state space is reconstructable, it enables the determination of dynamical insights from the embedded data.

In more complex scenarios where numerous factors influence the motion of the observed system, achieving an adequate deterministic description becomes challenging. When most of these effects are relatively small, the system can be treated as having a limited number of degrees of freedom with noisy observations, as is typical in the deterministic systems referenced earlier. However, in other instances where unknown effects are significant, they must be considered as integral components of the dynamics, leading to the classification of such systems as *stochastic systems*. While stochasticity plays an important role in mechanics [8], it is inevitable in economics [9].

Similar to the deterministic case, stochastic systems can be approached by pre-defined models. Due to the difficulty of reliable data acquisition, generic models are commonly used in this field. The autoregressive (AR) models are built on linear equations with Gaussian stochastic components [10], while various other models incorporate non-Gaussianity [11–13]. In this case, model fitting involves determining the model parameters.

It would be desirable to construct AI methods that could reconstruct the model even if it has stochastic dynamics. However, Takens' theorem does not apply to stochastic dynamic systems, such as Markov chains (MCs), due to severe limitations (see [14, 15]). The point is that deterministic motion forms a subspace in the multi-dimensional space of the embedded variables $(x_t, x_{t-\Delta t}, \dots, x_{t-N\Delta t})$. For example, Newton's equation claims that the embedded data lie on a two-dimensional surface for any N, allowing the revelation of $x_{t+\Delta t}$ from the knowledge of $(x_t, x_{t-\Delta t})$. However, in a stochastic case, there is no confinement to a submanifold, and such recovery is not possible.

In a recent paper [16], the authors employed the full power of topological theory in deterministic dynamical systems, including time delay embedding, Takens' theorem on state space reconstruction, Stark's generalization to forced systems, and Sauer's results on embedding ensuring invariance of counting and information dimensions against perturbed observations. The work also addresses the identification of an optimal embedding dimension alternative to Kennel, Brown, and Abarbanel's method [17]. All the mentioned works focus on deterministic dynamical systems, excluding noise or stochastic components from their scope. While they can handle a limited amount of noise, essentially stochastic processes are not considered.

Kantz and Ragwitz [16] highlighted that, from the perspective of time series analysis, the most challenging task is to develop a suitable model for a specific phenomenon. According to their work, several studies have reconstructed Fokker–Planck and Langevin equations [18, 19] from observed data, with Friedrich and Peinke's work considered pioneering in this regard. However, this approach requires observations of the entire state space of the system. Implicit or explicit reconstruction of unobserved variables has not yet been achieved.

In Kantz's and Ragwitz's paper [16], the focus is on predicting a full Markov model based on transition probabilities from data. The approach utilizes discrete time delay embedding to approximate the continuous space and time Markov process, enabling predictions without explicitly reconstructing a discrete state space. This method differs from traditional state space reconstruction methods discussed in Stark *et al* [15], which outlines stringent conditions for handling stochastic processes. Kantz and Ragwitz combine the time delay embedding with constant prediction methods placing it within the broader category of predictions based on local averaging (see Györfi *et al* [20]).

In the present study, we develop a state space reconstruction method for finite, discrete state, stationary MCs. Our method can be used to determine the true number of states in a MC, even when only a function of the MC is observed.

This paper is organized as follows. Section 2 introduces the theory and methodology used to restore the state space of a hidden Markov process. In section 3, we demonstrate our method on synthetic data, including examples with special symmetries in the problem. Section 4 discusses the results, and finally, section 5 concludes and presents future research directions.

2. Theory and methodology

In this section, we first introduce some definitions and notations. Then, we prove several theorems that can help restore the state space of hidden MCs. Finally, we present the main theorem and demonstrate the method for applying it to actual problems.

2.1. Definitions and notations

We consider an ensemble of $X:\mathbb{N}\to B$ series, where B is a (finite) set. We denote the probability of having $b\in B$ at the n^{th} step as $P(X_n=b)$. We have a Markov process, if this probability depends solely on the value at the $(n-1)$th step. The conditional probability is denoted by

$$P(X_n=b|X_{n-1}=a)=T^{(n)}_{ab}, \tag{1}$$

and is called a transfer matrix. We will consider time-homogeneous processes, where the transfer matrix does not depend on n, and we omit the upper index in the sequel.

Definition 1. A time-homogeneous MC is defined by the pair $X=(B,T)$ (the state space and the transfer matrix) and by the initial distribution $P(X_0=a)$.

Properties of the transfer matrix:

(i) $T_{ab}\in[0,1]$;
(ii) $\sum_{b\in B}T_{ab}=1$;
(iii) $\forall\lambda$eigenvalue of $T:\ |\lambda|\in[0,1]$;
(iv) from (ii) there exists at least one $\lambda=1$ eigenvalue;
(v) $P(X_n=b)=\sum_a P(X_{n-1}=a)T_{ab}$;
(vi) from (iv): $P(X_n=b)=\sum_a P(X_{n-k}=a)T^k_{ab}$.

We assume that our finite state MC is irreducible and aperiodic, and it always converges to a unique equilibrium distribution (the eigenvalue 1 has multiplicity one). We will assume in the whole sequel that the initial distribution is already the equilibrium one and hence the process is stationary. The stationary (or equilibrium) probability distribution is denoted by $P_a=P_{eq}(a)$.

2.2. State space reconstruction

Inspired by Takens' theorem, we consider the observations of a MC. Let $X:\mathbb{N}\to B$ be a MC, and g is our observation function $g:B\longrightarrow B_g$, where B_g is the set of possible observations with $|B_g|\leqslant|B|$. We denote the observations by $Y_n=g(X_n)$. One can consider a more complex observation like $Y_n=g^{(k)}(X_n,\ldots,X_{n+k})$ with a proper $g^{(k)}$, but that makes no difference in our investigation since $Z_n=(X_n,\ldots,X_{n+k})$ is also a MC.

