Università degli Studi di Macerata

Dipartimento di Istituzioni Economiche e Finanziarie

Optimal Dimension of Transition Probability Matrices for Markov Chain Bootstrapping

Roy Cerqueti, Paolo Falbo, Cristian Pelizzari

Quaderno di Dipartimento n. 53

Aprile 2009

Optimal Dimension of Transition Probability Matrices for Markov Chain Bootstrapping

Roy Cerqueti, Paolo Falbo, Cristian Pelizzari

Abstract

While the large portion of the literature on Markov chain (possibly of order higher than one) bootstrap methods has focused on the correct estimation of the transition probabilities, little or no attention has been devoted to the problem of estimating the dimension of the transition probability matrix. Indeed, it is usual to assume that the Markov chain has a one-step memory property and that the state space could not to be clustered, and coincides with the distinct observed values. In this paper we question the opportunity of such a standard approach. In particular we advance a method to jointly estimate the order of the Markov chain and identify a suitable clustering of the states. Indeed in several real life applications the "memory" of many processes extends well over the last observation; in those cases a correct representation of past trajectories requires a significantly richer set than the state space. On the contrary it can sometimes happen that some distinct values do not correspond to really "different states" of a process; this is a common conclusion whenever, for example, a process assuming two distinct values in t is not affected in its distribution in $t+1$. Such a situation would suggest to reduce the dimension of the transition probability matrix. Our methods are based on solving two optimization problems. More specifically we consider two competing objectives that a researcher will in general pursue when dealing with bootstrapping: preserving the similarity between the observed and the bootstrap series and reducing the probabilities of getting a perfect replication of the original sample. A brief axiomatic discussion is developed to define the desirable properties for such optimal criteria. Two numerical examples are presented to illustrate the method.

Roy Cerqueti, Università degli Studi di Macerata.
E-mail: roy.cerqueti@unimc.it.
Paolo Falbo, Università degli Studi di Brescia.
E-mail: falbo@eco.unibs.it.
Cristian Pelizzari, Università degli Studi di Brescia.
E-mail: pelizcri@eco.unibs.it.

1 Introduction

The interest to bootstrap methods has largely expanded after the seminal paper of [13]. Important extensions to the original method have appeared in the literature to solve some relevant problems. In particular major difficulties arise when the distribution of observed data is estimated under the hypothesis that data are dependent in some way.

It is possible to group different contributions to the data dependency problem, based on the technical approach adopted to preserve most of the information on the dependency structure. A first major category is that based on the construction of time blocks. Following this approach a series of stationary data is divided into blocks of length l (usually constant) of consecutive observations; bootstrap samples are then generated joining some of those intervals selected at random. The seminal idea appears first in [14] for spatial data, is applied to time series by [10], but has been fully developed starting with [18] and [21]. A crucial point in this method, largely analyzed in the literature, is that of determining an "optimal" value for the parameter l. In [15], [8], [28], and [20] the selection of the block size is driven by the observed data.

Many variants of the block bootstrap method exist by now; standard references include [26] for the *blocks-of-blocks bootstrap*; [27] for the *stationary bootstrap*; [22] and [24], for *tapered block bootstrap*. A survey on block bootstrap methods with a comparison of their performances is presented in [19]. Despite the block based bootstrap methods have been developed to get over the problem of dependence disruption, they only partially succeed in their goal. Indeed they pass from the loss of dependency between data to that between blocks.

A second important family of bootstrap methods explicitly developed to recognize and maintain the original data dependency is concerned with Markov chains (or processes). Earlier approaches to bootstrap Markov chains were initially advanced by [17], [3], and [2] and further investigated in [11]; these methods are based on a nonparametric estimation of the transition probabilities to capture the structure dependency of the driving Markov chain. This second group is more closely related to our work since it focuses on the transition probabilities of a stationary Markov chain (Markov process) as we also do here. Also in this case it is useful to distinguish some different approaches. The *sieve (Markov) bootstrap* method was first advanced by [6]; it consists of fitting Markovian models (such as an AR) to a data series and resampling randomly from the residuals. This idea has been further developed in [7], where the *variable length Markov chain sieve bootstrap* method is advanced. This is an intriguing approach since in nature it happens that only "some" sequences of states (i.e. paths) tend to reappear in an observed sequence more than others and to condition significantly the process evolution. However this method proceeds in a hierarchical way searching for the relevant paths, which can be a severe limitation when time dependence is not monotonically decreasing.

Still in the framework of Markov processes, [29] and [16] estimate the transition density function of

a Markov process using kernels. The idea of using kernels is adopted also by [25] and [23], which advance the so called *local bootstrap* method; at each step the next bootstrap value is extracted from the set of all the observed values based on a kernel probability estimate. In particular such probability favors those values which have the most similar previous path (i.e. sequence of values preceding them) to that of the bootstrap series built up to that step. This method rests on the (intuitive) assumption that similar trajectories will tend to show similar transition probabilities in the future. However it is not uncommon to observe empirical contradiction to such hypothesis. Besides the number of time lags to be observed to compare trajectories has to be chosen arbitrarily.

Another approach which is relevant to compare with our work is that of [1], who propose a method (*Markov chain bootstrap*) based on a finite state discrete Markov chain. Similarly to what we search here, the authors partition the state space of the series into I sets (bins). While the authors show some interesting estimation properties of their bootstrap method, the bins are formed observing no criteria, but simply distributing the ordered values evenly in each of them. Besides an arbitrary number of time lags is also fixed to bound the relevant path length.

The approach called *regenerative (Markov chain) block bootstrap* has been initially developed by [2] and [12] and further analyzed by [4] and [5]. It focuses on the times when a regenerative Markov process passes by a recurring state (atom). The consecutive observations between departure and return to the atom form a cycle (or block). Bootstrapping is then accomplished sampling at random from the observed cycles. This method reconciles the gap between Markov Chain bootstrapping methods and block bootstrapping, with the important difference that the cutting points (used to form the blocks) in the Markov Chain approach are not chosen at random, but are data driven. Besides it does not need to explicit estimate the transition probabilities of the observed process. However this relies heavily on the identification of the atom, which is unfortunately unknown.

Our work contributes to the literature in various ways.

We focus on the estimation of the transition probability matrix of a Markov chain. As previously discussed, methods which have focused on this problem have solved it only in partial way and in large portion unsatisfactorily. The most serious attempt in this sense is represented by the *variable length Markov chain* approach ([7]), as it explicitly tries to size the order of the Markov chain comparing the discriminatory power of transition probabilities of paths with variable length. Close to the spirit of this work we estimate the order of a Markov chain as the time beyond which it has no value to separate states in different groups. However our contribution is not in the direction of presenting an alternative search method, but in presenting some optimal search criteria. Under these criteria the cases of Markov chains where time relevance does not decay monotonically with time can be identified and their transition probability matrix correctly estimated. On the contrary the *variable length Markov chain* method can not account for processes where time dependency drops at some time lags and reappears at farther lags.

Another aspect of novelty that we adopt in assessing similarity between states is that we define distance on their transition probabilities rather than on their values. The kernel methods cited previously (as well as other works on scenario generation such as [9]) approximate transition probabilities by means of the paths most similar to a given trajectory, where similarity is measured on the observed values. However in these applications the closeness in the values is used as a proxy of the closeness in the transition probabilities, so their direct use should be preferred. The goodness of the estimation of transition probabilities obtained through kernel methods is heavy, and requires asymptotic tests after that a full bootstrap is obtained. We therefore seek to group states having similar transition probabilities.

However the single most relevant point of our work is the attempt of estimating the true dimension of a Markov chain operating jointly in the spaces of states and times. Optimal dimensioning of the transition probability matrix is defined here by pursuing two different objectives, which are somewhat competing:

- a *similarity* objective to join a Markov chain states into homogeneous classes,
- a *multiplicity* or *non replicability* objective to reduce the occurrence that during the bootstrap generation the original series is largely replicated.

These optimal criteria are chosen to be compliant with respect to a couple of axiomatic properties, that we advance here to introduce some minimal principles in the problem of the optimal dimensioning of the transition probability matrix of a Markov chain. The decision variables of the optimization problems are two: the order of the Markov chain and the partitions of the state space. As a result, such a delicate parameters choice is obtained through predefined criteria and is therefore subtracted from the subjectivity of the researchers. Though the solution space of this kind of problems can get very large, it is discrete and bounded, and can be clearly solved (at least theoretically).

The paper is organized as follows. Section 2 introduces the settings of the problem. Section 3 advances two axioms about the problem of partitioning and specifies the criteria used here to select the optimal dimension of a Markov chain transition probability matrix. Section 4 discusses some methodological issues. In Section 5 the criteria are applied to two examples. Section 6 concludes.

2 The Model

Let us consider an evolutive observable phenomenon. We suppose that N realizations homogeneously spaced in time are available and we introduce the set of the time-ordered observations of the phenomenon, $E = (y_1, \ldots, y_N)$. More explicitly, the state of the phenomenon is y_j at time j, for each $j = 1, \ldots, N$. We define the set of indexes $I = \{1, \ldots, N\}$. Moreover, there exist an integer number $J_N \in I$, J_N states $a_1, \ldots, a_{J_N} \in E$, and J_N subsets of I, named $I_1, \ldots, I_{J_N}$, which are a

partition of I defined by

$$I_h = \{i \in I | y_i = a_h\}, \qquad h = 1, \ldots, J_N. \tag{1}$$

In short I_h is the set of the times where the process has assumed the value a_h. We define a corresponding family of subsets of E, denoted as $E_1, \ldots, E_{J_N}$, such that

$$E_h = \{y_i \in E | i \in I_h\}, \qquad h = 1, \ldots, J_N. \tag{2}$$

E_h is therefore an equivalence class of E, whose representative element is a_h. Therefore the sets E's satisfy these properties:

$$\bigcup_{h=1}^{J_N} E_h = E;$$

$$E_{h'} \cap E_{h''} = \emptyset, \qquad \forall h', h'' = 1, \ldots, J_N, \text{ with } h' \neq h''.$$

The definition of the subsets E_h's provides a clustering of the observations y's. Fixed $h = 1, \ldots, J_N$, then the frequency of the value a_h in the observed series E is the cardinality of the set E_h. There are $J_N \leq N$ different observed values a_h for the phenomenon: let $A = \{a_1, \ldots, a_{J_N}\}$ be the range of the observed series.

We now consider a Markov chain of order $k \geq 1$, denoted as $\{X(t); t \geq 0\}$, with state space A. The k-lag memory of the Markov chain implies that the transition probability matrix should account for conditioning to trajectories of length k. Therefore, we refer hereafter to a *k-path transition probability matrix*.

We deal in our paper with a couple of questions:

- Which is the optimal value of k such that $X(t)$ provides a theoretical approximation of the observed phenomenon?
- Which is the optimal time-dependent clustering of the state space, in order to estimate the k-path transition probability matrix of the Markov chain?

Letting a_z the state of the Markov chain at time t, we introduce an empirical probability measure P as follows:

$$P(a_z | \mathbf{a}_{h,\mathbf{k}}) = P(X(t) = a_z | X(t-1) = a_{h_1}, \ldots, X(t-k) = a_{h_k}), \tag{3}$$

where $a_z, a_{h_1}, ..., a_{h_k} \in A$ and $\mathbf{a}_{h,\mathbf{k}} = (a_{h_1}, \ldots, a_{h_k}) \in A^k$. The frequency in the original observed series of the transition between path $\mathbf{a}_{h,\mathbf{k}}$ and element a_z drives the computation of the empirical transition probabilities in Eq. (3). The k-path transition probability matrix of $X(t)$ is generated by Eq. (3).

We now introduce Λ as the set of the partitions of A. $\lambda \in \Lambda$ if and only if $\lambda = \{A_{\lambda,1}, \ldots, A_{\lambda,s(\lambda)}\}$, where $s(\lambda)$ is the cardinality of λ, with $1 \leq s(\lambda) \leq J_N$, and $\{A_{\lambda,w}\}_{w=1,\ldots,s(\lambda)}$ is a partition of nonempty subsets of A.

Extending Λ to a multidimensional context, we define Λ_k as the set of partitions of the k-dimensional set A^k. However, for our purposes, we restrict our analysis to a smaller set $\Lambda^k \subset \Lambda_k$. Λ^k consists of the k replications of the set Λ and describes the time dependent partitions of A. More precisely Λ^k is defined as:

$$\Lambda^k = \{\lambda = (\lambda_1, \ldots, \lambda_k) | \lambda_w \in \Lambda, \ \forall\, w = 1, \ldots, k\}.$$

If $\lambda = (\lambda_1, \ldots, \lambda_k) \in \Lambda^k$ then, for each $w = 1, \ldots, k$, we can write

$$\lambda_w = \{A_{\lambda_w,1}, \ldots, A_{\lambda_w,s(\lambda_w)}\}, \tag{4}$$

with $1 \leq s(\lambda_w) \leq J_N$, and such that $\{A_{\lambda_w,w}\}_{w=1,\ldots,s(\lambda_w)}$ is a partition of nonempty subsets of A. We refer to the probability law P introduced in Eq. (3) and define

$$P(a_z|\mathbf{A}_{\lambda,h,\mathbf{k}}) = P(X(t) = a_z | X(t-1) \in A_{\lambda_1,h_1}, \ldots, X(t-k) \in A_{\lambda_k,h_k}), \tag{5}$$

where $(h_1, \ldots, h_k) \in \prod_{w=1}^{k} \{1, \ldots, s(\lambda_w)\}$,

$$\mathbf{A}_{\lambda,h,\mathbf{k}} = A_{\lambda_1,h_1} \times \cdots \times A_{\lambda_k,h_k} \subseteq \prod_{w=1}^{k} \lambda_w, \tag{6}$$

and $a_z \in A$. The quantity in Eq. (5) is the transition probability to reach state a_z at time t after the process has been in the classes $A_{\lambda_1,h_1}, \ldots, A_{\lambda_k,h_k}$ in the previous k times.
The probabilities in Eq. (5) generate a new transition matrix, which depends on the considered partition λ. To keep the notation as simple as possible, we continue to refer to this matrix as k-path transition probability matrix.

