Reconstruction and analysis of the state space for the identification of dynamical states in real-world time series

vorgelegt von

Iliusi Donaji Vega del Valle

Dissertation zur Erlangung des Grades

Doktor der Naturwissenschaften (Dr. rer. nat.)

am

Fachbereich Mathematik und Informatik

der

Freie Universität Berlin

Dezember, 2015

Betreuer:
Prof. Dr. Christof Schütte
Konrad-Zuse-Zentrum für Informationstechnik Berlin.
Takustrasse 7,
14195 Berlin.

Dr. Péter Koltai
Institut für Mathematik,
Arnimallee 6,
14195 Berlin.

Prof. Giovanni Ciccotti
Gastprofessor,
Fachbereich Physik, La Sapienza,
Roma, Italien.

Tag der Disputation:
7. November 2016

Reconstruction and analysis of the state space for the identification of dynamical states in real-world time series

Selbstständigkeitserklärung

Ich, Iliusi Donaji Vega del Valle, erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst habe. Ich habe alle Hilfsmittel, Hilfen und Publikationen angeben und versichere, auf dieser Grundlage die Arbeit selbständig verfasst zu haben. Diese Arbeit ist nicht schon einmal in einem früheren Promotionsverfahren eingereicht worden.

Iliusi Donaji Vega del Valle.
Dezember, 2015.

Abstract

One of the main goals of analyzing a high-dimensional time series is to identify structures in it. Some of these structures correspond to important dynamical features in the underlying system, like different dynamical states and the transitions between these.

In this thesis we introduce two new methodologies for the identification of different dynamical features in a system from the analysis of a real-world time series. We focus in the dynamical features corresponding to the different dynamical metastable states (in a system with multiple and well distinguished time scales, these can be understood as the attractors associated to each of the different time scales) in a system and the transitions between dynamical regimes in a system.

Our first method is designed for the identification of different dynamical metastable states, and takes a recurrence analysis approach. The results provided by this method seem to be robust to the introduction of noise and missing points.

Our second method is designed for the identification of transitions between different dynamical regimes, and takes an algebraic topological approach. It seems that our second method is, by construction, also robust to the noise and outliers in the data. However, it is still not sensitive enough to identify dynamical transitions where the shape of the attractors in a system suffer small changes.

Given that both methods introduced in this thesis rely on the geometrical analysis of the state space, another issue treated in this thesis is the reconstruction of the state space from a complex time series.

In this thesis, our criteria for an adequate state space reconstruction are given in terms of the gain or loss of geometrical information. These criteria are specifically developed for each of the approaches taken for every method: recurrence analysis and persistent homology.

Zusammenfassung

Die Identifizierung der Strukturen ist ein der Hauptziele der Analyse einer hochdimensionalen Zeitreihe. Einige dieser Strukturen entsprechen wichtigen dynamischen Eigenschaften in einem System, wie verschiedene dynamische Zustände und die Übergänge zwischen denen.

In dieser Dissertation stellen wir zwei neue Methoden für die Analyse des Real-World-Zeitreihen dar, die verschiedene dynamische Eigenschaften in einem System identifizieren. Wir fokussieren uns auf die Identifizierung der unterschiedlichen dynamischen metastabilen Zuständen in einem System (in einem System mit mehreren, unterscheidbaren Zeitskalen, kann jede Zeitskala mit verschiedenen Attraktoren verbunden sein) und auf die Übergänge zwischen verschiedenen dynamischen Regimen in einem System.

Unsere erste Methode identifiziert verschiedene dynamische metastabile Zustände. Diese Methode ist in dem Rekurrenz-Analyse-Ansatz geankert. Die Ergebnisse dieser Methode sind scheinbar gegen die Rauscheneinführung und fehlende Datenpunkte robust.

Unsere zweite Methode identifiziert, die Übergänge zwischen verschiedenen dynamischen Regimen. Diese Methode ist auf einer algebraischen topologischen Ansatz basiert. Es scheint, dass unsere zweite Methode gegen die Rauschen-Einführung und Ausreisser in den Daten robust ist. Es ist jedoch immer noch nicht empfindlich genug, dynamische Übergängen, wo die Form der Attraktoren in einem System kleine Änderungen ausweist, zu identifizieren.

Da beide in dieser Arbeit vorgestellte Methoden auf die geometrische Analyse des Phasenraums beruhen, wird in dieser Arbeit des Weiteren die Rekonstruktion des Phasenraums von komplexen Zeitreihen behandelt.

In dieser Dissertation, beziehen sich unsere Kriterien für eine angemessene Phasenraumrekonstruktion auf den Gewinn oder Verlust der geometrischen Informationen. Diese Kriterien sind speziell für jeden Ansatz bei den jeweiligen Methoden entwickelt worden: Rekurrenz-Analyse und persistente Homologie.

Acknowledgments

This dissertation marks the end of a very important period in my life. I want to dedicate this section to acknowledge the support of the several institutions, colleagues, friends and family who have accompanied me during this period.

Special thanks to my supervisors, Prof. Dr. Christof Schütte and Dr. Tim Conrad, for giving me the freedom to follow my curiosity and research interests, and for providing me with interesting ideas and questions during all these years.

To the Max Planck Research School on Computational Biology and Scientific Computing (IMPRS-CBSC), for giving me the opportunity to start this adventure. Especially, I want to thank Kirsten Kelleher, for all her moral support and academic advice.

To my funding institutions, who allowed me to research all these exciting problems during this time: The IMPRS-CBSC, the Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB), and the Freie Universität Berlin.

To the Biocomputing group, for all the scientific talks and jokes. In particular, I want to thank Victor Mireles, who has helped me every time my computer decided not to talk to me anymore, and who dedicated time to reading my thesis during critical moments in his research. I also want to thank Pooja and Nada, who have always procured me with a friendly and loving environment at work.

To my friends who, no matter if close by or at the distance, have kept their ears and hearts open to me. You are the space-time and matter of my joy. I love you for being those who can escape the temptations of individualism.

To my family, who never stopped supporting me despite not understanding not what nor why I was living so far away and working in a way that sometimes caused my heart and head so much pain.

To Amrit, who has been the most important person in my life since my arrival to Berlin. There are not enough words to express my gratitude and love for you. You are the counterpoint and syncopation.

Finally, I want to thank my parents, Estela and Prócoro, who have been an example of strength and courage, and have helped me grow dignified and brave. I dedicate this work to you.

On Time

And an astronomer said, "Master, what of Time?"

And he answered:

You would measure time the measureless and the immeasurable.

You would adjust your conduct and even direct the course of your spirit according to hours and seasons.

Of time you would make a stream upon whose bank you would sit and watch its flowing.

Yet the timeless in you is aware of life's timelessness,

And knows that yesterday is but today's memory and tomorrow is today's dream.

And that that which sings and contemplates in you is still dwelling within the bounds of that first moment which scattered the stars into space.

Who among you does not feel that his power to love is boundless?

And yet who does not feel that very love, though boundless, encompassed within the centre of his being, and moving not form love thought to love thought, nor from love deeds to other love deeds?

And is not time even as love is, undivided and spaceless?

But if in your thought you must measure time into seasons, let each season encircle all the other seasons,

And let today embrace the past with remembrance and the future with longing.

Khalil Gibran, *The Prophet*

Contents

1. Introduction

One of the main goals of analyzing a time series is to identify structures in it. These structures usually correspond to important dynamical features of the underlying system, like different dynamical states and the transitions between these.

There are several methods for the identification of dynamical features in a system. Each of these makes different assumptions about the underlying dynamical system: stationarity, determinism, the existence of a certain number of attractors, a certain dominant time scale in the dynamics, and so on. However, given that the dynamical properties of an analyzed system are often not known *a priori* when analyzing real-world[1] (or complex) time series, many of the assumptions of these methods are not satisfied.

Some of the most robust methods used to extract dynamical features from complex high-dimensional data, also known as methods for dimensionality reduction or manifold learning methods, are PCA [63], Isomap [117], Classical Multi-dimensional Scaling [75] (MSD), Stochastic Neighbor Embedding [59] (SNE), Locally Linear Embeddings [107] (LLE) and Hessian eigenmaps [75]. These methods aim to find a reduced representation of the data and are widely used for the visualization and classification of complex data. For further information about these methods, we refer the reader to the book of Wang on high-dimensional data and dimensionality reduction [123].

Among the main assumptions of these methods are that the data lies in a manifold (a locally Euclidean or smooth closed surface without self-intersections) and that the sampling of the state space is sufficiently dense and respects the local feature size of the state space. For example, PCA assumes that the state space is a linear subspace; Isomap, that it is intrinsically flat but isometrically embedded; and the method of Hessian eigenmaps, that it is isometrically embedded but has more relaxed topological restrictions.

According to van der Maaten et al. [120], each of these assumptions implies some drawbacks. For example, not preserving both the local and the global scale properties of complex data. Additionally, many of these methods depend on many undetermined parameters, leaving large part of the analysis open to subjective interpretation.

Despite the assumptions and drawbacks of these methods, they have provided important insights to many research fields. By mentioning these, we simply want to emphasize the importance of selecting an adequate method of analysis that better adapts to the characteristics of our time series.

Much of the attention in the development of new methods for the analysis of time series is

[1] Real-world time series may be high-dimensional, non-linear, noisy, sparse or have different time scales ruling its dynamics.

focused on reducing the number of their working assumptions and on making them robust to the presence of noise, outliers and other features common in real-world time series.

But in order to identify the different dynamical states or regimes in a time series, one must first remember that a time series does not show directly the features of the dynamical system governing its behavior. Actually, a time series is the result of applying an unknown measurement function to the states of the dynamical system. This measurement function is typically a projection of the space containing all the states of a dynamical system, or state space, into a subspace of lower dimension.

Therefore, in order to identify the different dynamical states in a time series, one must first adequately reconstruct the state space of the underlying dynamical system.

A commonly used reconstruction of the state space from a time series is provided by a delay-coordinate map. This type of mapping depends on two embedding parameters: the time delay and the embedding dimension. Many geometric, algebraic and topological tests, based on the embedding theorems of Takens [115] and Whitney [124], have been developed to determine these embedding parameters [106] for deterministic time series. For time series coming from some types of stochastic systems, Stark, Broomhead, Davies and Huke [112, 113, 114] have extended Takens' theorems. However, the implementation of these theorems is hard and requires knowing *a priori* the sequence of stochastic influences acting on the analyzed system.

In this thesis, we introduce two new methods for the identification of different dynamical features in complex time series. The first method is designed for the identification of different dynamical metastable states (in a system with multiple and well distinguished time scales, these can be understood as the attractors associated to each of the different time scales), and the second is designed for the identification of transitions between different dynamical regimes. Both methods rely on the geometrical analysis of the space containing all the states of a dynamical system, or state space. However, each of these is based on a different approach depending on its aim.

Given that the relevance of the results obtained by both of our methods depends on an adequate reconstruction of the state space, we discuss the many criteria used to obtain an adequate state space reconstruction form a complex time series in many parts of this thesis. We cover the general criteria in Chapter 2. And we put especial attention to this discussion within the frameworks of recurrence analysis and persistent homology analysis in Sections 3.2 and 4.4.

Our first method, for the identification of metastable states in a complex time series, is based on recurrence analysis. In Chapter 3 we first introduce the recurrence analysis approach and then present this method in Section 3.3.

In general terms, a recurrence is defined as the return of a state space trajectory to a state arbitrarily close to a previously visited state. The closeness between state space vectors is determined by a parameter called the recurrence threshold. Recent studies have provided ways to set the recurrence thresholds for uniformly distributed vector spaces [34] but for different cases, its selection remains a subjective matter. The graph representation of the recurrences in state space for a given recurrence threshold is called a recurrence network. Every module in a recurrence network can be associated to a different dynamical state.

Thus, our method for the identification of different dynamical states in a time series can be summarized as follows. Given an adequate state space reconstruction, we create a filtration of recurrence thresholds and construct the recurrence networks associated to these set of

parameters. Then, we identify those recurrence thresholds that produce a sub-collection of networks with similar modular structures. That is, those with the same number of modules and similar number of nodes in each module. Assuming that such range of parameters captures the main dynamics of the analyzed system, any recurrence threshold in such range is adequate, but we select the middle value in the identified range as the 'final' recurrence threshold. The modules identified in the recurrence network associated to this final recurrence threshold are said to correspond to the different metastable states in the time series.

We provide an adequate state space reconstruction within the framework of recurrence analysis by repeating the previous steps for different embedding parameters and selecting those parameters that obey the criteria on Shannon entropy and recurrence rate[2] mentioned in Section 3.1.3. This methodology is fully described in Section 3.4.

The results provided by this method seem to be robust to the introduction of a considerable amount of noise and missing points.

Among the advantages of using recurrence networks to identify the different dynamical states in a system are that these networks give information about the local, medium and global scales in high-dimensional, non-linear time series [35]. Besides, these depend on one parameter only, and require making few assumptions about the time series analyzed. Additionally, several studies on climate, financial and medical data [86], suggest the recurrence networks analysis approach is as a good candidate to deal with non stationary time series. This approach has also shown remarkable robustness for the analysis of time series with noise and missing data points.

However, we face two problems related to the assumptions made to construct recurrence networks. First, that there is one dominant scale in the analyzed time series. This assumption justifies the selection of a single recurrence threshold for its analysis, but might not be true for many real-world time series. And second, that all the relevant dynamical information of the analyzed system is contained in the graph structure of the associated recurrence network. Additionally, for analyzing a recurrence network, one usually has to provide a guess on the number of attractors in the dynamical system.

Since these assumptions might be too restrictive for many real-world time series, we developed a different method, following a path that allowed us to overcome these problems.

This way, our second method, which identifies transitions between different dynamical regimes, is based on persistent homology, an algebraic topological approach. In Chapter 4 we introduce the persistent homology approach and then present our second method of analysis in Section 4.3.

The main idea of persistent homology is to create a filtration from a data set depending on a parameter of proximity between data points, ϵ, and to identify the topological features (homology; see Section 4.1.3) surviving over different ranges of such parameter. Intuitively, those topological features surviving longer ranges may correspond to interesting signals, whereas short persisting features may correspond to noise or indicate inadequate sampling. This way, the second method does not assume the existence of a main scale in the dynamics of the system and seems to be robust, by construction, to the presence of noise and outliers in the data.

This method for the identification of transitions between different dynamical regimes can

[2]Shannon entropy and recurrence rate are two recurrence quantification analysis measures (RQA). The RQA measures quantify the information contained in the adjacency matrix of a recurrence network and are introduced in Section 3.1.3

be summarized as follows. Given an adequate state space reconstruction, we divide a time series into several windows of measurements of equal length, procuring that each window of measurements has enough data points as for capturing the dynamics of the system in the time elapsed by the window. We compute a topological summary for each window (using persistent homology) and then compare the results between different windows. We say that there is a change in the dynamics of a system if the topological summaries of two windows of measurements are significantly different in a scale of time related to the length of the windows of measurements.

Within the framework of persistent homology, we also introduce another method for the selection of a pair of embedding parameters that provide an adequate state space reconstruction, in Section 4.5. We look for embedding parameters that create a state space reconstruction whose persistent homology analysis produces similar results when analyzing different subsamples from a same time series.

The use of persistent homology allows the incorporation of higher-order geometrical information to the analysis of data, which constitutes an advantage when analyzing data with higher-dimensional or multiple attractors. In principle, this approach is also suitable for data where the use of a specific metric is not fully justified [9, 44]. Put in these terms, this approach seems to be the solution to all our problems. However, problems appear during its implementation, since it is generally very costly in computational terms. For this reason, we incorporated in our method the use of sampling and statistics to estimate the persistence homology of data [8, 18].

Nevertheless, our method is not sensitive enough to identify dynamical transitions where the shape of the attractors in a system suffer small changes. And even more challenging is the analysis of stochastic systems. However, in Conclusions (Chapter 5) we make some suggestions for future work, aiming to overcome these problems.

2. The state space of a dynamical system: states and evolution

When we analyze a time series data set, $u(t) \in \mathbb{R}^{d'}$, one of our aims is to identify the different states of the dynamical system underlying it.

Any dynamical system can be defined by its state space (or phase space) and an evolution operator. The state space of a dynamical system is the space containing all the states available for that system. A state for a dynamical system at a given time t, $\boldsymbol{\xi}(t) \in \mathbb{R}^d$, is commonly defined as a vector in a d-dimensional **manifold** usually assumed compact and smooth, $A \subseteq \mathbb{R}^d$.

Therefore, it is indispensable to understand that time series measurements do not constitute direct observations of the states of a dynamical system. What we observe in a time series $u(t)$ are the effects of a measurement function, $h : \mathbb{R}^d \to \mathbb{R}^{d'}$, where usually $d' < d$, applied to the states of a dynamical system described by $\boldsymbol{\xi}(t)$. Thus, $u(t) = h(\boldsymbol{\xi}(t))$. Depending on the properties of such measurement function, a time series might contain, or not, all the dynamical information of a system.

Definition 2.0.1 (Manifold). *In the context of topology, a manifold is a topological space provided with a family of pairs $\{(U_i, \phi_i)\}$, where $\{U_i\}$ is a family of open sets covering M and ϕ_i is a homeomorphism from U_i onto an open subset U_i' of $\mathbb{R}^m$. It satisfies that, given U_i and U_j such that $U_i \cap U_j \neq \emptyset$, the map $\psi_{ij} = \phi_i \circ \phi_j^{-1}$ from $\phi_j(U_i \cap U_j)$ to $\phi_i(U_i \cap U_j)$ is infinitely differentiable.*

In this thesis, we introduce two methods, one for the identification of metastable states in a complex time series, and another for the identification of transitions between dynamical regimes in a complex time series. The relevance of the results provided by these methods relies on providing an adequate reconstruction of the state space from a time series. This means, a reconstruction that "hides" the effects of the measurement function and "shows" the states and dynamics of a system.

The difficulty of reconstructing the state space of a system from a time series depends on the dynamical properties of the system. Whether a system is deterministic or stochastic, is of major importance.

When the future states of a system are uniquely determined by an initial state $\boldsymbol{\xi}(0)$, then the system is said to be deterministic. For a deterministic system, the evolution operator is a function of time which determines the change of state and dynamics of the system. It can be defined as a continuous or as a discrete function. When the evolution operator is continuous, one can identify trajectories in the state space. Deterministic systems described by continuous evolution operators have been analyzed widely, and the reconstruction of their state spaces is fairly understood.

On the other hand, if the future states of a system are not uniquely determined by an initial state, such system is said to be stochastic. The reconstruction of the state space of stochastic systems from a time series is not always possible. Such reconstruction is possible for some types of stochastic systems, as specified by the theorems of Stark [113] and Stark, Broomhead, Davies and Huke [112, 114]. However, the applicability of these theorems requires knowing *a priori* the sequence of stochastic influences acting on the system [101].

In this Chapter we will first review, in Section 2.1, the theorems and methods most commonly used to reconstruct the state space from a time series, assuming that the underlying dynamical system is deterministic. Then, in Section 2.2 we will comment on the considerations necessary to reconstruct the state space from specific types of stochastic systems. In Section 2.3 we will discuss the current criteria used to determine whether a state space reconstruction is adequate or not. And finally, in Section 2.4, we will briefly motivate our focus on two approaches for the analysis of the state space: recurrence analysis and persistent homology analysis.

2.1 State space reconstruction from a time series – *Deterministic systems*

When the time series of a dynamical system is mapped into a space of adequate dimension, this space contains all the dynamical information of the system, preserves determinism and creates a diffeomorphism for the attractors [106].

Then, the state space of a dynamical system defined on $\mathbb{R}^d$ can be reconstructed by finding an appropriate mapping of a time series. This reconstruction can be understood as finding an embedding delay-coordinate map , $\phi : U \to \mathbb{R}^d$, where U is the space of measurements (see Section 2.1.1).

However, it may occur that U is only locally d'-dimensional. This means that its **topological dimension** might be different to d'. In this case, U it is not necessarily a subset of $\mathbb{R}^{d'}$. Instead, one should assume that $U \subseteq \mathbb{R}^{\tilde{d}}$, where $\tilde{d} \geq d'$. And then, the reconstructed vectors, $\mathbf{x} \in \mathbb{R}^{\tilde{d}}$, are given by $\mathbf{x}(t) = \phi'(h(\boldsymbol{\xi}(t)))$, where $\phi' : U \to \mathbb{R}^{\tilde{d}}$.

Topological dimension

For topological spaces, there are different definitions of topological dimension, like the small inductive dimension, the large inductive dimension and the **Lebesgue** or topological dimension. For separable metrizable spaces, the three take the same value.

Definition 2.1.1 (Lebesgue dimension)**.** *The Lebesgue dimension of a topological space X is the minimum value d such that any open cover in X has a refinement (second open cover where each of its sets is a subset of a set in the first open cover) in which no point is included in more than $d + 1$ elements.*

How do we know whether an embedding map is adequate for reconstructing the state space or not?

The research conducted with the aim of replying to this question is vast. In this section we will simply summarize the main results used for the development of the delay-coordinate maps, the type of state space reconstruction that we later use in the development of the methods for the identification of different dynamical states or regimes in a complex time series. Let us begin stating some of the restrictions a map ϕ has in order to be an embedding.

One of these is that the map must preserve the differential information. This means that it must not collapse any point or tangent direction. Thus, an embedding, ϕ, of a compact

smooth ($\mathcal{C}^1$) differentiable manifold, A, is a smooth diffeomorphism from the manifold onto its image. This definition of an embedding is equivalent to ask ϕ to be a one-to-one **immersion**.

In order to find the dimension of an embedding of a compact smooth manifold, we can use a theorem proved by Whitney in 1936 [124]. This theorem states that if A is a smooth manifold of dimension d, then the set of maps into $\mathbb{R}^{2\tilde{d}+1}$ that are embeddings of A is an open and dense set in the $\mathcal{C}^1$-topology of maps. This theorem derives from the fact that two hyperplanes of dimensions d_1 and d_2 embedded in an $\tilde{d}$ dimensional space will typically intersect if $d_1 + d_2 \geq \tilde{d}$.

Definition 2.1.2 (Immersion)**.** *A smooth ($\mathcal{C}^1$) map ϕ on a compact smooth differentiable manifold A is an immersion if the derivative map $D\phi(x)$ represented by the Jacobian matrix of ϕ at x is one-to-one at every point x of A.*

Given that such set of embeddings of A is open, any arbitrarily small perturbation of an embedding will still be an embedding. And given that it is dense, every smooth map (wether an embedding or not) is arbitrarily near to an embedding.

However, open dense subsets can be thin in terms of probability and this theorem is not sufficient to find the embedding dimension.

In 1980, Packard et al. [96] proved that the geometry of a dynamical system can be obtained from the time series of one of its observables. Therefore, the state space can be reconstructed using other measurements taken at different times. This discovery could then be used to overcome the fact that a time series might not represent all the degrees of freedom of the manifold containing all the states of the dynamical system.

Just a year later, Takens [115] proved a theorem that would later allow to reconstruct the state space of a system in more general situations.

Theorem 2.1.3 (Takens embedding theorem)**.** *Let A be a d-dimensional manifold. For pairs (ϕ, h), where $\phi : A \to A$ is a smooth diffeomorphism and $h : A \to \mathbb{R}$ a smooth function, it is a generic property that the observation map $\Phi[\phi, h] : A \to \mathbb{R}^{2d+1}$ defined by:*

$$x \to \Big(h(x), h(\phi(x)), ..., h(\phi^{2d}(x))\Big) \tag{2.1}$$

is an immersion.

This theorem can be applied to time series when ϕ is a time τ map on the underlying dynamical system. In this case, it holds when there are no periodic orbits of period τ or 2τ and at most finitely many orbits of higher period. It also assumes that the measurements are taken with infinite precision.

In the presence of periodic orbits of period 2τ, $\Phi(h, f, \tau)$ cannot be one-to-one for any observation function h. When A is a periodic orbit of period 3τ, or any period not equal to τ or 2τ, there is no such problem. In this case, the delay-coordinate map is an embedding for almost every h, as long as the reconstruction dimension is at least $2d + 1$.

This theorem is the foundation for the reconstruction of the state space of deterministic (and stochastic) systems, but it gives no estimate on the lowest dimension for which almost every map is an embedding. The search for an estimation of an adequate embedding dimension dominated the following years.

Definition 2.1.4 (Prevalence)**.** *A Borel subset S of a normed linear space V is prevalent if there is a finite-dimensional subspace E of V (called the* probe space*) such that for each v in V, $v + e$ belongs to S for (Lebesgue) almost every $e \in E$.*

In 1991, Sauer et al. [106] introduced the concept of **prevalence** and strengthened Whitney's theorem, in the aim of finding boundaries for the embedding dimension.

They found that if a subset of a finite-dimensional vector space is prevalent, then the complement of this subset has zero measure. This also holds for a union or intersection of a finite number of prevalent subsets. Taking prevalence into account, they proved Theorem 2.1.5, which implies that almost all maps near an embedding are also embeddings.