Next, we introduce a bi-variate function $h:B_g\times B_g\longrightarrow\mathbb{R}$ which provides the correlation function. In fact we will deal with the observed pairs via $f=h\circ g$, that is $f(X_n,X_{n+k})=h(Y_n,Y_{n+k})=h(g(X_n),g(X_{n+k})$.

Definition 2. The $C_f(k)$ autocorrelation function associated to a function $f:B\times B\to\mathbb{R}$ is defined as

$$C_f(k)=\langle f(X_0,X_k)\rangle=\sum_{a,b\in B}P(X_0=a)P(X_k=b|X_0=a)f(a,b),$$

where $P_a=P_{eq}(a)=P_{eq}(X_0=a)$ is the equilibrium probability distribution.

Using the transfer matrix we can also write

$$C_f(k)=\langle f(X_0,X_k)\rangle=\sum_{a,b\in B}P_aT^k_{ab}f(a,b),$$

with the remark that for $k=0$ we have $T^0=\mathbf{1}$. With a little abuse of terminology, we will call both f and $C_f(k)$ as correlation functions where it does not cause confusion.

Here we should stress that in real-world scenarios g is typically given, we can not choose it, while h can be chosen. The property that f has no special symmetry will be referred to in short as f is generic. The proper definition of a generic observation and the discussion of symmetries in f is postponed to section 4.

We recall that the characteristic polynomial of a matrix M is defined as

$$p_M(x)=\det(M-x\boldsymbol{I})=\sum_{k=0}^{|B|}w_kx^k. \tag{2}$$

Each matrix satisfies its own characteristic polynomial $p_M(M)=0$.

We remark that, since T has a unit eigenvalue, $\sum_k w_k=0$.

The following lemma forms the basis of the state space reconstruction method.

Lemma 1. *With the above notations and assumptions for all l*

$$W_f^{|B|}(l)\overset{!}{=}\sum_{k=0}^{|B|}w_kC_f(l+k)=0$$

where w_k-s are the coefficients of the characteristic polynomial p_T.

In other words, all shifted correlation functions satisfy a linear equation of order $|B|$, which is independent of f and the shift l.

Proof. Let

$$Q^l_{b,c}=\sum_{a\in B}P_af(a,b)\,T_{a,c}{}^l,$$

then we obtain

$$W_f^{|B|}(l) = \sum_{b,c\in B} Q^l_{b,c} \left(\sum_{k=0}^{|B|} w_k T^k\right)_{c,b} = 0,$$

because T satisfies its own characteristic polynomial. □

Apart from the formal proof, the main reason behind lemma 1 is that the correlation function $C_f(k)$ consists of a sum of geometric series where the quotients are the eigenvalues of the transfer matrix. Indeed, if the transfer matrix is diagonalizable: $T = U^{-1}DU$ where $D = \mathrm{diag}(\lambda_1,\ldots,\lambda_{|B|})$ is diagonal, then

$$C_f(k) = \sum_{a,b,c\in B} P_a U^{-1}_{ac} \lambda^k_c U_{cb} f(a,b) = \sum_{c\in B} R_c \lambda^k_c, \tag{3}$$

where $R_c = \sum_{a,b\in B} P_a U^{-1}_{ac} U_{cb} f(a,b)$. Since each λ satisfies the characteristic polynomial, thus does the complete correlation function, too. We remark that for $k=0$ the condition $T^0 = \mathbf{1}$ is equivalent to $\lambda^0 = 1$, even for $\lambda = 0$.

Because of this fact, the correlation function provides information about the eigenvalues of the transfer matrix, in particular the number of the eigenvalues, which is $|B|$, the dimension of the state space. A slight problem may arise, however, if the transfer matrix has multiple roots, which can be present in special, very symmetric processes: in this case, the correlation functions contain a smaller number of terms[4]. We remark that a single zero eigenvalue influences the correlation functions, but only in the $k=0$ element (using the $\lambda^0 = 1$ rule).

It is another problem if some of the R_c coefficients are zero; that can be the result of using a special f function to calculate the correlations. This is an example of a specific symmetry in the system, which we exclude (cf discussion at the end of this subsection.

These pathological cases will be discussed further in section 4. Now we will assume that we consider a generic enough process and generic enough correlation function that such problems do not occur. In this generic situation, the number of states of the Markov process equals the order of the characteristic polynomial.

Usually, we use lemma 1 in the reverse direction, which means that we observe a finite chunk of the time series $Y^n_{i=0}$, and we want to give an estimate for $|B|$. If the correlation function f is generic, then we have the following corollary of lemma 1:

Corollary 1. *Assume that the underlying process has no special symmetries (T has no multiple eigenvalues). If for a given $f: B\times B \to \mathbb{R}$ we find a linear relation of $N+1$ order satisfying*

$$W_f^N(l) = \sum_{k=0}^{N} v_k C_f(k+l) = 0, \qquad \forall\, l$$

for some v, and there is no smaller-order linear relation with this property, then

$$N \leqslant |B|.$$

If f is generic then

$$N = |B|.$$

[4] In this case, the minimal polynomial is not equal to the characteristic polynomial.

Proof. With this notation of lemma 1 we have

$$\mathrm{Tr}\, Q^{(l)} \left(\sum_{k=0}^{N} v_k T^k \right) = 0.$$

This implies

$$\left. \sum_{k=0}^{N} v_k T^k \right|_S = 0, \tag{4}$$

where S is the subspace spanned by all $Q^{(l)}$ matrices.