2.1 Partition λ and k-path transition probability matrices

It is worth to explore how the k-path transition probability matrix of $X(t)$ modifies with the lag k and the particular time-dependent clustering of the state space. If we consider a partition λ, then we will associate to λ a k-path transition probability matrix of dimension $|\lambda| \times J_N$, where $|\lambda|$ indicates the cardinality of λ. Each row of this matrix corresponds to a class $\mathbf{A}_{\lambda,h,\mathbf{k}} \in \lambda$ of process paths of length k.
For a sufficiently high k, we can find a partition λ removing the randomness of transitions between paths and single states. Indeed, the longer the paths are the more the empirical observation of the phenomenon drives transition probabilities to be trivially equal to 0 or 1. More precisely, each row of the k-path transition probability matrix would consists of probabilities equal to 0, with the (possible) exception of one cell (equal to 1) corresponding to the value that is historically observed after the path (provided that such a value exists). We explain our concern with an example.

Example 1. *Consider a Markov chain of order k* $\{X(t); t \geq 0\}$*, with state space* $A = \{1, 2\}$*.*
The process is represented through different k-path transition probability matrices depending of the number of time lags. The transition probabilities are driven empirically by the observation of an evolutive phenomenon. In particular, we assume the following set of time-ordered observations of the phenomenon:

$$E = \{1, 2, 1, 1, 2, 2, 1\}.$$

We initially consider two time lags ($k = 2$*); the possible process paths* $\mathbf{a}_{h,\mathbf{k}} = (a_{h_1}, a_{h_2}) \in A^2$ *are:*

$$\mathbf{a}_{1,\mathbf{k}} = (1, 1),\ \mathbf{a}_{2,\mathbf{k}} = (1, 2),\ \mathbf{a}_{3,\mathbf{k}} = (2, 1),\ \textit{and}\ \mathbf{a}_{4,\mathbf{k}} = (2, 2).$$

We denote with $\mathcal{M}_2^{(0)}$ *the 2-path transition probability matrix of the Markov chain approximating the observed phenomenon. By visual inspection of E, we have*

$\mathcal{M}_2^{(0)} =$

	a_z	
$\mathbf{a}_{h,\mathbf{k}}$	1	2
$(1,1)$	0	1
$(1,2)$	0.5	0.5
$(2,1)$	1	0
$(2,2)$	1	0

It is easy to observe that there is not a partition $\lambda = (\lambda_1, \lambda_2)$ *such that the randomness of the transitions is completely removed.*
To get to "deterministic paths", we therefore extend k from 2 *to* 3*, i.e. we observe three time lags to form the process paths; we have* $\mathbf{a}_{h,\mathbf{k}} = (a_{h_1}, a_{h_2}, a_{h_3}) \in A^3$*,* $h = 1, ..., 8$*. We construct empirically the matrix* $\mathcal{M}_3^{(0)}$ *as*

$\mathcal{M}_3^{(0)} =$

	a_z	
$\mathbf{a}_{h,\mathbf{k}}$	1	2
$(1,1,1)$	0	0
$(1,1,2)$	0	1
$(1,2,1)$	1	0
$(1,2,2)$	1	0
$(2,1,1)$	0	1
$(2,1,2)$	0	0
$(2,2,1)$	0	0
$(2,2,2)$	0	0

It is totally evident that the partition of singletons removes the randomness of transitions to states 1 *and* 2*. Consider also partition* $\lambda = (\lambda_1, \lambda_2, \lambda_3)$*, with* $\lambda_1 = \{\{1, 2\}\}$*,* $\lambda_2 = \{\{1\}, \{2\}\}$*, and* $\lambda_3 = \{\{1, 2\}\}$*; the partition includes the following multidimensional classes:*

- $\mathbf{A}_{\lambda,1,\mathbf{k}} = \{1,2\} \times \{1\} \times \{1,2\} = \{(1,1,1),(1,1,2),(2,1,1),(2,1,2)\}$,
- $\mathbf{A}_{\lambda,2,\mathbf{k}} = \{1,2\} \times \{2\} \times \{1,2\} = \{(1,2,1),(1,2,2),(2,2,1),(2,2,2)\}$.

Such a partition removes randomness. The corresponding 3*-path transition probability matrix is*

$$\mathcal{M}_3^{(1)} = \begin{array}{c|cc} & a_z & \\ \mathbf{A}_{\lambda,h,\mathbf{k}} & 1 & 2 \\ \hline \mathbf{A}_{\lambda,1,\mathbf{k}} & 0 & 1 \\ \mathbf{A}_{\lambda,2,\mathbf{k}} & 1 & 0 \end{array}$$

Observe that, by passing from $k = 2$ *to* $k = 3$*, there are partitions identifying unique paths, that is paths for which there is a deterministic evolution. In this case, starting from an initial path, the evolution of the process continues in a deterministic way.*

Conversely, other partitions do not provide a deterministic information on the future evolution of the Markov chain. In addition to the partition of singletons associated to matrix $\mathcal{M}_2^{(0)}$*, the most evident ones are the partitions with unitary cardinality. In this case, we have the highest degree of uncertainty, independently of the considered time lag, and the corresponding* k*-path transition probability matrix always consists of a single row.*

In our example, for $k = 2$*, the all-comprehensive set partition* λ *is*

$$\lambda = \{\mathbf{A}_{\lambda,1,\mathbf{k}}\},$$

with $\mathbf{A}_{\lambda,1,\mathbf{k}} = \{\mathbf{a}_{1,\mathbf{k}}, \mathbf{a}_{2,\mathbf{k}}, \mathbf{a}_{3,\mathbf{k}}, \mathbf{a}_{4,\mathbf{k}}\}$**,** *and the corresponding* 2*-path transition probability matrix is*

$$\mathcal{M}_2^{all} = \begin{array}{c|cc} & a_z & \\ \mathbf{A}_{\lambda,h,\mathbf{k}} & 1 & 2 \\ \hline \{(1,1),(1,2),(2,1),(2,2)\} & 0.625 & 0.375 \end{array}$$

When $k = 3$*, the single set partition* λ *is*

$$\lambda = \{\mathbf{A}_{\lambda,1,\mathbf{k}}\},$$

with $\mathbf{A}_{\lambda,1,\mathbf{k}} = \{\mathbf{a}_{1,\mathbf{k}}, ..., \mathbf{a}_{8,\mathbf{k}}\}$*, and the* 3*-path transition probability matrix is*

$$\mathcal{M}_3^{all} = \begin{array}{c|cc} & a_z & \\ \mathbf{A}_{\lambda,h,\mathbf{k}} & 1 & 2 \\ \hline \{(1,1,1),...,(2,2,2)\} & 0.5 & 0.5 \end{array}$$

In the case of the all-comprehensive set partitions, each path realization has no influence on the evolution of the process. The corresponding k*-path transition probability matrices translate the empirical evidence that the process is all driven by chance.*

3 Optimal Criteria

The aim of this section is to present some optimal criteria for choosing the order k of the Markov chain and clustering A^k. As already mentioned in the introduction, our optimization problems are based on two competing guidelines: statistical similarity and multiplicity.

3.1 A brief theory about partitions as a tool to understand phenomena

There are a lot of approaches developed in the literature which have been applied to provide a partition of vectors of values observed about a given phenomenon. Just to mention an example, consider the many proposals advanced for the statistical methodology of cluster analysis. We need to fix some general principles to address the problem of partitioning in a coherent way.
Let us start by observing that all the possible partitions of a set can not be ordered in a "natural" way. There are however two clear opposite partitions, which can be somehow placed at two extremes: the partition of singletons and the all-comprehensive set partition. All other partitions can be placed somewhere in between, referring to some criteria (e.g. the number of equivalence classes generated). Partitioning a set can be seen as an exercise of separating "signal" from "noise". In general every element has unique features, making it different from all others. However, by comparing several elements, it can be possible to observe that some features are shared among some elements whereas other features among other elements. The opportunity to compare a single element to many others drives us to the problem of deciding whether its features are "close" enough to those of some groups or not. To this purpose, many would agree that they like to disregard chance or noise differences and rather prefer considering the significant ones. Such a wise stance is however tough to maintain, because in general nobody knows exactly how the differences are really generated in the elements and the boundary between "law" and noise is difficult to be placed.

Axiom 2 (double indeterminacy)**.** *A vector of values observed about a phenomenon is the result of an unknown combination of an unknown deterministic law and noise (rephrased: any observed element is the result of an unknown combination of chance and an unknown signal).*

When the features of an element are judged sufficiently close to those of a group, the element is joined to that group. This decision solves the two indeterminacies of the axiom. The common part of the features is taken as evidence of the signal carried by the element. At the same time the level of "idiosyncrasy" in the element is accepted as evidence of the noise level typical of the phenomenon analyzed.
The double indeterminacy axiom is tricky. If a researcher makes an hypothesis about the level of noise indeterminacy, then he will be able to compare observations, conclude when they can be joined, and possibly infer some law in the end. If he alternatively advances an hypothesis to tackle the law indeterminacy, he will be able to measure the noise carried by each element and from this determine

the level of noise in the phenomenon. In both cases a solution is found, although it depends on a hypothesis which could be wrong.
A researcher willing to avoid this risk has a remainder way: to hope that the observations themselves will reveal robust evidence to solve jointly the two indeterminacies. He can expect that any partition of his data will stay within two extreme solutions. If the phenomenon is completely determined by a law, every unique observation has to be taken as fully informative. The only partition compliant with such a hypothesis is of course that of letting every element to form a singleton. At the opposite side, if there is no law governing a phenomenon, no signal can be found in any element and similarity among observations will be attributed to chance. Joining all elements in a single cluster is the consequent partition that must be adopted.

Axiom 3 (noise-information duality)**.** *Given a set of vectors of values observed about a phenomenon, the partition of singletons attributes full informative value to each element (no noise or chance). The partition consisting of one all-comprehensive set treats elements as the result of noise or chance (no informative value).*

Following the previous reasoning, we can derive some conclusions about which criteria are necessary and sufficient to optimally solve the task of partitioning a set of observations about a phenomenon (unknown in the sense of the first axiom). A similarity criterion is required to join or distinguish the elements (observations) and infer the law or signal implied by a partition. A multiplicity criterion is required to measure the level of noise or chance inherent in a partition. Jointly used, these two criteria supply all the possible knowledge about the two indeterminacies of the first axiom.
The second axiom supplies two fundamental properties that both criteria should respect:

- the similarity criterion assumes its maximum value corresponding to the partition of singletons and its minimum with the all-comprehensive set partition.
- the multiplicity criterion assumes its minimum value corresponding to the partition of singletons and its maximum with the all-comprehensive set partition.

As we will see in the following sections, the criteria proposed in this work are compliant with the requirements discussed so far.
In the following we adopt distance as a similarity criterion. Of course, a problem of maximum similarity can be represented as a problem of minimum distance, so here we seek to form classes where the distance among the elements is low.

3.2 First distance indicator: Absolute difference of k-path transition probabilities

The first distance measure focuses on the absolute difference between the elements of the k-path transition probability matrix. Fixed a value for k, we can define a distance $d_{i,j}$ between two paths $\mathbf{a}_{i,\mathbf{k}}$ and $\mathbf{a}_{j,\mathbf{k}}$ as follows:

$$d_{i,j} := \sum_{z=1}^{J_N} \left| P(a_z|\mathbf{a}_{i,\mathbf{k}}) - P(a_z|\mathbf{a}_{j,\mathbf{k}}) \right|. \tag{7}$$

In order to preserve similarity, we notice that $\mathbf{a}_{i,\mathbf{k}}$ and $\mathbf{a}_{j,\mathbf{k}}$ should be grouped together when their distance is close to zero. Indeed if $d_{i,j}$ is exactly zero, we have no reason to distinguish the paths involved $\mathbf{a}_{i,\mathbf{k}}$ and $\mathbf{a}_{j,\mathbf{k}}$. By extending this argument, we stress that it is desirable that the elements composing the classes of a suitable partition are close enough to each other, at least on average. We formalize this point. Let us consider a partition $\lambda \in \Lambda^k$ such that $\lambda = (\lambda_1, \ldots, \lambda_k)$ and $\mathbf{A}_{\lambda,h,\mathbf{k}}$ as in Eq. (6). The distance in $\mathbf{A}_{\lambda,h,\mathbf{k}}$ is defined as:

$$d_{\mathbf{A}_{\lambda,h,\mathbf{k}}} := \max_{i,j:\mathbf{a}_{i,\mathbf{k}},\mathbf{a}_{j,\mathbf{k}} \in \mathbf{A}_{\lambda,h,\mathbf{k}}} d_{i,j}. \tag{8}$$

We can finally characterize the distance d_λ of partition λ with the average value of its classes distances. More precisely, we have:

$$d_\lambda := \frac{1}{\sum_{h=1}^{|\lambda|} |\mathbf{A}_{\lambda,h,\mathbf{k}}|} \cdot \sum_{h=1}^{|\lambda|} d_{\mathbf{A}_{\lambda,h,\mathbf{k}}} \cdot |\mathbf{A}_{\lambda,h,\mathbf{k}}|, \tag{9}$$

where $|\mathbf{A}_{\lambda,h,\mathbf{k}}|$ is the cardinality of partition class $\mathbf{A}_{\lambda,h,\mathbf{k}}$ and $|\lambda|$ is the cardinality of partition λ, i.e.

$$|\lambda| = \prod_{w=1}^{k} s(\lambda_w).$$

Proposition 4. $d_\lambda \in [0, 2]$.

Proof. See Appendix A. □

Remark 5. *The partition composed by the all-comprehensive set takes the maximum value of d_λ (not necessarily 2).*
The opposite case, represented by the partition of singletons, is associated (with certainty) to $d_\lambda = 0$, since any singleton has zero distance from itself.

Remark 6. *Observe that if we defined the distance indicator by interchanging the calculations of Eq. (8) and (9), we would obtain a contradiction. Indeed, define*

$$\tilde{d}_{\mathbf{A}_{\lambda,h,\mathbf{k}}} := \frac{1}{|\mathbf{A}_{\lambda,h,\mathbf{k}}| \times [|\mathbf{A}_{\lambda,h,\mathbf{k}}| - 1]} \sum_{i,j:\mathbf{a}_{i,\mathbf{k}},\mathbf{a}_{j,\mathbf{k}} \in \mathbf{A}_{\lambda,h,\mathbf{k}}} d_{i,j}$$

as the (simple) average distance of partition class $\mathbf{A}_{\lambda,h,\mathbf{k}}$. Define then

$$\tilde{d}_\lambda := \max_{\mathbf{A}_{\lambda,h,\mathbf{k}} \subseteq \lambda} d_{\mathbf{A}_{\lambda,h,\mathbf{k}}}$$

as the distance indicator of partition λ.

It is easy to show that such a defined distance indicator causes the all-comprehensive set partition to take a value strictly less than other partitions; such indicator contradicts the request of a similarity (distance) criterion to exhibit its minimum (maximum) value when all the elements are groped together (see the discussion after Axiom 3 at the end of the previous subsection).