Theorem 2.1.5 (Whitney Embedding Prevalence Theorem)**.** *Let A be a compact smooth manifold of dimension d contained in $\mathbb{R}^k$. Then almost every smooth map $\mathbb{R}^k \to \mathbb{R}^{2d+1}$ is an embedding of A.*

Additionally, they proved Theorem 2.1.6, an extension of the prevalence theorem to fractal sets that are not smooth manifolds. This was achieved by introducing the box-counting dimension (see definition in Section 2.1.2), or capacity dimension, and requiring the embedding maps to be immersions.

Theorem 2.1.6 (Fractal Whitney Embedding Prevalence Theorem)**.** *Let A be a compact subset of $\mathbb{R}^k$ of box-counting dimension d, and let n be an integer greater than $2d$. Then, for almost every smooth map $\phi : \mathbb{R}^k \to \mathbb{R}^n$, a) ϕ is one-to-one on A, and b) ϕ is an immersion on each compact subset C of a smooth manifold contained in A.*

However, one can only have an immersion on compact subsets of a smooth manifold contained in the compact fractal set A. And even when one knows the box-counting dimension of an attractor A, this theorem still does not give an estimate of the lowest embedding dimension.

Nevertheless, Sauer et al. later expanded Theorem 2.1.6 and this led to the development of delay-coordinate maps, which are explained in the following section. This mappings depend on only two parameters, the embedding dimension and the embedding delay, and have some nice properties. For example, there exists, with probability one, a delay-coordinate map which is an embedding. Eventually, some measures were developed in order to set the two embedding parameters and, finally, there was a way to create a reliable reconstruction of the state space of a dynamical system from a time series.

2.1.1 Delay-coordinate maps

One of the maps most commonly used to reconstruct the state space of a deterministic system from a time series is a delay-coordinate map. This is also the type of map we use for the reconstruction of the state space in the two new methods presented in this thesis.

Delay-coordinate maps, or time delay embeddings, are based on Takens' result stated in Theorem. 2.1.3. For these maps, only two *embedding parameters* have to be set: the embedding dimension, m, and the time delay τ. Later on, in Section 2.1.2, we comment on the selection of the parameters it depends on.

A delay-coordinate map for the reconstruction of the state space of a dynamical system, $\phi(h, f, \tau) : A \to \mathbb{R}^m$, is given by:

$$\phi(h, f, \tau) = \big(h(\boldsymbol{\xi}), h(f^{\tau}(\boldsymbol{\xi})), h(f^{2\tau}(\boldsymbol{\xi})), ..., , h(f^{(m-1)\tau}(\boldsymbol{\xi}))\big)^{\top} = \mathbf{x} \tag{2.2}$$

where h is a measurement function, $\tau > 0$ is called the delay and f is a flow in the manifold A. Ideally, the delay-coordinate map would be an embedding for almost every $h : A \to \mathbb{R}$.

This map can also be created with lagged observations, considering that f is a diffeomorphism on an open subset U such that A is a compact subset of U. Then, according to Eq. 2.2, the reconstructed vectors $\mathbf{x}(t)$ for a time series $u(t)$ are obtained by gathering m adjacent measurements delayed by τ:

$$\mathbf{x}(t) = \big(u(t), u(t+\tau), u(t+2\tau), ..., u(t+(m-1)\tau)\big)^\top \tag{2.3}$$

This means that if we have N measurements in our time series, we would obtain $N^* = N - \tau(m-1)$ reconstructed vectors. These vectors are no longer localized in time, but somewhere in between a time interval of length $\tau(m-1)$.

The state space vectors reconstructed via a delay-coordinate map have some interesting properties which imply that delay-coordinate maps fold the smooth manifold hiding all deterministic behavior on length scales larger than the typical lengths of the foldings. This might in turn make the dynamics of the attractor seem stochastic when it actually is not.

One of such properties is that, since any two consecutive reconstructed vectors will differ in only one component, the time evolution in an delay embedding space is trivial but in one component. This is a rotation in $m-1$ components and contains only one nontrivial scalar function that contains all nonlinearities in the flow and produces all entropy: $u(t+\tau) = F^\tau(\mathbf{x}(t))$ [68]. This way, entropy must increase with τ. Another property is that this mapping uses the same measurements to reconstruct all the directions of the state space. This might introduce an artificial isotropy [53].

According to Vlachos and Kugiumtzis [122] and Palit et al. [97], to overcome such induced artificial isotropy, one may use *non-uniform delay-coordinate map*, or non-uniform time-delay embeddings. This type of embedding was introduced by Judd and Mees [67], who used it to obtain global reduced autoregressive models.

Considering a time series $u(t)$ and m different embedding delays, $\tau_1, \tau_2, ..., \tau_m$, a non-uniform delay-coordinate map of $u(t)$ is given by:

$$\mathbf{x}(t) = \big(u(t), u(t+\tau_1), u(t+\tau_2), ..., u(t+\tau_{m-1})\big)^\top \tag{2.4}$$

Additionally, one might wonder whether all functions h in Eq. 2.2 guarantee that the delay-coordinate map $\phi(h, f, \tau)$ from A into $\mathbb{R}^m$ is an embedding.

To answer this question, one can refer to Theorem 2.1.7, a stronger version of Whitney's embedding theorem also proven by Sauer et al. in [106].

Theorem 2.1.7 (Fractal Delay Embedding Prevalence Theorem (taken from [106]))**.** *Let f be a flow on an open subset $U \subset \mathbb{R}^{\tilde{d}}$ and A be a compact subset of U of box-counting dimension d_C, where $d_C < \tilde{d}$. Let $n > 2d_C$ be an integer and $T > 0$. Assume that A contains at most a finite number of equilibria, no periodic orbits of f of period T or $2T$, at most finitely many periodic orbits of period $3T, 4T, ...nT$, and that the linearizations of those periodic orbits have distinct eigenvalues. Then for almost every smooth function h on U, the delay-coordinate map $\phi(h, f, T) : U \to \mathbb{R}^n$ is one-to-one on A and an immersion on each compact subset C of a smooth manifold contained in A.*

Theorem 2.1.7 allows the transit from delay-coordinate maps generically giving embeddings of smooth manifolds of dimension d, to delay-coordinate maps being prevalent on giving embeddings on compact sets of box-counting dimension d_C. In other words, this theorem states that, with probability one, there is a delay-coordinate map $\phi(h, f, \tau) : U \to \mathbb{R}^{\text{ceil}(2d_C)}$ which is an embedding.

In the following section, now that we have revisited the main theorems around delay-coordinate maps, we will expand on some of the measures, based in such theorems, used to set the embedding parameters necessary for an adequate state space reconstruction of a deterministic dynamical system. We will distinguish between the two types of embedding mentioned above: uniform and non-uniform.

2.1.2 Embedding parameters

As stated in the theorems mentioned above, for a uniform embedding of a deterministic dynamical system, one must set the time delay τ and the embedding dimension m. And for a non-uniform embedding, one must select a set of delays $\tau_1, ..., \tau_{m-1}$.

Setting the embedding parameters for either of these embeddings implies the geometrical, dynamical and topological analysis of the time series analyzed [79]. In the following paragraphs we will review the most commonly used measures to set these parameters.

Embedding delay for uniform embedding

An adequate embedding delay, τ, is the one that guarantees that the components of the state space vectors are as uncorrelated or independent as possible. This means, that the vector built with all the i-th entries of the state space trajectories is linearly independent from the vector built with all the j-th entries of the state space trajectories, for all $i \neq j$.

Typically, the embedding delay, τ, is chosen either as the first minimum of linear autocorrelation function or as the first minimum of the average mutual information [52]. In the two following sections we briefly review these measurements.

Autocorrelation. For a time series $u(t)$ with N time points and zero mean and a time delay τ, the amount of linear correlation within this time series is given by the variation from zero of the product $u(t)u(t+\tau)$, measured in units of sampling time, on average over the entire time series. Thus, the second-order autocorrelation function [53], or autocovariance function, is given by

$$A(\tau) = \frac{1}{N-\tau-1} \sum_{t=0}^{N-\tau-1} u(t)u(t+\tau) \tag{2.5}$$

For smoothly decaying autocorrelation functions, the autocorrelation time τ_a can be considered a measure of the time scale for which there are significant linear correlations within a time series. This time is given by $A(\tau_a) = \frac{A(0)}{e}$.

The linear autocorrelation function cannot be applied to any given time series because we would need to know in advance that the time series is not nonlinear. However, one can use nonlinear autocorrelation functions, higher-order correlation functions or a generalized approach based in mutual information[53].

Average mutual information. A non-linear generalization of the linear autocorrelation is considered to be given by the average mutual information (AMI). This measurement tells us how much information about $u(i+\tau)$ we get when we observe $u(i)$.

For a time series $u(t)$, time delay τ and $\hat{p}(u(t), u(t+\tau))$ the estimated joint probability distribution of the bivariate time series $(u(t), u(t+\tau)$, the average mutual information is given by:

$$AMI(\tau) = \sum_{i=0}^{N} \hat{p}\big(u(i), u(i+\tau)\big) \log_2 \frac{\hat{p}\big(u(i), u(i+\tau)\big)}{\hat{p}\big(u(i)\big)\hat{p}\big(u(i+\tau)\big)} \tag{2.6}$$

Since any two measurements are completely independent when the mutual information is zero, the time delay τ can be chosen as the one for which we obtain the first minimum in average mutual information. However, for some systems, the average mutual information might not have a minimum. In these cases, a deeper analysis is required. For an extended discussion on this topic, see the work of H. D. I. Abarbanel [1].

Even when the average mutual information can be applied to nonlinear time series, it still has some drawbacks. One of them is that, since it is computed in terms of two variables only, it fails on giving information about high-dimensional relations between all attractor values [97]. Another drawback is that it provides a unique time delay in terms of two-dimensional information, which is problematic when we analyze a system with multi-dimensional attractors. A solution to these problems would imply selecting a different time delay for every variable in the reconstructed state space that contains the attractors.

Embedding delays for non-uniform embedding

There are several methods to set adequate delays for a non-uniform embedding. In general, the first methods focused on the use of high-dimensional mutual information estimators. Among these are the methods of Judd and Mees [67], Boccaletti et al. [5], Kraskov et al. [72], and Simon and Verleysen [111]. Unfortunately, these estimators did not vary with changes on the dynamical system analyzed and depended on the dimension of the time series. For these reasons, these could not be used to reconstruct the state space of general dynamical systems.

More recently, the methods focused on being system dependent and on depending less on the dimension of the time series. Among these is the method of I. Vlachos and D. Kugiumtzis [122], which introduced the use of conditional mutual information. And the method of Palit et al. [97], which introduced the use of high-dimensional cross auto-correlation, which apart from being system dependent, varies for uniform and non-uniform embedding delays. In these two approaches, the delay (or non-uniform delays) is chosen as the one providing the cross auto-correlation value closest to zero.

In Section 4.3 we will come back to the use of non-uniform embeddings for the reconstruction of the state space. There, we will introduce a methodology, which uses topological measurements and the idea of finding the lowest cross-correlation, for setting the non-uniform embedding delays.

Embedding dimension

In order to set the embedding dimension, m, one can perform different geometrical, dynamical or topological tests. The geometrical tests indicate the variations in distance between two close points when the embedding dimension increases. These typically involve the computation of the *box-counting dimension* (often called the *fractal dimension*) or of the false nearest neighbors (FNN) [69].

The dynamical tests look for an embedding dimension that provides a unique future for every data point. These involve the implementation of predictability tests or the estimation of Lyapunov exponents.

The topological tests look for an m that avoids the intersections of stable periodic orbits. Generally, for an n-dimensional dynamical systems with an attractor of fractal dimension d_A, the embedding dimension is $m > 2d_A$. Additionally, according to Whitney et al. [124], $m < 2n$. One-dimensional chaotic data, for example, have embedding dimension $m \geq 3$.

Box-counting dimension. The box counting dimension gives information about the "fractal" dimension of the attractors and therefore can be used to estimate the embedding dimension using Theorem 2.1.7.

For a positive number ϵ, let A_ϵ be the set of all points within distance ϵ of A, such that $A_\epsilon = \{x \in \mathbb{R}^n : |x - a| \leq \epsilon \text{ for some } a \in A\}$. And let $\text{vol}(A_\epsilon)$ denote the n-dimensional outer volume of A_ϵ. Then the box-counting dimension of A is given by:

$$\text{boxdim}(A) = \lim_{\epsilon \to 0} \frac{\log \text{vol}(A_\epsilon)}{\log 1/\epsilon} \tag{2.7}$$

If $d = \text{boxdim}(A)$ exists, then $vol(A_\varepsilon) \approx \varepsilon^{n-d}$.

False nearest neighbors. The false nearest neighbors approach is the test most common tool for setting the embedding dimension.

For a given vector $\mathbf{y}(t) \in \mathbb{R}^n$ and $\mathbf{y}^{(k)}(t)$ a k-th nearest neighbor, the square of the Euclidean distance between them is given by

$$d_n^2(\mathbf{y}(t), \mathbf{y}^{(k)}(t)) = \sum_{j=0}^{n-1} y(t + j\tau) - y^{(k)}(t + j\tau)^2. \tag{2.8}$$

When moving to a space of dimension $n+1$ via a delay-coordinate embedding with embedding parameters τ and m, we add a coordinate to each vector. In this case, $d_{n+1}^2(\mathbf{y}(t), \mathbf{y}^{(k)}(t)) = d_n^2(\mathbf{y}(t), \mathbf{y}^{(k)}(t)) + [x(t + n\tau) - x^{(k)}(t + n\tau)]^2$.

Then, a way to identify that the embedding in a lower dimension did not capture all the dynamical information in the state space is by finding a large increase in $d_n^2(\mathbf{y}(t), \mathbf{y}^{(k)}(t))$ for a large percentage of k-th nearest neighbors when moving to a higher dimension. The ratio of false nearest neighbors (FNN) is a measurement of such increase in distance. This way, a correct selection of the embedding dimension would eliminate all false nearest neighbors.

According to Kennel et al. [69], for $\mathbf{y}(t) \in \mathbb{R}^n$, $\mathbf{y}^{(k)}(t)$ its k-th nearest neighbor and $d_n^2(\mathbf{y}(t), \mathbf{y}^{(k)}(t))$ the square of the Euclidean distance between them, the rate of false nearest neighbors is given by:

$$\left[\frac{d_{n+1}^2(\mathbf{y}(t), \mathbf{y}^{(k)}) - d_n^2(\mathbf{y}(t), \mathbf{y}^{(k)}(t))}{d_n^2(\mathbf{y}(t), \mathbf{y}^{(k)}(t))} \right]^{1/2} = \frac{x(t + n\tau) - x^{(k)}(t + n\tau)}{d_n(\mathbf{y}(t), \mathbf{y}^{(k)}(t))} \tag{2.9}$$

Given an arbitrary threshold R_{tol}, a false nearest neighbor is one for which this rate is larger than R_{tol}.

A drawback of using the FNN approach to set the embedding dimension has been recently exposed by C. Nichkawde [94]: for uniform time delay embeddings, this does not constitute a minimal approach. However, this drawback could be overcome by using a non-uniform time delay approach.

Lyapunov exponents. The use of Lyapunov exponents is not widely used for setting the embedding dimension of real-world time series, given that their computation is very sensitive to noise. However, it is an important theoretical reference for the study of nonlinear systems and therefore we will comment on it.

Being $f : \mathbb{R}^n \to \mathbb{R}^n$ a dynamical system and $\mathbf{x}_{t+1} = f(\mathbf{x}_t)$, where $t = 0, 1, ...$, a trajectory of the system, there are n Lyapunov exponents, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, that measure the average rate of divergence of a trajectory.

If f is continuous, the Lyapunov exponents can be related to the long term evolution of an infinitesimal n-sphere in its state space. An n-sphere defined with given initial conditions will transform in time into an n-ellipsoid due to the deformation induced by the flow. Then, the i-th Lyapunov exponent (out of n) can be defined in terms of the length of the i-th ellipsoidal principal axis at time t, $p_i(t)$, by:

$$\lambda_i = \lim_{t\to\infty} \frac{1}{t} \log_2 \frac{p_i(t)}{p_i(0)} \tag{2.10}$$

This way, Lyapunov exponents can give information about the rate of divergence or convergence of nearby trajectories in state space [42].

The Jacobian of the reconstructed attractor can also be used to compute the Lyapunov exponents of a system with unknown dynamics [43]. However, for this, the exponents must be computed in the tangent space of the attractor in the embedding space and the dimension of the embedding space must be equal to the dimension of the original dynamics [27]. These restrictions can clearly become troublesome but one must stick to them because, otherwise, "spurious" Lyapunov exponents [27] appear when the reconstructed state space has dimension larger than the dimension of the attractor. These spurious exponents can be even larger that the largest exponent of the true system.

As mentioned by J. P. Eckmann and Ruelle [43], other quantities related to the Lyapunov exponents are the entropy (or Kolmogorov-Sinai invariant) and the amount of information. These relationships occur because the entropy is related to how chaotic a system is. And how chaotic a system is, is in turn related to the degrees of freedom involved in the dynamics of a system and on their sensitivity to initial conditions.

There are linear and nonlinear methods to compute the Lyapunov exponents from a time series. Nonlinear methods produce more accurate results for negative exponents and when the time series are noisy, but pose bigger computational problems especially when the system has a high embedding dimension [126]. However, there is a growing number of ways to overcome these problems, like the one provided by Yang, Wu and Zhang [127].

2.2 State space reconstruction from a time series – *Stochastic systems*

Previously, in Section 2.1, we show Takens' theorem in its most common formulation: that applicable to nonlinear deterministic dynamical systems where the dynamics and the data are both autonomous (i.e. independent of time and any other external force) [112], and data have no noise. However, there are some types of stochastic systems for which the state space can be reconstructed using extended versions of this theorem.

In a series of articles between 1997 and 2003, Stark, Broomhead, Davies and Huke [112, 113, 114] extended Takens' theorem to deterministically, arbitrarily and stochastically forced systems. In addition, the authors also analyzed iterated function systems and noisy data. In the following paragraphs we will summarize the main results of these articles.

In a discrete time dynamical system, the forcing at time $i \in \mathbb{Z}$ can be given by a variable $\omega_i \in N$, for N an appropriate space (we will specify it later on). Then, the state of the system at time i can be denoted by $x_i \in M$, for M a smooth compact manifold [112] (M can also be noncompact, as shown in [65, 115]). The state in the next time step is thus given by

$$x_{i+1} = f(x_i, \omega_i) \tag{2.1}$$

Considering ω_i a parameter, then Eq. 2.1 describes a noisy system or a system with forcing on the parameters. For $f = f_{\omega_i}(x_i)$ denoting the application of a different map (evolution) at a different time step, one can describe a deterministic system with additive dynamical noise by

$$f_{\omega_i}(x_i) = f(x_i) + \omega_i \tag{2.2}$$

In noisy systems, ω_i is randomly chosen w.r.t. a probability measure μ on N.

For the general case where M is compact, N can be the space of all maps on M. But N can also be a discrete space consisting of a finite number of points, such that f_{ω_i} is selected from a finite set of maps. Or N can be a compact manifold.

To deal with arbitrarily forced systems, one possible approach to enlarge the state space of non-autonomous systems like those described by Eq. 2.1 is to use shift spaces.

Let $\Sigma = N^{\mathbb{Z}}$ be the space of bi-infinite sequences, $\omega = (..., \omega_{-1}, \omega_0, \omega_1, ...)$, of elements in N with the **product topology**. Assuming that N is compact, then Σ is also compact due to the Tychonoff theorem.

Definition 2.2.1 (Product topology). *Let $(X_\lambda)_{\lambda \in L}$ be a family of topological spaces indexed by $\lambda \in L$, $p_\lambda : X \to X_\lambda$ a **canonical projection** and $X = \prod_\lambda X_\lambda$ the cartesian product of $(X_\lambda)_{\lambda \in L}$. Then, a product topology is the topology with the fewest open sets such that all p_λ are continuous $\forall \lambda$.*

Definition 2.2.2 (Canonical projection). *Let Y be a subspace of a vector space X and $f \in X$. Then, the canonical projection of X onto X/M, $\pi : X \to X/M$, is given by*

$$\pi(f) = f + M$$

Theorem 2.2.3 (Tychonoff theorem). *If $(X_\lambda)_{\lambda \in L}$ is a family of compact spaces, then the **cartesian product** $X = \prod_\lambda X_\lambda$ is compact.*

Then, one can define a shift map $\sigma : \Sigma \to \Sigma$, $[\sigma(\omega)]_i = \omega_{i+1}$, where ω_i is the i-th component of $\omega \in \Sigma$. Then, the evolution of $x_i \in M$ (Eq. 2.1) can be expressed by a skew product $T : M \times \Sigma \to M \times \Sigma$, where

$$T(x, \omega) = (f(, \omega_0), \sigma(\omega)) \tag{2.3}$$

Definition 2.2.4 (Cartesian product). *Let X and Y be two sets. Then, the Cartesian product $X \times Y$ is:*

$$X \times Y = \{(x, y) \mid x \in X,\, y \in Y\}$$

This describes a general model of systems driven by arbitrary sequences.

This approach can also be applied to irregularly sampled data. In this case, ω_i represents the time between two consecutive samples x_i and x_{i+1}.

Random dynamical systems are a type of arbitrarily forced systems. For T' a delay reconstruction of a random dynamical system T, then T' is equivalent to the original dynamical

system T under an invertible coordinate change $H : M \times \Sigma \to M \times \Sigma$, if

$$T' = H \circ T \circ H^{-1}. \tag{2.4}$$

This equivalence might not hold for all $\omega \in \Sigma$. In this case, one may ask Eq. 2.4 to hold for μ_Σ-almost every ω, in the probabilistic setting where μ_Σ is a σ-invariant measure. One may also ask for only generic ω in a topological setting.

Since, in general, the space defined by $M \times \Sigma$ is infinitely dimensional, Stark et al. [112] restrict the reconstruction of the dynamical system to the reconstruction of M only, which can be understood like asking the Σ component of H to be the identity. This means, considering $H = (h, Id)$, where Id is the identity map and $h : M \times \Sigma \to M$.

Additionally, Eq. 2.4 might hold for only typical ω. This means, for μ_Σ-almost every forcing ω in the probabilistic setting or for generic ω in the topological setting. In these cases, $h_\omega = h(\cdot, \omega) : M \to M$ is an invertible map and this coordinate change is called a *bundle conjugacy*.

In the standard Takens' theorem, a delay map $\Phi : M \to \mathbb{R}^d$ is generically an embedding and it thus defines a coordinate change $F = \Phi \circ f \circ \Phi^{-1}$ on $\Phi(M)$. For forced systems, a delay map Ψ depends on ω and thus, for every ω, $\Psi_\omega : M \times \Sigma \to \mathbb{R}^d$. This means that Ψ is a bundle embedding and Ψ_ω is an embedding for typical ω. in this case, the range of $H = (\Psi, Id)$ is the reconstruction space $\mathbb{R}^d \times \Sigma$ and T (Eq. 2.4) is defined on $H(M \times \Sigma)$, which is bundle diffeomorphic to $M \times \Sigma$.

If observing $T(x, \omega)$ (Eq. 2.3) with a measurement function $\psi : M \to \mathbb{R}$, such that $\psi_i = \psi(x_i)$, then a delay map is given by

$$\Psi_{f,\psi}(x, \omega) = \big(\psi(x), \psi(f_{\omega_0}(x)), \psi(f_{\omega_1\omega_0}(x)), ..., \psi(f_{\omega_{d-2}...\omega_0}(x))\big)^\dagger, \tag{2.5}$$

where $f_{\omega_i...\omega_0} = f_{\omega_i} \circ \cdot \circ f_{\omega_0}$ and $\Psi_{f,\psi}(x, \omega) : M \times \Sigma \to \mathbb{R}^d$, is a bundle embedding for typical ω in the case where we have finitie-dimensional deterministic forcing.

Let $\mathcal{D}^r(M \times N, M)$ be the space of maps $f : M \times N \to M$ such that $f_y : M \to M$ is a $\mathcal{C}^r$ diffeomorphism of M for any y and $f_y(x) = f(x, y)$. Then, the Takens' theorem for stochastic systems and the Takens' theorem for iterated function systems can be formulated as follows.