First, we consider the case when S is the complete space. Since for a matrix without multiple eigenvalues, the minimal polynomial satisfied by the transfer matrix is the characteristic polynomial, then we find $N \geqslant |B|$. If there is no smaller order linear relation with this property, then $N = |B|$. If S is smaller, then we have less constraint, and so $N < |B|$. For generic f we have a generic Q^l and the space S is complete and the equality should hold. □

We remark that if $|B| = |B_g|$, then the observations Y_n form a MC themselves. In this case, we can fully determine the transition matrix purely from observations, and our method is not needed. Unfortunately, if only a projection of the MC is observed, that straightforward estimate of the transition matrix is not possible. The reason is that the observed time series is no longer Markovian. The transition probabilities $P_{n\to n+k}$, although they provide $|B_g| \times |B_g|$ dimensional transition matrices, do not arise from powers of a single $|B_g| \times |B_g|$ dimensional generator transition matrix. Our method aims precisely at such a situation and provides the reconstruction of the discrete, finite state space of the MC, assuming the observation lacks any special symmetry.

As an illustration of degenerate cases where the dimension of the state space cannot be reconstructed due to symmetry, consider a hypothetical scenario with a 2-dimensional vector-valued discrete space for the MC. Let $d = a - b$ and $e = a + b$, but our observation is $f = d + \alpha e$. It is evident that the time delay embedding of f cannot reconstruct the process if $\alpha = 1$; however, in all other cases, it can. This specific case is excluded in Sauer's paper [21] under the condition termed 'prevalent', indicating that the subset of degenerate cases within the function space of observations has zero measure or probability.

2.3. Methodology

Building upon the theory described in the previous subsection, we now outline the methodology that can be applied to actual data series.

In the reality we have (time) series $y : \mathbb{N} \to B$, and we can measure the equilibrium autocorrelation functions $C_f(x)$

$$\begin{aligned} C_f(k) &= \lim_{M\to\infty} \frac{1}{M} \sum_{n=0}^{M-1} h\left(y_n, y_{n+k}\right) \\ &= \lim_{M\to\infty} \frac{1}{M} \sum_{n=0}^{M-1} h\left(g\left(x_n\right), g\left(x_{n+k}\right)\right) \\ &= \lim_{M\to\infty} \frac{1}{M} \sum_{n=0}^{M-1} f(x_n, x_{n+k}) \,. \end{aligned} \tag{5}$$

In practice, of course, M is finite, and there are precision issues; these will be discussed later. In this subsection, we deal with the problem, of how to determine the minimal length linear relation that is satisfied by C_f.

So let us assume that C_f satisfies a linear relation

$$\sum_{k=0}^{s} u_k C_f(k+l) = 0. \tag{6}$$

In fact, we expect from lemma 1 that it is indeed true. For the estimate of the number of states we need the smallest s possible (cf corollary 1).

To determine the u_k coefficients, we prepare a matrix

$$F^{(f)}_{lk} = C_f(k+l)\,, \qquad k \in \{0,\ldots,s\}\,,\ l \in \{0,\ldots,L\}\,. \tag{7}$$

The linear relation of equation (6) implies

$$0 = \sum_{k=0}^{K} F^{(f)}_{lk} u_k. \tag{8}$$

In matrix notation, with $F^{(f)}$ matrix we write

$$F^{(f)} v = 0. \tag{9}$$

In the first method, we choose the maximal l to be s, then F is a symmetric, $s+1$ order square matrix. Then (9) implies that $F^{(f)}$ has at least one *zero eigenvalue*[5]. The eigenvectors belonging to the zero eigenvalues give the coefficients of the linear equation for the correlation functions. But usually, there are several such relations, and we need the one with the minimal order.

We recall that if there are two linear equations with length s, then we can prepare a single linear relation with length $s-1$, by simply expressing the last variable from the first equation, and substituting it into the second.

To generalize this observation, let us denote the non-zero subspace of $F^{(f)}$ by:

$$\mathrm{Sp}^{(f)}_{+} = \mathrm{Span}\left\{ v \mid F^{(f)} v = \lambda v,\, \lambda > 0 \right\}, \qquad N^{(f)}_{+} = \dim\left(\mathrm{Sp}^{(f)}_{+}\right).$$

Then $F^{(f)}$ has $N_0 = s+1-N_+$ zero eigenvalues. This means that there are N_0 linear relations, each of order $s+1$, that are satisfied by the correlation functions. But it follows that there is also a single $s+2-N_0$ order linear equation which is satisfied by $C_f(k)$. So the minimal order is $s_{min}+1 = N_+ + 1$, implying $s_{min} = N_+$.

For another method, we do not require $s = L$. Then we start from (9) to write

$$|F^{(f)} v|^2 = v^T {F^{(f)}}^T F^{(f)} v = 0. \tag{10}$$

Since the left-hand side is always positive, its zero value means a minimum. Normalizing $\sum_{k=0}^{K} v_k^2 = 1$ we arrive at the conditional minimization problem:

$$v^T {F^{(f)}}^T F^{(f)} v = \text{minimal},\ v^T v = 1. \tag{11}$$

With the notation

$$S^{(f)} = {F^{(f)}}^T F^{(f)} \tag{12}$$

[5] To avoid confusion we remark that the eigenvalues of F and those of T are different. In particular, a zero eigenvalue of T implies no restriction on the spectrum of F.

we arrive at the condition, using Lagrange multipliers

$$v^T S^{(f)} v - \lambda v^T v = \text{minimal}. \tag{13}$$

This leads to

$$S^{(f)} v = \lambda v \tag{14}$$

eigenvalue equation. The condition $v^T v = 1$ implies

$$0 = |F^{(f)} v|^2 = v^T S^{(f)} v = \lambda v^T v = \lambda. \tag{15}$$

Therefore we need the zero eigenvalues of the $S^{(f)}$ matrix. All eigenvectors belonging to a zero eigenvalue of the matrix $S^{(f)}$ give a linear relation.