3.3 Second distance indicator: Variance-type measure of k-path transition probabilities

The second distance indicator is constructed by taking into account the average error made within the classes of a partition. Let us consider a partition $\lambda \in \Lambda^k$ such that $\lambda = (\lambda_1, \ldots, \lambda_k)$ and $\mathbf{A}_{\lambda,h,\mathbf{k}}$ as in Eq. (6).

We then proceed by defining a variance-type measure of the multidimensional class in Eq. (6) as follows:

$$v_{\mathbf{A}_{\lambda,h,\mathbf{k}}} := \frac{1}{J_N \cdot |\mathbf{A}_{\lambda,h,\mathbf{k}}|} \cdot \sum_{z=1}^{J_N} \left\{ \sum_{i:\mathbf{a}_{i,\mathbf{k}} \in \mathbf{A}_{\lambda,h,\mathbf{k}}} \left[P(a_z|\mathbf{a}_{i,\mathbf{k}}) - P(a_z|\mathbf{A}_{\lambda,h,\mathbf{k}}) \right]^2 \right\}. \tag{10}$$

In this case, we preserve the similarity by imposing that the classes of a suitable partition have a low value of the indicator defined in Eq. (10). More generally, the entire partition should have a low value of the variance-type measure. To this end, we introduce a weighted average of variance-type measures of partition classes: given λ, we define its associated variance-type measure as the weighted average of the $v_{\mathbf{A}_{\lambda,h,\mathbf{k}}}$'s:

$$v_\lambda := \frac{1}{\sum_{h=1}^{|\lambda|} |\mathbf{A}_{\lambda,h,\mathbf{k}}|} \cdot \sum_{h=1}^{|\lambda|} v_{\mathbf{A}_{\lambda,h,\mathbf{k}}} \cdot |\mathbf{A}_{\lambda,h,\mathbf{k}}|. \tag{11}$$

Proposition 7. $v_\lambda \in [0, 0.25]$.

Proof. The proposition is stated without a rigorous proof. Appendix A shows some supporting arguments. □

Remark 8. *The all-comprehensive set partition identifies the minimum level of similarity, i.e. the maximum value of v_λ (not necessarily 0.25).*

It is easily observed that $v_\lambda = 0$ when the k-path transition probability matrix shows uniformly distributed columns within each class $\mathbf{A}_{\lambda,h,\mathbf{k}}$. The partition of singletons clearly verifies such condition.

Both similarity criteria in Eq. (9) and (11) favor those partitions joining any two states observed at time $t - w$ which have the same pattern of transition probabilities and penalize those partitions keeping these states separate. Under such partitions, the two states will form a single state at time lag w. If all states at time lag w show no difference of their transition probabilities, the partitions joining them will be preferred and time lag w will be irrelevant. (k-path) transition probabilities are

therefore crucial to allow the joint estimation of the order and of the relevant number (partition) of the states of a Markov chain.

3.4 Multiplicity indicator

The multiplicity indicators we propose are based on the size of the partition classes.
Let us define l_λ an *absolute multiplicity indicator* of the partition λ:

$$l_\lambda := \sum_{h=1}^{|\lambda|} |\mathbf{A}_{\lambda,h,\mathbf{k}}|^2. \tag{12}$$

The following result holds:

Proposition 9. *It results*

$$(J_N)^k \leq l_\lambda \leq (J_N)^{2k}. \tag{13}$$

Proof. See Appendix A. □

We can also define a *relative multiplicity indicator* m_λ, related to a partition λ, by normalizing l_λ as follows:

$$m_\lambda := \frac{l_\lambda - (J_N)^k}{(J_N)^{2k} - (J_N)^k}. \tag{14}$$

By Proposition 9 and arguments above, we have $m_{\ \lambda} \in [0,1]$, being

$$\begin{cases} m_\lambda = 0, & \text{for } |\lambda_w| = J_N, \ \ \forall\, w = 1,\dots,k; \\ m_\lambda = 1, & \text{for } |\lambda_w| = 1, \ \ \forall\, w = 1,\dots,k. \end{cases} . \tag{15}$$

3.5 Two optimization problems

We now present two optimization problems based on the similarity and multiplicity criteria developed so far. Solving them will provide a way to answer the questions addressed in this paper.
The first one is based on the distance defined in Eq. (9).

Definition 10. *Let us consider* $\gamma \in [0,1]$, $k^* \in \mathbb{N}$, *and* $\lambda^* = (\lambda_1^*,\dots,\lambda_{k^*}^*) \in \Lambda^{k^*}$.
We say that the couple (k^*,λ^*) *is d-m-optimal if and only if it is the solution of the following minimization problem:*

$$\begin{aligned} &\min_{(k,\lambda)\in\mathbb{N}\times\Lambda^k} d_\lambda \\ &\text{s.t.} \quad m_\lambda \geq \gamma. \end{aligned} \tag{16}$$

The second optimization problem involves the variance-type measure defined in Eq. (11).

Definition 11. *Let us consider* $\gamma \in [0,1]$, $k^* \in \mathbb{N}$, *and* $\lambda^* = (\lambda_1^*, \ldots, \lambda_{k^*}^*) \in \Lambda^{k^*}$.
The couple (k^*, λ^*) *is said to be v-m-optimal if and only if it the solution of the following minimization problem:*

$$\min_{(k,\lambda)\in\mathbb{N}\times\Lambda^k} v_\lambda \quad \text{s.t.} \quad m_\lambda \geq \gamma. \tag{17}$$

In both Definition 10 and 11, we have that k^* is the optimal order of a Markov chain describing the evolutive phenomenon. Moreover, λ^* provides the optimal time-dependent clustering of the state space, in order to have an approximation of the k^*-path transition probability matrix.
According to the definitions of d_λ, v_λ, and m_λ, we can briefly discuss the two optimization problems. Letting the multiplicity indicator reach its minimum ($\gamma = 0$) is equivalent to allow for the partition of singletons, which ensures the minimum distance ($d_\lambda, v_\lambda = 0$). Letting $\gamma = 1$ corresponds to forcing the maximum level of multiplicity. This boundary in our case is satisfied only by the all-comprehensive set, in which case the two distance indicators take their maximum value.

4 Methodological Issues

To perform the optimization procedures, a researcher faces several technical problems; an important computational problem is the restriction of the space of admissible solutions. In particular, we present in the following two methods/concepts that could help identifying which time lags "count" to determine the evolution of a process at time t.
We initially introduce the concept of *longest-memory* k in the following:

Definition 12. *Let us consider a* k*-dimensional partition* $\lambda = (\lambda_1, \ldots, \lambda_k)$. *The longest-memory* k *for* λ, *call it lm-*k_λ, *is a time lag such that:*

- *lm-*$k_\lambda \in \{1, ..., k\}$;
- $s(\lambda_{lm\text{-}k_\lambda}) > 1$ *(the cardinality of partition* $\lambda_{lm\text{-}k_\lambda}$ *is greater than* 1*);*
- $s(\lambda_j) = 1$ *for each* $j \in \{lm\text{-}k_\lambda + 1, k\}$.

An *lm-*k_λ represents the maximum number of time lags that can be considered in building up a partition without loosing information: indeed, the time series values are grouped all together before that time lag (third condition of the previous definition).

We discuss now some important properties of partitions and similarity indicators depending on the previous definition of *longest-memory* k. Let us consider the partitions λ and λ' with $\lambda := \lambda_1 \times \cdots \times \lambda_{lm\text{-}k_\lambda} \times \cdots \times \lambda_k$ and $\lambda' := \lambda_1 \times \cdots \times \lambda_{lm\text{-}k_\lambda}$. It is easily seen that the two partitions have

the same number of classes; in addition, the existence of *lm-k_λ* implies that the similarity indicators should yield the same value for both the partitions λ and λ'. We only give a hint of the proof based on the observation of the transition probability matrices associated to the two partitions. In particular, each matrix block associated to a partition class $\mathbf{A}_{\lambda',h,\mathbf{k}}$ of λ' will be exactly replicated J_N times to build the matrix block associated to the corresponding partition class $\mathbf{A}_{\lambda,h,\mathbf{k}}$ of λ. This is a consequence that a transition probability matrix must respect if we want that time lags greater than *lm-k_λ* do not condition the evolution of a process at time t.
We can extend the properties of partitions and similarity indicators to a generic time lag (not necessarily a *longest-memory* k). More precisely, the following theorem states immediately.

Theorem 13. *Consider a partition $\lambda = (\lambda_1, \ldots, \lambda_k)$. Define the w-penalized partition $\lambda_{-w} := (\lambda_1, \ldots, \lambda_{w-1}, \lambda_{w+1}, \ldots, \lambda_k)$, with $w \in \{1, \ldots, k\}$. Assume that*

a $s(\lambda_w) = 1$;

b given two any paths $\mathbf{a}_{i,\mathbf{k}} = (a_{i_1}, \ldots, a_{i_k})$ and $\mathbf{a}_{j,\mathbf{k}} = (a_{j_1}, \ldots, a_{j_k})$, the transition probabilities associated to these paths are equal if $a_{i_h} = a_{j_h}$ for $h = 1, ..., w-1, w+1, ..., k$.
Then

1. *$|\lambda| = |\lambda_{-w}|$ (partitions λ and λ_{-w} have the same cardinality);*
2. *$d_\lambda = d_{\lambda_{-w}}$ and $v_\lambda = v_{\lambda_{-w}}$.*

The theorem holds not only for a generic time lag w, but also for a set of generic time lags $\mathbf{w}$, with $|\mathbf{w}| > 1$.
We now introduce the important concept of *ε-active time lag.*

Definition 14. *Given $\varepsilon \in [0, 1]$ and $w \in \{1, \ldots, k\}$, a time lag w is said ε-active if and only if, for any $a_z \in A$, the following conditions are fulfilled:*

- *$|P(a_z|\mathbf{a}_{i,\mathbf{k}}) - P(a_z|\mathbf{a}_{j,\mathbf{k}})| < \varepsilon$, where $\mathbf{a}_{i,\mathbf{k}}$ can differ from $\mathbf{a}_{j,\mathbf{k}}$ in all times but $t - w$,*
- *$|P(a_z|\mathbf{a}_{i,\mathbf{k}}) - P(a_z|\mathbf{a}_{j,\mathbf{k}})| \geq \varepsilon$, where $\mathbf{a}_{i,\mathbf{k}}$ can be equal to $\mathbf{a}_{j,\mathbf{k}}$ in all times but $t - w$,*

for any couple i, j.

In other words, the observation of the process in $t - w$ brings a "key information" to determine its evolution at time t.
This definition can be extended to combinations of several *ε-active time lags* as follows:

Definition 15. *Given $\varepsilon \in [0, 1]$ and ρ indexes $w_1, \ldots, w_\rho \in \{1, ..., k\}$, the time lags $w_1, \ldots, w_\rho$ are said joint ε-active if and only if, for any $a_z \in A$, the following conditions are fulfilled:*

- *$|P(a_z|\mathbf{a}_{i,\mathbf{k}}) - P(a_z|\mathbf{a}_{j,\mathbf{k}})| < \varepsilon$, where $\mathbf{a}_{i,\mathbf{k}}$ can differ from $\mathbf{a}_{j,\mathbf{k}}$ in all times but $t - w_1, \ldots, t - w_\rho$,*

- $|P(a_z|\mathbf{a}_{i,\mathbf{k}}) - P(a_z|\mathbf{a}_{j,\mathbf{k}})| \geq \varepsilon$, *where* $\mathbf{a}_{i,\mathbf{k}}$ *can be equal to* $\mathbf{a}_{j,\mathbf{k}}$ *in all times but* $t-w_1, \ldots, t-w_\rho$,

for any couple i, j.

Remark 16. *It does not make sense to extend the search for active* ρ*-tuples whose size is greater than* $k-1$*, where* k *is the order of the Markov chain* $\{X(t); t \geq 0\}$*. Verifying that all the* k *time lags are* ε*-active is equivalent to find that none time is of particular importance over the others for the analysis at time* t *of the phenomenon described by* $X(t)$.

We now see how we can jointly use the definitions of *longest-memory* k and *joint* ε*-active time lags*. Consider the time lags which are less than or equal to the *longest-memory* k, i.e. the set $\{1, ...,lm\text{-}k_\lambda\}$. If we know which time lags in $\{1, ...,lm\text{-}k_\lambda\}$ are *joint* ε*-active*, we can neglect all the others and avoid to evaluate the corresponding partitions.
To conclude this section, we detail the conditions for selecting the non-dominated (efficient) solutions.

Definition 17. *Let us consider a couple of partitions* $\lambda_1, \lambda_2 \in \Lambda^k$*; we say that* λ_1 *is* $\natural$*-m-non-dominated by* λ_2 *if*

$$\begin{cases} \natural_{\lambda_1} \geq \natural_{\lambda_2} \\ m_{\lambda_1} \geq m_{\lambda_2} \end{cases} \quad \text{or} \quad \begin{cases} \natural_{\lambda_1} \leq \natural_{\lambda_2} \\ m_{\lambda_1} \leq m_{\lambda_2} \end{cases}, \tag{18}$$

where $\natural = d, v$ *and at least one inequality in each system is strict.*

According to the previous definition, dominated partitions will be discarded in our analysis; basically, the rejected partitions show no lower similarity (d_λ, or v_λ) and no higher multiplicity (m_λ), with at least a strict inequality holding.
The procedure to solve the two optimization problems in Definitions 10 and 11 can be synthesized in the following points:

1. initially the researcher orders the admissible solutions by increasing values of their similarity indicator (v or d);
2. starting from the solution with the lowest value of similarity, he scans for the next solution with a higher similarity and a higher value of multiplicity (m) and discards the intermediate solutions (dominated in the sense of Eq. (18));
3. step 2. is repeated until the worst value of similarity is reached.

The partitions remaining after step 3. constitute the *set of efficient solutions.*

5 Some Numerical Examples

To test the effectiveness of our method we devise the following experiment:

1. we fix a k-path transition probability matrix, describing some relevant features of the corresponding process (e.g. spikes, switching regimes, etc.);

2. we apply our optimization procedure as explained at the end of the previous section;

3. we analyze the solution space to check if it describes correctly the features initially imposed in the k-path transition probability matrix.