Definition 2.2.5 (Lebesgue measure). *For $E \subset \mathbb{R}$, the Lebesgue measure of E is equal to*

$$\mu(E) = \inf\left\{\sum_{j=1}^{\infty} |I_j| : E \subset \bigcup_{j=1}^{\infty} I_j\right\},$$

where I_j are bounded intervals, if for every $A \subset \mathbb{R}$:

$$\mu(A) = \mu(A \cap E) + \mu(A \cap E^C)$$

Theorem 2.2.6 (Takens' Theorem for Stochastic Systems [114]). *Let M be a compact manifold of dimension $m \leq 1$ and N a compact manifold of dimension n. Suppose that $d \leq 2m + 1$ and let μ_Σ be an invariant measure on $\Sigma = N^{\mathbb{Z}}$ which is absolutely continuous w.r.t. the* ***Lebesgue measure*** *on N^{d-1}. Then, for $r \leq 1$, there is a residual set of $(f, \psi) \in \mathcal{D}^r(M \times N, M) \times \mathcal{C}^r(M, \mathbb{R})$ such that for any (f, ψ) in this set, $\Psi_{f,\psi,\omega}$ is an embedding for μ_Σ-almost every ω.*

Theorem 2.2.7 (Takens' Theorem for Iterated Function Systems [114]). *Let M and N be compact manifolds of dimension $m \leq 1$ and $n = 0$ respectively. If $d \leq 2m + 1$ and $r \leq 1$, then there exists an open dense set of $(f, \psi) \in \mathcal{D}^r(M \times N, M) \times \mathcal{C}^r(M, \mathbb{R})$ such that for any (f, ψ) in this set, $\Psi_{f,\psi,\omega}$ is an embedding for $\omega \in \Sigma$.*

In summary, just as in the standard Takens' theorem, all of these results are only valid for generic f and ψ. In theory, it is possible to reconstruct the dynamics of a random dynamical system using successive observations of ψ. Although, in these cases, it is necessary to know ω, which makes the reconstruction mostly impossible.

For more information about the stochastic extensions of Takens' theorem, see [112, 113, 114]. For an extension of Takens' theorem to non-uniformly sampled dynamical systems, see [64].

2.3 Current criteria for a good state space reconstruction

Let us consider the scenario in which we observe partial information of a system and we want to use it to recover the full dynamics of the system. In this case, we may use delay-coordinate maps to obtain a reconstruction of the state space from the partial information.

To illustrate this situation, we will use the widely studied Lorentz system, a non-linear, deterministic system introduced by Edward Lorentz in 1963 in order to model the phenomenon of atmospheric convection. This simple system may show chaotic or stationary behavior depending on the parameters and initial conditions selected. It is described by the following system of ordinary differential equations:

$$\frac{dx}{dt} = \sigma(y - x), \qquad \frac{dy}{dt} = x(\rho - z) - y, \qquad \frac{dz}{dt} = xy - \beta z \tag{2.1}$$

Taking $\sigma = 10$, $\beta = 8/3$ and $\rho = 28$ in Eq. 2.1, the system shows chaotic behavior. In Figure 2.1 we show a time series generated using these parameters and taking some random initial conditions (x, y_0, z_0) with values between 0 and 1. This time series contains all the dynamical information of the system.

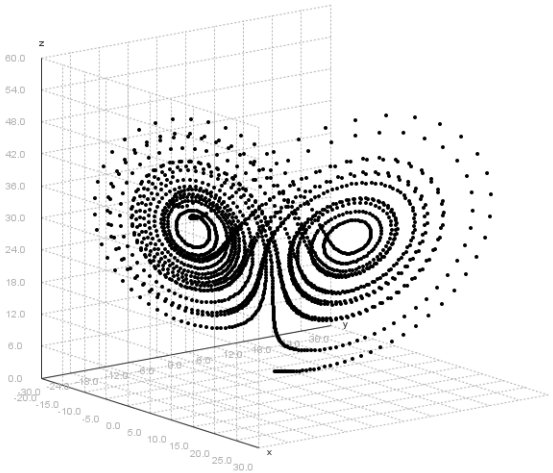


Figure 2.1: A three-dimensional time series for the Lorentz system given by Eq. 2.1, taking parameters $\sigma = 10$, $\beta = 8/3$ and $\rho = 28$. For these parameters, the system has chaotic solutions.

Now, consider the case in which we only observe the x-component of the time series, shown in Fig 2.2.

One may use delay-coordinate maps to create a state space reconstruction that recovers the dynamical information contained in the three-dimensional time series. Let us for example consider the state space reconstruction from the x-component time series using embedding delay $\tau = 15$ and embedding dimension $m = 3$, shown in Fig. 2.3.

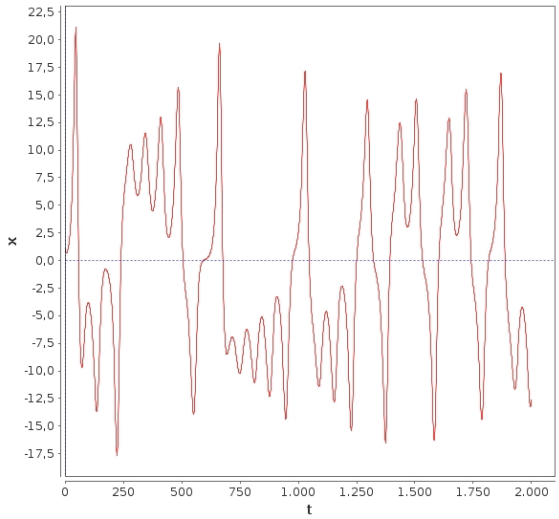


Figure 2.2: x-component of the time series shown in Fig. 2.1, which describes of a Lorentz system showing chaotic behavior.

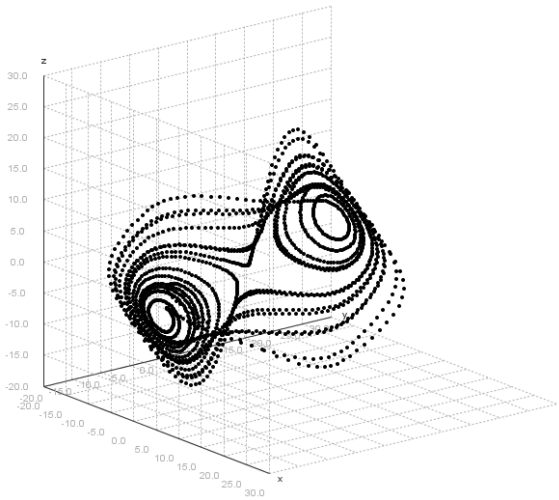


Figure 2.3: A state space reconstruction of the Lorentz system described in Fig. 2.1. This was done using the x-component time series, shown in Fig. 2.2, and a delay-coordinate map with embedding delay $\tau = 15$ and embedding dimension $m = 3$. This reconstruction seem to be adequate since it recovers the two attractors of the system, shown in Fig. 2.1.

This reconstruction seems to be adequate, since the two attractors contained in the original time series, shown in Fig. 2.1, are recovered. But how can we assess how good this reconstruction is?

There are several criteria to identify adequate embeddings. Some of the initial criteria were developed considering the simplest scenario in which the true state space of a system is known and we want to measure how good a state space reconstruction is.

In 2009, Cross and Gilmore [24] introduced a topological notion of equivalence of representations, such that two embeddings are equivalent if they are isotopic. This means that one embedding must be a smooth deformation of the other and thus go to the other through a continuous sequence of embeddings. If two embeddings are not equivalent (non-isotopic), they will not provide equivalent representations of the attractor.

Formally, given two embeddings of a same manifold $M \in \mathbb{R}^N$, $f_0(x)$ and $f_1(x)$, these embed-

dings are isotopic if there is a smooth map $F(x, s)$ on $M \times [0, 1]$ such that $F(x, 0) = f_0(x)$ and $F(x, 1) = f_1(x)$, and $F(x, s) = f_s(x)$ is an embedding for each fixed s [24].

The smoothness of the transformation between an embedding and the true state space can been measured in different ways. For example, Rulkov et al. [102] developed the "mutual false nearest-neighbor statistic", a measurement of smoothness between the reconstructions of two synchronized chaotic time series.

In 2013 Nichkawde [94] stated that a transformation is smooth if both the Lyapunov exponent and the fractal dimension are well evaluated. However, one may not look at these quantities when the true state space is unknown because, as Lettelier et al. [79] showed in 2008, not the Lyapunov exponent nor the fractal dimension can predict whether a mapping is an embedding or not. Additionally, one could face the problem of having spurious Lyapunov exponents.

In the same year, Palit et al. [97] measured how well a state space reconstruction is using a shape distortion parameter. This way, when the true state space is known, a good reconstruction will be the one providing the lowest shape distortion. In general terms, they determine the best reconstruction as the one for which the reconstructed attractor is more dense and has less outliers.

However, the most common case is not knowing the true state space. In these cases, to the best of our knowledge, the best attempt to set criteria for a good state space reconstruction is the one provided by Uzal et al. [119] in 2011. They propose a cost function, L_k, which measures the conditional probability of the future state given the present state. Their measure has the advantage of not needing to know the true state space. This idea originated from the proposition of Casdagli et al. [13] from 1991, which stated that for a good reconstruction, given an original state s and a reconstructed state $\mathbf{x}(t)$, the conditional probability $p[s|\mathbf{x}(t)]$ should be well defined.

In this thesis, our criteria for an adequate state space reconstruction are given in terms of the gain or loss of geometrical information. This is measured in different ways, depending on the approach followed for the reconstruction of the state space: recurrence analysis (see Section 3.4) or topological data analysis (see Section 4.4).

2.4 Two approaches for the analysis of the state space: recurrence analysis and persistent homology

In this thesis we introduce two methods for the identification of different dynamical states or regimes in a complex time series. Both methods rely on the geometrical analysis of different state space reconstructions in order to select an adequate state space reconstruction and later identify the different dynamics of a system.

Given our focus on the geometric analysis of the state space, we first follow the recurrence analysis approach in order to identify different dynamical states in a time series.

A recurrence denotes the return of a state space trajectory, given sufficiently long time, to a state arbitrarily close to a former state. The study of recurrences in state space can be traced back to Poincaré's recurrence theorem, which states the conditions of a system for recurrences to occur. The regions in state space to which a system recurs the most can be associated to the existence of attractors or different dynamical states.

The first method we introduce in this thesis is designed to the identify different metastable states in a complex time series. This method is robust to noise and missing points. In this method we make a state space reconstruction from a time series and then analyze the recurrences in this reconstruction for a filtration of proximity values for the reconstructed state space vectors. We say that a state space reconstruction is adequate if there is a large range of proximity values for which the geometrical structure of the state space is similar. For the analysis of the structure of the state space we make use of network representations and module finding algorithms.

Later, in our second method, we follow the topological analysis approach of persistent homology. In this, we aim to identifying transitions between different dynamical regimes in a system. We introduce this method following the idea of analyzing the state space for different proximity values but overcoming the use of module finding algorithms for the identification of the geometrical structure of the state space.

Persistent homology is an algebraic topological approach theoretically suitable for the analysis of time series where the use of a specific metric and coordinates is not fully justified. It consists on creating a filtration depending on a parameter of proximity between points in a data set, ϵ, and identifying the topological features that survive over different ranges in the filtration.

The main idea for using this approach is that, for a system showing metastability, there will be different topological features of the state space surviving over different ranges of proximity between state space vectors. And those features surviving for longer ranges of proximity may correspond to large scale geometric features, or interesting signals, whereas short ranges may correspond to noise or inadequate sampling.

Within this framework, we say that a state space reconstruction is adequate if the persistent homology results of different subsamples from a same time series are similar. We say that this guarantees the silencing of noise effects and having dense state space reconstructions.

Both of these approaches, as well as the criteria developed for obtaining adequate state space reconstructions within each approach, will be properly introduced in the following chapters.

3. A new method for identifying metastable dynamical states in real-world time series via recurrence networks

In this chapter we introduce a new method for the identification of different metastable dynamical states in a complex time series; this means, in a time series with at least one of the following features: different time scales in its dynamics (or metastability), noise and missing data.

Our method is based on the recurrence analysis of the state space reconstructed from a time series. We follow the idea that the regions in state space to which state space vectors recur the most, correspond to different attractors in the system. Therefore, we construct a *recurrence network* associated to a given proximity parameter that measures the closeness between reconstructed state space vectors−the *recurrence threshold*−, and finally we use a fuzzy algorithm to identify the modules in this network. We suggest that these modules correspond to different metastable dynamical states in the system (see Section 3.2.3 for an extended discussion on metastability).

Before presenting our method, we first need to introduce the concepts in which this is based. In Section 3.1 we introduce the concepts of recurrence analysis, recurrence plots and recurrence networks. We also comment on two recurrence quantification analysis measures we use in our method: Shannon entropy and recurrence rate. Then, in Section 3.2, we comment on the additional considerations needed within the framework of recurrence analysis for analyzing complex time series.

Once a theoretical foundation has been provided, we introduce our method, in Section 3.3. This is divided into three main steps: 1) given a pair of embedding parameters for the state space reconstruction, set an adequate recurrence threshold for the construction of a recurrence network, 2) identify the different modules in such network, and 3) repeat the two previous steps for different embedding parameters in order to select those that provide an adequate state space reconstruction, following the criteria and algorithm mentioned in Section 3.4.

To test the robustness of our method, we analyze the difference in results for time series with different levels of noise and missing points. The methodologies followed for these tests are described in Section 3.5.

Finally, we show the performance of our method with two examples. The first time series analyzed, in Section 3.6, describes a double-well potential system to which a one-dimensional diffusion process has been added. For this example we include the robustness tests and a few comments on modularity. The second example, in Section 3.7, consists on the analysis of a time series describing the changes in the molecular configurations of a molecule of trialanine at low temperature.

As the results of these examples suggest, our method is robust for the analysis of complex time series with considerable levels of noise and missing data.

The method introduced in this chapter, the analysis of the two examples that illustrate it and the tests on robustness, are part of [121], an article written in collaboration with T.O.F. Conrad and Ch. Schütte.

3.1 Introduction to recurrence analysis

The study of recurrences in measure preserving dynamical systems dates back to the phase space studies of Poincaré. Recurrence analysis can be summarized as the study of the regions that the state space trajectories reconstructed from a time series (see Section 2.1.1) cross the most.

As mentioned in Section 1, any dynamical system can be defined by its state space and an evolution operator. When the system underlying a time series is deterministic, even when it might be very difficult, it is feasible to determine its evolution equations and, given an initial point (state space vector), all its future states. In a dissipative dynamical system, on the other hand, any small perturbation of the state at a given time can cause an exponential divergence in its future state. However, as stated by Poincaré in his famous recurrence theorem from 1890 [99], for sufficiently long time, the system will return to a state arbitrarily close to a former state and this return is called a *recurrence*.

Theorem 3.1.1 (Poincaré's recurrence theorem [3, 99])**.** *If a flow preserves volume and has only bounded orbits then, for each open set, exist orbits that intersect the set infinitely often.*

This way, to identify different dynamical states in a system, one may analyze the regions in state space that the dynamical trajectories of a system cross the most. These regions are called *recurrence regions*.

The recurrence analysis approach for studying the state space is interesting because it can help distinguish between deterministic and stochastic systems, transitions to chaos and number of attractors, among other features.

3.1.1 Recurrence plots

Inspired by the work of Poincaré and with the aim of understanding the dynamics of complex data sets, Eckmann et al. [41] developed in 1987 the concept of *recurrence plot*.

A recurrence plot is a tool useful to represent recurrences of state space trajectories to the neighborhood of a set of states. It is defined by a square binary matrix where rows represent each of the state space vectors associated to a time series, and every entry j of row i represents

the closeness between state space vectors $\mathbf{x}(i)$ and $\mathbf{x}(j)$:

$$\mathbf{R}_{ij}(\varepsilon) = \Theta\left(\varepsilon - d\big(\mathbf{x}(i), \mathbf{x}(j)\big)\right) - \delta_{ij} \tag{3.1}$$

In this expression, $\Theta(\cdot)$ is a Heaviside function and $d\big(\mathbf{x}(i), \mathbf{x}(j)\big) = d_{ij}$ is a metric. Along this document, we will use a Euclidean metric. For a detailed explanation of the effects of choosing a different metric, see [35].

The parameter ε is called the *recurrence threshold*. It determines the size of a recurrence neighborhood in state space. This way, choosing different recurrence thresholds may reveal different scales of structure in the state space.

In Fig. 3.1(a) we show a recurrence plot associated to the x-component time series of a Lorentz attractor shown in Fig. 2.2 (for details about this system see Section 2.3). This is computed from the state space vectors reconstructed using embedding parameters $m = 3$ and $\tau = 5$, and recurrence threshold $\varepsilon = 5$.

3.1.2 Recurrence networks

The concept of a recurrence network was introduced in 2008, seemingly independently, by Krishnan et al. [73, 74], Xu et al. [125] and Yang and Yang [128], as the graph representation of a recurrence matrix. Since then, many different definitions of a recurrence network have emerged, differing on the way the recurrence matrix is constructed [37].

The development of recurrence networks introduced the tools of graph theory to the study of recurrences and helped analyze complex time series.

We define a recurrence network, $G(\varepsilon)$, as the graph associated to a recurrence plot $\mathbf{R}_{ij}(\varepsilon)$ constructed using Eq. 3.1 to which the diagonal line has been removed. According to this definition, a recurrence network is unweighted and undirected.

Consider a time series with N data points and a state space reconstruction from it using a delay-coordinate map, with embedding delay τ and embedding dimension m. Then, the recurrence network associated to this time series will have $N^* = N - \tau(m-1)$ nodes, each representing a state space vector reconstructed from the time series. And each of its edges will represent the belonging of a pair of state space vectors to a same recurrence neighborhood.

The structure of a recurrence network depends on the closeness between state space vectors, which is measured by the recurrence threshold. This way, variations in this parameter will produce changes in the network connectivity, characterized by the size and number of its dense groups of interacting nodes, or modules [62]. We assume that modules in a recurrence network correspond to recurrence regions in the state space.

A path in a recurrence network can be interpreted in state space as a trajectory. It means, a sequence of mutually overlapping balls of radius ε, where each ball is defined as $B_\varepsilon(\mathbf{x}(i)) = \{\mathbf{y} \in \mathbb{R}^m : \|\mathbf{x}(i) - \mathbf{y}\| < \varepsilon\}$, for m the dimension of the state space.

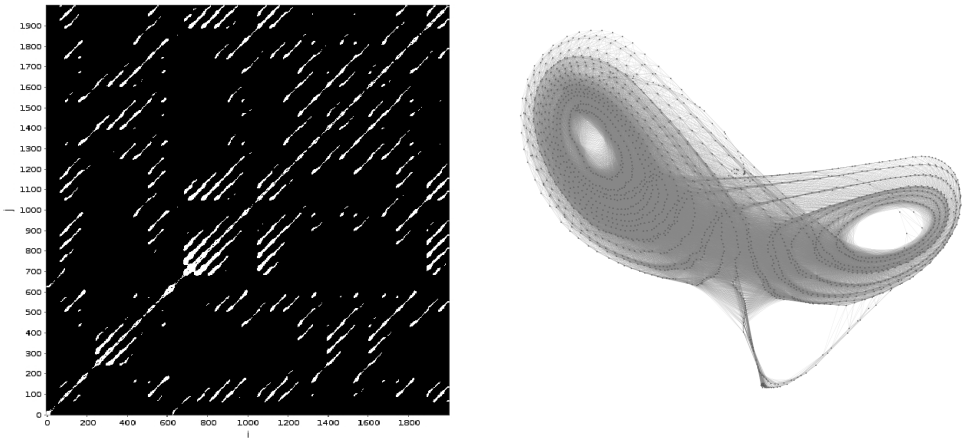
The local, medium and global geometric information of a system can also be recovered from a recurrence network. Donner et al. [36] have provided the definition and meaning of different path- and neighborhood-based measures for recurrence networks.

A drawback of the recurrence networks approach is that all complex network approaches based on the proximity of different parts of the trajectory do not preserve information about

the temporal order of the respective state vectors. This means that complex network approaches are invariant with respect to random permutations of state space vectors. Additionally, considering only the information of a system provided by a recurrence network, one is blinded to higher order topological features of the state space (this problem is addressed in Chapter 4).

However, Donges et al. have suggested that recurrence networks can be adequate for the analysis of non stationary real-world time series [31, 32, 33]. This way, we insist on analyzing them and present in Chapter 3 a method that uses them for identifying different dynamical states in complex time series data.

In Fig. 3.1(b) we show the recurrence network representation of the recurrence plot mentioned in Section 3.1.1, associated to the x-component time series of a Lorentz attractor shown in Fig. 2.2. Here it is interesting to see the similarity between the structure of the recurrence network and the complete three-dimensional time series of the system shown in Fig. 2.1.



(a) Recurrence plot (b) Associated recurrence network

Figure 3.1: (a) Recurrence plot associated to the x-component time series of the Lorentz attractor shown in Fig. 2.2 (for details and parameters, see Section 2.3). This is computed with embedding parameters $m = 3$, $\tau = 5$ and $\varepsilon = 5$. (b) Recurrence network representation of such recurrence plot.

3.1.3 Recurrence quantification analysis (*RQA*)

The study of recurrence plots started being mostly qualitative. One of the main subjects of study was the recovery of dynamical information of a system through the existence of diagonal, vertical or horizontal lines in its recurrence plot. It turned out that it is possible to know wether a system has limit cycles, is stochastic or chaotic from its visual analysis.

One of the results from these qualitative studies is that the diagonal lines in a recurrence plot are associated with the divergence of phase space trajectory segments. A diagonal line of length l in a recurrence plot indicates that two state space trajectories, corresponding to different evolution times, run almost parallel for a time lapse of l time units. This means, $\mathbf{x}_i \approx \mathbf{x}_j, \mathbf{x}_{i+1} \approx \mathbf{x}_i \approx \mathbf{x}_{j+1}, \ldots$

In Fig. 3.2, we show the different diagonal line structures of recurrence plots associated to a periodic, a chaotic and a stochastic dynamical system. There, we represent the stochastic system by pure white noise; the periodic, by a sinusoidal time series with period 5.6; and the chaotic, by a humped **Mackey Glass process**.

Eventually, the study of recurrences became more quantitative and in 1992, J. P. Zbilut and C. L. Webber Jr. [130] introduced the recurrence quantification analysis (RQA) measures. These were either based on the amount of recurrences in a recurrence plot−*recurrence density-dependent measures*−, or on the length and width of diagonal and vertical lines in it−*path-dependent measures*−. The RQA measures broadened the concept of recurrence and opened the door to the potential analysis of high-dimensional time series [35, 86].

Mackey Glass process The Mackey Glass equation is given by:

$$\frac{dx}{dt} = \beta \frac{x_\tau \theta^m}{x_\tau^n + \theta^n} - \gamma x \tag{3.2}$$

where $x_\tau = x(t-\tau)$ and $\tau, \gamma, n > 0$.

This nonlinear time delay differential equation describes a process in which the state at a given time depends on the state at a previous time delayed by time τ. It produces either periodic or chaotic time series depending on the values chosen for the parameters.

Path-dependent RQA measures based on the diagonal lines of a recurrence plot are computed using the histogram of diagonal lines of length l [83], given by

$$P(\varepsilon, l) = \sum_{i,j=1}^{N} \left(1 - \mathbf{R}_{i-1,j-1}(\varepsilon)\right) \left(1 - \mathbf{R}_{i+l,j+l}(\varepsilon)\right) \prod_{k=0}^{l-1} \mathbf{R}_{i+k,j+k}(\varepsilon) \tag{3.3}$$

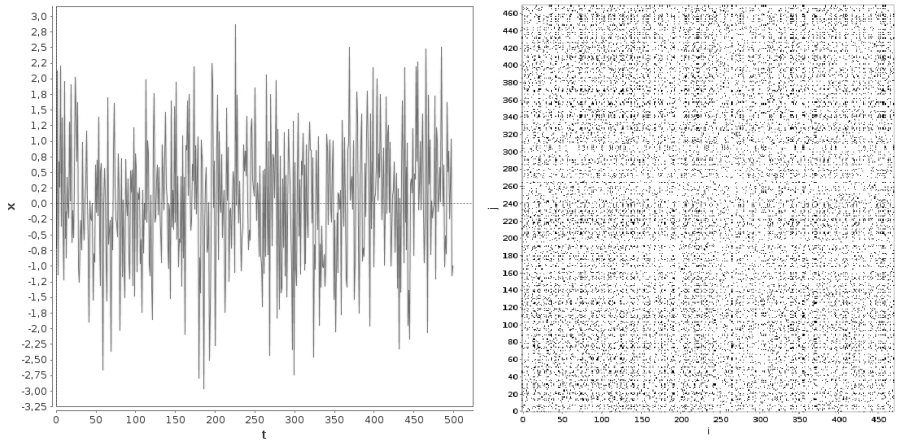
If one decided to use vertical lines instead, the RQA measures are computed using the histogram of vertical lines of length ν [83], given by

$$P(\varepsilon, \nu) = \sum_{i,j=1}^{N} \left(1 - \mathbf{R}_{i,j}(\varepsilon)\right) \left(1 - \mathbf{R}_{i,j+\nu}(\varepsilon)\right) \prod_{k=0}^{\nu-1} \mathbf{R}_{i,j+k}(\varepsilon) \tag{3.4}$$

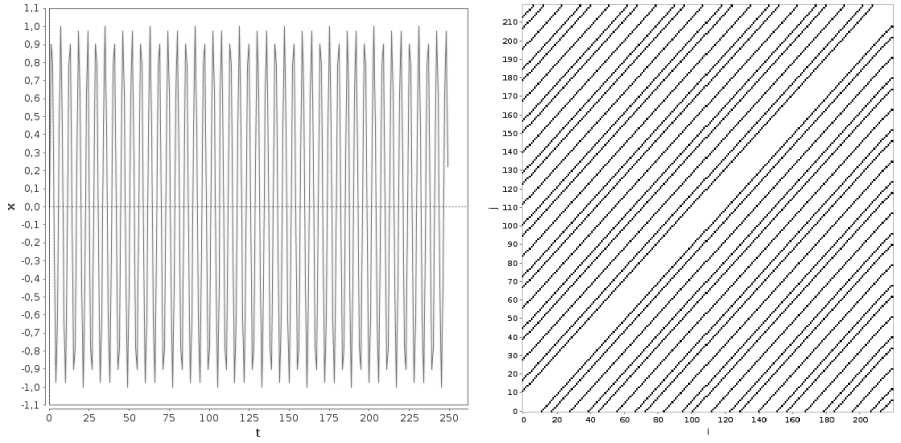
An interesting characteristic of RQA measures based on vertical lines is that these are able to identify chaos-chaos transitions [85].