Now we are in the situation that lowers s until $S^{(f)}$ has only a single zero eigenvalue. Or, just as in the case of a square matrix F, we use the dimension of the nonzero eigenspace to determine $|B|$.

2.4. Finite size analysis

The theoretical analysis above was valid for infinite data length and infinite precision calculations. In practice, however, computer number representation accuracy is finite, and we have also a finite, sometimes just a limited number of data. The question is, what can we say in these constrained possibility cases?

First, let us deal with the problem of finite data size. What we propose is to exploit the fact that at the infinite number of data limits we in fact find a zero eigenvalue either for F or S. Thus we expect that if we increase the number of data, we find smaller and smaller eigenvalues.

So we propose to prepare the plot $\lambda_i(M)$, the ith eigenvalue as a function of the number of data, for the complete domain of M, not just considering the spectrum at the maximal available M.

We expect that in the case of a finite eigenvalue, there is a critical M_i, which is needed to find that finite value. For $M > M_i$ we safely find the λ_i value in the spectrum, and so the $\lambda_i(M)$ shows a saturation.

The other type of eigenvalues, which are zero in the infinite number of data limit, show no saturation; they decrease for large M. We prove the following lemma:

Lemma 2. *If the data consist of a large number of independent pieces, then the would-be-zero eigenvalues of the embedded correlation matrix decrease as* $\sim 1/M$ *for M data points.*

Proof. Let X_M denote the MCs with M data points. We assume that the correlation length L of the series is much smaller than M, hence we have approximately $n_M = M/L \gg 1$ quasi-independent observations for the correlation length. Let $\{C_a \mid a = 1 \ldots n_M\}$ be the ensemble of correlation functions, and $\{F_a \mid a = 1 \ldots n_M\}$ the corresponding embeddings. The embedding F_M for the complete length M is their average $F_M = \frac{L}{M} \sum_{a=1}^{n_M} F_a$. We denote $F = F_\infty$. Clearly, $F = \lim_{M\to\infty} F_M$.

We can express F_a as $F_a = F + \delta F_a$. If these matrices are independent for different a, then δF_a is a random variable with zero mean. Therefore, $F_M = F + \delta F_M$, where

$$\delta F_M = \frac{L}{M} \sum_{a=1}^{n_M} \delta F_a \tag{16}$$

is the average of approximately M i.i.d random variables. By the central limit theorem, δF_M is a normally distributed random variable with zero mean and standard deviation $1/\sqrt{M}$. For the quadratic form, we have

$$F_M^T F_M = F^T F + \delta F_M^T F + F^T \delta F_M + \delta F_M^T \delta F_M = F^T F + \delta A. \tag{17}$$

Now, consider an eigenvector v_M of $F_M^T F_M$ corresponding to a would-be zero eigenvalue (any of them). Let $v = \lim_{M\to\infty} v_M$, which satisfies $F^T F v = 0$. Introducing the difference $\delta v_M = v - v_M$, and for the corresponding eigenvalue, we write $\lambda_M = \lambda + \delta\lambda_M$ (in this case $\lambda = 0$). The eigenvalue equation becomes

$$F_M^T F_M v_M = \left(F^T F + \delta A\right)(v + \delta v_M) = \lambda_M v_M = \delta\lambda_M (v + \delta v_M). \tag{18}$$

To leading order in the perturbation, we find

$$\delta A v + F^T F \delta v = \delta\lambda_M v. \tag{19}$$

Multiplying this equation by v^T from the left, and using $v^T F^T F = (F^T F v)^T = 0$, we obtain

$$\delta\lambda_M = v^T \delta A v. \tag{20}$$

Using the form $\delta A = \delta F_M^T F + F^T \delta F_M + \delta F_M^T \delta F_M$, and noting that if $v^T F^T F v = 0$, then $Fv = 0$, it follows that only the last term remains:

$$\delta\lambda_M = |\delta F_M v|^2 \sim \mathcal{O}\left(\frac{1}{M}\right). \tag{21}$$

□

Thus, we expect that with a finite number of data points, the would-be-zero eigenvalues approach zero as $\lambda_i(M) \sim 1/M$, without saturation. Numerical evidence supports this analysis (cf section 3).

If we have a finite number of available data, then we see some (N_{sat}) eigenvalues that are saturated already and some (N_{decr}) that are decreasing. Then the best we can say, based on this observation, is that the number of states $|B|$ is at least N_{sat}, and that the remaining N_{decr} states fulfill an approximate linear law.

3. Results

Let us consider some simulated data to demonstrate, how our method works.

3.1. Generic 2-state MC

Our first, simplest example intends to show all the practical details. Later on, we use the same methodology for other cases.

Let us have a 2-state MC with states $B = \{a, b\}$ and with the transfer matrix

$$T = \begin{pmatrix} 0.3 & 0.7 \\ 0.6 & 0.4 \end{pmatrix}. \tag{22}$$

These numbers are meant to be general enough.

This transfer matrix has two eigenvalues: $(1, -0.3)$. The unit eigenvalue is the consequence of the fact that the sum of each row is one. The left eigenvalue belonging to $\lambda = 1$ is the equilibrium distribution, in this case, we have $P_{\mathrm{eq}} = (6/13, 7/13)$.