5.1 k-path transition probability matrix design

We develop an application of our methodology for two simple processes:

I. a process on three different states, with range $A = \{1, 2, 3\}$. We also assume that the search domain of the time lags is limited to a small set of values $\{1, ..., \bar{k}\}$ and we fix $\bar{k} = 5$. The set of partitions associated to A is Λ_A and $|\Lambda_A| = 5$; consequently, the set of 5-dimensional partitions[1] on A is Λ_A^5 and $|\Lambda_A^5| = |\Lambda_A|^5 = 5^5 = 3,125$. $\mathcal{A}$ denotes the 5-path transition probability matrix of the Markov chain $\{X(t); t \geq 0\}$ approximating the observed phenomenon; $\mathcal{A}$ has dimensions $3^5 \times 3 = 243 \times 3$;

II. a process on five different states, whose range is $B = \{1, 2, 3, 4, 5\}$. Moreover, we consider $\bar{k} = 3$. The set of partitions on B is Λ_B, with $|\Lambda_B| = 52$, and the cardinality of Λ_B^3, the set of 3-dimensional partitions on B, is $|\Lambda_B^3| = |\Lambda_B|^3 = 52^3 = 140,608$. In this case, the 3-path transition probability matrix is denoted with $\mathcal{B}$ and has dimensions $5^3 \times 5 = 125 \times 5$.

Remark 18. *The fast growing behavior of the Bell numbers with the cardinality of the range of an observed phenomenon increases dramatically the computational complexity of our optimization problems*[2]*. This fact explains why, in our didactic applications* ***I.*** *and* ***II.****, the cardinality of the ranges A and B is small.*

We have allocated the numbers of the k-path transition probability matrices based on the concept of *joint ε-active time lags* (see Definition 15). In matrix $\mathcal{A}$ time lags 3 and 2 are *joint* 0.23-*active* (singularly considered, $t-5$, $t-4$, $t-3$, $t-2$, and $t-1$ are ε-*active*, with ε between 0.83 and 0.84). Further extensions to triples and quadruples do not generate a relevant reduction of the value of ε, and so they are not suitable for our analysis. In matrix $\mathcal{B}$ time lags 2 and 1 are *joint* 0.10-*active* (singularly considered, $t-3$, $t-2$, and $t-1$ result 0.47-*active*, 0.34-*active*, and 0.45-*active*

[1] The number of the partitions that can be extracted from a discrete set with cardinality N is the Bell number $B(N)$. The Bell numbers satisfy the following recursion formula:

$$B(N) = \sum_{k=0}^{N-1} \binom{N-1}{k} B(k),$$

with $B(0) = 1$.

[2] As an example, we have that $B(10) = 115,975$ and $B(14) = 190,899,322$.

respectively). Tables 1 and 2 in Appendix B show the k-path transition probability matrices $\mathcal{A}$ and $\mathcal{B}$. The definition of the *joint ε-active time lags* implies that the two original k-path transition probability matrices can be reduced to a lower row dimension, keeping alive only the states at critical times; Tables 3 and 4 show the reduced k-path transition probability matrices and their aim is to help have a fast view of the relevant process "mechanics".

Insert Table 3 here

Insert Table 4 here

In Table 3, which refers to case **I.**, each row represents a sequence of process states at time lags 2 and 3. Since there are other three "non critical" times ($t-5$, $t-4$, and $t-1$) and the series can take 3 values, we have 27 (i.e. 3^3) alternative sequences for each row. The transition probabilities in Table 3 are obtained averaging the 27 corresponding rows of the original matrix $\mathcal{A}$. The horizontal lines drawn in Table 3 help visualize possible classes of partitions built on time lags 2 and 3; the lines group the process paths with similar average transition probabilities by jointly considering only the time lags 3 and 2.

For case **II.**, in Table 4 each row shows the average transition probabilities from the 5 possible sequences at time lags 2 and 1 (irrespective of time $t-3$). If we focus our attention on time lag 2, we can see that partition $\lambda_2 = \{\{1,2\},\{3,4,5\}\}$ appears immediately as a plausible one to minimize the distance for this time lag. A partition of time series paths reflecting such similarity for time lag 2 should have reasonable chances to be a solution to our optimization problem. Notice that the horizontal lines drawn in Table 4 group the process paths with similar average transition probabilities by jointly considering both time lags 2 and 1.

5.2 Optimization procedure

To solve the two optimization problems (16) and (17), we have calculated the similarity and multiplicity indicators for every partition associated to $\mathcal{A}$ and $\mathcal{B}$, the k-path transition probability matrices of order 243×3 and 125×5 introduced in the previous subsection: we have calculated d_λ, v_λ, and m_λ (see Eq. (9), (11), and (14)) for the $3,125$ partitions of $\mathcal{A}$ and the $140,608$ partitions of $\mathcal{B}$.

The two cases analyzed here have required approximately 3 and 70 minutes of computing time, on an Intel Pentium M-processor at 2.8 Ghz.

We have then applied the 3 step-algorithm presented at the end of Section 4 to compute the set of efficient solutions.

5.3 Analysis of results

In this section we analyze the results of the optimization procedures applied to cases **I.** and **II.**.

Tables 5-6 and Fig. 1-2 show the set of efficient solutions for case **I.** respectively for problems (16)

and (17). Tables 7-8 and Fig. 3-4 show the set of efficient solutions for case **II.** of the same two problems.

Insert Table 5 here

Insert Table 6 here

Insert Table 7 here

Insert Table 8 here

All the tables report a column named "Partition times" showing for which time lag the corresponding (unidimensional) partition λ has cardinality greater than 1. It is possible to observe that for case **I.** the time lags 3 and 2 have been included in the large majority of efficient solutions (2 actually in all of them). Partition times of case **II.** show a situation almost equal to that of case **I.**, with the role of time lag 3 played here by time lag 1 (time lag 2 is included in all efficient solutions). We observe that in both cases the efficient solutions are largely composed with partitions under the expected "active times", for both optimization problems. This is the major result of our application.

The figures let us better appreciate which solution, among the efficient frontier, is relatively preferable. Some points in the plane are "better positioned" than others along the efficient frontier: in particular it can be advisable choosing those solutions showing a better trade-off between multiplicity and similarity.

We comment at first on Fig. 2; starting from the origin, we observe that the first eight solutions (which include partitions on non-active times) can be easily discarded in favor of solution 9, which includes only partitions on the *joint* 0.23-*active time lags* 3 and 2, as it offers a large improvement in the multiplicity at a very low cost in the similarity (measured by v_λ). Partitions based on time lag 2 only reach higher multiplicity at the expense of a remarkable jump in similarity.

Fig. 1 shows an efficient frontier almost similar to that of Fig. 2, with partitions obtained on the active times expected for case **I.** in the better position.

Insert Figure 1 here

Insert Figure 2 here

Moving on to case **II.**, we consider initially Fig. 4. The first solutions of Fig. 4 convey some increase in the multiplicity at low cost of the similarity (measured by v_λ). Partitions based on time lags 2 and 1 largely dominate the central part of the graph. Contrary to case **I.**, partitions based only on time lag 2 offer improvements in the multiplicity criteria at a low cost of the similarity; indeed, the third efficient solution counted from the right corresponds to the time lag 2 partition $\{\{1,2\},\{3,4,5\}\}$, which has been indicated as a plausible one at the end of Subsection 5.1.

Contrary to Fig. 4, Fig. 3 shows that partitions on both active time lags 2 and 1 score a better

trade-off d_λ-m_λ than partitions based only on time lag 2. The third point counted from the right corresponds to the time lag 2 partition $\{\{1,2\},\{3,4,5\}\}$, however the trade-off of this solution does not drive for its choice.

Insert Figure 3 here

Insert Figure 4 here

6 Conclusions

This paper proposes an optimization method to the problem of estimating the dimension of the transition probability matrix of a Markov chain. Several aspects were to be addressed to this purpose. We first advanced a couple of axioms to establish the necessary properties required by the criteria which are adopted to search the solutions of a partitioning problem.

Based on such axioms, we then formalized our problem as a search of the partition of the states and the order of a Markov chain which minimize the distance inside each class, for a given level of noise. Two alternative distance indicators were proposed, which use the transition probabilities. The noise measure is based on the cardinality of the partitions of the states.

Several benefits originate from this approach.

Since the solution of this optimal problem is completely data driven, the optimal partition of the states and order of a Markov chain emerge without any arbitrary choice on the side of the researcher. The bootstrap methods based on the explicit estimation of the transition probabilities can therefore adopt an objective choice about the optimal dimension of the transition probability matrix.

The introduction of a small set of necessary axiomatic properties will help the development of other comparable criteria, both optimal or sub-optimal. Indeed it is important noticing that the full search over the set of admissible solutions is not computationally feasible if the states of the Markov chain reach a number of some tens. So the introduction of heuristic methods to restrict the search among the admissible solutions is welcome. The axiomatic properties however will help avoiding the development of inappropriate trials. Such a risk is not vague. Consider, for example, our first distance indicator (i.e. the absolute difference of transition probabilities between the paths comprised in a partition class): it can be easily shown (see Remark 6 at page) that if we compute, for each partition, the maximum of the *average class distances* instead of the weighted average of the *maximum class distances* then the all-comprehensive set partition fails to take the lowest value of similarity and Axiom 3 is violated.

Finally the availability of the two full search methods advanced here will ease to compare the effectiveness of new alternatives.

Appendix A - Proofs of Propositions 4, 7, and 9

Proof of ***Proposition 4****.* The proposition is rather obvious, so we just sketch the proof. Let $\mathbf{a}_{i,\mathbf{k}}$ and $\mathbf{a}_{j,\mathbf{k}}$ any two paths of a transition probability matrix. Since $d_{i,j}$ in Eq. (7) is a distance, its minimum value is 0 which indeed occurs iff $\mathbf{a}_{i,\mathbf{k}} = \mathbf{a}_{j,\mathbf{k}}$. Letting $\mathbf{a}_z$ and $\mathbf{a}_y$ two vectors partitioning A, the maximum value for $d_{i,j}$ occurs iff $P(\mathbf{a}_z|\mathbf{a}_{i,\mathbf{k}}) = 1$ and $P(\mathbf{a}_y|\mathbf{a}_{j,\mathbf{k}}) = 1$. In that case, $d_{i,j} = 2$. The conclusion follows by Eq. (8) and Eq. (9) iff every class of partition λ is formed by at least one couple of paths, say i and j, whose distance, $d_{i,j}$, is equal to 2. □

Proof of ***Proposition 7****.* We show that any transition probability matrix $\mathcal{M}$ is associated to the maximum value of v_λ iff it is organized in the following way:

$$\mathcal{M} = \begin{array}{r|cc} & a_1 & a_2 \\ \hline \mathbf{a}_{1,\mathbf{k}} & 0 & 1 \\ ... & 0 & 1 \\ \mathbf{a}_{M,\mathbf{k}} & 0 & 1 \\ \mathbf{a}_{M+1,\mathbf{k}} & 1 & 0 \\ ... & 1 & 0 \\ \mathbf{a}_{2M,\mathbf{k}} & 1 & 0 \end{array}$$

and is partitioned by the all-comprehensive set, i.e. $\lambda = \{\mathbf{A}_{\lambda,1,\mathbf{k}}\} = \{\{\mathbf{a}_{1,\mathbf{k}}, ..., \mathbf{a}_{M,\mathbf{k}}, \mathbf{a}_{M+1,\mathbf{k}}, ..., \mathbf{a}_{2M,\mathbf{k}}\}\}$. Observe in particular that $\mathcal{M}$ is of type $2M \times 2$, with $M \in \mathbb{N}$, and shows transition probabilities equally distributed between states a_1 and a_2. By Eq. (10), the variance of the all-comprehensive set partition associated to $\mathcal{M}$ is equal to the variance of its unique class:

$$v_\lambda = v_{\mathbf{A}_{\lambda,1,\mathbf{k}}} = \frac{1}{2} \cdot (0.25 + 0.25) = 0.25.$$

There are three ways to modify matrix $\mathcal{M}$:

1. modifying the 1's of a column into smaller values (or the 0's into greater values),
2. distributing the 1's and 0's of a column unequally,
3. introducing more columns,
4. introducing a finer partition.

In all four cases, it is easy to see that v_λ decreases.

1. By means of obvious arguments, the maximum variance of each column of $\mathcal{M}$ is achieved when the values consist of 0's and 1's (extreme distribution).

2. Depending on the proportion of 0's and 1's, the variance is

$$Var(M) = \frac{1}{N}\left[M \times \left(1 - \frac{M}{N}\right)^2 + (N - M) \times \left(0 - \frac{M}{N}\right)^2\right],$$

where we consider M ones and $N - M$ zeros.

Calculating $\frac{dVar(M)}{dM}$ and putting it equal to 0, we get

$$\frac{dVar(M)}{dM} = \frac{N - 2M}{N^2},$$

$$\frac{N - 2M}{N^2} = 0 \Rightarrow M = \frac{N}{2}.$$

Observe that $\frac{d^2Var(M)}{dM^2} = -\frac{2}{N^2} < 0$. In conclusion, $Var(\frac{N}{2}) = 0.25$.

3. Suppose now we expand our matrix $\mathcal{M}$ and consider more than two columns (states); it is easily observed that, for the added columns to show extreme distributions of 1's and 0's (point 1.), we should allocate some 1's to these new columns, thus all the columns would no longer have equally distributed numbers (point 2.); as a result, the variance of the all-comprehensive set partition will decrease.

4. It is easy to see that each partition λ takes a value of v_λ less than or equal to the value of the corresponding all-comprehensive set partition. This fact is easily explained by observing that Eq. (11) is a weighted average of the variances inside the classes of partition λ and does not consider the variance between these classes.

□

Proof of ***Proposition 9****.* The absolute multiplicity indicator l_λ attains its minimum value when, for each $w = 1, \ldots, k$, it results $|\lambda_w| = J_N$. In this case, the unidimensional partitions λ_w are composed by singletons, i.e. $\lambda_w = \{\{a_1\}, \ldots, \{a_{J_N}\}\}$, and have maximum cardinality. The multidimensional partition is $\lambda = \prod_{w=1}^{k}\{\{a_1\}, \ldots, \{a_{J_N}\}\}$ and consists of $(J_N)^k$ multidimensional singletons; Eq. (12) becomes

$$l_\lambda = \sum_{h=1}^{(J_N)^k} |\mathbf{A}_{\lambda,h,\mathbf{k}}|^2 = \sum_{h=1}^{(J_N)^k} 1^2 = (J_N)^k.$$

Conversely, l_λ attains its maximum value when, for each $w = 1, \ldots, k$, it results $|\lambda_w| = 1$, i.e. $\lambda_w = \{\{a_1, \ldots, a_{J_N}\}\} = \{A\}$. The multidimensional partition is $\lambda = \prod_{w=1}^{k}\{A\}$ and consists of the multidimensional set A^k; in this case, $|\lambda| = 1$ and we have

$$l_\lambda = \sum_{h=1}^{1} |\mathbf{A}_{\lambda,h,\mathbf{k}}|^2 = |A^k|^2 = [(J_N)^k]^2.$$

□

7 Appendix B - k-Path Transition Probability Matrices $\mathcal{A}$ and $\mathcal{B}$

We present here the k-path transition probability matrices $\mathcal{A}$ and $\mathcal{B}$ used for our application.