Around a decade after the introduction of the RQA measures, Marwan et al. [84, 87] had provided a comprehensive geometrical interpretation of these in phase space. Their results were later used to show that recurrence plots were a convenient tool to analyze non-linear [85] and non-stationary [19, 31, 82] time series as well.

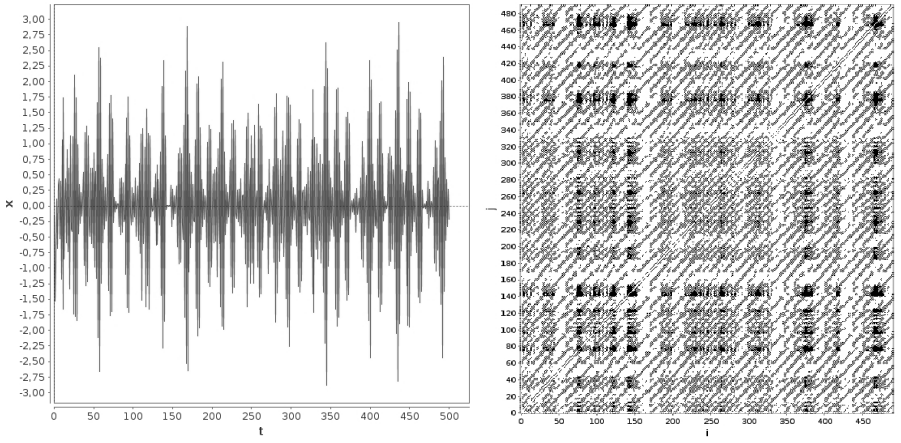
A selection of the most commonly computed RQA measures, their definition and interpretation in state space, is given in table 3.1. However, for more measures we refer the reader to the PhD Thesis of N. Marwan [82]. Two of the RQA measures included in Table 3.1 are necessary for the understanding of the following chapters: the recurrence rate and the Shannon entropy. The former is a recurrence density dependent measure and the latter depends on the diagonal lines of a recurrence plot. We will expand on these in the following sections.



(a) Stochastic: Pure white noise (b) RP associated to (a)



(c) Periodic: Sinusoidal time series (d) RP associated to (c)



(e) Chaotic: Mackey Glass system (f) RP associated to (e)

Figure 3.2: Three different time series and their associated recurrence plots (RP). These represent a periodic, a stochastic and a chaotic system. In (a) we see a time series corresponding composed of pure white noise and in (b) its associated RP. (c) shows a periodic sinusoidal time series, with period $p = 5.6$ and (d) its associated RP. Finally, in (e) and (f) we see, respectively, the time series and RP of a Mackey-Glass humped system showing chaotic dynamics (see text), where we set the parameters in Eq. 3.2 to $\gamma = 0.8$, $\beta = 1.2$, $n = 10$, $\theta = 1$ and $\tau = 7$.

Table 3.1: RQA measures more commonly analyzed

Measure	Expression	Interpretation in state space
Determinism	$DET = \frac{\sum_{l=l_{min}}^{N} lP(l,\varepsilon)}{\sum_{i,j=1}^{N} \mathbf{R}_{ij}(\varepsilon)}$	Percentage of diagonal points creating diagonal hyper-surfaces.
Average diagonal length	$L = \frac{\sum_{l=l_{min}}^{N} lP(l,\varepsilon)}{\sum_{l=l_{min}}^{N} P(l,\varepsilon)}$	Average time two segments of state space trajectory are close. Mean prediction time.
Divergence	$DIV = \frac{1}{L_{max}}$	Inverse of the maximum diagonal length, L_{max}. Exponential divergence go a state space trajectory.
Laminarity	$LAM = \frac{\sum_{\nu=\nu_{min}}^{N} \nu P(\nu,\varepsilon)}{\sum_{\nu=1}^{N} \nu P(\nu,\varepsilon)}$	Percentage of diagonal points creating vertical hyper-surfaces.
Trapping time	$TT = \frac{\sum_{\nu=\nu_{min}}^{N} \nu P(\nu,\varepsilon)}{\sum_{\nu=\nu_{min}}^{N} P(\nu,\varepsilon)}$	Mean time the system will stay (be trapped) at a specific state.
Recurrence rate	$RR(\varepsilon) = \frac{1}{N^2} \sum_{i,j=1}^{N} \mathbf{R}_{i,j}(\varepsilon)$	Fraction of recurrences of a state to its ε-neighborhood in a recurrence plot.
Shannon entropy	$S(\varepsilon) = -\sum_{l=l_{min}}^{N^*} \frac{P(l,\varepsilon))}{N_l} \ln \frac{P(l,\varepsilon)}{N_l}$	Measure of complexity that characterizes distributions of statistical variables[a].

[a] For details about this measure, see text in subsection "Shannon entropy".

Recurrence rate

One of the two RQA measures we will use in our method, described in Section 3.3, is the recurrence rate.

The recurrence rate indicates the fraction of recurrences of a state to its ε-neighborhood in a recurrence plot $\mathbf{R}_{i,j}(\varepsilon)$. It is given by

$$RR(\varepsilon) = \frac{1}{N^2} \sum_{i,j=1}^{N} \mathbf{R}_{i,j}(\varepsilon) \tag{3.5}$$

As can be seen from Eq. 3.5, when the number of time points N tends to infinity, the recurrence rate approximates the probability that a state recurs to a neighborhood of radius ε in state space. Therefore, it has be thought as an estimation of the correlation integral [57].

In terms of the recurrence network, the recurrence rate indicates the average contribution of a node to the relative frequency of edges [36]. This way, higher values in this measure indicate that the nodes are more connected. Or, in other words, that a larger number of state space vectors are inside a same state space neighborhood.

Shannon entropy

The second RQA measures we will use in our method, described in Section 3.3, is the Shannon entropy. As classically defined, it is a measure of complexity expected to increase

with the development of chaotic behavior in a system. Therefore, it is expected to indicate the transitions between non-chaotic and chaotic regimes.

There are various definitions of entropy besides the Shannon entropy, for example the Kolmogorov-Sinaï (KS) entropy and the second order Rènyi entropy (K_2). For more information, we refer the reader to references [2, 49, 77, 78, 129]. In geometric terms, entropy indicates the rate with which nearby trajectories diverge. This way, one can see that it should be closely related to the largest Lyapunov exponent.

The KS entropy is an invariant measure of complexity and was defined in terms of recurrence plots by Baptista et al in 2010. [2]. However, one of the aims of entropy measures is for these to positively correlate with the larges Lyapunov exponent, and the KS entropy only constitutes a lower boundary for the sum of all positive Lyapounov exponents. Additionally, calculating the KS entropy for complex systems is sometimes an extremely complicated task.

The second order Rènyi entropy, or correlation entropy K_2, has the advantage of being easier to calculate, since it can be computed from the correlation integral of a time series data [49]. It indicates the existence of periods of time in which some trajectories evolve almost parallel in a tube of radius equal to the recurrence threshold, or ε-tube, and can thus be related to the cumulative frequency distribution of the lengths of the diagonal lines in a recurrence plot. However, the K_2 entropy is a lower bound to the KS entropy and, therefore, is also related to the lower limit of the sum of the positive Lyapunov exponents [85], instead as to the largest Lyapunov exponent.

The original definition of Shannon entropy from a recurrence plot, $\mathbf{R}_{ij}(\varepsilon)$, is given by

$$S(\varepsilon) = -\sum_{l_{min}}^{N^*} p(l) \log p(l), \tag{3.6}$$

where $N_l = \sum_{l \geq l_{min}} P(l, \varepsilon)$ and $p(l) = P(l, \varepsilon)/N_l$, for $P(l, \varepsilon)$ the histogram of diagonal lines of length l, as defined in Section 3.1.3.

The length l_{min} is a lower boundary for the diagonal lines. The introduction of this boundary is intended to exclude all diagonal lines formed by short-time correlations between state space trajectories. A good selection of l_{min} should preserve only those diagonal lines created by correlations due to the geometry of the attractor in the state space. In addition, it is intended to remove the effects of noise, since in words of Marwan et al. [85], noisy time series produce recurrence plots with many short and thin diagonal lines and single points. For simplicity, we define l_{min}, as the first local maximum in the frequency distribution of diagonal line lengths.

According to Eckman et al. [41], the lengths of the diagonal lines in a recurrence plot should be related with the inverse of the largest positive Lyapunov exponent. However, the Shannon entropy as defined in Eq. 3.6 does not always increase in the presence of chaotic behavior. In addition, it does not show a consistent behavior in its correlation with the largest Lyapunov exponent: it tends to be positively correlated but in many cases it is negatively correlated. This makes this measure unreliable to indicate transitions between different dynamical regimes. Additionally, having a finite number of data points in a time series largely affects the probability of occurrence of diagonal line segments of different lengths. These issues are largely discussed in [47].

To overcome the problem of anti-correlation with the Lyapunov exponent, Letellier [78] introduced a new definition of entropy. Instead of using the relative frequency of occurrence of diagonal lines of length l of recurrent points in a recurrence plot in Eq. 3.6, Letellier redefined

this relative frequency in terms of the "non-recurrent points". However, this definition does not correspond to the classical statistical physics definition of entropy for physical systems.

Therefore, Eroglu et al. [47] introduced another definition based on weighted recurrence plots that do not require the selection of a recurrence threshold. For discrete systems, this definition correlates positively with the largest Lyapunov exponent and seems to adequately identify transitions between different dynamical regimes. For continuous systems, it shows better results than Letellier's entropy for the identification of transitions between different dynamical regimes, and similar results to those obtained using the classically defined Shannon entropy, although in this case, Eroglu's entropy is better correlated to the behavior of the largest Lyapunov exponent.

3.1.4 Recurrence threshold selection

The problem of selecting the recurrence threshold that originates the recurrence plot or network providing more dynamical information of a system has been largely studied. When the recurrence threshold is not well set, the fine geometry of the dynamical system is not well represented neither by the neighborhood- nor by the path-measures. A summary of the problems associated to the selection of the recurrence threshold is given by Donner et al. in [38]. An example of such sensitivity is shown in Fig. 3.3.

Initially, the recurrence threshold was set "using rules of thumb" [38] based on the diameter of the state space [90, 130], the recurrence rate [131] or the derivative of the recurrence rate [54]. However, the recurrence structure obtained with these approaches was very sensitive to small variations of the threshold.

In [35, 36], Donner et al. select a recurrence threshold based on previous studies over dynamical measures such as the correlation integrals [132], correlation dimensions [57] or second order Rènyi entropy [109, 118], and attractor dimensions [38]. They arrive to some restrictions for the recurrence threshold which can be summarized as requesting it to be as small as possible while preserving a low edge density in its associated recurrence network (even in the presence of noise), given that higher edge density values tend to hide important dynamical features.

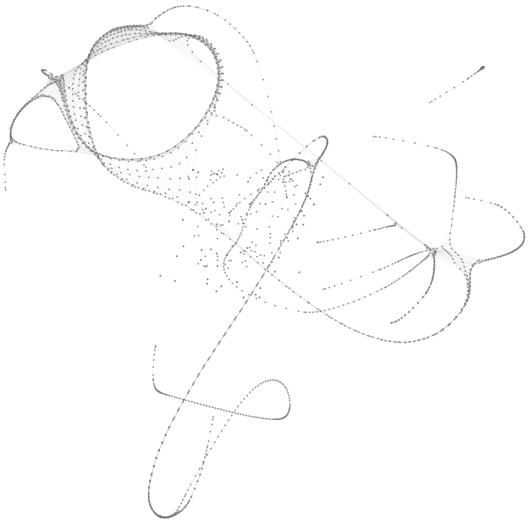
Feldhoff et al. [50] also selected recurrence thresholds that produced recurrence networks with low edge densities, but additionally selected recurrence thresholds that guaranteed that a small variation in the recurrence threshold did not produce noticeable differences in the dynamical analysis results. These restrictions led to use lower recurrence thresholds since they implied using smaller neighboring distances for the analysis of the state space.

Random geometric graph (RGG)

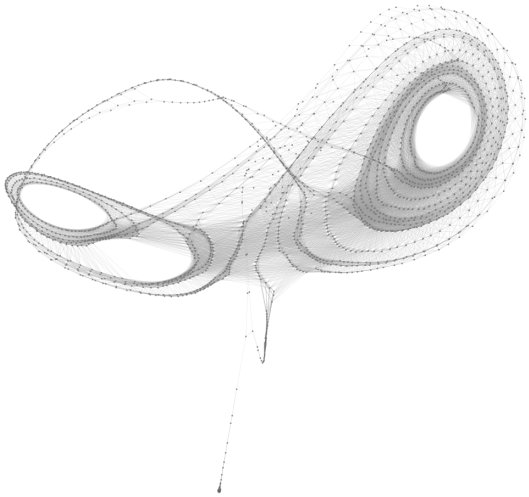
According to J. Dall and M. Christensen [25], a random geometric graph, $\mathcal{G}(\mathcal{X}_n^{(n)}; r)$, is a graph whose nodes' locations are denoted by independent and identically distributed (i.i.d.) variables in $\mathbb{R}^n$, with common probability density f. The undirected links or edges between nodes in a random graph are determined by geometric proximity $r > 0$, measured with a particular norm.

The giant component of a random geometric graph is its unique largest connected component containing a constant fraction of the nodes.

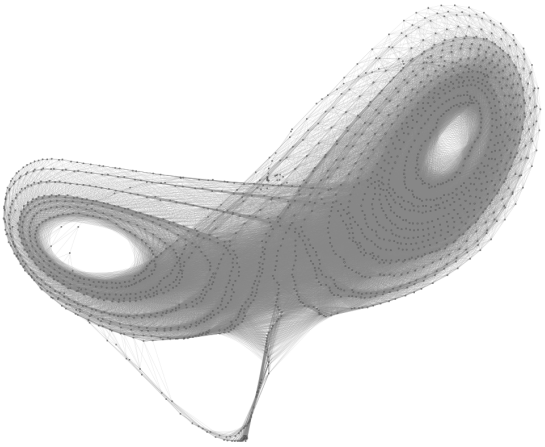
In 2012, Donges et al. [34] introduced an analytical framework for the study of recurrence networks based on **random geometric graphs** (RGG) theory. This way, they addressed the problem of selecting an appropriate recurrence threshold for one-dimensional, non-noisy time series with uniform probability density distribution. By introducing RGG theory, they found that the



(a) $\varepsilon = 1.5$



(b) $\varepsilon = 3$



(c) $\varepsilon = 5$

Figure 3.3: Difference in structure between the recurrence networks produced with different recurrence thresholds for a same state space reconstruction of the Lorentz system described by Eq. 2.1 in Section 2.3. The embedding parameters used for the state space reconstruction are $m = 3$ and $\tau = 5$. An adequate recurrence threshold selection should have an associated recurrence network resembling the structure of the three-dimensional time series shown in Fig. 2.1, which capture the entire dynamics of the system. In these terms, we see that the recurrence threshold used in Fig. 3.3(a) is not adequate, since its associated recurrence network has no structure. The recurrence threshold used in Fig. 3.3(b) is better, but the one used in Fig. 3.3(c) is much more similar to the structure of the time series shown in Fig. 2.1.

recurrence threshold, ε, can be set in terms of the the percolation threshold ϵ_c of the associated recurrence network.

The percolation threshold, ϵ_c, points out the limit in which the network's giant component breaks down and makes impossible to recover information about mesoscopic and path-based measures [98]. For $\varepsilon \ll \epsilon_c$ the recurrence network becomes too dense, and for $\varepsilon < \epsilon_c$, the recurrence network's giant component breaks down into smaller disconnected components.

Donges et al. also approached the problem of the metric selection for the construction

of a recurrence plot. For this, they used the average path length of a RGG, which is a global path-based measure that gives the geodesic distance (in ε units) between two state space vectors chosen randomly according to a specific probability distribution p. Despite the promising results provided by the introduction of RGG theory, this could only be used to analyze one-dimensional, non-noisy time series with uniform probability density distribution. Unfortunately, since there are still no exact analytical results for d-dimensional random geometric graphs of arbitrary d, Donges et al. [34] had to return to the results of Donner et al. [36, 38] for general cases.

A possible extension of the work of Donges et al. could be provided by considering the results of Kong and Yeh [70]. These authors have investigated the problem of characterizing the critical density and critical mean degree of random geometric graphs with non-uniform probability distributions and, based on probabilistic methods and clustering analysis, have provided lower bounds for the critical density of a Poisson RGG in an m-dimensional Euclidean space.

Other approaches to the analysis of networks computed from time series with non-uniform distributions are the study of changes in connectivity of recurrence networks by Hsing and Rootzé [60], and more recently by Cooper and Frieze [23].

However, despite the studies mentioned above, the problem of analyzing time series with non-uniform probability distribution has not been fully addressed. This, in turn, implies that selecting an appropriate recurrence threshold for real-world time series is still an open problem [38, 108], since real-world time series do not tend to have uniform probability distributions.

3.2 Considerations for the analysis of complex time series – *With recurrence analysis*

Real-world time series tend to have noise, miss some measurement points and show metastability. In the following sections we will comment on the consequences of these features on the recurrence analysis of dynamical systems.

3.2.1 Noise

According to J. P. Zbilut and C. L. Webber Jr. [130], having noise in a time series tends to inflate the embedding dimension when reconstructing the state space. We suggest that this could in turn derive in an increase of the recurrence threshold adequate to analyze a dynamical system.

In an article of 2007, Marwan et al. [85] proposed some conditions to choose the threshold when analyzing one-dimensional noisy time series. Their reasoning was based on the distinction of false positive and false negative recurrences.

A false negative is a recurrence point which is not identified as such with probability $1 - p_b$, where p_b is the percentage of recurrence points that are properly identified in the presence of noise. A false positive is a non-recurrence point which is recognized as a recurrence point with probability $1 - p_w$, where p_w is the percentage of properly recognized non-recurrence points. This way, Marwan et al. proposed that an adequate recurrence threshold, ε, should

maximize p_b and p_w simultaneously. According to numerical simulations, they concluded that ε should be at least five times the standard deviation of the observational noise.

Later, in an article from 2013, Marwan et al. [86] established some boundaries for the recurrence threshold when analyzing high-dimensional time series data with noise.

3.2.2 Missing points

This problem can also be understood as the one of not having uniform sampling when obtaining a time series. In general, the construction of recurrence networks for this type of data follows the typical procedures of omitting the missing measurements or, when the time series is long, subsampling the time series in order to produce a uniformly measured process.

To this respect, Donges et al. [32, 33] have analyzed the ability to detect nonlinear dynamical transitions in time series with irregular sampling with recurrence networks.

3.2.3 Metastable dynamical states

When a system or physical phenomena has multiple and well distinguished time scales, it is said to show metastability. This means, that for short time scales, the system appears to be in equilibrium, which makes possible to identify the so called *metastable states*. And for longer time scales, it seems to transit between different metastable states, which makes possible to identify the so called *transition region*.

The study of metastability as a phenomenon occurring in stochastic processes was introduced by H. A. Kramer in 1940 [71]. The time series he used to analyze metastability in the context of chemical reactions, corresponds to a double-well potential to which a one-dimensional diffusion process is added (we analyze this system in Section 3.6). For more information on metastability, see the work of A. Bovier [6, 7].

One can expect metastability in the systems underlying several real-world time series. However, the analysis of this characteristic in recurrence networks – to the best of our knowledge – had not been addressed yet.

A natural question when analyzing metastability with recurrence networks is how to distinguish between trajectory bundles and metastable states. To answer this, we refer to the analysis of the topological structure of a network via random walks done by Maila and Shi [81], Deuflhard and Weber [28] and Sarich et al. [30, 104], among others.

According to this approach, by running random walks in a modular network, it is possible to analyze both its local and its global topological structure. Considering that modules in a network correspond to metastable states of a random walk, then the metastable states can be identified via the spectral properties of the Markov processes.

This way, to identify metastable states in a time series, we analyze the modular structure of the constructed recurrence network using the fuzzy clustering method developed by Sarich et al. [30, 62, 104, 105].

Module finding algorithm of Sarich et al. [104].

This algorithm is based on the spectral analysis of time-continuous random walks on modular networks and uses transition rules of the random walk which increase the spectral gaps.

It also considers modules as groups of densely connected nodes that do not partition the network completely and do not overlap, which gives rise to the identification of a transition region with contains all nodes which are not assigned to a specific module. Contrary to other density based clustering methods, this module finding method distinguishes between dense state space regions corresponding to metastable states and dense state space regions corresponding to the passage between them.

The quality of the results of this method for module identification rely on the quality of the approximation of the dominant eigenvalues. Its error is analytically defined and therefore the reproducibility of its results guarantee the reproducibility of our method for selecting embedding parameters and identifying metastable states in a time series.

In computational terms, this algorithm scales linearly with the size of the network, making it also useful for analyzing large networks. In order to keep the sum of nodes assigned to all modules equal to N^*, we assign all nodes identified as part of the transition region to an additional module. One consideration when using this method is that it does not work for disconnected or fully connected networks.

This method does not work for disconnected or fully connected networks. However, this fact was taken into account for defining the recurrence threshold values in the filtration.

3.3 *A new method for identifying metastable states in real-world time series*

As we have seen in the previous sections, recurrence analysis has been used in some cases for the identification of different dynamical regimes−from chaotic to non-chaotic or chaotic-chaotic transitions−in complex time series. However, given the constrains of sensitivity of the RQA measures, this approach has not been widely used for the identification of the different dynamical states within a same regime.

In this section we introduce our method for selecting adequate embedding parameters and recurrence threshold that, we suggest, allows the better identification of metastable states in real-world time series data.

Assuming we have already reconstructed the state space from a time series data (we use a new method described in Section 3.4, based on the methods described in Section 2.1.2), we first create a filtration defined by the recurrence threshold. This filtration ranges from the 50th to the 95th percentiles of the distances between state space vectors in the reconstructed state space.

We analyze the modular structure of the recurrence networks associated to the recurrence thresholds in the filtration. The modular structure in the recurrence networks is identified with the method of Sarich et al. [104], which performs a fuzzy clustering of the state space with respect to metastability (for details see [30, 62, 105]).

Then, we assume that an adequate recurrence threshold should belong to a region in the filtration for which the associated recurrence networks are less dissimilar. Therefore, we identify those networks whose modular structure, meaning their number and size of modules, is within some boundaries of similarity given by Eq. 3.2, Eq. 3.3 and Eq. 3.4.

We set an adequate recurrence threshold as the average of recurrence thresholds belonging to the region of the filtration satisfying such expressions.

3.3.1 Setting an appropriate recurrence threshold

It is known that even small variations in the recurrence threshold can lead to very different modular structure in its associated recurrence network. For this reason, some recurrence threshold selections can hide important dynamical features in a system.

We suggest that an adequate recurrence threshold should lie in a region of values producing recurrence networks with less dissimilar modular structures. Since this region of values varies according to the distribution of our data, we propose a methodology, summarized in Algorithm 1 to set the recurrence threshold that adapts to it.

Our approach is inspired by the analysis of data with persistent homology [9, 12]. Persistent homology has been briefly explained in Sections 2.4 and 4.1.4, and is presented in depth later on, in Chapter 4. This approach arose from the problem of computing the homotopy type of an underlying topological space from point cloud data that does not uniformly sample the space. Thus, persistence homology captures the persistence of topological entities in a filtration. The connections between persistent homology and delay-coordinate embeddings have been recently investigated, for example by Emrain et al. [46].

In our case, the state space can be understood as the topological space of interest and the recurrence threshold as the one-parameter defining the filtration. The modular structure of the recurrence networks can be thought as the topological entity whose persistence is analyzed.

This way, in order to set an appropriate recurrence threshold, we first construct a filtration: a set of recurrence thresholds and its associated networks. Then, we analyze the modular structure of each network in the filtration. Finally, we identify a subset of thresholds for which the modular structure of their associated recurrence networks is the least dissimilar. The similarity in modular structure depends on the number and size of the modules identified in every recurrence network. Finally, we select a recurrence threshold equal to the average value of the recurrence thresholds in such subset of thresholds.