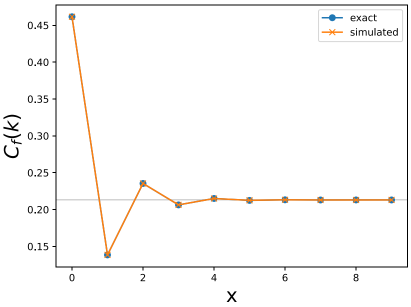

Figure 1. The correlation function in the exact case compared with the numerical one from an MC containing 10^6 data.

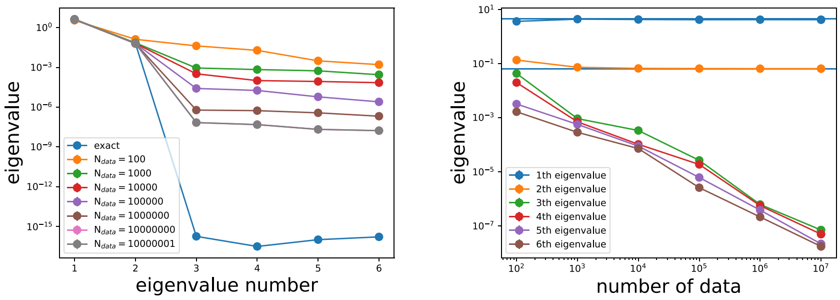

Figure 2. The data number dependence of the spectrum of S matrix.

Let us consider a simple observable associated with $f(x,y) = \delta_{xa}\delta_{ya}$. The corresponding $C_f(k)$ is

$$C_f(k) = P_a \left(T^k\right)_{aa}.$$

We expect generically that at $k = 0T^0 = 1$ and so $C_f(0) = P_a$; in this case it is $6/13$. For $k \to \infty$ the probability of having a at zero and at k becomes independent, and so $C_f(k \to \infty) = P_a^2$; in our case it is $(6/13)^2$. The exact $C_f(k)$ curve is shown in the left panel of figure 1. In this plot we also see the numerical result coming from an MC containing one million data. As we see, the exact and the numerically determined autocorrelation functions are very close to each other.

For a numerical analysis, we prepared the MC based on the given transfer matrix, starting from state a. On this chain, we determined the autocorrelation function up to $k = 12$. Then, using $L = 6$, we determine the F matrix and the corresponding $S = F^T F$ matrix. The number of its nonzero eigenvalues will determine the number of states $|B|$, which, in this case, shall be 2.

In the left panel of figure 2 we see the spectrum for different numbers of data. As we see, in the exact case there are two large eigenvalues and four small ones in the order of 10^{-16}. This is the machine accuracy of the eigenvalue determination in our case. The exact eigenvalue smaller than this number can not be found even in the limit when we collect an infinite number of data.

On the right panel, we find the data number dependence of the different eigenvalues, where horizontal lines show the exact values. As was promised, the nonzero eigenvalues are stabilized

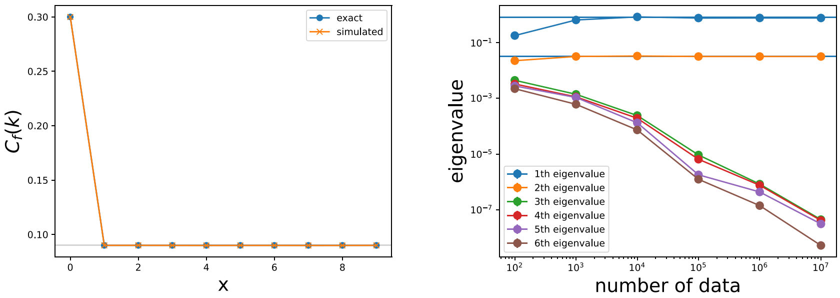

Figure 3. The data number dependence of the spectrum of S matrix.

after a certain M, which is rather small in this 2×2 case. The would-be zero eigenvalues decrease as $1/M$ (a straight line in the log-log plot).

As we have seen, the dimension of the state space can be determined from the number of nonzero eigenvalues. In the present case, it was enough to collect a mere 1000 data to see the two states. If we collect 10^7 data, we can state that with the precision of 10^{-7}, there is no third state in the system.

To go on, we shall determine the smallest linear relation that maps the data to zero. To this end, we take $L=3$ and repeat the procedure. We obtain the eigenvalues $6.79, 0.252, 1.04\cdot 10^{-16}$. The eigenvector corresponding to the third, machine-zero eigenvalue, normalized in a way that the last element is one, reads $(-0.3,-0.7,1)$. If we multiply this vector with F, then we indeed obtain (a numerical) zero. This eigenvector yields the polynomial

$$\lambda^2-0.7\lambda-0.3=0.$$

This is indeed the characteristic polynomial of T, giving the roots 1 and -0.3, as we have already seen.

3.2. 2-state MC with zero eigenvalue

Let us repeat that same analysis, with the matrix

$$T=\begin{pmatrix}0.3 & 0.7\\ 0.3 & 0.7\end{pmatrix}. \tag{23}$$

This matrix has two eigenvalues: $(1,0)$. The equilibrium distribution is $(0.3,0.7)$.

Despite the T-matrix having a zero eigenvalue, this does not pose any problem for our method. In figure 3, in the left panel we can see the correlation function. It starts with 0.3, and in the second step, it reaches 0.09. On the right panel, we see the spectrum of the S matrix. It has two finite eigenvalues, and, down to 10^{-7} we do not have a third one. This suggests, correctly, that the number of states is 2.