Insert Table 1 here

Insert Table 2 here

Tables

Table 1: Transition probability matrix $\mathcal{A}$. Each row shows the transition probabilities from states at times $t-5$, $t-4$, $t-3$, $t-2$ and $t-1$ (y_{t-5}, y_{t-4}, y_{t-3}, y_{t-2} and y_{t-1}) to states at time t (y_t). States are labeled 1, 2, 3, 4 and 5.

					y_t								y_t		
y_{t-5}	y_{t-4}	y_{t-3}	y_{t-2}	y_{t-1}	1	2	3	y_{t-5}	y_{t-4}	y_{t-3}	y_{t-2}	y_{t-1}	1	2	3
1	1	1	1	1	0.075	0.212	0.713	1	1	2	1	1	0.107	0.797	0.096
1	1	1	1	2	0.131	0.134	0.734	1	1	2	1	2	0.070	0.788	0.142
1	1	1	1	3	0.099	0.214	0.687	1	1	2	1	3	0.117	0.778	0.105
1	1	1	2	1	0.779	0.141	0.081	1	1	2	2	1	0.773	0.094	0.132
1	1	1	2	2	0.795	0.116	0.090	1	1	2	2	2	0.793	0.082	0.125
1	1	1	2	3	0.782	0.111	0.107	1	1	2	2	3	0.769	0.112	0.119
1	1	1	3	1	0.829	0.126	0.045	1	1	2	3	1	0.735	0.105	0.159
1	1	1	3	2	0.781	0.137	0.081	1	1	2	3	2	0.772	0.122	0.106
1	1	1	3	3	0.765	0.147	0.087	1	1	2	3	3	0.799	0.072	0.129
1	1	3	1	1	0.139	0.754	0.107	1	2	1	1	1	0.131	0.134	0.734
1	1	3	1	2	0.119	0.786	0.095	1	2	1	1	2	0.086	0.207	0.707
1	1	3	1	3	0.094	0.856	0.050	1	2	1	1	3	0.199	0.101	0.701
1	1	3	2	1	0.811	0.129	0.060	1	2	1	2	1	0.776	0.124	0.100
1	1	3	2	2	0.778	0.089	0.133	1	2	1	2	2	0.768	0.104	0.129
1	1	3	2	3	0.759	0.151	0.090	1	2	1	2	3	0.799	0.126	0.075
1	1	3	3	1	0.143	0.747	0.110	1	2	1	3	1	0.769	0.084	0.147
1	1	3	3	2	0.100	0.782	0.118	1	2	1	3	2	0.814	0.109	0.078
1	1	3	3	3	0.097	0.866	0.037	1	2	1	3	3	0.770	0.070	0.160
1	2	2	1	1	0.098	0.755	0.147	1	2	3	1	1	0.099	0.790	0.111
1	2	2	1	2	0.104	0.771	0.125	1	2	3	1	2	0.100	0.822	0.078
1	2	2	1	3	0.104	0.782	0.114	1	2	3	1	3	0.070	0.782	0.148
1	2	2	2	1	0.819	0.069	0.112	1	2	3	2	1	0.778	0.115	0.107
1	2	2	2	2	0.765	0.149	0.086	1	2	3	2	2	0.758	0.150	0.091
1	2	2	2	3	0.745	0.113	0.142	1	2	3	2	3	0.771	0.062	0.167
1	2	2	3	1	0.787	0.114	0.099	1	2	3	3	1	0.141	0.775	0.084
1	2	2	3	2	0.838	0.066	0.096	1	2	3	3	2	0.143	0.765	0.093
1	2	2	3	3	0.765	0.113	0.122	1	2	3	3	3	0.135	0.782	0.084
1	3	1	1	1	0.060	0.226	0.714	1	3	2	1	1	0.092	0.782	0.127
1	3	1	1	2	0.240	0.080	0.680	1	3	2	1	2	0.074	0.772	0.155

Table 1: (continued)

					y_t								y_t		
y_{t-5}	y_{t-4}	y_{t-3}	y_{t-2}	y_{t-1}	1	2	3	y_{t-5}	y_{t-4}	y_{t-3}	y_{t-2}	y_{t-1}	1	2	3
1	3	1	1	3	0.096	0.204	0.700	1	3	2	1	3	0.130	0.785	0.086
1	3	1	2	1	0.777	0.096	0.127	1	3	2	2	1	0.765	0.109	0.127
1	3	1	2	2	0.745	0.159	0.096	1	3	2	2	2	0.753	0.112	0.135
1	3	1	2	3	0.775	0.143	0.082	1	3	2	2	3	0.757	0.148	0.095
1	3	1	3	1	0.810	0.138	0.052	1	3	2	3	1	0.750	0.093	0.157
1	3	1	3	2	0.785	0.083	0.132	1	3	2	3	2	0.772	0.105	0.123
1	3	1	3	3	0.798	0.147	0.055	1	3	2	3	3	0.824	0.080	0.095
1	3	3	1	1	0.116	0.782	0.102	2	1	1	1	1	0.254	0.073	0.673
1	3	3	1	2	0.098	0.779	0.123	2	1	1	1	2	0.072	0.242	0.687
1	3	3	1	3	0.123	0.800	0.077	2	1	1	1	3	0.270	0.065	0.665
1	3	3	2	1	0.763	0.138	0.099	2	1	1	2	1	0.798	0.139	0.064
1	3	3	2	2	0.788	0.104	0.108	2	1	1	2	2	0.780	0.111	0.110
1	3	3	2	3	0.807	0.135	0.058	2	1	1	2	3	0.751	0.151	0.098
1	3	3	3	1	0.128	0.779	0.093	2	1	1	3	1	0.774	0.058	0.168
1	3	3	3	2	0.133	0.772	0.095	2	1	1	3	2	0.872	0.082	0.046
1	3	3	3	3	0.142	0.798	0.061	2	1	1	3	3	0.771	0.063	0.166
2	1	2	1	1	0.099	0.783	0.118	2	1	3	1	1	0.096	0.797	0.107
2	1	2	1	2	0.154	0.781	0.066	2	1	3	1	2	0.132	0.802	0.066
2	1	2	1	3	0.093	0.801	0.107	2	1	3	1	3	0.093	0.773	0.134
2	1	2	2	1	0.763	0.095	0.142	2	1	3	2	1	0.768	0.127	0.104
2	1	2	2	2	0.787	0.093	0.120	2	1	3	2	2	0.815	0.114	0.071
2	1	2	2	3	0.809	0.080	0.111	2	1	3	2	3	0.760	0.081	0.158
2	1	2	3	1	0.778	0.150	0.073	2	1	3	3	1	0.075	0.766	0.159
2	1	2	3	2	0.813	0.066	0.121	2	1	3	3	2	0.097	0.831	0.072
2	1	2	3	3	0.787	0.107	0.107	2	1	3	3	3	0.140	0.771	0.089
2	2	1	1	1	0.039	0.236	0.725	2	2	2	1	1	0.142	0.761	0.096
2	2	1	1	2	0.222	0.089	0.689	2	2	2	1	2	0.037	0.870	0.094
2	2	1	1	3	0.059	0.235	0.706	2	2	2	1	3	0.129	0.771	0.100
2	2	1	2	1	0.769	0.113	0.118	2	2	2	2	1	0.767	0.134	0.099
2	2	1	2	2	0.747	0.140	0.113	2	2	2	2	2	0.824	0.063	0.113
2	2	1	2	3	0.765	0.083	0.152	2	2	2	2	3	0.753	0.150	0.097
2	2	1	3	1	0.735	0.152	0.113	2	2	2	3	1	0.751	0.088	0.160

Table 1: (continued)

					y_t								y_t		
y_{t-5}	y_{t-4}	y_{t-3}	y_{t-2}	y_{t-1}	1	2	3	y_{t-5}	y_{t-4}	y_{t-3}	y_{t-2}	y_{t-1}	1	2	3
2	2	1	3	2	0.762	0.092	0.147	2	2	2	3	2	0.764	0.133	0.103
2	2	1	3	3	0.852	0.105	0.042	2	2	2	3	3	0.738	0.099	0.163
2	2	3	1	1	0.129	0.774	0.097	2	3	1	1	1	0.150	0.125	0.725
2	2	3	1	2	0.084	0.775	0.141	2	3	1	1	2	0.090	0.211	0.698
2	2	3	1	3	0.107	0.813	0.080	2	3	1	1	3	0.178	0.111	0.711
2	2	3	2	1	0.813	0.110	0.077	2	3	1	2	1	0.828	0.107	0.065
2	2	3	2	2	0.762	0.098	0.140	2	3	1	2	2	0.776	0.057	0.166
2	2	3	2	3	0.851	0.094	0.055	2	3	1	2	3	0.781	0.129	0.090
2	2	3	3	1	0.140	0.800	0.060	2	3	1	3	1	0.768	0.080	0.152
2	2	3	3	2	0.084	0.766	0.150	2	3	1	3	2	0.802	0.130	0.068
2	2	3	3	3	0.120	0.801	0.079	2	3	1	3	3	0.782	0.076	0.142
2	3	2	1	1	0.113	0.731	0.157	2	3	3	1	1	0.075	0.785	0.141
2	3	2	1	2	0.159	0.765	0.076	2	3	3	1	2	0.114	0.795	0.091
2	3	2	1	3	0.109	0.755	0.135	2	3	3	1	3	0.097	0.774	0.129
2	3	2	2	1	0.752	0.098	0.150	2	3	3	2	1	0.778	0.059	0.163
2	3	2	2	2	0.766	0.129	0.106	2	3	3	2	2	0.783	0.143	0.074
2	3	2	2	3	0.787	0.087	0.126	2	3	3	2	3	0.770	0.061	0.169
2	3	2	3	1	0.776	0.102	0.121	2	3	3	3	1	0.088	0.781	0.131
2	3	2	3	2	0.786	0.098	0.116	2	3	3	3	2	0.148	0.760	0.092
2	3	2	3	3	0.768	0.156	0.075	2	3	3	3	3	0.113	0.785	0.101
3	1	1	1	1	0.107	0.224	0.669	3	1	2	1	1	0.147	0.752	0.101
3	1	1	1	2	0.192	0.104	0.704	3	1	2	1	2	0.095	0.798	0.107
3	1	1	1	3	0.078	0.212	0.710	3	1	2	1	3	0.143	0.762	0.095
3	1	1	2	1	0.773	0.111	0.116	3	1	2	2	1	0.772	0.107	0.121
3	1	1	2	2	0.792	0.146	0.062	3	1	2	2	2	0.838	0.041	0.121
3	1	1	2	3	0.776	0.133	0.092	3	1	2	2	3	0.777	0.153	0.070
3	1	1	3	1	0.862	0.093	0.046	3	1	2	3	1	0.822	0.073	0.105
3	1	1	3	2	0.788	0.090	0.122	3	1	2	3	2	0.762	0.129	0.109
3	1	1	3	3	0.786	0.113	0.101	3	1	2	3	3	0.800	0.070	0.130
3	1	3	1	1	0.130	0.772	0.097	3	2	1	1	1	0.230	0.085	0.685
3	1	3	1	2	0.100	0.777	0.123	3	2	1	1	2	0.062	0.231	0.706
3	1	3	1	3	0.128	0.786	0.086	3	2	1	1	3	0.190	0.105	0.705

Table 1: (continued)

					y_t								y_t		
y_{t-5}	y_{t-4}	y_{t-3}	y_{t-2}	y_{t-1}	1	2	3	y_{t-5}	y_{t-4}	y_{t-3}	y_{t-2}	y_{t-1}	1	2	3
3	1	3	2	1	0.848	0.091	0.061	3	2	1	2	1	0.789	0.132	0.079
3	1	3	2	2	0.791	0.095	0.114	3	2	1	2	2	0.775	0.054	0.171
3	1	3	2	3	0.795	0.148	0.056	3	2	1	2	3	0.810	0.139	0.051
3	1	3	3	1	0.143	0.799	0.058	3	2	1	3	1	0.756	0.093	0.150
3	1	3	3	2	0.084	0.773	0.143	3	2	1	3	2	0.791	0.122	0.087
3	1	3	3	3	0.143	0.749	0.108	3	2	1	3	3	0.778	0.102	0.120
3	2	2	1	1	0.088	0.807	0.105	3	2	3	1	1	0.092	0.769	0.139
3	2	2	1	2	0.141	0.768	0.091	3	2	3	1	2	0.093	0.830	0.077
3	2	2	1	3	0.082	0.781	0.138	3	2	3	1	3	0.068	0.767	0.165
3	2	2	2	1	0.766	0.097	0.137	3	2	3	2	1	0.793	0.099	0.109
3	2	2	2	2	0.771	0.153	0.076	3	2	3	2	2	0.739	0.165	0.097
3	2	2	2	3	0.793	0.071	0.136	3	2	3	2	3	0.776	0.065	0.159
3	2	2	3	1	0.777	0.126	0.096	3	2	3	3	1	0.105	0.777	0.118
3	2	2	3	2	0.856	0.037	0.108	3	2	3	3	2	0.092	0.861	0.047
3	2	2	3	3	0.768	0.157	0.075	3	2	3	3	3	0.068	0.769	0.163
3	3	1	1	1	0.089	0.236	0.674	3	3	2	1	1	0.144	0.776	0.080
3	3	1	1	2	0.202	0.099	0.699	3	3	2	1	2	0.101	0.797	0.102
3	3	1	1	3	0.102	0.228	0.670	3	3	2	1	3	0.092	0.785	0.123
3	3	1	2	1	0.774	0.112	0.114	3	3	2	2	1	0.774	0.106	0.120
3	3	1	2	2	0.814	0.132	0.054	3	3	2	2	2	0.787	0.080	0.133
3	3	1	2	3	0.765	0.070	0.164	3	3	2	2	3	0.769	0.142	0.089
3	3	1	3	1	0.705	0.173	0.122	3	3	2	3	1	0.825	0.068	0.107
3	3	1	3	2	0.786	0.101	0.113	3	3	2	3	2	0.790	0.127	0.082
3	3	1	3	3	0.855	0.087	0.059	3	3	2	3	3	0.801	0.077	0.122
3	3	3	1	1	0.136	0.794	0.071								
3	3	3	1	2	0.106	0.788	0.106								
3	3	3	1	3	0.104	0.840	0.056								
3	3	3	2	1	0.824	0.126	0.050								
3	3	3	2	2	0.761	0.105	0.135								
3	3	3	2	3	0.786	0.130	0.084								
3	3	3	3	1	0.120	0.799	0.082								
3	3	3	3	2	0.123	0.765	0.112								