Algorithm 1 - SETTING AN APPROPRIATE RECURRENCE THRESHOLD

1. Construct set of recurrence networks, $\{G_\nu\}$:
   for $\nu = 0$ **to** $\nu = \nu_f$ **do**
   ▷ Compute recurrence threshold ε_ν according to Eq. 3.1
   ▷ Compute associated recurrence plot $R(\varepsilon_\nu)$ and recurrence network $G_\nu = G(\varepsilon_\nu)$.
   end for
2. Find modular structure in all recurrence networks in $\{G_\nu\}$: $\{C(G_\nu)\}$ and $\{|C_k(G_\nu)|\}$
   for $\nu = 0$ **to** $\nu = \nu_f$ **do**
   ▷ Perform modular structure analysis of associated recurrence network $G_\nu = G(\varepsilon_\nu)$.
   ▷ Compute number of modules, $C(G_\nu)$, and number of nodes in each module, $|C_k(G_\nu)|$, on G_ν.
   end for
3. Select subset of networks $\{G_\nu\}^-$ with the same number of cluster, satisfying Eq. 3.2, $\{\varepsilon_\nu\}^*$:
   for $\chi_j = \chi_0$ **to** $\chi_j = \chi_{j^*}$ (defined in Eq. 3.4) **do**
   for all $\varepsilon_\lambda \in \{\varepsilon_\nu\}^-$ **do**
   if $|C_k(G_{\lambda+1})| - |C_k(G_\lambda)| < \chi_j$ **then**
   Add recurrence threshold ε_λ to subset $\{\varepsilon_\nu\}^{\chi_j}$.
   end if
   end for
   if $\{\varepsilon_\nu\}^{\chi_j} = \emptyset$ **then**
   $\chi_{j^!} = \chi_{(j-1)}$ and $\{\varepsilon_\nu\}^* = \{\varepsilon_\nu\}^{\chi_{j^!}}$
   else $\{\varepsilon_\nu\}^* = \{\varepsilon_\nu\}^{\chi_j}$
   end if
   end for
4. Set final recurrence threshold, ε^*, as the average value of thresholds in $\{\varepsilon_\nu\}^*$.

Constructing a filtration of recurrence networks

The initial step of our method consists on constructing a filtration defined by the recurrence threshold and its associated recurrence networks.

We want this set of recurrence thresholds to span a large set of distances in state space, so that we can see different structures in the associated recurrence networks. However, we want to avoid negligible recurrence rates or disconnected recurrence networks.

This way, even at the cost of having very high recurrence rates for some recurrence thresholds in the filtration, we compute the initial recurrence threshold, ε_0, as the 95th percentile of the distances between state space vectors. The smaller scale information we analyze is the one visible when the recurrence threshold is set as the 50th percentile of the distances between state space vectors, denoted by ε_f.

Let us denote the set of recurrence thresholds in the filtration by $\{\varepsilon_\nu\}$, where $\nu = [0, \nu_f]$ and $\nu_f = 14$. Then, the ν-th element of $\{\varepsilon_\nu\}$ is given by

$$\varepsilon_\nu = \varepsilon_0 + \nu \left(\frac{\varepsilon_f - \varepsilon_0}{\nu_f} \right) \tag{3.1}$$

The set of recurrence thresholds given by Eq. 3.1 is a suggestion. The range and number of thresholds in the set could be modified according to information on the distribution of the state space vectors. In particular, if the data is uniformly distributed, the initial threshold could be given in terms of the five percent of the standard deviation, as proposed by Marwan et al. [85]. For multidimensional time series, we propose to use the largest standard deviation. In that case, we propose to construct the set of recurrence thresholds as $\varepsilon_\nu = (1.5 - 0.1\nu)\,\varepsilon_0$.

Finally, we compute a recurrence plot for every recurrence threshold in $\{\varepsilon_\nu\}$, $\mathbf{R}_\nu = \mathbf{R}_{ij}(\varepsilon_\nu)$, as well as the associated recurrence networks, $G_\nu = G(\varepsilon_\nu)$. We denote the set containing these recurrence networks by $\{G_\nu\}$.

Analyzing modular structure of the filtration

Every recurrence network in $\{G_\nu\}$ may have a different modular structure. We analyze the number and size of their modules with the aim of finding a subset with similar structure.

The problem of finding modules (or clusters) in complex networks has been approached in several ways and many clustering algorithms exist for this purpose [62]. However, we use the algorithm of Sarich et al. [104] (see Section 3.2.3) because it is specifically developed for the case in which a system shows metastability.

The differences in modular structure between every recurrence network in $\{G_\nu\}$ can be represented with a flow diagram called the Sankey diagram. A Sankey diagram is a visualization tool we use to show the number of clusters and the nodes' distribution for each of the different recurrence networks computed from the set $\{\varepsilon_\nu\}$.

In a Sankey diagram, every network is represented by a column and every column is divided into blocks. The number of blocks in a column represents the number of modules identified in a network. The size of a block in a column corresponds to the number of nodes such module contains.

Let G_ν and $G_{\nu+1}$ be two consecutive recurrence networks. Then, if a group of nodes initially assigned to module A in G_ν is assigned to module B in $G_{\nu+1}$, this *flux* will be represented as an arrow in a Sankey diagram, with a thickness determined by the number of nodes *flowing*. In Fig. 3.7 we show the Sankey diagram for a filtration of recurrence networks from the double-well potential time series analyzed in Section 3.6.

The first similarity requirement on $\{G_\nu\}$ is to have the same number of modules. The subset of recurrence networks satisfying this restriction is denoted by $\{G_\nu\}^-$ and $\{\varepsilon_\nu\}^-$ is the subset of recurrence thresholds generating these networks.

Let $G_\mu \in \{G_\nu\}$ have $C(G_\mu)$ modules. Then, this network will satisfy the restriction on similarity in number of modules if, given three consecutive networks $G_{\mu-1}$, G_μ, $G_{\mu+1} \in \{G_\nu\}$, the following holds

$$C(G_{\mu+1}) = C(G_{\mu-1}) = C(G_\mu) > 1 \tag{3.2}$$

The next similarity requirement is applied on $\{G_\nu\}^-$. It consists on asking the recurrence networks to have modules of similar size, where the level of similarity is expressed by a tolerance value, $\chi_j \in [\chi_0, \chi_{j^*}]$. The subset satisfying this restriction is denoted by $\{G_\nu\}^{\chi_j}$ and the subset of recurrence thresholds producing these networks is denoted by $\{\varepsilon_\nu\}^{\chi_j}$.

Let $C_k(G_\lambda)$ be the k-th module of $G_\lambda \in \{G_\nu\}^-$, and $|C_k(G_\lambda)|$ the number of nodes in such module. Then, the size of the k-th module in a pair of consecutive recurrence networks $G_\lambda, G_{\lambda+1} \in \{G_\nu\}^-$ varies less than χ_j if

$$|C_k(G_{\lambda+1})| - |C_k(G_\lambda)| < \chi_j \tag{3.3}$$

The tolerance value depends on the number of nodes in the recurrence networks, N^*. Initially, we say that two modules have similar size if the number of nodes they contain is different in no more than ten percent of N^*. This means that $\chi_0(N^*) = 0.1N^*$.

By decreasing the tolerance value, we strengthen the condition of similarity between modules. We say that complete similarity is reached when the number of nodes in two modules is different in no more than one percent of N^*, which means that $\chi_{j^*}(N^*) = 0.01N^*$. This way, we define a ten steps procedure, where the tolerance value for each step is given by

$$\chi_j = \chi_0(1 - j/10), \text{ for } j = [0, j^*] \tag{3.4}$$

If the subset of recurrence networks satisfying the maximum decrease of tolerance is not empty, it is denoted by $\{G_\nu\}^* = \{G_\nu\}^{\chi_{j^*}}$. Then, the subset of recurrence thresholds producing these networks is denoted by $\{\varepsilon_\nu\}^* = \{\varepsilon_\nu\}^{\chi_{j^*}}$. However, it is possible that no subset of $\{G_\nu\}^-$ satisfies the maximum tolerance decrease and that $\{G_\nu\}^{\chi_j}$ is empty for a certain $\chi_j > \chi_{j^!}$. In this case, we define $\{G_\nu\}^* = \{G_\nu\}^{\chi_{j^!}}$ and $\{\varepsilon_\nu\}^* = \{\varepsilon_\nu\}^{\chi_{j^!}}$.

Finally, we assume that the modular structure of the recurrence networks associated to all values in the range of $\{\varepsilon_\nu\}^*$ is the least dissimilar, not only the values that we tested. This way, we set the *final recurrence threshold*, ε^*, as the average value of the recurrence thresholds in $\{\varepsilon_\nu\}^*$. Alternatively, ε^* could be set as the minimum threshold in $\{\varepsilon_\nu\}^*$, in order to avoid irregularities in case the average value of $\{\varepsilon_\nu\}^*$ does not belong to such set.

It is worth mentioning that this procedure is equivalent to finding the minimum in the differences of Eq. 3.3 for all lambda, but introducing a tolerance. This tolerance would allow us to include a larger range of values within the recurrence thresholds filtration for the computation of the final recurrence threshold.

3.3.2 Identifying metastable states

Once that the final recurrence threshold ε^* has been set, we generate the recurrence network associated to it, $G_* = G(\varepsilon^*)$. The analysis of the modular structure of this network will lead to the identification of metastable states (and transition region) in the time series. The methodology followed to assign every data point in a time series into different metastable states (or transition region), is summarized in Algorithm 2.

For simplicity, if node i of G_* has been assigned to a specific module C_k, we will assign the data point u_i in the first component of state space vector $\mathbf{x}(i)$ to the k-th metastable state.

This metastable state assignment approach is naïve because, when using the time delay embedding method to construct the state space, every data point appears in a different number of state space vectors. Let $M(u_i)$ denote the number of state space vectors in which data point u_i appears, τ and m be the embedding parameters, and α be an integer such that $0 \leq \alpha < m - 1$. Then, $M(u_i) = \alpha + 1$ if $\alpha\tau \leq i \leq (\alpha + 1)\tau$ or if $N - (\alpha + 1)\tau < i \leq N - \alpha\tau$, and $M(u_i) = m$ for any other data point.

Alternatively, the metastable state a data point u_i is assigned to, could be determined by its *dominant module*. This means, the module to which u_i has been assigned the most. Let $m_k(u_i)$ denote the number of state space vectors in which u_i appears and that are assigned to module $C_k(G_*)$. Then u_i belongs to

$$\underset{C_k(G_*)}{\arg\min} \Big(M(u_i) - m_k(u_i) \Big) \tag{3.5}$$

If there are more than one module satisfying Condition 3.5 for this data point, then there is no dominant module for u_i and we consider it part of the transition region.

Algorithm 2 - Identifying metastable states in a time series

1. Perform modular structure analysis of recurrence network $G_* = G(\varepsilon^*)$. This is done with the fuzzy clustering method of Sarich et al. [104].

2. Classify every data point in the time series into metastable states or transition region, according to any of the following two criteria:
   (a) According to the assignment of the state space vector for which the data point constitutes the first component, or
   (b) According on the dominant module, as given in Eq. 3.5 (see text).

3.4 *A new method for state space reconstruction – Based on recurrence analysis*

In this section we show a new methodology, based on recurrence analysis, for choosing a pair of embedding parameters that provide an adequate state space reconstruction.

As mentioned in the introduction to this chapter, this is intertwined with the method described in Section 3.3–which selects an adequate recurrence threshold for a state space reconstruction–.

This methodology, summarized in Algorithm 3, consists on the following steps. First, for a given embedding dimension, we select the embedding delay that has a simultaneous first local minima in Shannon entropy (defined with respect to diagonal lines, as in Eq. 3.6) and first local maxima in recurrence rate (defined as in Eq. 3.5). Then we choose the pair of parameters which additionally has the lowest Shannon entropy.

Our two main ideas for selecting the embedding parameters are the following. First, that a minima in Shannon entropy indicates the recovery of more dynamical features from a recurrence network. And second, that higher recurrence rate values indicate that the nodes in the recurrence network are more connected, or that a larger number of state space vectors fall inside a same state space neighborhood.

Algorithm 3 - Reconstructing the state space

1. Given a state space reconstruction and a recurrence threshold, compute $RR(\varepsilon)$ and $S(\varepsilon)$:
   for $\tau = \tau_0$ **to** $\tau = \tau_F$ **do**
   for $m = m_0$ **to** $m = m_F$ **do**
   ▷ For embedding parameters τ and m, construct $N^* = N - (m-1)\tau$ state space vectors using the time delay embedding method (see Eq. 2.3) from the normalized time series.
   ▷ Compute recurrence threshold, ε, as explained in Section 3.3.1.
   ▷ Compute its associated recurrence plot $\mathbf{R}_{ij}(\varepsilon)$.
   ▷ Compute $RR(\varepsilon)$ and $S(\varepsilon)$ for the associated recurrence plot, as given in Eqs. 3.5 and 3.6 respectively.
   end for
   end for

2. Select m and τ that first provide a simultaneous local minima in entropy and local maxima in recurrence rate, and that also give the lowest minimum in entropy.

3.5 Robustness tests

We define robustness as the similarity between the metastable states (or modules in the associated recurrence network) identified in two time series: the *original* and a *modified* time series. The modified time series is created by adding artifacts (noise or missing data points) to the original time series.

As we mentioned before, we analyze the modular structure of a recurrence network using the algorithm developed by Sarich et al. [104] (see Section 3.2.3), which divides the nodes into a fuzzy partition consisting of modular and transition regions. The assignment of every state space vector into a different module or to the transition region constitutes a partition.

Then, we measure the similarity between the partitions associated to the original and the modified time series with a modified version to the Adjusted Rand Index [61, 103] (ARI).

The ARI was developed by Hubert and Arabie in 1985 to measure the agreement between two partitions. When the partitions are not similar at all it is equal to zero, and when the partitions are equivalent it is equal to one. This index offers the advantage of being computable even when the number of modules in the two partitions compared is not the same. Additionally, its results are meaningful even if the labeling of the partitions is switched. For more details see Appendix A.

In order to account for the division into modular and transition regions, we use the modification to the ARI proposed by Hueffner et al. [62]. This modified ARI assigns every state space vector identified as part of the transition region into an independent module in order to create a full partition. This can be measured either considering only the modules or considering the modules together with the transition region.

However, this is not the only index that can be used to measure the similarity between partitions. For fuzzy (or soft) partitions in which every object is assigned to various clusters with different weight values, one may use adaptations of the normalized mutual information (NMI) or of the Jaccard index to fuzzy partitions. These measures could potentially substitute the use of the modified ARI of Hueffner et al. in our method.

We measure the robustness of our method for identifying metastable states in two scenarios: when a percentage of noise is added and when a percentage of time points is removed from a time series.

3.5.1 Noise

We use the definition of noise given by Hassona [58] for the analysis of variations in RQA measures. According to this, a noisy time series is created by adding Gaussian white noise, with mean equal to zero and standard deviation equal to one, to the time series.

The amplitude of the noise added is equal to a percentage μ_N of the amplitude of the original time series. This means that the amplitude of the noise is $\mu_N\%$ the amplitude of the time series. We vary μ_N from 0 to 20 in intervals $\Delta\mu_N = 1$.

We generate 50 different noisy time series for every μ_N and calculate the ARI as the average of the ARI obtained for every noisy time series with a same amplitude of noise. The aim of this procedure is to remove the bias induced by the selection of noise.

3.5.2 Missing points

One of the typical features of real-world time series is having observations irregularly taken. This irregularities can be understood as if a percentage of measurement points, randomly distributed in the time series, had been removed from a time series containing a set of measurements regularly taken.

We produce the original time series, with regularly spaced measurements. The modified time series is then obtained by assigning a "null" value to a percentage μ_R of randomly distributed data points in the original time series. We vary μ_R from 0 to 19, in intervals $\Delta\mu_N = 1$. Since we are not ignoring time points but only assigning a new value to some time points, the length of the original and the modified time series is the same.

Again, in order to remove the bias induced by the selection of data points to remove, we analyze 50 different time series with the same number of missing points.

3.6 Example 1: Double-well potential

To illustrate the ability of our method to identify metastable states in complex time series, we analyze the time series shown in Fig. 3.5. It describes the motion of a particle in a heat bath with temperature T, under the gradient of a double-well potential and a random force.

3.6.1 The system

The double-well potential model was proposed by Kramer in 1949 [71], during his studies on chemical reactions and is one of the first models for metastability. It is described by

$$dX_t = -\nabla V(x)dt + \sqrt{2\epsilon}dB_t, \tag{3.1}$$

where B_t is a Brownian motion, $\nu > 0$ is a friction parameter and $\epsilon = \nu T$.

Due to its formulation, this model corresponds to one of the stochastic systems that can be analyzed with the extended Takens theorem proposed by Stark. [112]. But instead of going through the procedure that such analysis requires, we analyze this system with our proposed method.

The potential, given by $V(x) = (x^2 - a^2)^2$, has two local minima at $x_1 = a$ and $x_2 = -a$. Fig. 3.4 shows a representation of the double-well potential we use, where $a = 1$: $V(x) = (x^2 - 1)^2$. In this figure, ΔV is the trap depth difference between the potential wells which controls how metastable the system is.

Our one-dimensional time series, shown in Fig. 3.5, results integrating the system's Langevin dynamical equations. For this, we use the Euler-Maruyama integrator with friction $\nu = 0.001$, 7500 iterations, initial positions $q_{init} = (0, 1)$ and temperature $T = 100K$. Additionally, we sample this time series every 10 time points. Therefore, the length of our time series is $N = 750$ data points.

In this time series, we expect to find two metastable dynamical states and their transition region. Every metastable state corresponds to each of the wells in the potential. Besides the potential wells, we expect to identify a transition region. This indicates the moments when the system is neither in one well nor in the other.

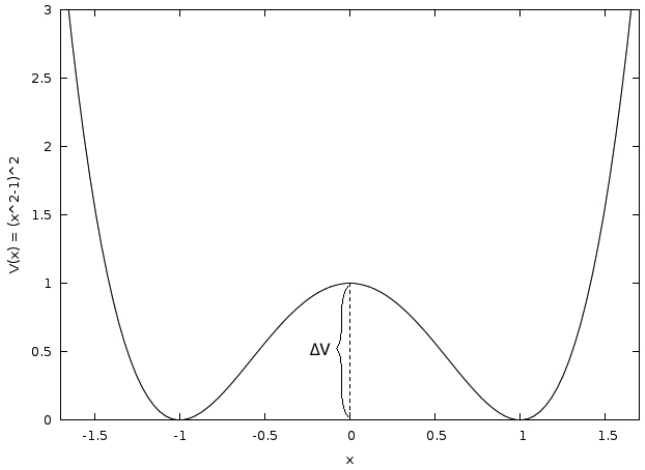


Figure 3.4: Scheme representing a double-well potential $V(x) = (x^2 - 1)^2$, with two wells centered in $x = -1$ and $x = 1$. $\Delta V = 1$ is the trap depth difference between wells.

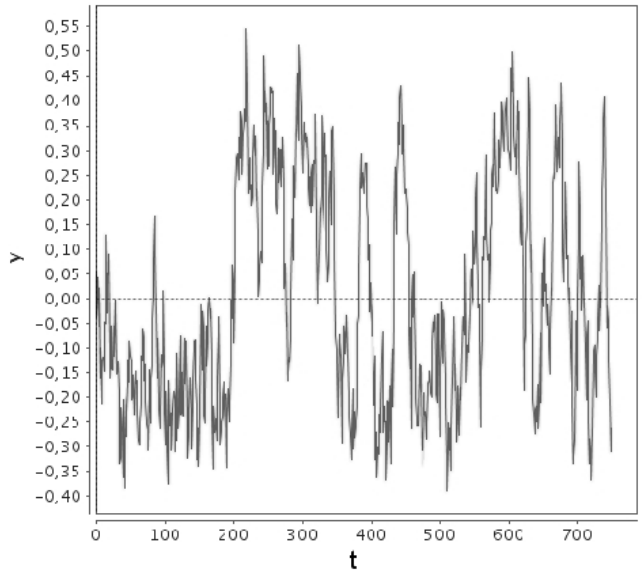


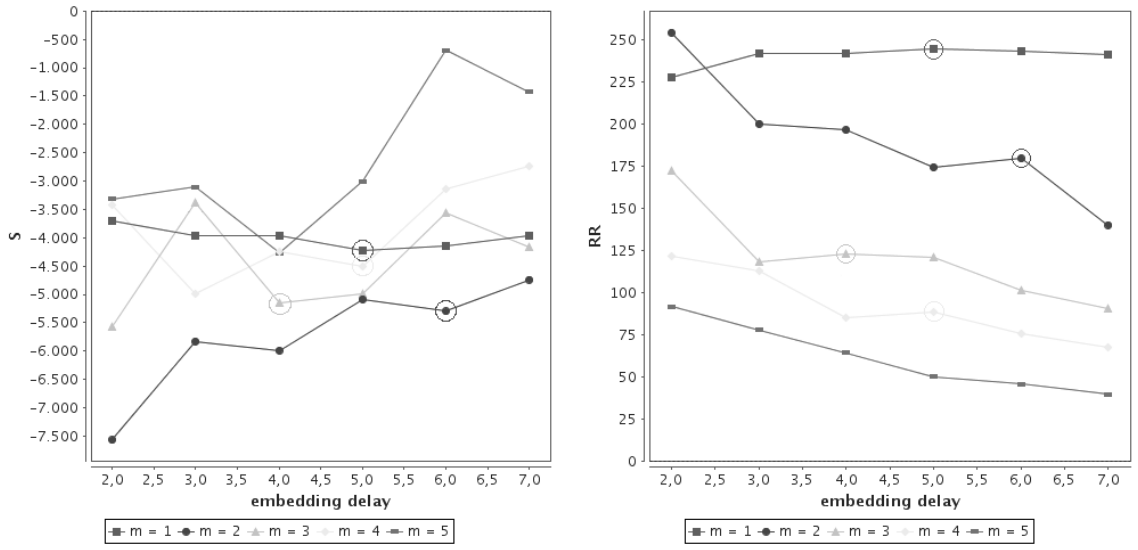
Figure 3.5: Time series for a particle in a double-well potential. This is obtained by normalizing the time series computed by integration of the Langevin equations 3.1 and sampling every 10 data points. For the integration we use an Euler Maruyama integrator with friction $\nu = 0.001$, 7500 iterations, initial positions $q_{init} = (0, 1)$ and temperature $T = 100$.

3.6.2 Analysis results

As mentioned in Sec. 3.3.1, our method starts by constructing the state space from the time series. In this case, we use embedding parameters $\tau = 7$ and $m = 2$. These were determined as explained in Algorithm 3 (Section 3.4) taking $\tau_0 = 2$, $\tau_F = 10$, $m_0 = 2$ and $m_F = 8$.

In Fig. 3.6 we show in circles the embedding parameters that first provide a simultaneous local minima in Shannon entropy and recurrence rate, for embedding delay $\tau \in [2, 7]$ and embedding dimension $m \in [1, 5]$.

The next step consists on defining a set of recurrence thresholds, using Eq. 3.1, and analyzing the modular structure of the recurrence networks associated to each of the recurrence thresh-



(a) Shannon entropy (b) Recurrence rate

Figure 3.6: Variation in Shannon entropy (Eq. 3.6) and recurrence rate (Eq. 3.5) for different embedding parameters, obtained from the recurrence plots associated to a time series describing the dynamics of a double-well potential (this system is introduced in Section 3.6). Circles indicate the embedding parameters that first provide a simultaneous local minima in Shannon entropy and recurrence rate. The threshold values used are set according to the methodology described in Section 3.3.1.

olds in this set. The visualization of the results from the analysis of the modular structure of a filtration (constructed with the embedding parameters mentioned above), is shown in the Sankey diagram in Fig. 3.7.

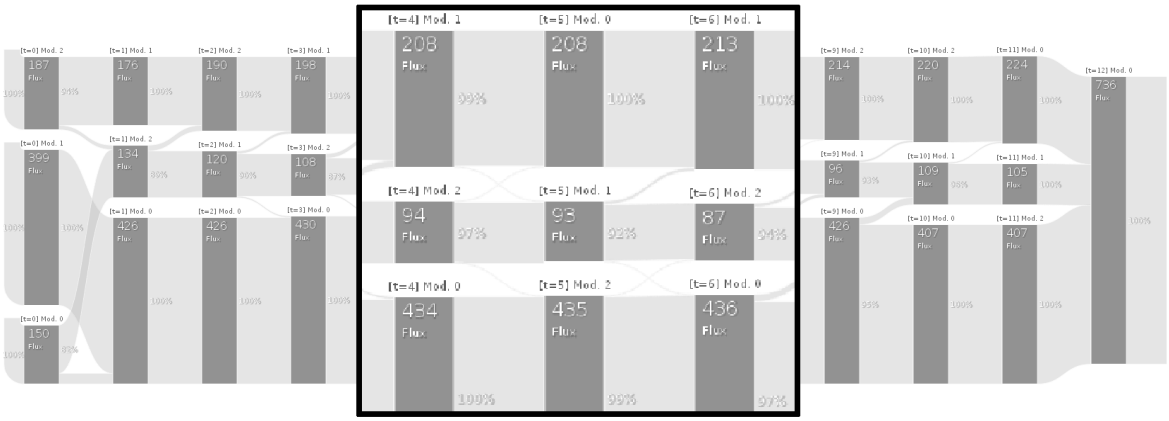


Figure 3.7: Sankey diagram showing the subgroup of recurrence networks (columns) with the same number of modules (see Eq. 3.2) and similar number of nodes (see Eq. 3.3). Networks are computed from tuning set $\{\varepsilon_\nu\}$ (see Eq. 3.1) on the state space constructed from a two well potential time series and embedding parameters $\tau = 7$ and $m = 2$. We suggest that this group of networks determines the recurrence threshold giving robust results about the dynamics of the time series analyzed.