We can reconstruct the characteristic polynomial, if we choose $L=3$, and then take the eigenvector belonging to the only zero eigenvalue. It reads $(0,-1,1)$, corresponding to the polynomial $\lambda-\lambda^2$, which is indeed the correct characteristic polynomial of the above transfer matrix.

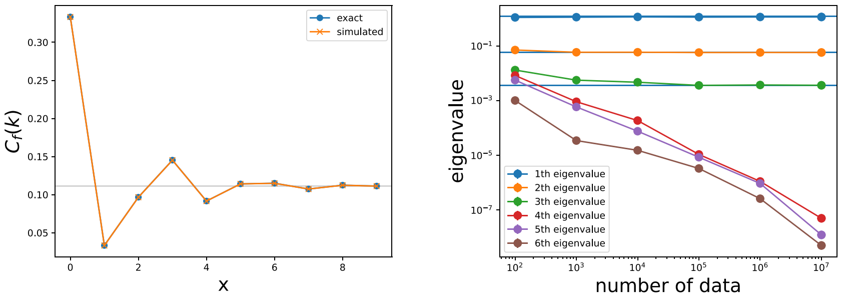

Figure 4. (left) The autocorrelation function of a 3-state Markov chain. (right) The data number dependence of the eigenvalues of the S matrix. We can observe three stabilized eigenvalues, corresponding to the number of states $|B| = 3$.

3.3. Generic 3-state MC

We can apply the same strategy for a 3-state MC, too. Let us choose for the states $B = \{a, b, c\}$ and for the transfer matrix:

$$T = \begin{pmatrix} 0.1 & 0.2 & 0.7 \\ 0.7 & 0.1 & 0.2 \\ 0.2 & 0.7 & 0.1 \end{pmatrix}. \tag{24}$$

A special feature of this matrix is that it has complex eigenvalues $-0.35 \pm 0.433i$ besides the unit eigenvalue.

In figure 4 the left panel shows the autocorrelation function. As a consequence of the complex eigenvalues of the transfer matrix, it exhibits oscillations.

On the right panel, the data number dependence of the spectrum is shown. As we can see, we must wait until we have collected approximately 10^5 data points before we can confidently claim that the third eigenvalue has stabilized. Nevertheless, the third eigenvalue deviates from the $1/M$ law much before the saturation occurs.

3.4. 3-state MC with degenerate eigenvalues of the transfer matrix

In a 3-state Markov process, we can have a transfer matrix with two identical eigenvalues. Consider

$$T = \begin{pmatrix} 1/2 & 1/4 & 1/4 \\ 1/4 & 1/2 & 1/4 \\ 1/4 & 1/4 & 1/2 \end{pmatrix}. \tag{25}$$

Its three eigenvalues are $(1, 1/4, 1/4)$.

If we study the MC generated by this matrix we find the results plotted in figure 5. The numerical procedure correctly identifies the two eigenvalues, however, we do not have information about the multiplicities. In the next section, we analyze this problem in detail.

4. Discussion

4.1. Multiple eigenvalues and perturbation

As we have demonstrated in the previous section, the state space of Markov processes generated by a transfer matrix with a non-degenerate spectrum can be completely characterized

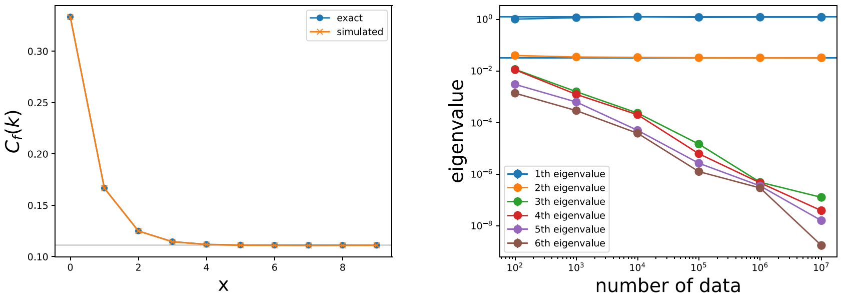

Figure 5. (left) The correlation function of a 3-state Markov chain where the transfer matrix has multiple eigenvalues. (right) The data number dependence of the eigenvalues of the S matrix. We can not obtain multiplicity information from the spectrum, and correspondingly we just observe two stabilized eigenvalues.

by the nonzero part of the spectrum of the corresponding S matrix. As we have seen, the zero and non-zero eigenvalues can be separated numerically only up to a certain precision that is proportional to the collected data.

A single topic remained to be discussed, and this is the case of the degenerate spectrum. Let us return for a moment to the last example of the previous subsection with transfer matrix (25), and study the evolution of the probability distribution. It is governed by the equation

$$P_{n+1} = P_n T. \tag{26}$$

The T matrix of (25) can be written as a sum of projectors:

$$T_{ab} = \Pi^{(\mathrm{eq})}_{ab} + \frac{1}{4}\left(\delta_{ij} - \Pi^{(\mathrm{eq})}_{ab}\right), \quad \text{where} \quad \Pi^{(\mathrm{eq})}_{ab} = \frac{1}{3}. \tag{27}$$

The initial condition can be written accordingly:

$$P_0 = P_{\mathrm{eq}} + \hat{P},$$

where $P_{\mathrm{eq}} = (1/3, 1/3, 1/3)$.

The point is that for any n we can write

$$P_n = P_{\mathrm{eq}} + \frac{1}{4^n}\hat{P},$$

which corresponds to a 2-state MC.

This means that the reason that in the actual MC, we can not discover the third eigenvalue is, that it does not appear in the process at all. All concrete MCs run in a 2-state subspace.