Table 1: (continued)

					y_t								y_t		
y_{t-5}	y_{t-4}	y_{t-3}	y_{t-2}	y_{t-1}	1	2	3	y_{t-5}	y_{t-4}	y_{t-3}	y_{t-2}	y_{t-1}	1	2	3
3	3	3	3	3	0.098	0.834	0.068								

Table 2: Transition probability matrix $\mathcal{B}$. Each row shows the transition probabilities from states at times $t-3$, $t-2$ and $t-1$ (y_{t-3}, y_{t-2} and y_{t-1}) to states at time t (y_t). States are labeled 1, 2 and 3.

			y_t								y_t				
y_{t-3}	y_{t-2}	y_{t-1}	1	2	3	4	5	y_{t-3}	y_{t-2}	y_{t-1}	1	2	3	4	5
1	1	1	0.400	0.471	0.129	0.000	0.000	1	2	1	0.440	0.438	0.122	0.000	0.000
1	1	2	0.450	0.450	0.100	0.000	0.000	1	2	2	0.350	0.500	0.120	0.030	0.000
1	1	3	0.500	0.401	0.099	0.000	0.000	1	2	3	0.500	0.405	0.095	0.000	0.000
1	1	4	0.400	0.404	0.096	0.100	0.000	1	2	4	0.550	0.304	0.126	0.020	0.000
1	1	5	0.320	0.460	0.200	0.000	0.020	1	2	5	0.280	0.550	0.130	0.000	0.040
1	3	1	0.150	0.457	0.323	0.000	0.070	1	4	1	0.120	0.504	0.276	0.000	0.100
1	3	2	0.220	0.452	0.328	0.000	0.000	1	4	2	0.220	0.505	0.275	0.000	0.000
1	3	3	0.100	0.502	0.328	0.070	0.000	1	4	3	0.100	0.554	0.276	0.070	0.000
1	3	4	0.150	0.402	0.328	0.100	0.020	1	4	4	0.150	0.453	0.277	0.100	0.020
1	3	5	0.150	0.401	0.369	0.050	0.030	1	4	5	0.150	0.452	0.318	0.050	0.030
1	5	1	0.100	0.357	0.323	0.000	0.220	2	1	1	0.420	0.454	0.126	0.000	0.000
1	5	2	0.210	0.358	0.322	0.000	0.110	2	1	2	0.450	0.450	0.100	0.000	0.000
1	5	3	0.200	0.402	0.328	0.070	0.000	2	1	3	0.500	0.406	0.094	0.000	0.000
1	5	4	0.250	0.305	0.325	0.100	0.020	2	1	4	0.400	0.405	0.095	0.100	0.000
1	5	5	0.250	0.300	0.370	0.050	0.030	2	1	5	0.350	0.460	0.150	0.000	0.040
2	2	1	0.430	0.448	0.122	0.000	0.000	2	3	1	0.130	0.452	0.328	0.000	0.090
2	2	2	0.350	0.507	0.133	0.010	0.000	2	3	2	0.220	0.457	0.323	0.000	0.000
2	2	3	0.550	0.401	0.049	0.000	0.000	2	3	3	0.100	0.502	0.328	0.070	0.000
2	2	4	0.530	0.245	0.125	0.100	0.000	2	3	4	0.150	0.402	0.328	0.100	0.020
2	2	5	0.250	0.560	0.140	0.000	0.050	2	3	5	0.150	0.405	0.365	0.050	0.030
2	4	1	0.130	0.504	0.276	0.000	0.090	2	5	1	0.100	0.351	0.329	0.000	0.220
2	4	2	0.220	0.504	0.276	0.000	0.000	2	5	2	0.220	0.355	0.325	0.000	0.100
2	4	3	0.100	0.552	0.278	0.070	0.000	2	5	3	0.200	0.407	0.323	0.070	0.000

Table 2: (continued)

			y_t								y_t				
y_{t-3}	y_{t-2}	y_{t-1}	1	2	3	4	5	y_{t-3}	y_{t-2}	y_{t-1}	1	2	3	4	5
2	4	4	0.150	0.450	0.280	0.100	0.020	2	5	4	0.250	0.305	0.325	0.100	0.020
2	4	5	0.150	0.455	0.315	0.050	0.030	2	5	5	0.250	0.301	0.369	0.050	0.030
3	1	1	0.420	0.446	0.124	0.010	0.000	3	2	1	0.410	0.441	0.119	0.000	0.030
3	1	2	0.450	0.450	0.100	0.000	0.000	3	2	2	0.350	0.500	0.130	0.020	0.000
3	1	3	0.500	0.401	0.099	0.000	0.000	3	2	3	0.500	0.351	0.099	0.050	0.000
3	1	4	0.400	0.403	0.097	0.100	0.000	3	2	4	0.540	0.254	0.146	0.060	0.000
3	1	5	0.340	0.420	0.160	0.060	0.020	3	2	5	0.270	0.530	0.150	0.000	0.050
3	3	1	0.120	0.450	0.330	0.000	0.100	3	4	1	0.110	0.507	0.273	0.000	0.110
3	3	2	0.220	0.456	0.324	0.000	0.000	3	4	2	0.220	0.503	0.277	0.000	0.000
3	3	3	0.100	0.502	0.328	0.070	0.000	3	4	3	0.100	0.557	0.273	0.070	0.000
3	3	4	0.150	0.401	0.329	0.100	0.020	3	4	4	0.150	0.453	0.277	0.100	0.020
3	3	5	0.150	0.400	0.370	0.050	0.030	3	4	5	0.150	0.451	0.319	0.050	0.030
3	5	1	0.100	0.353	0.327	0.000	0.220	4	1	1	0.420	0.458	0.122	0.000	0.000
3	5	2	0.230	0.352	0.328	0.000	0.090	4	1	2	0.450	0.456	0.094	0.000	0.000
3	5	3	0.200	0.401	0.329	0.070	0.000	4	1	3	0.500	0.401	0.099	0.000	0.000
3	5	4	0.250	0.306	0.324	0.100	0.020	4	1	4	0.400	0.405	0.095	0.100	0.000
3	5	5	0.250	0.300	0.370	0.050	0.030	4	1	5	0.330	0.520	0.140	0.000	0.010
4	2	1	0.420	0.457	0.123	0.000	0.000	4	3	1	0.140	0.457	0.323	0.000	0.080
4	2	2	0.350	0.507	0.103	0.040	0.000	4	3	2	0.220	0.455	0.325	0.000	0.000
4	2	3	0.500	0.403	0.097	0.000	0.000	4	3	3	0.100	0.502	0.328	0.070	0.000
4	2	4	0.560	0.317	0.093	0.030	0.000	4	3	4	0.150	0.406	0.324	0.100	0.020
4	2	5	0.230	0.580	0.110	0.000	0.080	4	3	5	0.150	0.405	0.365	0.050	0.030
4	4	1	0.120	0.502	0.278	0.000	0.100	4	5	1	0.100	0.353	0.327	0.000	0.220
4	4	2	0.220	0.507	0.273	0.000	0.000	4	5	2	0.220	0.353	0.327	0.000	0.100
4	4	3	0.100	0.556	0.274	0.070	0.000	4	5	3	0.200	0.401	0.329	0.070	0.000
4	4	4	0.150	0.455	0.275	0.100	0.020	4	5	4	0.250	0.304	0.326	0.100	0.020
4	4	5	0.150	0.456	0.314	0.050	0.030	4	5	5	0.250	0.304	0.366	0.050	0.030
5	1	1	0.420	0.451	0.129	0.000	0.000	5	2	1	0.420	0.450	0.130	0.000	0.000
5	1	2	0.450	0.454	0.096	0.000	0.000	5	2	2	0.350	0.506	0.094	0.050	0.000
5	1	3	0.500	0.407	0.093	0.000	0.000	5	2	3	0.500	0.400	0.100	0.000	0.000
5	1	4	0.400	0.403	0.097	0.100	0.000	5	2	4	0.570	0.310	0.060	0.060	0.000
5	1	5	0.320	0.520	0.130	0.000	0.030	5	2	5	0.280	0.540	0.130	0.000	0.050

Table 2: (continued)

			y_t								y_t				
y_{t-3}	y_{t-2}	y_{t-1}	1	2	3	4	5	y_{t-3}	y_{t-2}	y_{t-1}	1	2	3	4	5
5	3	1	0.150	0.458	0.322	0.000	0.070	5	4	1	0.120	0.502	0.278	0.000	0.100
5	3	2	0.220	0.457	0.323	0.000	0.000	5	4	2	0.220	0.501	0.279	0.000	0.000
5	3	3	0.100	0.503	0.327	0.070	0.000	5	4	3	0.100	0.551	0.279	0.070	0.000
5	3	4	0.150	0.404	0.326	0.100	0.020	5	4	4	0.150	0.456	0.274	0.100	0.020
5	3	5	0.150	0.401	0.369	0.050	0.030	5	4	5	0.150	0.456	0.314	0.050	0.030
5	5	1	0.100	0.358	0.322	0.000	0.220								
5	5	2	0.210	0.355	0.325	0.000	0.110								
5	5	3	0.200	0.403	0.327	0.070	0.000								
5	5	4	0.250	0.300	0.330	0.100	0.020								
5	5	5	0.250	0.303	0.367	0.050	0.030								

Table 3: Reduced 5-path transition probability matrix $\mathcal{A}$.

					y_t		
y_{t-5}	y_{t-4}	y_{t-3}	y_{t-2}	y_{t-1}	1	2	3
-	-	1	2	-	0.780	0.118	0.102
-	-	1	3	-	0.791	0.106	0.104
-	-	2	2	-	0.778	0.106	0.116
-	-	2	3	-	0.785	0.101	0.113
-	-	3	2	-	0.786	0.111	0.103
-	-	1	1	-	0.137	0.164	0.699
-	-	2	1	-	0.110	0.780	0.111
-	-	3	1	-	0.105	0.791	0.104
-	-	3	3	-	0.116	0.787	0.097

This table refers to case I. (matrix $\mathcal{A}$).

Each row represents a sequence of process states

at active times t-3 and t-2.

Since there are other three "non critical" times (t-5, t-4, and t-1)

and the series can take 3 values,

we have 27 (i.e. 3^3) alternative sequences for each row.

The transition probabilities are obtained averaging

the 27 corresponding rows of the original matrix $\mathcal{A}$.

Horizontal lines help visualize possible clusters of rows.

Table 4: Reduced 3-path transition probability matrix $\mathcal{B}$.

			y_t				
y_{t-3}	y_{t-2}	y_{t-1}	1	2	3	4	5
-	5	1	0.100	0.354	0.326	0.000	0.220
-	3	3	0.100	0.502	0.328	0.070	0.000
-	4	3	0.100	0.554	0.276	0.070	0.000
-	4	1	0.120	0.504	0.276	0.000	0.100
-	3	1	0.138	0.455	0.325	0.000	0.082
-	3	4	0.150	0.403	0.327	0.100	0.020
-	3	5	0.150	0.403	0.367	0.050	0.030
-	4	4	0.150	0.453	0.277	0.100	0.020
-	4	5	0.150	0.454	0.316	0.050	0.030
-	5	3	0.200	0.403	0.327	0.070	0.000
-	5	2	0.218	0.355	0.325	0.000	0.102
-	3	2	0.220	0.455	0.325	0.000	0.000
-	4	2	0.220	0.504	0.276	0.000	0.000
-	5	5	0.250	0.302	0.368	0.050	0.030
-	5	4	0.250	0.304	0.326	0.100	0.020
-	2	5	0.262	0.552	0.132	0.000	0.054
-	1	5	0.332	0.476	0.156	0.012	0.024
-	2	2	0.350	0.504	0.116	0.030	0.000
-	1	4	0.400	0.404	0.096	0.100	0.000
-	1	1	0.416	0.456	0.126	0.002	0.000
-	2	1	0.424	0.447	0.123	0.000	0.006
-	1	2	0.450	0.452	0.098	0.000	0.000
-	1	3	0.500	0.403	0.097	0.000	0.000
-	2	3	0.510	0.392	0.088	0.010	0.000
-	2	4	0.550	0.286	0.110	0.054	0.000

This table refers to case II. (matrix $\mathcal{B}$).

Each row represents a sequence of process states

at active times t-2 and t-1.

Since there is another "non critical" time (t-3)

and the series can take 5 values,

we have 5 alternative sequences for each row.

The transition probabilities are obtained averaging

the 5 corresponding rows of the original matrix $\mathcal{B}$.

Horizontal lines help visualize possible clusters of rows.

Table 5: Non dominated partitions of 5-path transition probability matrix $\mathcal{A}$ according to similarity indicator d_λ and multiplicity indicator m_λ.

		Efficient solutions $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$					Partition
m_λ	d_λ	λ_1	λ_2	λ_3	λ_4	λ_5	times
0	0	{{1},{2},{3}}	{{1},{2},{3}}	{{1},{2},{3}}	{{1},{2},{3}}	{{1},{2},{3}}	5,4,3,2,1
0.019946897	0.04906	{{1},{2},{3}}	{{1},{2},{3}}	{{1},{2},{3}}	{{1,3},{2}}	{{1},{2},{3}}	5,4,3,2,1
0.045698229	0.08856	{{1},{2},{3}}	{{1},{2},{3}}	{{1},{2},{3}}	{{1,3},{2}}	{{1,3},{2}}	5,4,3,2,1
0.078943057	0.12332	{{1,3},{2}}	{{1},{2},{3}}	{{1},{2},{3}}	{{1,3},{2}}	{{1,3},{2}}	5,4,3,2,1
0.084729177	0.18522	{{1},{2},{3}}	{{1},{2},{3}}	{{1},{2},{3}}	{{1},{2,3}}	{{1,2,3}}	4,3,2,1
0.129331794	0.20846	{{1,3},{2}}	{{1},{2},{3}}	{{1},{2},{3}}	{{1},{2,3}}	{{1,2,3}}	4,3,2,1
0.137094688	0.22067	{{1},{2},{3}}	{{1},{2},{3}}	{{1},{2},{3}}	{{1,2,3}}	{{1,2,3}}	3,2,1
0.196935378	0.23826	{{1,3},{2}}	{{1},{2},{3}}	{{1},{2},{3}}	{{1,2,3}}	{{1,2,3}}	3,2,1
0.287635104	0.27774	{{1,2,3}}	{{1},{2},{3}}	{{1},{2},{3}}	{{1,2,3}}	{{1,2,3}}	3,2
0.391282219	0.56537	{{1,2,3}}	{{1},{2},{3}}	{{1,2},{3}}	{{1,2,3}}	{{1,2,3}}	3,2
0.525090069	0.879	{{1,2,3}}	{{1,3},{2}}	{{1,2},{3}}	{{1,2,3}}	{{1,2,3}}	3,2
0.548378753	1.17143	{{1,2,3}}	{{1},{2},{3}}	{{1,2,3}}	{{1,2,3}}	{{1,2,3}}	2
0.727900822	1.19585	{{1,2,3}}	{{1,3},{2}}	{{1,2,3}}	{{1,2,3}}	{{1,2,3}}	2
1	1.67032	{{1,2,3}}	{{1,2,3}}	{{1,2,3}}	{{1,2,3}}	{{1,2,3}}	-

Column "m_λ" lists the values of the multiplicity indicator (see Eq. (14)).