In this particular diagram, we observe that the sizes of the metastable modules (and transition region) of the recurrence networks (columns) associated with recurrence thresholds $\varepsilon_4 < \varepsilon < \varepsilon_6$, vary the least. This means, these have the same number of modules (size of sections of a column) and the number of nodes in each module is almost the same (low flux of nodes from one column to another), given that these satisfy the dissimilarity restrictions

of Eq. 3.3 for $\chi^* = 0.01N^* \approx 7$. This is the set of thresholds from which we compute the final recurrence threshold used for the identification of metastable states in the two well potential time series: $\varepsilon^* \simeq 0.29$.

Every module identified in the recurrence network constructed using $\varepsilon^* \approx 0.29$ indicates a different metastable state in the time series. These metastable states are indicated in Fig. 3.8 by modules 0 and 1, and may correspond to each of the two potential wells. The transition region, on the other hand, is indicated by 'Module -1'.

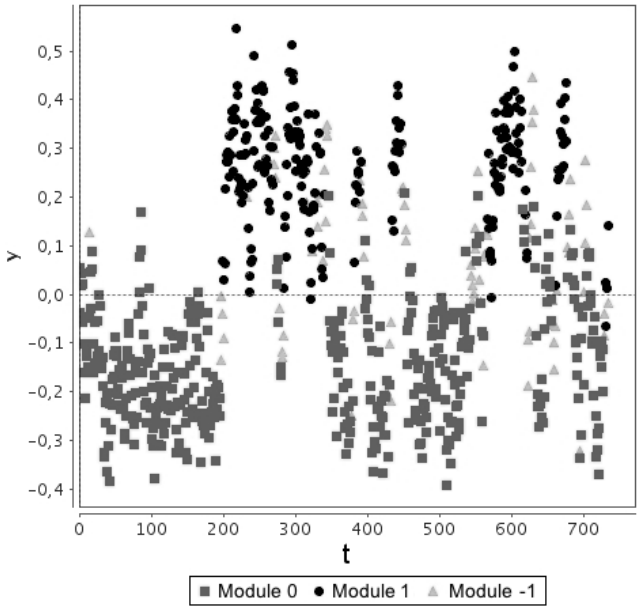


Figure 3.8: Metastable states identified on the time series in Fig. 3.5. The state space is reconstructed using a delay mapping with embedding parameters $\tau = 7$ and $m = 2$. And the recurrence network is associated to recurrence threshold $\varepsilon^* \simeq 0.29$. The grayscale color code shows the different metastable states in a time series, corresponding to the different modules in the associated recurrence network.

3.6.3 A note on modularity

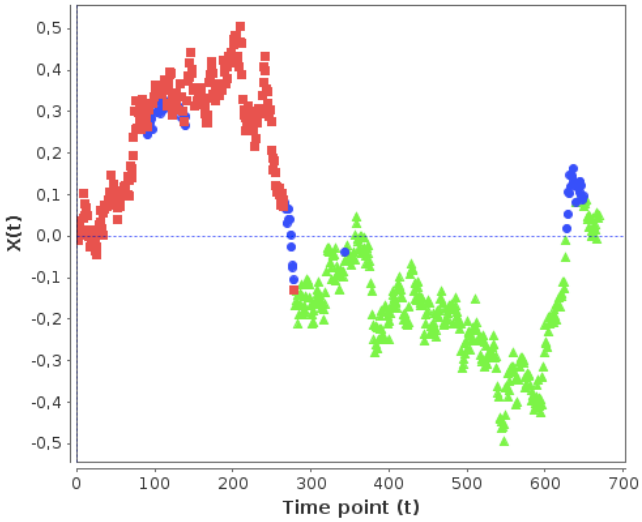
In this section we will show an example that supports our suggestion that our recurrence threshold selection coincides with having an associated recurrence network with higher modularity.

To illustrate this, we take the time series shown in Fig. 3.9(a), which describes the dynamics of a double-well potential (this dynamical system is described in Section 3.6).

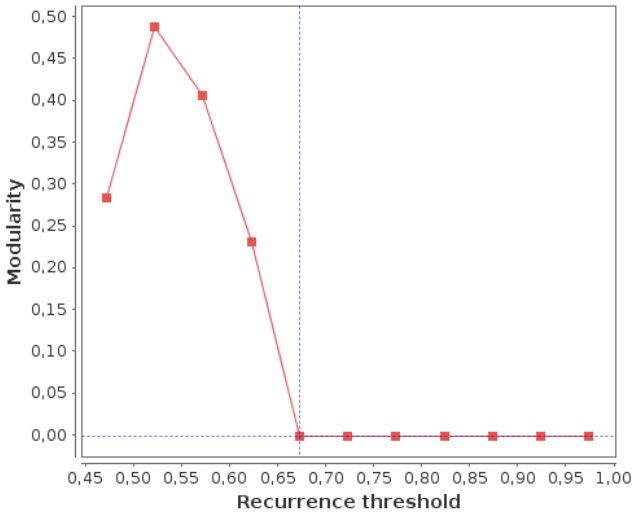
This time we reconstruct the state space of this time series using the delay-coordinate method with classical methods: average mutual information for the embedding delay and false nearest neighbors for the embedding dimension.

Then we construct a filtration on recurrence threshold, which ranges from the 50th to the 95th percentile statistic of the distances between state space vectors. Then, in Fig. 3.9(b) we show the modularity for the different recurrence thresholds in this filtration.

The recurrence threshold selected with the method described in this section is $\varepsilon \simeq 0.52$. The recurrence network associated to this recurrence threshold has the highest modularity from all the networks associated to the filtration.



(a) Different metastable states identified, indicated by different colors.



(b) Modularity for the recurrence thresholds in the filtration.

Figure 3.9: (a) Time series analyzed, describing the dynamics of a double-well potential. To reconstruct its state space, we use a delay-coordinate method with embedding parameters $\tau = 166$ and $m = 3$. Such state space vectors were reconstructed using classical methods for setting embedding parameters: false nearest neighbors and average mutual information. (b) Modularity of every recurrence network associated to the recurrence thresholds in a filtration that ranges from the 50th to the 95th percentile statistic of the distances between state space vectors. Observe that the peak in modularity occurs for $\varepsilon \simeq 0.52$, which is the recurrence threshold set with our methodology.

3.6.4 Robustness tests

In this section we perform the robustness tests described in Section 3.5 on the double well potential time series.

As mentioned before, we define robustness as the similarity between two partitions, one corresponding to the module identification on the recurrence network associated to the original time series, and the other corresponding to the module identification in the recurrence network associated to a modified time series. A modified time series is created either by adding noise or by removing a percentage of data points to the original time series.

The similarity between partitions is measured using the modified version of the Adjusted Rand Index (ARI) proposed by Hueffner et al. [62]. The parameters we use to analyze all modified time series are $\varepsilon^* \approx 0.39$, $\tau = 3$ and $m = 2$.

Noise

As mentioned in Section 3.5.1, we define noise as a percentage of the amplitude of the original time series. We vary the amplitude of noise from 0 to 20% in intervals of 1%.

Our results, in Figs. 3.10(a) and 3.10(b), show that our method is robust (ARI of around 0.6) to noise with amplitude of up to 6% the amplitude of the original time series when the ARI is measured only in the modules, and to noise with up to 2% the amplitude of the original time series when the ARI is measured in the modules and transition region.

As mentioned by Zbilut in 1992 [130], having noise in a time series has an effect of inflation of the embedding dimension when reconstructing the state space. Therefore, we could expect our method to be robust for noise with larger amplitudes if we considered different recurrence threshold and embedding parameters for the analysis of every noisy time series. However, this analysis is not done in this thesis.

Missing points

For this test, we create the modified time series by removing from 0% to 19% of the data points, in intervals of 1%.

Our results, in Figs. 3.11(a) and 3.11(b), show that our method is robust (ARI of around 0.6) to the removal of up to 7% of randomly distributed data points when the ARI is measured only in the modules, and up to 3% of randomly distributed data points when the ARI is measured in the modules and transition region.

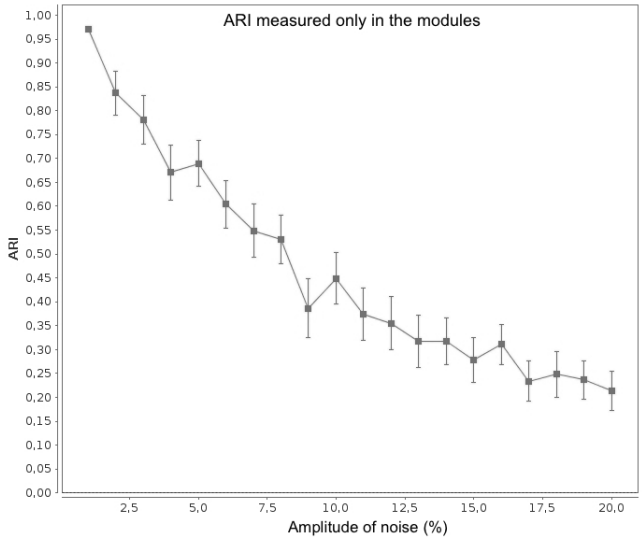
Interpreting the case of missing data points as another case of noise, we could expect our method to be robust for a larger percentage of missing points if we considered different recurrence threshold and embedding parameters for the analysis of every time series with missing points. This analysis is not done in this thesis.

3.7 Example 2: Molecular configurations of trialanine

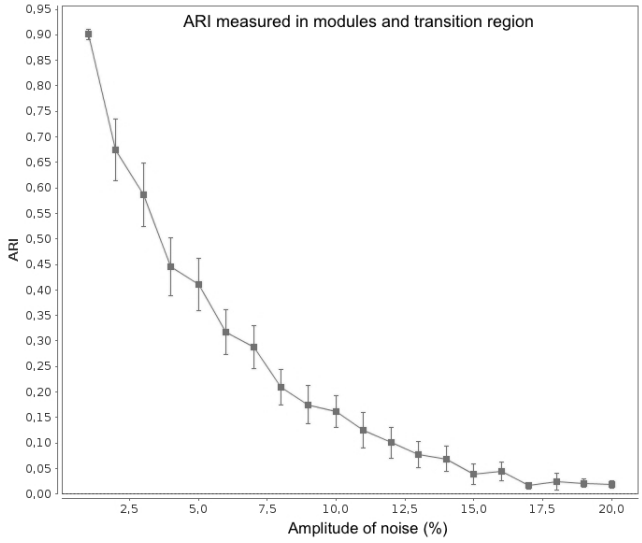
In this section we analyze a time series describing the changing molecular configurations of a molecule of trialanine at low temperature. In other words, the variation of two of the three torsion angles describing its conformation, shown in Fig. 3.12.

The conformation of a molecule is a mean geometric structure which is conserved on a large time scale compared to the fastest molecular motions, such that the associated subset of configurations is metastable.

The aim of this analysis is to identify the main molecular conformations of trialanine from a time series. For this, we analyze the time series using the method described in Section 3.3, which identifies metastable states in real-world data using recurrence networks.



(a) ARI measured only in the modules.



(b) ARI measured in the modules and transition region.

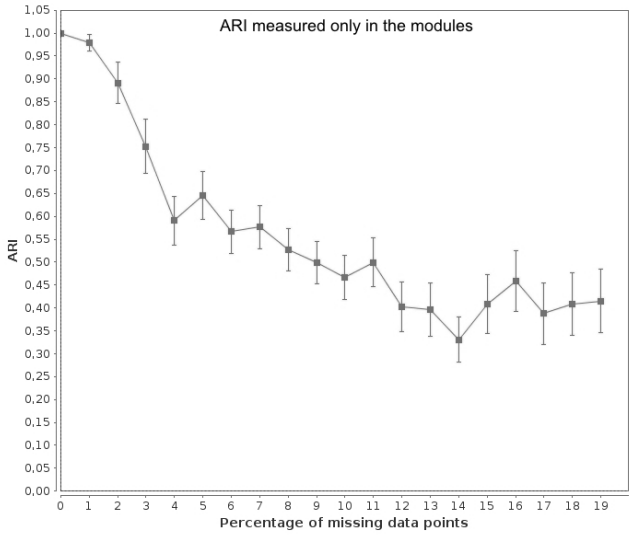
*Figure 3.10: Robustness to **noise**. Similarity between the metastable states identified in a time series and in its modified version, where noise has been added. The noisy time series is created by adding white Gaussian noise with amplitude equal to a percentage, μ_N, of the amplitude of the original time series. (a) Shows the similarity measured with the modified Adjusted Rand Index (ARI) considering only the modules. (b) Shows the similarity measured considering both modules and transition region. Error bars show confidence interval of 90%.*

3.7.1 The system

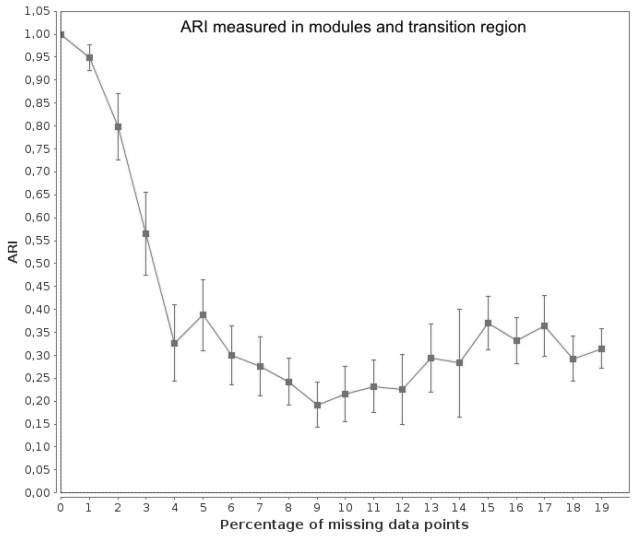
Trialanine is one of the simplest systems exhibiting the typical feature of biomolecules: having a backbone with various stable conformations.

Characterizing a molecule with its central peptide dihedral angles, or torsion angles, has the advantage of producing a reference system invariant to translations and rotations of the molecule, reducing this way the dimensionality of the description.

At low temperatures, for example $T = 300K$, the different molecular conformations of tri-



(a) ARI measured only in the modules.



(b) ARI measured in the modules and transition region.

*Figure 3.11: Robustness to **missing data points**. Similarity between the metastable states identified in a time series and in its modified version, where a percentage of randomly distributed data points has been removed. (a) Shows the similarity measured with the modified Adjusted Rand Index (ARI) considering only the modules. (b) Shows the similarity measured considering both modules and transition region. Error bars show confidence interval of 90%.*

alanine can be sufficiently characterized by the two central peptide dihedral angles, ϕ and ψ. At higher temperatures, for example $T = 700K$, one should also consider the changes in the peptide bond angle, Ω. A ball-and-stick diagram of trialanine and its three torsion angles is shown in Fig. 3.12.

According to Prei et al.[100] and Metzner, Putzig and Horenko[89], clustering the state space of trialanine at high temperatures results in the identification of five metastable states. This is the number of modules we will guess in the module finding algorithm when analyzing the modular structure of our recurrence networks.

We simulate a time series for the torsion angles of trialanine at low temperature, $T = 300K$,

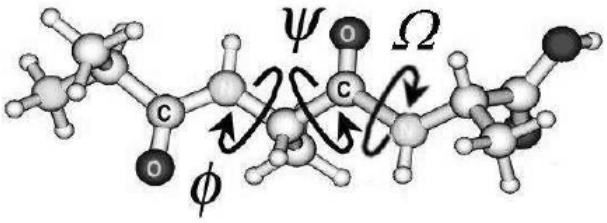


Figure 3.12: Ball-and-stick representation of a trialanine dipeptide molecule and its torsion angles ϕ, ψ and Ω. At low temperatures, its stable molecular conformations can be sufficiently characterized by the central peptide dihedral angles, ϕ and ψ. At higher temperatures, one should also consider the peptide bond angle, Ω.

and in vacuum using JGromacs [93], in which trialanine is represented by 21 united atoms. This simulation consists of 5000 steps. The resulting time series can be considered stationary. Finally, we obtain a time series of 500 data points by sampling the simulated time series with rate $\Delta t = 10$. This sampling rate does not hide transitions between states for any torsion angle. For more details about this type of simulation, see the article of Prei et al. from 2004 [100].

Since the time series is simulated at low temperature, the following analysis considers only the two central peptide dihedral angles, ϕ and ψ. Thus, the molecular conformations of trialanine can be shown in a two-dimensional plot, called the Ramachandran plot, which contains the dependency between ϕ and ψ only and is shown in Fig. 3.13.

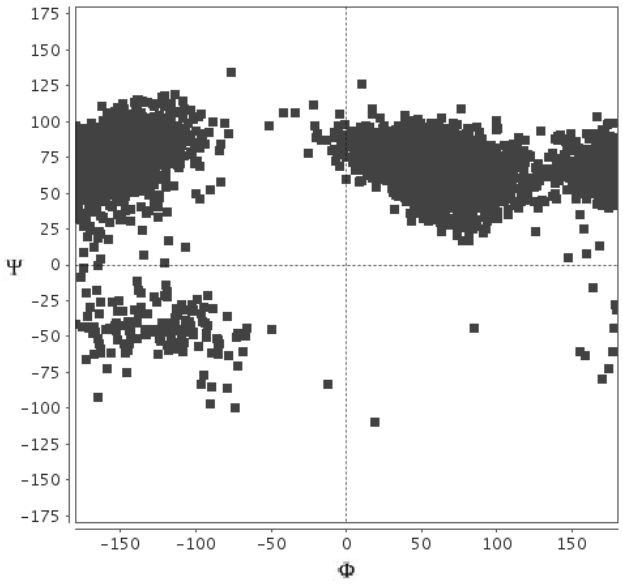


Figure 3.13: Ramachandran plot containing a sample of the molecular conformations of trialanine, simulated in vacuum at $T = 300K$ (for details go to the text). Conformations are given by the dependency between torsions angles ϕ and ψ.

3.7.2 Analysis results

The state space associated to trialanine's molecular conformations is constructed using the time delay embedding, with embedding parameters $m = 2$ and $\tau = 7$. The final recurrence network is computed with final recurrence threshold $\varepsilon^* \simeq 0.2796$.

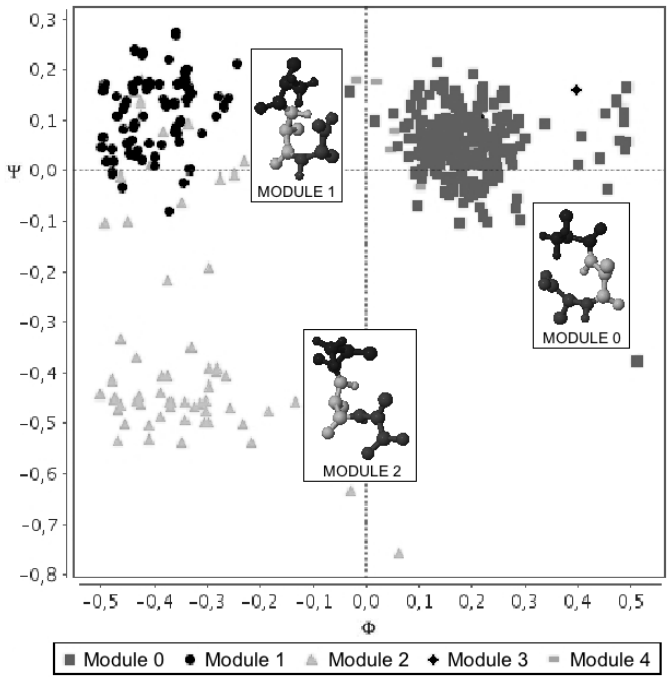


Figure 3.14: Ramachandran plot containing a sample of the normalized molecular conformations of trialanine shown in Fig. 3.13. The grayscale color code identifies the five different metastable states (or main molecular conformations) identified in the final recurrence network with $\varepsilon^* \simeq 0.2796$, $\tau = 7$ and $m = 2$. For each module we show an example of molecular conformation of trialanine belonging to it.

The analysis of the modular structure of the final recurrence network, performed analogously to the analysis of the double well potential in Section 3.6, leads to the identification of the five modules. These modules are shown in Fig. 3.14.

Due to the location in the Ramachandran plot of the three larger metastable states identified, one can identify the three main sets with the three main molecular conformations for trialanine mentioned by Fischer et al. in 2006 [51]. The two smaller sets could be a consequence to the way we assign data points to a metastable state (Section 3.3.2).

3.8 Final remarks

In this Chapter, we introduced our new method for the identification of metastable states in real-world time series. We analyzed two examples with it and obtained results robust to the introduction of noise and missing data.

This suggest that the geometrical analysis of a filtration of recurrence networks, constructed from the reconstructed state space, is adequate for the identification of an adequate recurrence threshold. In addition, as shown in the modularity analysis of the double-well potential example in Section 3.6.3, our selected recurrence threshold produces the recurrence network with highest modularity.

However, this method cannot be followed as such when there is more than one main scale in the state space of the analyzed system. In those cases, one should identify all the ranges in the recurrence networks filtration satisfying the restrictions in Eqs. 3.2 and 3.3.

We also introduced a new method for the reconstruction of the state space from a time

series, based on the analysis of filtrations of recurrence networks constructed from the state space vectors of different state space reconstructions via delay-coordinate maps with different embedding parameters.

This method consists on the geometrical analysis of several recurrence networks. Therefore, its results are heavily dependent on the specific module finding algorithm selected for the analysis of the recurrence networks.

For dynamical systems whose attractors have high-order topologies, due to module finding algorithms generally considering only low-order topological information, this method for the reconstruction of the state space may not be sufficient to identify all the metastable states.

These drawbacks motivate the method presented in the following chapter, where we introduce an algebraic topological approach to the analysis of complex data.

4. A new method for identifying transitions between dynamical regimes in real-world time series via persistent homology

Previously, in Chapter 3, we introduced a method for the identification of metastable states in real-world time series. In such method we make use of recurrence networks analysis to obtain an adequate state space reconstruction (see discussion in Section 2.3) via a delay-coordinate map. Later, we set an adequate recurrence threshold imposing some restrictions on two recurrence quantification measures−recurrence rate and entropy− and using a fuzzy clustering algorithm on a recurrence network constructed with such threshold over the reconstructed state space vectors, we identify modules that correspond to different metastable states in the underlying system.

Despite the first method being robust to the introduction of low levels of noise or missing data points (see Section 3.6 for details), it has some drawbacks. One of these is the implicit assumption of having one main scale in the state space of the analyzed system. Another drawback arises from the heavy dependence on the specific module finding algorithm selected for the analysis of recurrence networks. In general, module finding algorithms consider only low-order topological information. Therefore, the results of our analysis are only reliable for time series whose reconstructed state spaces have low-order topology.

In order to overcome these limitations and analyze time series whose reconstructed state spaces have high-order topologies, in this chapter we introduce a new method for topological data analysis (TDA) that uses an algebraic topological approach: persistent homology (introduced in Section 4.3). Persistent homology can be understood as the analysis of topological features (homology; see Section 4.1.3) that persist in a set of coverings with different fixed radius ϵ (a filtration depending on ϵ).

The main aim of performing topological data analysis (TDA) over the reconstructed state space vectors is to find a representation of such vectors that allows the robust computation of topological invariants (see Section 4.1). Our particular selection of the persistent homology approach is based in the fact that this is, in principle, suitable for the analysis of time series data where the use of a specific type of metric and coordinates is not fully justified, or when one wants to study the behavior of a system for a wide range of parameters and not only for a single selection of parameters [9, 44].

Another advantage of this approach is the existence of several theorems on the stability of

persistence (see Section 4.1.5), which provide equivalent persistent homology results regardless of the filtration function and provide some confidence boundaries when the data has noise or when persistence is computed from a sample of the entire data. In Section 4.2 we present some of the results on how to estimate persistence when considering the existence of noise or outliers in data.

This way, we suggest that this new method is adequate for the identification of transitions between different dynamical regimes, given that it includes higher-order topological information into the analysis of a time series. Additionally, due to the existence of stability theorems, the confidence in the results it provides is bounded, and these results are robust to noise and missing data. In Section 4.6 we illustrate these results by identifying dynamical transitions in the time series of a two-dimensional double-well potential with this method.