In practice, it is very unlikely that we encounter an exact degenerate spectrum. First of all, multiple eigenvalues very rarely occur, as it is not connected directly to explicit symmetries of the system, therefore a noisy environment most probably will break this symmetry, and lift the degeneracy. Thanks to that and with the aid of corollary 1 we can determine $|B|$ if the observation is generic.

Anyway, if it is suspicious that the system has multiple eigenvalues we may use a simple perturbation to reveal the hidden eigenvalues and states. We need to generate random uniform number(s) and use them as additional stay probability for the observed state(s). For instance, if $q_a \in [0,1]$ is the generated stay rate for state a, for each observed $x_i = a$ we shall insert an additional a into the observed sequence with probability q_a until the random trial stops. (Of

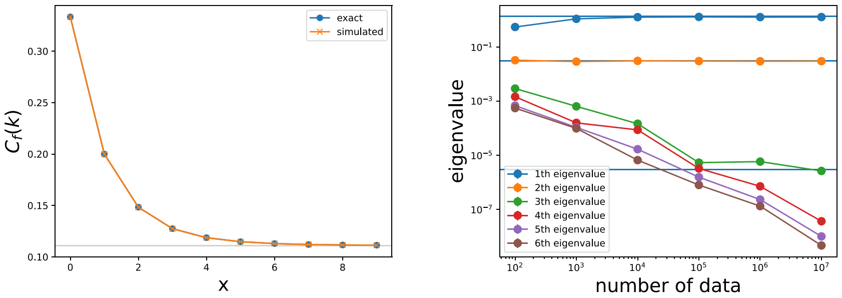

Figure 6. (left) The correlation function of a 3-state Markov chain coming from a perturbed transfer matrix. (right) The data number dependence of the eigenvalues of the S matrix. The perturbation lifted the degeneracy, and we can observe three stabilized eigenvalues corresponding correctly to the number of states $|B| = 3$.

course, other perturbations may work as well.) The observed perturbed sequence is also coming from a MC but the multiple eigenvalues are decoupled as the next example demonstrates.

One should note that if there are no multiple eigenvalues, such perturbation does not change the number of eigenvalues.

Let us take the same T matrix as in (25), and the stay rates are incorporated using a perturbation matrix:

$$R = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix}, \tag{28}$$

such as the perturbed transfer matrix is

$$T_{\text{pert}} = T + pR, \tag{29}$$

where p is a tuning parameter (in our example $p = 0.1$). The autocorrelation function and the spectrum are shown in figure 6. As we can see, the autocorrelation function changes only slightly, still, the spectrum of the S matrix exhibits a third eigenvalue, demonstrating that the true number of states is 3.

4.2. *Nearby eigenvalues*

Even if the spectrum is not degenerate, nearby eigenvalues may be difficult to identify. In this subsection, we try to give a hint about the numerical nature of this problem.

Let us assume that our original correlation function contains two geometric series, but the q_1, q_2 quotients are close $q_1 \approx q_2$. Let us introduce the notations

$$q_1 = q + \frac{\Delta q}{2}, \quad q_2 = q - \frac{\Delta q}{2}. \tag{30}$$

The original series is

$$C_k = a_0 + a_1 q_1^k + a_2 q_2^k. \tag{31}$$

We approximate it with a single quotient series:

$$C'_k = a_0 + (a_1 + a_2)\, q^k. \tag{32}$$

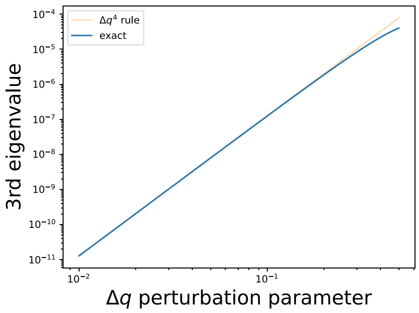

Figure 7. The 3rd eigenvalue as a function of the perturbation parameter.

This latter correlation function satisfies the law $C'_{k+1} - qC'_k = 0$. What happens, if we apply this law to the original series? To characterize the error we introduce the corresponding χ^2 value as

$$\chi^2 = \sum_{n=0}^{\infty} \left(C_{k+1} - qC_k\right)^2. \tag{33}$$

After a short algebraic manipulation, we obtain the leading order in Δq

$$\chi^2 = \frac{\Delta q^4}{4} \frac{1+q^2}{\left(1-q^2\right)^3} + \mathcal{O}\left(\Delta q^6\right) > \frac{\Delta q^4}{4}. \tag{34}$$

Another characterization of the problem comes from the study of the third eigenvalue of the S matrix. Here we shall choose actual numerical values, for example, $a_0 = a_1 = a_2 = 1$, and $q = 0.5$. The third eigenvalue is zero for $\Delta q = 0$, and it grows with Δq. Numerically we find the plot in figure 7.

In this log-log plot, we compared the third eigenvalue and the $\sim \Delta q^4$ law. As we see, somewhat surprisingly, it works for almost $\Delta q = 0.5$.

This means that the smallest eigenvalue of the S matrix is proportional to the fourth power of the splitting of the eigenvalues of the T-matrix. Since the number of data necessary to resolve an eigenvalue is $1/\lambda$, we find that we need to collect at least $M \sim 1/\Delta q^4$ data to see the different eigenvalues.

This is bad news at one hand, and good news at another. It is bad since the autocorrelation function of an MC reveals very slowly the true number of states if the eigenvalues of the T matrix are close.

In reality, however, few-state Markov processes are very rare. In nature, everything is interconnected; for example, the position of Mars can influence the period of a clock, even if only in a very subtle way. Therefore, in real processes, many states are typically involved, sometimes even continuously (as in hydrodynamics). But the above evidence suggests that we can approximate these cases with a finite number of states since the difference in observables like the correlation function will be small. If we have a finite amount of data M, then we shall examine the spectrum of the S matrix with the finite size analysis described in this paper, and use so many states that correspond to the number of nonzero (would-be) eigenvalues. In most practical applications there will be no difference between the two models.