Column "d_λ" lists the values of the similarity indicator "Absolute difference of 5-path transition probabilities" (see Eq. (9)).

The 5 columns labeled "Efficient solutions" show explicitly the partition sets of the time series values - 1, 2, and 3 - for each of the $\bar{k} = 5$ time lags.

Column "Partition times" lists the time lags whose partitions of states are different from $\lambda = \{\{1, 2, 3\}\}$.

Table 6: Non dominated partitions of 5-path transition probability matrix $\mathcal{A}$ according to similarity indicator v_λ and multiplicity indicator m_λ.

		Efficient solutions $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$					Partition
m_λ	v_λ	λ_1	λ_2	λ_3	λ_4	λ_5	times
0	0	{{1},{2},{3}}	{{1},{2},{3}}	{{1},{2},{3}}	{{1},{2},{3}}	{{1},{2},{3}}	5,4,3,2,1
0.019946897	0.00018	{{1},{2},{3}}	{{1},{2},{3}}	{{1},{2},{3}}	{{1,3},{2}}	{{1},{2},{3}}	5,4,3,2,1
0.045698229	0.0003	{{1,3},{2}}	{{1},{2},{3}}	{{1},{2},{3}}	{{1},{2},{3}}	{{1,3},{2}}	5,4,3,2,1
0.078943057	0.00038	{{1,3},{2}}	{{1},{2},{3}}	{{1},{2},{3}}	{{1,3},{2}}	{{1,3},{2}}	5,4,3,2,1
0.084729177	0.0009	{{1,2,3}}	{{1},{2},{3}}	{{1},{2},{3}}	{{1},{2},{3}}	{{1,3},{2}}	5,4,3,2
0.129331794	0.00094	{{1,2,3}}	{{1},{2},{3}}	{{1},{2},{3}}	{{1,3},{2}}	{{1,3},{2}}	5,4,3,2
0.137094688	0.00102	{{1,2,3}}	{{1},{2},{3}}	{{1},{2},{3}}	{{1,2,3}}	{{1},{2},{3}}	5,3,2
0.196935378	0.00103	{{1},{2,3}}	{{1},{2},{3}}	{{1},{2},{3}}	{{1,2,3}}	{{1,2,3}}	3,2,1
0.287635104	0.00106	{{1,2,3}}	{{1},{2},{3}}	{{1},{2},{3}}	{{1,2,3}}	{{1,2,3}}	3,2
0.391282219	0.01451	{{1,2,3}}	{{1},{2},{3}}	{{1,2},{3}}	{{1,2,3}}	{{1,2,3}}	3,2
0.525090069	0.02631	{{1,2,3}}	{{1},{2,3}}	{{1},{2,3}}	{{1,2,3}}	{{1,2,3}}	3,2
0.548378753	0.04196	{{1,2,3}}	{{1},{2},{3}}	{{1,2,3}}	{{1,2,3}}	{{1,2,3}}	2
0.727900822	0.04727	{{1,2,3}}	{{1},{2,3}}	{{1,2,3}}	{{1,2,3}}	{{1,2,3}}	2
1	0.08246	{{1,2,3}}	{{1,2,3}}	{{1,2,3}}	{{1,2,3}}	{{1,2,3}}	-

Column "m_λ" lists the values of the multiplicity indicator (see Eq. (14)).

Column "v_λ" lists the values of the similarity indicator "Variance-type measure of 5-path transition probabilities" (see Eq. (11)).

The 5 columns labeled "Efficient solutions" show explicitly the partition sets of the time series values - 1, 2, and 3 - for each of the $\bar{k} = 5$ time lags.

Column "Partition times" lists the time lags whose partitions of states are different from $\lambda = \{\{1, 2, 3\}\}$.

Table 7: Non dominated partitions of 3-path transition probability matrix $\mathcal{B}$ according to similarity indicator d_λ and multiplicity indicator m_λ.

		Efficient solutions $\lambda = (\lambda_1, \lambda_2, \lambda_3)$			Partition
m_λ	d_λ	λ_1	λ_2	λ_3	times
0	0	{{1},{2},{3},{4},{5}}	{{1},{2},{3},{4},{5}}	{{1},{2},{3},{4},{5}}	3,2,1
0.017997037	0.00708	{{1},{2},{3},{4},{5}}	{{1},{2},{3},{4},{5}}	{{1},{2},{3},{4,5}}	3,2,1
0.047467934	0.01766	{{1},{2},{3},{4},{5}}	{{1},{2},{3},{4},{5}}	{{1,4,5},{2},{3}}	3,2,1
0.060160226	0.02965	{{1},{2},{3},{4},{5}}	{{1},{2},{3},{4},{5}}	{{1,4,5},{2,3}}	3,2,1
0.082895945	0.03208	{{1},{2},{3},{4},{5}}	{{1},{2},{3},{4},{5}}	{{1,2,4,5},{3}}	3,2,1
0.121417162	0.05284	{{1},{2},{3},{4},{5}}	{{1},{2},{3},{4},{5}}	{1,2,3,4,5}	2,1
0.161659761	0.09464	{{1},{2},{3},{4},{5}}	{{1},{2},{3,4},{5}}	{1,2,3,4,5}	2,1
0.196457095	0.15868	{{1},{2},{3},{4},{5}}	{{1,2},{3,4},{5}}	{1,2,3,4,5}	2,1
0.209275446	0.16166	{{1,2},{3},{4},{5}}	{{1},{2},{3,4},{5}}	{1,2,3,4,5}	2,1
0.227558688	0.21209	{{1,2,3},{4},{5}}	{{1},{2},{3},{4},{5}}	{1,2,3,4,5}	2,1
0.250448207	0.22161	{{1},{2},{3,4},{5}}	{{1,2},{3,4},{5}}	{1,2,3,4,5}	2,1
0.287248108	0.25329	{{1,2,3},{4},{5}}	{{1},{2},{3,4},{5}}	{1,2,3,4,5}	2,1
0.29713373	0.28486	{{1,2},{3,4},{5}}	{{1,2},{3,4},{5}}	{1,2,3,4,5}	2,1
0.306778129	0.30733	{{1,2,3,4},{5}}	{{1},{2},{3},{4},{5}}	{1,2,3,4,5}	2,1
0.338860896	0.31512	{{1},{2,3,4},{5}}	{{1,2},{3,4},{5}}	{1,2,3,4,5}	2,1
0.380981815	0.35215	{{1,2,3,4},{5}}	{{1},{2},{3,4},{5}}	{1,2,3,4,5}	2,1
0.384992013	0.36943	{{1,2,3},{4},{5}}	{{1},{2},{3,4,5}}	{1,2,3,4,5}	2,1
0.445144929	0.42507	{{1,2,3,4},{5}}	{{1,2},{3,4},{5}}	{1,2,3,4,5}	2,1
0.502493433	0.45761	{{1,2,3,4},{5}}	{{1},{2},{3,4,5}}	{1,2,3,4,5}	2,1
0.554825096	0.53052	{{1,2,3,4},{5}}	{{1,2},{3,4,5}}	{1,2,3,4,5}	2,1
0.560708581	0.59216	{1,2,3,4,5}	{{1,2},{3,4},{5}}	{1,2,3,4,5}	2
0.630253858	0.59453	{1,2,3,4,5}	{{1},{2},{3,4,5}}	{1,2,3,4,5}	2
0.693715321	0.6887	{1,2,3,4,5}	{{1,2},{3,4,5}}	{1,2,3,4,5}	2
0.807393913	0.89526	{1,2,3,4,5}	{{1,3,4,5},{2}}	{1,2,3,4,5}	2
1	1.06	{1,2,3,4,5}	{1,2,3,4,5}	{1,2,3,4,5}	-

Column "m_λ" lists the values of the multiplicity indicator (see Eq. (14)).

Column "d_λ" lists the values of the similarity indicator "Absolute difference of 3-path transition probabilities" (see Eq. (9)). The 3 columns labeled "Efficient solutions" show explicitly the partition sets of the time series values - 1, 2, 3, 4, and 5 - for each of the $\bar{k} = 3$ time lags.

Column "Partition times" lists the time lags whose partitions of states are different from $\lambda = \{\{1, 2, 3, 4, 5\}\}$.

Table 8: Non dominated partitions of 3-path transition probability matrix $\mathcal{B}$ according to similarity indicator v_λ and multiplicity indicator m_λ.

		Efficient solutions $\lambda = (\lambda_1, \lambda_2, \lambda_3)$			Partition
m_λ	v_λ	λ_1	λ_2	λ_3	times
0	0	{{1},{2},{3},{4},{5}}	{{1},{2},{3},{4},{5}}	{{1},{2},{3},{4},{5}}	3,2,1
0.017997037	0.00001	{{1},{2},{3},{4},{5}}	{{1},{2},{3},{4},{5}}	{{1},{2},{3},{4,5}}	3,2,1
0.047467934	0.00002	{{1},{2},{3},{4},{5}}	{{1},{2},{3},{4},{5}}	{{1,4,5},{2},{3}}	3,2,1
0.060160226	0.00004	{{1},{2},{3},{4},{5}}	{{1},{2},{3},{4},{5}}	{{1,4,5},{2,3}}	3,2,1
0.082895945	0.00005	{{1},{2},{3},{4},{5}}	{{1},{2},{3},{4},{5}}	{{1,2,4,5},{3}}	3,2,1
0.121417162	0.00008	{{1},{2},{3},{4},{5}}	{{1},{2},{3},{4},{5}}	{1,2,3,4,5}	2,1
0.161659761	0.00018	{{1},{2},{3},{4},{5}}	{{1},{2},{3,4},{5}}	{1,2,3,4,5}	2,1
0.196457095	0.00044	{{1},{2},{3},{4},{5}}	{{1,2},{3,4},{5}}	{1,2,3,4,5}	2,1
0.209275446	0.00047	{{1},{2,5},{3},{4}}	{{1},{2},{3,4},{5}}	{1,2,3,4,5}	2,1
0.250448207	0.00068	{{1},{2},{3,4},{5}}	{{1,2},{3,4},{5}}	{1,2,3,4,5}	2,1
0.287248108	0.00093	{{1,2,5},{3},{4}}	{{1},{2},{3,4},{5}}	{1,2,3,4,5}	2,1
0.29713373	0.00094	{{1,2},{3,4},{5}}	{{1,2},{3,4},{5}}	{1,2,3,4,5}	2,1
0.338860896	0.00114	{{1,2,5},{3},{4}}	{{1,2},{3,4},{5}}	{1,2,3,4,5}	2,1
0.376937773	0.00138	{{1,2,5},{3,4}}	{{1,2},{3,4},{5}}	{1,2,3,4,5}	2,1
0.380981815	0.00155	{{1,2,3,5},{4}}	{{1},{2},{3,4},{5}}	{1,2,3,4,5}	2,1
0.384992013	0.00158	{{1,2,5},{3},{4}}	{{1},{2},{3,4,5}}	{1,2,3,4,5}	2,1
0.445144929	0.00165	{{1,2,3,4},{5}}	{{1,2},{3,4},{5}}	{1,2,3,4,5}	2,1
0.472850297	0.00202	{{1,2,5},{3,4}}	{{1,2},{3,4,5}}	{1,2,3,4,5}	2,1
0.482899376	0.00213	{1,2,3,4,5}	{{1,2},{3},{4},{5}}	{1,2,3,4,5}	2
0.502493433	0.00215	{{1,2,3,5},{4}}	{{1},{2},{3,4,5}}	{1,2,3,4,5}	2,1
0.560708581	0.00224	{1,2,3,4,5}	{{1,2},{3,4},{5}}	{1,2,3,4,5}	2
0.630253858	0.0028	{1,2,3,4,5}	{{1},{2},{3,4,5}}	{1,2,3,4,5}	2
0.693715321	0.00281	{1,2,3,4,5}	{{1,2},{3,4,5}}	{1,2,3,4,5}	2
0.807393913	0.006	{1,2,3,4,5}	{{1},{2,3,4,5}}	{1,2,3,4,5}	2
1	0.00792	{1,2,3,4,5}	{1,2,3,4,5}	{1,2,3,4,5}	-

Column "m_λ" lists the values of the multiplicity indicator (see Eq. (14)).

Column "v_λ" lists the values of the similarity indicator "Variance-type measure of 3-path transition probabilities" (see Eq. (11)). The 3 columns labeled "Efficient solutions" show explicitly the partition sets of the time series values - 1, 2, 3, 4, and 5 - for each of the $\bar{k} = 3$ time lags.

Column "Partition times" lists the time lags whose partitions of states are different from $\lambda = \{\{1, 2, 3, 4, 5\}\}$.

Figures Captions

Figure 1: The points represent efficient solutions $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$ of case **I.** through the values of the multiplicity indicator m_λ (horizontal axis) and the similarity indicator d_λ (vertical axis). Each point also shows the partition times, i.e. for which time lags the corresponding partitions of states are different from $\lambda = \{\{1, 2, 3\}\}$.

Figure 2: The points represent efficient solutions $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5)$ of case **I.** through the values of the multiplicity indicator m_λ (horizontal axis) and the similarity indicator v_λ (vertical axis). Each point also shows the partition times, i.e. for which time lags the corresponding partitions of states are different from $\lambda = \{\{1, 2, 3\}\}$.