However, during the development of this method, we faced several difficulties associated to the state space reconstruction from a time series. This led to the review of the conditions that guarantee an adequate state space reconstruction (see Section 4.4) and to the development of a new method for the reconstruction of the state space from a real-world time series based on persistent homology (see Section 4.5). In Section 4.7 we use the two methods introduced in this chapter to analyze the time series of a logistic map whose dynamics vary from non-stochastic to stochastic, according to the parameters selected for the simulation of a segment of such time series.

4.1 Introduction to topological data analysis (TDA)

Whenever a metric, or distance function, and a coordinated system are introduced for the analysis of a system, and when the topological information of the system is of lower-order, it is possible to use geometric tools specific for the analysis of sets of data points, like clustering.

When a data set is clustered, it is partitioned into different subsets, or modules. Every module, given a similarity measure, is distinguishable from another. This way, the process of identifying the modular structure of data can be thought as "the statistical counterpart to the geometric construction of the path-connected components of a space"[9]. The methodology presented in Chapter 3 is based on this approach.

However, it is often the case that the intrinsic features of a data set do not justify the use of metrics and coordinated systems. In these cases one may perform topological data analysis (TDA).

Topology can be defined as the study of the geometric properties of data, without being sensitive to specific choices of metric, coordinates or curvature (unlike recurrences analysis). Fig. 4.1 illustrates the common saying that, in topological terms, a teacup and a bagel are the same.

Figure 4.1: "Topology is the branch of mathematics which cannot distinguish between a teacup and a bagel" [10].

The topological analysis of a finite data set sampled from an unknown topological space $\mathbb{X}$, $S \subset \mathbb{X}$, is meant to recover the topology of such space.

Definition 4.1.1 (Topological space)**.** *A topological space is a pair* $X = (\mathbb{X}, \mathcal{U})$ *where* $\mathbb{X}$ *is a set of points and* $\mathcal{U}$ *is a topology on* $\mathbb{X}$*. The topology* $\mathcal{U}$ *is defined as a collection of open sets of* $\mathbb{X}$ *such that:*

1. *If* $S_1, S_2 \in \mathcal{U}$*, then* $S_1 \cap S_2 \in \mathcal{U}$*.*
2. *For* $S_j \in \{S_j\}_{j \in J}$*, where* J *is an index set that can be infinite or uncountable, if* $S_j \in \mathcal{U}\, \forall j$ *then* $\bigcup_j S_j \in \mathcal{U}$*.*
3. $\emptyset, \mathbb{X} \in \mathcal{U}$*.*

A data set can be, for example, a point cloud data (PCD). This means, an unordered collection of points in a Euclidean d-dimensional space, $\mathbb{E}^d$. A PCD can also be a sample of points from a lower dimensional subset. In this case, the topological features of the space can be inferred by reconstructing it from the sample.

This reconstruction is frequently done by performing planar projections. However, this approach may not be adequate when the space of the system underlying the PCD is not a manifold; for example, when it is curved.

To study the topological features of a PCD, the first step usually consists on transforming the PCD into a family of simplicial complexes by defining a proximity parameter, ϵ. This step, as we will later see, resembles the construction of a recurrence plot by determining a metric and a recurrence threshold, ε.

The characteristics of these simplicial complexes are analyzed via the theory of persistent homology. Finally, the results from this analysis are summarized in either a so called barcode or a so called persistence diagram (defined below). Both of these summaries contain the same information as the Betti numbers.

4.1.1 Simplicial complexes

Let us consider every data point in a PCD as a vertex in a combinatorial graph, with edges indicating the proximity between data points, measured by a distance parameter, ϵ. Considering this graph as the structure for a higher dimensional object, intstead of proceeding to cluster it as in the previous chapter, we analyze its higher order topological features. For this, we first transform the graph into a collection of simplices in order to obtain a simplicial complex.

Definition 4.1.2 (Simplices)**.** *A* k*-simplex is the convex hull of* $k+1$ affinely independent *points,* $\sigma = conv\{u_o, u_1, ..., u_k\}$*.*

The diameter of a simplex $\sigma \in K$, $\mathrm{diam}(\sigma)$, is given by the maximum distance between the images of any two points $x, y \in \sigma$. This means, $\mathrm{diam}(\sigma) = \max_{x,y \in \sigma} d(f(x), f(y))$.

For k from zero to three, the k-simplices have particular names: a 0-simplex is called a vertex; a 1-simplex, an edge; a 2-simplex, a triangle; and a 3-simplex, a tetrahedron.

Therefore, a simplicial complex can be understood as a collection of vertices, edges, triangles, tetrahedron and polyhedra of higher orders.

Definition 4.1.3 (Simplicial complex)**.** *A simplicial complex is a finite collection of simplices,* K*, such that:*

- *If* $\sigma \in K$ *and* $\tau \leq \sigma$*, then* $\tau \in K$*,*
- *If* $\sigma, \sigma_0 \in K$*, then* $\sigma \cap \sigma_0$ *is either empty or a face of both.*

The underlying space of a simplicial complex, denoted by $|K|$, is the union of its simplices with the topology inherited from the Euclidean ambient space where the simplices are contained. The dimension of a simplicial complex is equal to the maximum dimension of its simplices.

Definition 4.1.4 (Topological space of a simplicial complex.)**.** *Associated to a simplicial complex, K, there is a topological space $|K| = |(V, \Sigma)|$, where V is a finite set, and Σ is a family of non-empty subsets of V. Let $\phi : V \to \{1, 2, ..., N\}$ be a bijection and $c(\sigma)$ be the convex hull of the set $\{e_{\phi(s)}\}_{s\in\sigma}$, where e_i denotes the i-th standard basis vector. Then, the topological space $|K|$ may be defined as the subspace of $\mathbb{R}^N$ given by the union $\bigcup_{\sigma\in\Sigma} c(\sigma)$.*

A topological space X is said to be triangulable if there exists a simplicial complex K and a homeomorphism $f : |K| \to X$. The pair (K, f) is called a triangulation of X. This way, a simplicial complex can serve as a simple combinatorial way to describe a triangulable topological space and a lot of effort has been put in approximating topological spaces with simplicial complexes [11, 9, 44].

The simplices in a simplicial complex can be constructed in different ways and the simplicial complexes can thus have a geometric realization or not. However, a desired property in any simplicial complex is to preserve the homotopy type of the underlying topological space.

In the case where a simplicial complex does not have a geometric realization, the concept of abstract simplicial complex is useful.

Definition 4.1.5 (Abstract simplicial complex.)**.** *An abstract simplicial complex is a pair (V, Σ), where V is a finite set, and Σ is a family of non-empty subsets of V such that $\sigma \in \Sigma$ and $\tau \subseteq \sigma$ implies that $\tau \in \Sigma$.*

The nerve is a construction of an abstract simplicial complex that, under certain conditions, has the important property of being homotopy equivalent to the underlying space.

Definition 4.1.6 (Nerve)**.** *Let X be a topological space and $\mathcal{U} = \{U_\alpha\}_{\alpha\in A}$ be an open covering of X, where the sets in the collection $\mathcal{U}$ are not necessarily convex. The nerve of $\mathcal{U}$, $N(\mathcal{U})$, is the abstract simplicial complex associated to $\mathcal{U}$. $N(\mathcal{U})$ has a vertex set $A = \{v_\alpha\}$, with a vertex for every U_α, such that $k + 1$ vertices span a k-simplex if and only if there are $k + 1$ sets of $\mathcal{U}$ whose intersection is non-empty; i.e. $U_{\alpha_0} \cap ... \cap U_{\alpha_k} \neq \emptyset$.*

As stated in the following theorem, if the sets in $\mathcal{U}$ are convex, then the nerve preserves the homotopy type.

Theorem 4.1.7 (Nerve theorem)**.** *Suppose that $\mathcal{U}$, the covering of X, consists of open sets and is numerable and that for all $\emptyset \neq S \subset A$, we have that $\bigcap_{s\in S} U_s$ is either contractible or empty. Then $N(\mathcal{U})$ is homotopy equivalent to X.*

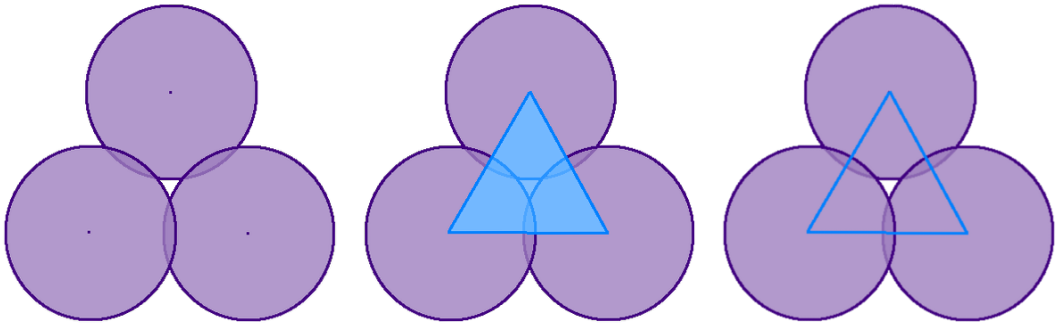
Interestingly, another way to guarantee that $N(\mathcal{U})$ preserves the homotopy type is by requiring $\bigcup_\alpha U_\alpha$ to be triangulable. This way, all sets in the collection are closed and all non-empty common intersections are contractible, and then $N(\mathcal{U}) \simeq \bigcup_\alpha U_\alpha$.

There are other constructions that create simplicial complexes with geometric realizations and also preserve the homotopy type of the underlying topological space.

Later on, when we introduce our method for obtaining an adequate state space reconstruction following the persistent homology approach, we will make use of the witness complex con-

struction. This construction can be applicable to real-world data since it is robust to noise and outliers. In the following paragraphs we will describe the Vietoris-Rips complex and the Čech complex, given that these constructions motivate the introduction to the witness complex construction.

In Fig. 4.2 we illustrate three different geometric realizations: the Nerve, the Čech and the Vietoris-Rips complex for a given collection of three points.



(a) Nerve (b) Čech complex (c) Vietoris-Rips complex

Figure 4.2: The nerve, Čech and Vietoris-Rips complexes for a collection of three data points.

Čech complex

When a topological space, X, is a metric space, a type of covering of X is given by $\mathcal{B}_\epsilon(X) = \{B_\epsilon(x)\}_{x\in X}$, where $\epsilon > 0$. This way, a Čech complex is an abstract simplicial complex (which not always has a geometric realization) created by considering the sets in the covering $\mathcal{U}$ as closed geometric balls with the same radius.

Definition 4.1.8 (Čech complex)**.** *Let $S = \{u_\alpha\}_{\alpha\in A}$ be a finite set of points in Euclidean space $\mathbb{E}^d$. Let $B_\epsilon(u_\alpha) = u_\alpha + \epsilon\mathbb{B}^d$ be a closed ball with center in u_α and radius ϵ. Then, the Čech complex of S is defined as the nerve of the collection of balls $\{B_\epsilon(u_\alpha)\}_{\alpha\in A}$ and substituting the center of each ball:*

$$\check{C}(\epsilon) = \left\{ \sigma \subseteq S \mid \bigcap_{u_\alpha\in\sigma} B_\epsilon(u_\alpha) \neq \emptyset \right\} \tag{4.1}$$

The Čech theorem states that a Čech complex has the homotopy type of the union of closed balls with radius equal to $\epsilon/2$ about $S = \{u_\alpha\}_{\alpha\in A}$, the finite set of points in Euclidean space $\mathbb{E}^d$. This implies that the Čech complex behaves exactly like $\mathbb{E}^d$, even when it can produce simplices in dimensions much higher than the dimension of the space.

However, the homology computation with these complexes is very sensitive to outliers and noise [4]. This way, it is desirable to use alternative constructions of simplices when the PCD is suspected to have these characteristics, for example, the Vietoris-Rips construction.

Vietoris-Rips complex

As mentioned before, the Čech construction is sensitive to outliers and noise. Additionally, computing the homotopy via the Čech construction is computationally expensive. An alternative construction to overcome these drawbacks is provided by the Vietoris-Rips complex.

Definition 4.1.9 (Vietoris-Rips complex.). *For X a metric space with metric $d(\cdot,\cdot)$, the Vietoris-Rips complex for X, attached to the parameter ϵ, is denoted by $VR(X,\epsilon)$. This $VR(X,\epsilon)$ is the simplicial complex whose vertex set is X and where $\{x_0, x_1, ..., x_k\}$ spans a k-simplex if and only if $d(x_i, x_j) \leq \epsilon$ for all $0 \leq i,\ j \leq k$.*

Even when the Vietoris-Rips complex is computationally less expensive than the Čech construction, it may have more simplices than a Čech complex, given that it is maximal among all simplicial complexes with a same 1-skeleton.

Definition 4.1.10 (k-skeleton of a simplicial complex). *The k-skeleton of a simplicial complex consists of all simplices of dimension k or less:*

$$K(k) = \{\sigma \in K | \dim \sigma \leq k\} \tag{4.2}$$

This way, a Vietoris-Rips complex can be recovered completely from a graph. This implies another computational advantage over the Čech complex, which requires to store the entire boundary operator. However, contrary to the Čech complex, the Vietoris-Rips does not have the same homotopy type of $\mathbb{E}^d$. Even more, it might not be similar to a d-dimensional space.

Witness complexes

A drawback of the Čech and the Vietoris-Rips complexes is that they depend on the precision with which the distances between data points are measured. These also tend to produce simplices in dimensions much higher than the dimension of the space [9].

Therefore, one would like to recover the topology of a space in a way such that it is robust to the introduction of noise in the measurements. With this idea in mind, and introducing a set of landmark points (see Definition 4.1.11), one can construct the so called witness complexes. These produce smaller simplicial complexes than the Čech or the Vietoris-Rips complexes and are divided into strong and weak witness complexes.

According to V. de Silva and G.Carlsson [11], some of the main advantages of constructing witness complexes are that these are adaptable to arbitrary metrics, do not suffer from curse of dimensionality and show good results for topological data analysis (TDA) with persistent homology, even for noisy data.

The computational cost of constructing a witness complex may be high. Silva and Carlsson introduced the lazy witness construction [10] in order to reduce it. These simplex are able to robustly reconstruct the same stream created using Vietoris-Rips complexes, but with lower computational cost. However, the quality of their reconstruction depends on selecting an appropriate number of landmark points [11]. Therefore, despite being less expensive than regular witness complexes, we did not use them in any of the methods introduced in this thesis.

Definition 4.1.11 (Strong witness complex). *Let X be any metric space and $\{l_0, ..., l_k\} = \mathcal{L} \subseteq X$ a finite set of points in X called the landmark set. For every point $x \in X$, denote by m_x the minimum distance from x to any point in $\mathcal{L}$.*
Then, given a parameter ϵ, the strong witness complex is the complex, $W^S(X, \mathcal{L}, \epsilon)$, whose vertex set is $\mathcal{L}$ and where $\{l_0, ..., l_k\}$ spans a k-simplex if and only if there is a point $x \in X$, called the witness, such that $d(x, l_i) \leq (m_x + \epsilon)\ \forall i$.

Definition 4.1.12 (Weak witness complex). *Let $\Lambda = \{l_0, ..., l_k\}$ be a finite subset of a metric space X. Then, a weak witness for Λ is a point $x \in X$ such that $d(x, l) \geq$*

$d(x, l_i)\,\forall i$ *and* $\forall l \notin \Lambda$. *And an* ϵ*-weak witness for* Λ *is a point* $x \in X$ *such that* $d(x, l) + \epsilon \geq d(x, l_i)\,\forall i$, $\forall l \notin \Lambda$.

Then, the weak witness complex $W^w(X, \mathcal{L}, \epsilon)$ *is obtained by declaring that a family* $\Lambda = \{l_0, ..., l_k\}$ *spans a* k*-simplex if and only if* Λ *and all its faces admit* ϵ*-weak witnesses.*

Strong and weak witness complexes have an **inclusion** property that makes them interesting when analyzing a filtration.

> **Definition 4.1.13** (Inclusion)**.** *An inclusion map* $A \mapsto B$ *is a map* $A \to B$ *where* $A \subset B$. *2 fig,*

Let $W(X, \mathcal{L}, \epsilon)$ and $W(X, \mathcal{L}, \epsilon')$ be two different strong (or weak) witness complexes originated from the same metric space and landmark set, but using different parameters ϵ and ϵ' such that $0 \leq \epsilon \leq \epsilon'$. Then, the following holds:

$$W(X, \mathcal{L}, \epsilon) \mapsto W(X, \mathcal{L}, \epsilon') \tag{4.3}$$

Landmark points. In witness complexes, the landmark points are assumed to be well distributed over the PCD and the remaining points are used to construct the simplicial complex. Under this assumption, if the set of landmark points is fixed, when the number of data points increases, the simplicial complex constructed should approximate an "ideal" complex in which every data point is a witness. However this will not be the case if the selection of the landmark points or the definition of a witness is not appropriate.

As can be understood from Definition 4.1.11, every landmark point determines a Voronoi cell in the graph metric. Ideally, each of these cells will correspond to convex, convexly intersecting regions in the underlying space, and their size is $\lambda = \|Z\|/\|L\|$, where Z is the vertex set and L is the set of landmark points. If the distribution of the landmark points is appropriate, every Voronoi cell should contain approximately λ points.

The two typical types of landmark points selection are the random and *sequential minmax* [10]. The random selection is self-explained so we will proceed to explain the second way of selection.

The *sequential minmax* landmark point selection, or *minmax* selection, starts by selecting a landmark point at random. Then, we continuing selecting landmark points until producing the desired collection, $L = \{\ell_1, \ell_2, ..., \ell_n\}$, where $\ell_i \in Z \setminus \{\ell_j\}_{j=0}^{i-1}$.

For the selection of additional landmark points, we need to consider a metric d, which may be the Euclidean or the shortest-paths metric[1]. Then, every additional landmark point is selected as the data point which maximizes the following function:

$$z \mapsto \min\{d(z, \ell_1), ..., d(z, \ell_{i-1})\} \tag{4.4}$$

According to G. Carlsson and V. de Silva [9], randomly selected landmark points tend to be selected from high-density regions of the data and *minmax* selected landmark points tend to be well separated. However, the *minmax* selector tends to take outliers as landmarks, which implies that in some cases, some preprocessing of the data should be done. Nevertheless, these authors report good results when using a *minmax* selector. Fig. 4.3 illustrates the differences in persistent homology results due to the selection of landmark points for constructing witness complexes, suggesting that the minmax selector provides better results.

[1]The shortest-paths metric is used for the generation of landmark points for the combinatorial Delaunay triangulation. This triangulation was introduced by Carlsson and de Silva in [11], and is defined as the nerve of a covering of X made with Voronoi cells. Such triangulation has the advantage of producing very small simplicial complexes of dimension usually equal to the one of the manifold of X. However, we do not cover it here since it tends to produce degenerate complexes for finite metric spaces [9].

4.1.2 Homotopy

One can use simplicial maps to provide a notion of equivalence between topological spaces. This can in turn be encoded into an equivalence relation. The first notion of equivalence we will review is the so called homotopy.

Definition 4.1.14 (Homotopy)**.** *Two continuous maps f, $g : \mathbb{X} \to \mathbb{Y}$ are* homotopic *if there is a continuous map $H : \mathbb{X} \times [0,1] \to \mathbb{Y}$ such that $H(x,0) = f(x)$ and $H(x,1) = g(x) \forall x \in \mathbb{X}$. This defines an equivalence relation, $f \simeq g$. One says that f and g are homotopic if there is a homotopy between them.*

Two topological spaces X and Y are said to be homotopy equivalent if there are two continuous maps $f : X \to Y$ and $g : Y \to X$ such that f and g are homotopy equivalences of homotopy inverses of each other.

That means that $f \circ g$ is equivalent to the identity map on Y ($f \circ g \simeq id_Y$) and that $g \circ f$ is equivalent to the identity map on X ($g \circ f \simeq id_X$).

The definition of homotopy equivalence provides an equivalence relation, $X \simeq Y$. This way, X and Y have the same homotopy type if they are homotopy equivalent.

4.1.3 Homology

Despite the insights that homotopy provides, it has high computational costs. For this reason, homology, another notion of equivalence between topological spaces, is more practical. This formalism, however, does not capture as much topological information as homotopy.

Homology describes quantitatively the connectivity of a topological space. The connectivity of a space consists of its number of k-dimensional chain complexes.

Definition 4.1.15 (Chain complex)**.** *A k-dimensional chain, c, is a formal sum of k-simplices in a simplicial complex K. Let $\{\sigma_i\}$ be the collection of k-dimensional simplices and $\{a_i\}$ the collection of coefficients, where a_i are typically modulo 2 coefficients, meaning that they take values 0 or 1. This way, c is defined by:*

$$c = \sum_i a_i \sigma_i \tag{4.5}$$

Given two chains, $c = \sum_i a_i \sigma_i$ and $c' = \sum_i b_i \sigma_i$, their addition is defined by $c + c' = \sum_i (a_i + b_i)\sigma_i$, where $a_i + b_i = 0$ if $a_i = 1$ and $b_i = 1$. Introducing the neutral element, $0 = \sum_i 0\sigma_i$, and the inverse of a k-chain, $-c = c$, the pair of k-chains and sum operation, $C_k(K) = (C_k, +)$, forms the Abelian group of k-chains.

Cycles and boundaries

The k-dimensional chain complexes divide into cycles and boundaries. The boundary of a k-chain is the sum of the boundaries of its simplices. A k-cycle is a k-chain with empty boundary.

Definition 4.1.16 (Boundary of a k-simplex)**.** *Given a k-simplex $\sigma = [u_0, u_1, ..., u_k]$, its boundary is the object that relates the groups of k-chains created for every $k < \dim(K)$. The boundary of σ, denoted by $\partial_k \sigma$, is defined as the sum of the $(k-1)$-dimensional faces*

of σ. If $[u_0, ..., \hat{u}_j, ..., u_k]$ denotes the simplex where u_j is omitted, then one can define linear operators ∂_k such that

$$\partial_k \sigma = \sum_{j=0}^{k} [u_1, ..., \hat{u}_j, ..., u_k] \tag{4.6}$$

Definition 4.1.17 (k-Cycle)**.** *A k-cycle is a k-chain, $c \in C_k$, where $\partial c = 0$. This means, it has an empty boundary.*

Definition 4.1.18 (k-Boundary)**.** *A k-boundary, c, is the boundary of a $(k+1)$-chain, d, where $c = \partial d$ for $d \in C_{k+1}$.*

From Def. 4.1.16, we see that $\partial_k : C_k \to C_{k-1}$. Besides, taking the boundary commutes with the addition. The k-cycles group inherits being Abelian from the group of k-chains. This way, one can also define the Abelian group of k-cycles, denoted by $Z_k = Z_k(K)$. The group of k-cycles is a subgroup of the group of k-chains and the kernel of the k-boundary homomorphism, $Z_k = \ker \partial_k$.

Since $\partial_k(c+c') = \partial_k c + \partial_k c'$, the boundary is a homomorphism and can therefore be called the boundary map for chains. This way, a chain complex can also be understood as a sequence of chain groups connected by boundary maps:

$$\cdots \xrightarrow{\partial_{k+2}} C_{k+1} \xrightarrow{\partial_{k+1}} C_k \xrightarrow{\partial_k} C_{k-1} \cdots \tag{4.7}$$

The group of k-boundaries, $B_k = B_k(K)$, is also Abelian and can be also understood as a subgroup of k-chains. It constitutes the image of the $(k+1)$-boundary homomorphism, $B_k = \operatorname{im} \partial_{k+1}$.

An important property of boundaries is that the boundary of a boundary is zero. This result constitutes the Fundamental lemma of homology.

Lemma 4.1.19 (Fundamental lemma of homology)**.** *For every $k \in \mathbb{Z}$ and $(k+1)$-chain*

$$\partial (\partial c_{k+1}) = 0 \tag{4.8}$$

Given a k-chain, the boundary ∂_k can be expressed as matrices which represent the boundary homomorphisms. These matrices can be reduced[2] and used to compute homology. However, this process can be computationally expensive.

Homology groups

From Lemma 4.1.19, one can derive that every k-boundary is a subgroup of a k-cycle group: $B_k \subseteq Z_k$. This property allows the construction of homology groups.

Definition 4.1.20 (Homology group)**.** *The k-homology group is the quotient of the k-cycle group modulo and the k-boundary group:*

$$H_k(K) = H_k = Z_k / B_k \tag{4.9}$$

[2]The reduction typically consists on bringing them to the so called *Smith normal form*, which we will not explain here. For details about the Smith normal form and its construction, see [40]

An homology group, H_k, is an Abelian group. It indicates the number of k-dimensional subspaces of the topological space X which have no boundary in X and, at the same time, which are not boundary of any $(k+1)$-dimensional subspace. Taking continuous maps between topological spaces will induce maps on homology as well.