5. Conclusions and future work

In this paper, we proposed a method that is capable of reconstructing the phase space of a Markov process and getting some information about the eigenvalues of the corresponding transfer matrix. For that, we needed to know a (generic enough) autocorrelation function.

The method is based on the observation that all autocorrelation functions satisfy a linear relation. With the embedding of the autocorrelation function, we can create a symmetric matrix (the embedding matrix itself, or the PCA matrix). In theory, the number of nonzero eigenvalues yields the number of states in the Markov process, while the minimal length zero eigenvalue reproduces the characteristic equation of the transfer matrix.

The practical application of the theoretical method is challenged by some problems, first of all, the finite number of available data, but also the finite precision calculations in the eigenvalue determination. We propose a finite-size scaling method to separate the established nonzero eigenvalues and the would-be zero eigenvalues, up to a given precision. We have shown that the value of the would-be zero eigenvalues scale by $1/M$ where M is the sample size, while the established eigenvalues reach a constant value.

We discussed also the problem of multiple, or (as it occurs in practice) nearby eigenvalues. We demonstrated that the necessary sample size is $M\sim 1/\Delta q^4$, to resolve eigenvalues of the T matrix that differ by Δq.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgments

The project is supported by the Hungarian National Research, Development and Innovation Office (NKFIH) under reference number RRF-2.3.1-21-2022-00004. A T was supported by Hungarian Research Network (formerly Eötvös Loránd Research Network), Grant SA-114/2021 and Hungarian Scientific Research Fund (OTKA/NRDI Office) under Contract Number K13583. Project No. PD142593 was implemented with the support provided by the Ministry of Culture and Innovation of Hungary from the National Research, Development, and Innovation Fund, financed under the PD_22 'OTKA' funding scheme.

Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could influence the work reported in this paper.

ORCID iDs

Antal Jakovác https://orcid.org/0000-0002-7410-0093
Marcell T Kurbucz https://orcid.org/0000-0002-0121-6781
András Telcs https://orcid.org/0000-0002-3205-3081

References

[1] Brenner M, Eldredge J and Freund J 2019 Perspective on machine learning for advancing fluid mechanics *Phys. Rev. Fluids* **10** 100501

[2] Snyder J C, Rupp M, Hansen K, Muller K-R and Burke K 2012 Finding density functionals with machine learning *Phys. Rev. Lett.* **108** 253002

[3] Desai S and Strachan A 2021 Parsimonious neural networks learn interpretable physical laws *Sci. Rep.* **11** 12761

[4] Jakovac A, Kurbucz M T and Pósfay P 2022 Reconstruction of observed mechanical motions with artificial intelligence tools *New J. Phys.* **24** 073021

[5] Takens F 2006 Detecting strange attractors in turbulence *Dynamical Systems and Turbulence, Warwick 1980: Proc. Symp. Held at the University of Warwick 1979/80* (Springer) pp 366–81

[6] Sugihara G, May R, Ye H, Hsieh C, Deyle E, Fogarty M and Munch S 2012 Detecting causality in complex ecosystems *Science* **338** 496–500

[7] Stippinger M, Bencze A, Zlatniczki Adám, Somogyvári Z and Telcs A 2023 Causal discovery of stochastic dynamical systems: a Markov chain approach *Mathematics* **11** 852

[8] De la Peña L and María Cetto A 1975 Stochastic theory for classical and quantum mechanical systems *Found. Phys.* **5** 355–70

[9] Shreve S 2004 *Stochastic Calculus for Finance 2, Continuous-Time Models* (Springer)

[10] Deistler M and Scherrer W 2022 *Time Series Models* (Springer International Publishing)

[11] Gatheral J, Jaisson T and Rosenbaum M 2014 Volatility is rough (arXiv:1410.3394)

[12] Heath D, Jarrow R and Morton A 1990 Bond pricing and the term structure of interest rates: a discrete time approximation *J. Financ. Quant. Anal.* **25** 419–40

[13] Hagan P S, Kumar D, Kesniewski A S and Woodward D E 2002 Managing smile risk *Wilmott* **1** 84–108

[14] Stark J 1999 Delay embeddings for forced systems. I. Deterministic forcing *J. Nonlinear Sci.* **9** 255–332

[15] Stark J, Broomhead D S, Davies M E and Huke J 2003 Delay embeddings for forced systems. II. Stochastic forcing *J. Nonlinear Sci.* **13** 519–77

[16] Kantz H and Ragwitz M 2004 Phase space reconstruction and nonlinear predictions for stationary and nonstationary markovian processes *Int. J. Bifurcation Chaos* **14** 1935–45

[17] Kennel M B, Brown R and Abarbanel H D I 1992 Determining embedding dimension for phase-space reconstruction using a geometrical construction *Phys. Rev.* A **45** 3403–11

[18] Friedrich R and Peinke J 1997 Description of a turbulent cascade by a Fokker-Planck equation *Phys. Rev. Lett.* **78** 863

[19] Ragwitz M and Kantz H 2002 Markov models from data by simple nonlinear time series predictors in delay embedding spaces *Phys. Rev.* E **65** 056201

[20] Györfi L'o, Kohler M, Krzyzak A and Walk H 2002 *A Distribution-Free Theory of Nonparametric Regression* (*Springer Series in Statistics*) (Springer)

[21] Sauer T, Yorke J A and Casdagli M 1991 Embedology *J. Stat. Phys.* **3–4** 11