Figure 3: The points represent efficient solutions $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ of case **II.** through the values of the multiplicity indicator m_λ (horizontal axis) and the similarity indicator d_λ (vertical axis). Each point also shows the partition times, i.e. for which time lags the corresponding partitions of states are different from $\lambda = \{\{1, 2, 3, 4, 5\}\}$.

Figure 4: The points represent efficient solutions $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ of case **II.** through the values of the multiplicity indicator m_λ (horizontal axis) and the similarity indicator v_λ (vertical axis). Each point also shows the partition times, i.e. for which time lags the corresponding partitions of states are different from $\lambda = \{\{1, 2, 3, 4, 5\}\}$.

Figures

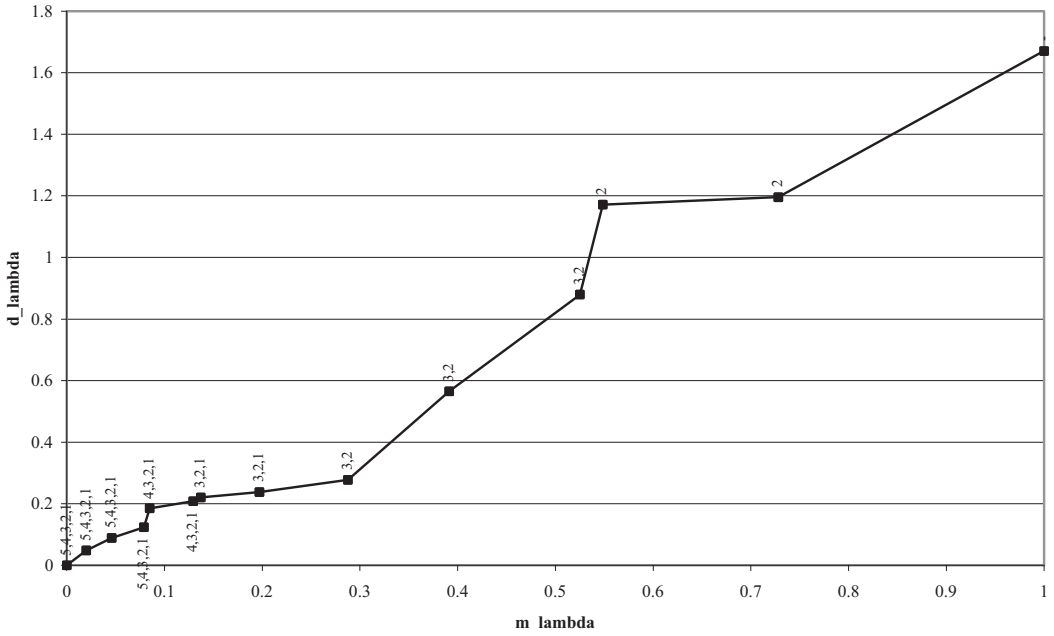

Figure 1:

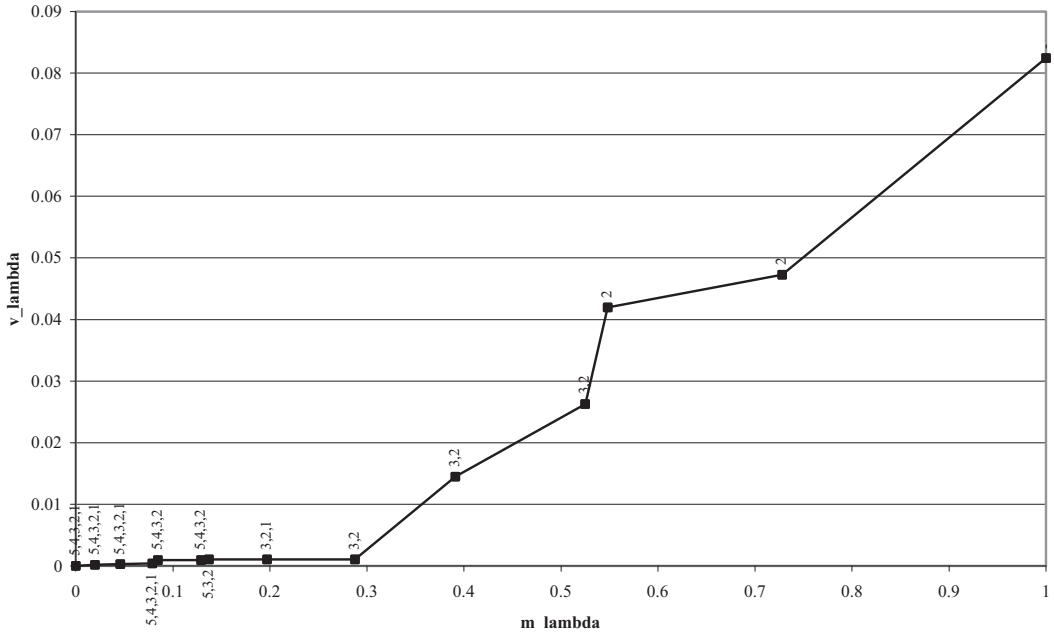

Figure 2:

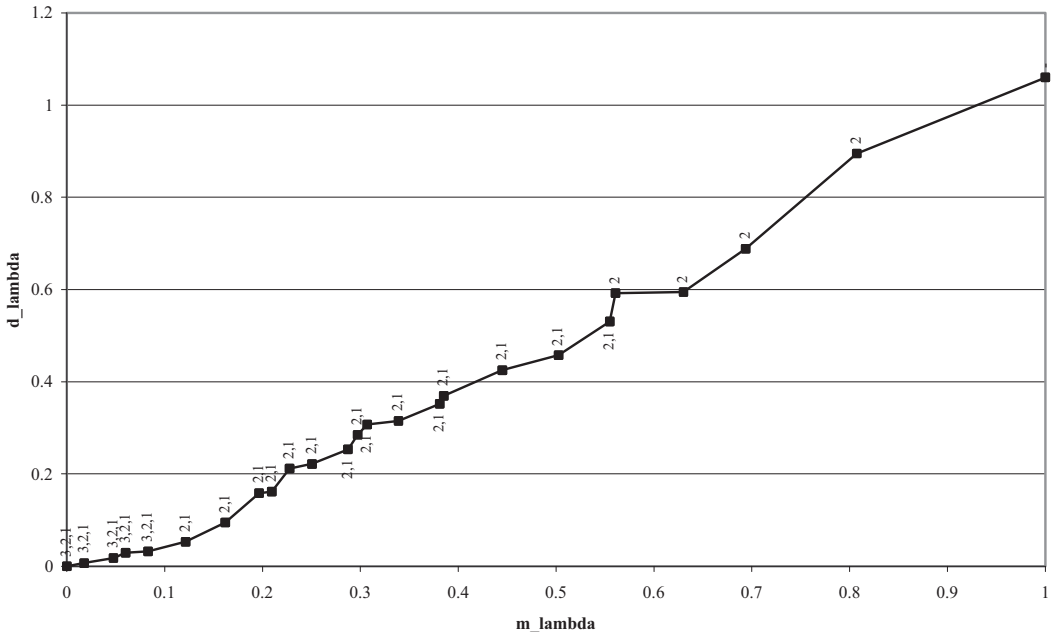

Figure 3:

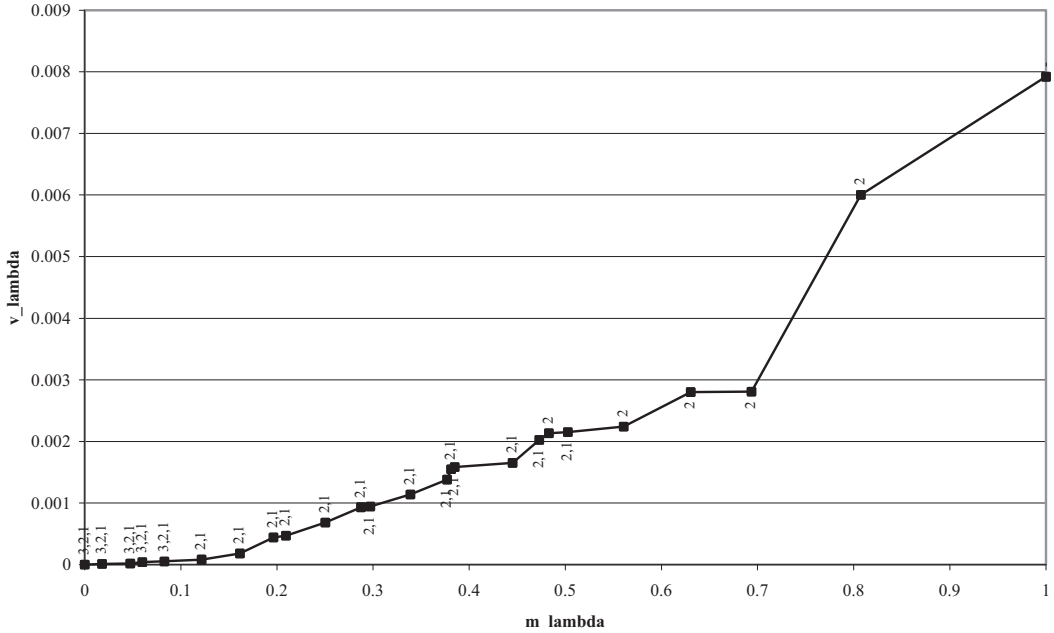

Figure 4:

References

[1] Anatolyev, S. and Vasnev, A.: 2002, Markov chain approximation in bootstrapping autoregressions, *Economics Bulletin*, **3** (19), 1–8.

[2] Athreya, K. B. and Fuh, C. D.: 1992, Bootstrapping Markov chains: Countable case, *Journal of Statistical Planning and Inference*, **33** (3), 311–331.

[3] Basawa, I. V., Green, T. A., McCormick, W. P. and Taylor, R. L.: 1990, Asymptotic bootstrap validity for finite Markov chains, *Communications in Statistics - Theory and Methods*, **19** (4), 1493–1510.

[4] Bertail, P. and Clémençon, S.: 2006, Regenerative block bootstrap for Markov chains, *Bernoulli*, **12** (4), 689–712.

[5] Bertail, P. and Clémençon, S.: 2007, Second-order properties ofregeneration-based bootstrap for atomic Markov chains, *Test*, **16**, 109–122.

[6] Bühlmann, P.: 1997, Sieve bootstrap for time series, *Bernoulli*, **3** (2), 123–148.

[7] Bühlmann, P.: 2002, Sieve bootstrap with variable-length Markov chains for stationary categorical time series, *Journal of the American Statistical Association*, **97** (458), 443–456.

[8] Bühlmann, P. and Künsch, H. R.: 1999, Block length selection in the bootstrap for time series, *Computational Statistics & Data Analysis*, **31**, 295–310.

[9] Canestrelli, E. and Giove, S.: 1999, Scenarios identification for financial modelling, *in* E. Canestrelli (ed.), *Current Topics in Quantitative Finance*, Physica-Verlag, Heidelberg, Germany, pp. 25–36.

[10] Carlstein, E.: 1986, The use of subseries values for estimating the variance of a general statistic from a stationary sequence, *The Annals of Statistics*, **14** (3), 1171–1179.

[11] Datta, S. and McCormick, W. P.: 1992, Bootstrap for a finite state Markov chain based on i.i.d. resampling, *in* R. LePage and L. Billard (eds), *Exploring the Limits of Bootstrap*, John Wiley & Sons, New York, NY, USA, pp. 77–97.

[12] Datta, S. and McCormick, W. P.: 1993, Regeneration-based bootstrap for Markov chains, *The Canadian Journal of Statistics*, **21** (2), 181–193.

[13] Efron, B.: 1979, Bootstrap methods: Another look at the jackknife, *The Annals of Statistics*, **7** (1), 1–26.

[14] Hall, P.: 1985, Resampling a coverage pattern, *Stochastic Processes and their Applications*, **20** (2), 231–246.

[15] Hall, P., Horowitz, J. L. and Jing, B.-Y.: 1995, On blocking rules for the bootstrap with dependent data, *Biometrika*, **82** (3), 561–574.

[16] Horowitz, J. L.: 2003, Bootstrap methods for Markov processes, *Econometrica*, **71** (4), 1049–1082.

[17] Kulperger, R. J. and Prakasa Rao, B. L. S.: 1989, Bootstrapping a finite state Markov chain, *Sankhya, The Indian Journal of Statistics*, **51** (2), 178–191.

[18] Künsch, H. R.: 1989, The jackknife and the bootstrap for general stationary observations, *The Annals of Statistics*, **17** (3), 1217–1241.

[19] Lahiri, S. N.: 2003, *Resampling Methods for Dependent Data*, Springer-Verlag, New York, NY, USA.

[20] Lahiri, S. N., Furukawa, K. and Lee, Y.-D.: 2007, A nonparametric plug-in rule for selecting optimal block lengths for block bootstrap methods, *Statistical Methodology*, **4**, 292–321.

[21] Liu, R. Y. and Singh, K.: 1992, Moving blocks jackknife and bootstrap capture weak dependence, *in* R. LePage and L. Billard (eds), *Exploring the Limits of Bootstrap*, John Wiley & Sons, New York, NY, USA, pp. 225–248.

[22] Paparoditis, E. and Politis, D. N.: 2001a, Tapered block bootstrap, *Biometrika*, **88** (4), 1105–1119.

[23] Paparoditis, E. and Politis, D. N.: 2001b, A Markovian local resampling scheme for nonparametric estimators in time series analysis, *Econometric Theory*, **17**, 540–566.

[24] Paparoditis, E. and Politis, D. N.: 2002a, The tapered block bootstrap for general statistics from stationary sequences, *The Econometrics Journal*, **5** (1), 131–148.

[25] Paparoditis, E. and Politis, D. N.: 2002b, The local bootstrap for Markov processes, *Journal of Statistical Planning and Inference*, **108**, 301–328.

[26] Politis, D. N. and Romano, J. P.: 1992, A general resampling scheme for triangular arrays of α-mixing random variables with application to the problem of spectral density estimation, *The Annals of Statistics*, **20** (4), 1985–2007.

[27] Politis, D. N. and Romano, J. P.: 1994, The stationary bootstrap, *Journal of the American Statistical Association*, **89** (428), 1303–1313.

[28] Politis, D. N. and White, H.: 2004, Automatic block-length selection for the dependent bootstrap, *Econometric Reviews*, **23** (1), 53–70.

[29] Rajarshi, M. B.: 1990, Bootstrap in Markov-sequences based on estimates of transition density, *Annals of the Institute of Statistical Mathematics*, **42** (2), 253–268.