An element of H_k can be obtained by adding all k-boundaries, B_k, to a given k-cycle, $c \in Z_k$. This cycle is called a *coset* of B_k in Z_k.

Let c' be another k-cycle such that $c' = c + c''$, where $c'' \in B_k$. Then, c' returns the same class: $c + B_k = c' + B_k$, because $c'' + B_k = B_k$. Such class is a coset of H_k and is called a *homology class*. Since c and c' are in the same homology class, they are said to be *homologous*, which is denoted by $c \sim c'$. Besides, the addition of two classes is also well defined, since $(c + B_k) + (c' + B_k) = (c + c') + B_k$.

The cardinality, or order, of a homology group H_k is determined by the number of cycles in it:

$$\text{ord}H_k = \text{ord}Z_k / \text{ord}B_k. \tag{4.10}$$

The rank of a k-homology group H_k, denoted by β_k, is called its Betti number. Intuitively, the k-th Betti number corresponds to the number of independent k-dimensional surfaces in H_k.

Equivalently to the computation of the order of H_k, the rank of a k-homology group is given by

$$\beta_k = \text{rank}H_k = \text{rank}Z_k - \text{rank}B_k \tag{4.11}$$

An important result on algebraic topology is that homology groups do not depend on the triangulation of a topological space.

Considering that homotopy equivalent spaces have isomorphic homology groups, then we have that the homotopy equivalence between two spaces can also be measured in terms of their Betti numbers. This means that all their Betti numbers are equal. For this reason, despite the loss of topological resolution, being able to compute homology groups using linear algebra methods constitutes a big advantage.

4.1.4 Persistent homology

It is often the case that a set of data points does not precisely recover the topology of the subspace $X \subseteq \mathbb{R}^n$. This may occur because the set of data points constitutes a sample, possibly with noise.

An important question is thus, how much of the homology of X (measured not only by its Betti numbers) can be obtained from the sample?

To answer this question, one might use methods of manifold learning, like the one of Niyogi et al. [95]. These authors assume working with Riemannian manifolds and consider that the Čech complex associated to a covering by balls of a fixed radius ϵ, is homotopy equivalent to the underlying manifold. However, this approach relies in the assumption that the data lies in a submanifold, which is often not true in experimental settings.

Another approach is analyzing the behavior of homology for several values of ϵ, the parameter used to construct the simplices (see section 4.1.1). This approach should provide a notion of the topological resolution of the data points. The main idea underlying this approach is

that some topological features will exist over ranges of ϵ of different length, the so called "surviving" features. Those features surviving longer ranges may correspond to large scale geometric features, or interesting signals, whereas short ranges may correspond to noise or inadequate sampling.

Persistent homology is a computational scheme that provides a summary of homology under the the entire range of ϵ.

Recalling that Čech and Vietoris-Rips complexes grow as ϵ grows, the chain maps of a k-persistence complex can be understood as inclusion maps.

Definition 4.1.21 (k-Persistence complex, according to R. Ghrist [55]). *A k-persistence complex is defined as a sequence of chain complexes, $C = \left(C_k^i\right)_i$, together with chain maps, $x : C_k^i \to C_k^{i+1}$.*

Then, the inclusion map going from the underlying space of K_i to the one of K_j, for $i \leq j$, induces an homomorphism $f_k^{i,j} : H_k(K_i) \to H_k(K_j)$ for every dimension k. This way, a **filtration of** f corresponds to a sequence of homology groups connected by homomorphisms.

Definition 4.1.22 (Filtration). *Let K be a simplicial complex and $f : K \to \mathbb{R}$ a monotonic function. Then, a subcomplex of K can be obtained from $K_a = f^{-1}(-\infty, a]$, for every $a \in \mathbb{R}$. And if K contains m simplices, then there are $n + 1 \leq m + 1$ subcomplexes, which can be arranged in a filtration of f. The filtration of f is the sequence:*

$$\emptyset = K_0 \subseteq K_1 \subseteq \cdots \subseteq K_n = K \tag{4.12}$$

The homomorphisms corresponding to the filtration of f given in expression 4.12 are:

$$0 = H_k(K_0) \mapsto H_k(K_1) \mapsto \cdots \mapsto H_k(K_n) = H_k(K) \tag{4.13}$$

These constritute the so called k-homology groups, which indicate the k-classes of K_i that survive in K_j.

Definition 4.1.23 (k-Persistent homology, according to [44]). *The k-persistent homology groups are the images of the homomorphisms induced by the inclusion $f_k^{i,j}$:*

$$H_k^{i,j} = \operatorname{im} f_k^{i,j} = Z_k(K_i)/(B_k(K_j) \cap Z_k(K_i)), \text{ for } 0 \leq i \leq j \leq k \tag{4.14}$$

Consider the birth point of a homology class as the point where it is created, and the death point as the point where it becomes trivial or identical to some class that was born before. The birth and death points of a homology class are called the homological **critical values**.

Definition 4.1.24 (Critical value). *Let $\mathbb{X}$ be a topological space and $f : \mathbb{X} \to \mathbb{R}$ an inclusion function. Then, a homological critical value is a number $a \in \mathbb{R}$ for $k \in \mathbb{Z}$, such that for all sufficiently small $\varepsilon > 0$, the map $H_k(K_{a-\varepsilon}) \to H_k(K_{a+\varepsilon})$ is not an isomorphism.*

According to Cohen-Steiner et al. [21], a homology class $\alpha \in H_k(K_i)$ is born in point i if α did not exist in the image of the filtration function $f_k^{(i-\delta,i)}$, for any $\delta > 0$. And it dies entering point j if $f_k^{(i,j-\delta)}(\alpha)$ is not in the image of $f_k^{(i-\delta,j-\delta)}$ for any $\delta > 0$ but $f_k^{(i,j)}(\alpha)$ is in the image of $f_k^{(i-\delta,j)}$.

Persistence diagram

From now on we will denote the birth point of the k-homology class α by $b(\alpha) = i$ and its death point by $d(\alpha) = j$. Then, the persistence of α is equal to the difference between the parameter value where it dies and the parameter value where it is born:

$$\text{Pers}(\alpha) = d(\alpha) - b(\alpha) = j - i \tag{4.15}$$

The k-persistent Betti numbers, $\beta_k^{i,j}$, are the ranks of the k-persistent homology groups. These contain all the information about the k-persistent homology groups.

Lemma 4.1.25 (Fundamental lemma of persistent homology)**.** *Given a filtration as in Eq. 4.12, for every $0 \leq l \leq m \leq n$ and every dimension k, one can express the k-persistent Betti numbers as:*

$$\beta_k^{l,m} = \sum_{i \leq l} \sum_{j > m} \mu_k^{i,j}, \tag{4.16}$$

where $\mu_k^{i,j}$ denotes the multiplicity, defined as the number of k-classes that are born with K_i and that die with K_j. The multiplicity is given by:

$$\mu_k^{i,j} = (\beta_k^{i,j-1} - \beta_k^{i,j}) - (\beta_k^{i-1,j-1} - \beta_k^{i-1,j}) \quad \forall i < j, \forall k \tag{4.17}$$

The k-persistent diagram, $\mathrm{D}_k = \mathrm{D}_k(f_k)$, is a visualization of the k-persistent Betti numbers, developed by Edelsbrunner, Letscher and Zomorodian [45]. It consists on a plot of the birth- versus the death- points of the independent k-homology classes along a filtration.

Definition 4.1.26 (Persistence diagram, according to Cerri et al. [15])**.** *The k-persistence diagram, D_k, is the set of all points $\{(i,j) \in \mathbb{R} \times \mathbb{R} : i < j\}$ such that $\mu_k^{i,j} > 0$ (counted with their own multiplicity), union the set of all points $\{(i,j) \in \mathbb{R} \times \mathbb{R} : i = j\}$ (counted with infinite multiplicity).*

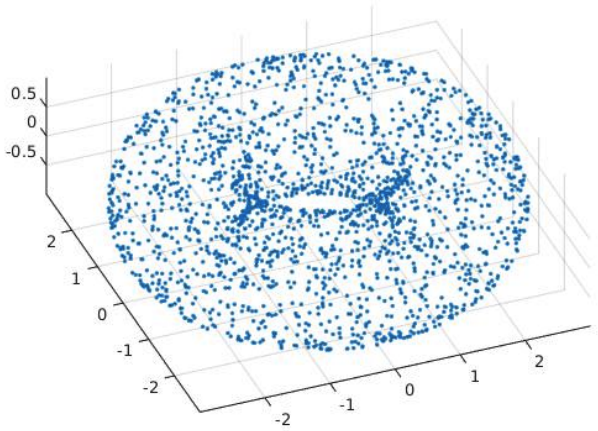
In Fig. 4.3 we show two different persistence diagrams for a same PCD obtained from sampling a torus. This PCD is included in the default examples of the JavaPlex library [116]. The different persistence diagrams illustrate how the differences that the selection of landmark points can produce, in the case of computing persistent homology using witness complexes. It suggests that the *minmax* selector is better to recover the homology of the PCD.

Barcodes

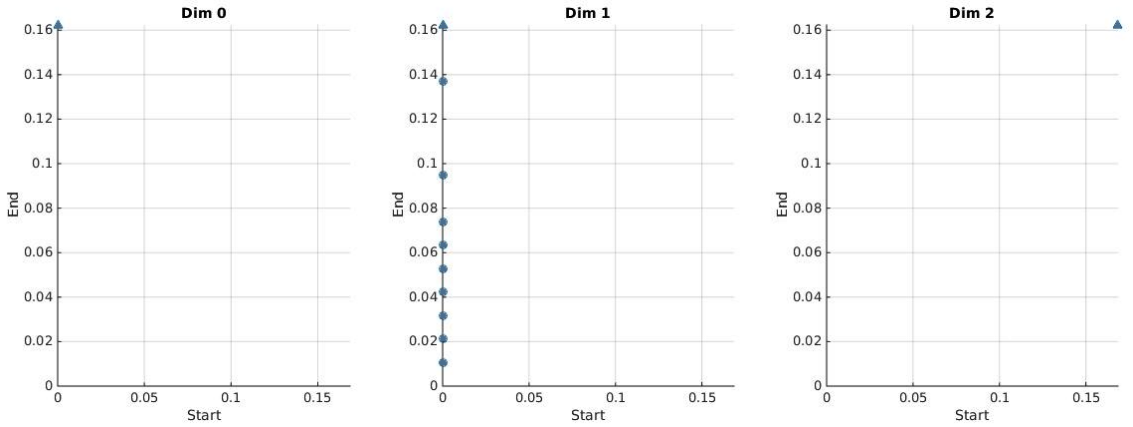
A barcode of a k-homology group $H_k^{i,j}$ is a plot containing a collection of horizontal line segments, where the x-axis indicates the values of parameter ϵ and the y-axis indicates an ordering of the homology generators. This visualization tool was developed by A. Collins, A. Zomorodian, G. Carlsson and L. J. Guibas in 2004 [22].

In a barcode, the number of intervals between two given ϵ values, i and j, is equal to the rank of $H_k^{i,j}$. This way, it provides the same homology information as a Betti number.

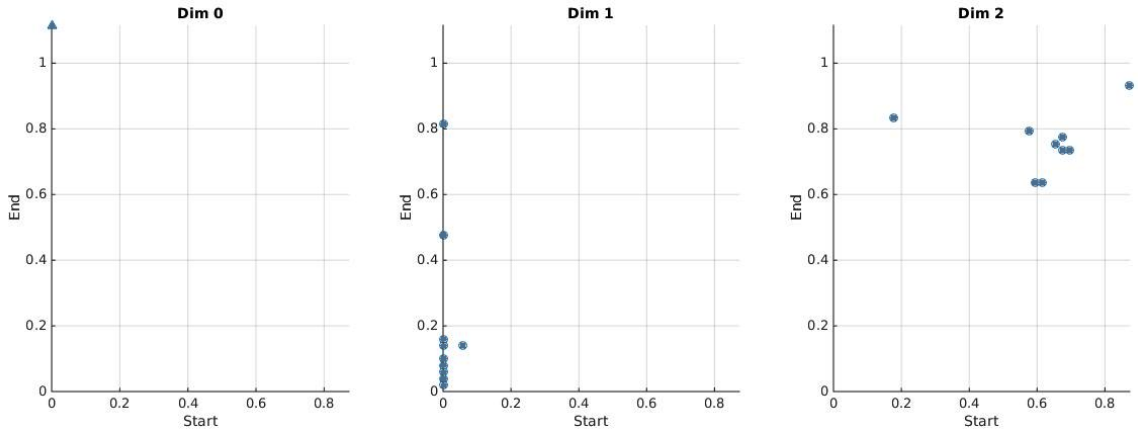
The length of every interval in a barcode can be intuitively related to the persistence of topological features. Again, long intervals would correspond to large scale geometric features and short intervals to noise. However, whether an interval is considered long or short depends on the system analyzed.



(a) PCD: 2000 points sampling a torus.



(b) Persistence diagram produced with *minmax* selected landmark points.

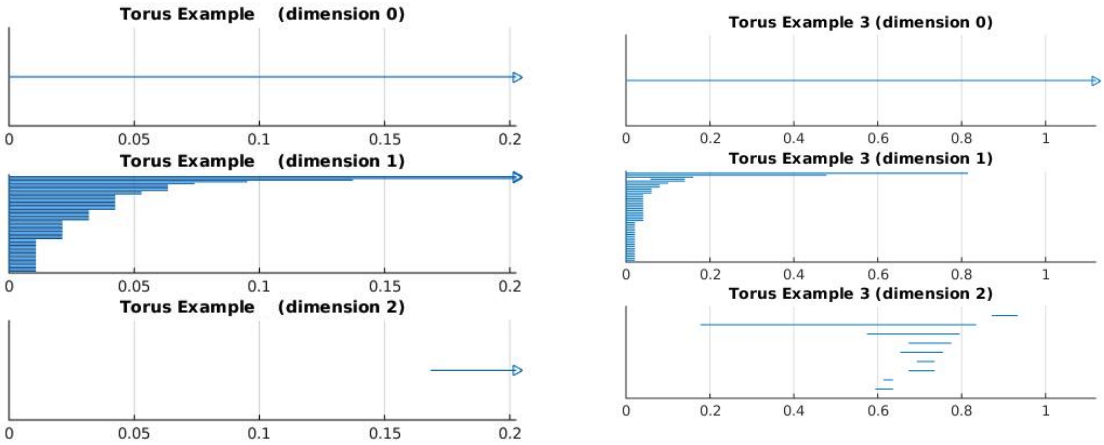


(c) Persistence diagram produced with randomly selected landmark points.

Figure 4.3: (a) A point cloud data (PCD) of 2000 points sampling a torus, following a uniform distribution. (b) Persistence diagram associated to this PCD when using a minmax *landmark point selector. (c) Persistence diagram associated to this PCD when using randomly selected landmark points. Both persistence diagrams were computed for maximum homology group 2, selecting 150 landmark points to construct a witness complex (see Section. 4.1.1), and considering 50 intervals for the filtration. Every circle in the plot indicates the birth (x-axis) and death point (y-axis) of a topological feature in a homology group (Dim). The triangular points indicate features whose death point is at infinity.*

This way, the information of a barcode is more meaningful when we are interested on analyzing different scales of representation of a space.

In Fig. 4.4 we show two different barcodes for the same PCD shown in Fig. 4.3(a). This illustrates the differences originated by the different selection of landmark points used to compute witness complexes.



(a) Barcode produced using *minmax* selector.

(b) Barcode produced using random selector.

Figure 4.4: Consider the same point cloud data and parameters of Fig. 4.3. Then, (a) shows the barcode associated to this PCD when using a minmax *landmark point selector and (b) shows the barcode associated to this PCD when using randomly selected landmark points. Intervals with a triangular end are those that die at infinity. We can see that the random landmark points selector produces clearer intervals in H_1 but several additional intervals in H_2. Considering the PCD that originates them, the minmax selector seems to be more adequate.*

4.1.5 Stability thorems

An important question in the study of persistence is how much the computations are affected by the presence of noise and approximations (inherent to the measurement process). For persistent homology to be stable, perturbations in the data should not produce big variations in its computations.

As proven by Cohen-Steiner, Edelsbrunner and Harer [20], persistence diagrams are stable. This means that a small variation in the filtration function will produce small changes in the points out of the diagonal in the persistence diagrams.

Proving stability is in general a complicated task. To deepen into this topic, we refer the reader to the articles of A. Cerri et al. [14, 16]; Cohen-Steiner, Edelsbrunner, Harer and Mileyko [20, 21]; P. Bendich, T. Gulkovskyi and J. Harer [4]; B. T. Fasy et al. [48]; or P. Niyogi, S. Sale and S. Weinberger [95].

In order to know how much a perturbation in the data affects the persistent homology results, it is necessary to measure the similarity between two given persistence diagrams or barcodes. There are different measures to compute similarity between distributions; for example, the Kullback-Leibler divergence, the matching distance, the bottleneck distance and the Wasserstein distance.

The Kullback-Leibler divergence measures the loss of information when transiting from one distribution to the other. It is commonly used to measure the similarity between two probability distributions. However, this measure is generally non-symmetric and therefore is not

a true metric [88]. A common approach to overcome this, consists on symmetrizing it [110].

In the following sections we will comment further on the matching, the bottleneck and the Wasserstein distances. These measurements are relevant in the persistent homology literature; see for example [17]. Following the chronological order of their development, we start describing the matching distance. Later, the bottleneck distance, which reduced the computational complexity of the matching distance computations, while maintaining the same boundaries in accuracy. Since this measure shows no sensitivity to differences in bijection beyond the furthest pair of corresponding points in the persistence diagram, the Wasserstein distance was developed. However, since there is no general stability theorem for the Wasserstein distance, the bottleneck distance is still the most widely used measure of similarity between persistence diagrams or barcodes.

Later on, in Section 4.2, where we analyze additional considerations for the analysis of complex time series, we will refer again to the bottleneck and the Wasserstein distances to define confidence sets [48], the stability of "persistence landscapes" [18]−a tool introduced by P. Bubenik [8] in order to reduce the computational cost of persistent homology calculations−, and the Kernel density estimation (KDE) approach [48].

The KDE approach is used in our new method for the identification of transitions between dynamical regimes in real-world data (see Section 4.3), where we estimate the "mean persistence" (Eq. 4.1) from multiple samples of a data set in the same fashion as in [48].

Matching distance

Given two k-persistence diagrams D_k and D'_k, the matching distance $d_{match}(D_k, D'_k)$ measures the minimum between the cost of moving a point in D_k to a point in D'_k and the cost of moving these two points onto the diagonal.

Let us define the domain of points in a persistence diagram, $\overline{\Delta^+}$, in terms of those points lying in the diagonal and those out of this. This way, $\overline{\Delta^+} = \Delta^+ \cup \Delta = \{(i,j) \in \mathbb{R} \times \mathbb{R} : i < j\} \cup \{(i,j) \in \mathbb{R} \times \mathbb{R} : i = j\}$.

Then, given two points (i,j), $(i',j') \in \overline{\Delta^+}$, the distance between them is given by

$$d\big((i,j),(i',j')\big) = \min\left\{\max\left\{|i-i'|, |j-j'|\right\}, \max\left\{\frac{j-i}{2}, \frac{j'-i'}{2}\right\}\right\}$$

Considering σ a function that varies over all bijections between D_k and D'_k, the matching distance can be expressed by

$$d_{match}(D_k, D'_k) = \min_{\sigma} \max_{p \in D_k} d(p, \sigma(p)) \tag{4.18}$$

As proved by Cohen-Steiner, Edelsbrunner and Harer [20], the matching distance between two persistence diagrams computed with different filtration functions is bounded by the L_∞-norm between these functions. This way, the matching distance preserves the stability of persistence diagrams: a small variation in the filtration function will produce small changes in the persistence diagrams and in the matching distance. For this reason, this distance is a very good descriptor of stability.

However, if M is the total number of proper points in a persistence diagram, the computational complexity of the matching distance is $O(M^{2.5})$ [26]. This makes it hard to compute for large data sets. Coarse approximations to this distance have been provided, for example by Cerri et al. in [15], for cases where not too much refinement is needed.

Bottleneck distance

Let X and Y be two barcodes (or persistence diagrams) computed with different filtration functions. Then, the bottleneck distance is the infimum over all bijections $\eta : X \to Y$, of all the distances measured with L_∞-norm between birth and death points in every bar, x, in the barcodes (for a given homology group) associated to any two steps in the process. In other words, it is given by

$$W_\infty(X, Y) = \inf_{\eta: X \to Y} \sup_{x \in X} \|x - \eta(x)\|_\infty \tag{4.19}$$

As can be seen from Eq. 4.19, the bottleneck distance is a metric because $W_\infty(X, Y) = 0$ iff $X = Y$, $W_\infty(X, Y) = W_\infty(Y, X)$ and $W_\infty(X, Z) \leq W_\infty(X, Y) + W_\infty(Y, Z)$.

The stability theorem for filtrations, states that for a dimension k, the bottleneck distance between the persistence diagrams generated with two different **tame** filtering functions is upper bounded by the L_∞ norm of the difference between the two filtering functions [44].

Definition 4.1.27 (Tame function)**.** *A function $f : \mathbb{X} \to \mathbb{R}$ is tame if it has a finite number of homological critical values and the homology groups, $H_k(f^{-1}(-\infty, a]), are finite dimensional for all $k \in \mathbb{Z}$ and $a \in \mathbb{R}$.*

Theorem 4.1.28 (Bottleneck stability for persistent diagrams [44])**.** *Let X be a triangulable space with continuous, tame functions $f, g : X \to \mathbb{R}$. And let $D_k(f)$ and $D_k(g)$ their respective associated k-persistence diagrams. Then the following is satisfied:*

$$W_\infty(D_k(f), D_k(g)) \leq \|f - g\|_\infty \tag{4.20}$$

This stability theorem can also be thought in terms of finite point clouds.

Let $\mathbb{X}$ and $\mathbb{Y} \in \mathbb{R}^d$ be two point clouds. These have induced tame distance functions, $d_\mathbb{X}$ and $d_\mathbb{Y}$, where $d_\mathbb{X}(a) = \inf_{x \in \mathbb{X}} \|x - a\|$ and $d_\mathbb{X}^{-1}(-\infty, a)$ is homotopic to the filtration of Čech complexes of $\mathbb{X}$. Then, thanks to Theorem 4.1.28, one obtains that [92]:

$$W_\infty(D(d_\mathbb{X}), D(d_\mathbb{Y})) \leq \|d_\mathbb{X} - d_\mathbb{Y}\|_\infty \tag{4.21}$$

Now, considering that $H(\mathbb{X}, \mathbb{Y}) = \|d_\mathbb{X} - d_\mathbb{Y}\|_\infty$ and Theorem 4.1.28, one obtains the following corollary [92].

Corollary 4.1.29. *Given two finite point clouds, $\mathbb{X}, \mathbb{Y} \in \mathbb{R}^d$:*

$$W_\infty(D(d_\mathbb{X}), D(d_\mathbb{Y})) \leq H(\mathbb{X}, \mathbb{Y}) \tag{4.22}$$

Wasserstein distance

Despite its stability and being easily computable, a drawback of the bottleneck distance is that it is not sensitive to small differences in the bijection beyond the furthest pair of corresponding points in the persistence diagram. The Wasserstein distance overcomes this drawback.

For a given degree q, persistence diagrams X and Y and bijections $\eta : X \to Y$, the q-Wasserstein distance between X and Y is given by:

$$W_q(X, Y) = \left[\inf_{\eta: X \to Y} \sum_{x \in X} \|x - \eta(x)\|_\infty^q \right]^{1/q} \tag{4.23}$$

Intuitively, the Wasserstein distance can be understood as the physical work needed to transform a pile of earth with shape equal to the filtration function f, to a pile of earth with the shape of g. This interpretation motivates the other name by which this distance is known for $q = 1$: the *earth movers distance* (EMD).

The Wasserstein distance, $W_q(X, Y)$, also constitutes a true metric. And, as can be seen from Eq. 4.23 the bottleneck is equal to the Wasserstein distance in the limit when $q \to \infty$.

A stability theorem as the one proven for the bottleneck distance is, in general, not possible for the q-Wasserstein distance. However, Cohen-Steiner et al. [21] have proven that there exist constants, $k \leq q$ and C, in the case when the filtration functions f and g are Lipschitz, such that:

$$W_q(f, g) \leq C\|f - g\|_\infty^{1-k/q} \tag{4.24}$$

These authors have also proven that, for every $q \geq k+1$, the degree-q total persistence which expresses the sum of q-th powers of persistences, denoted by $\mathrm{Pers}_q(f)$, satisfies

$$|\mathrm{Pers}_q(f) - \mathrm{Pers}_q(g)| \leq C\|f - g\|_\infty \tag{4.25}$$

In practice, given the constrains needed to state a stability theorem for the Wasserstein distance, the Bottleneck distance is more commonly used for topological data analysis.

4.2 Considerations for the analysis of complex time series – *With persistent homology*

As mentioned before, some of the problems found when analyzing real-world data are its length, noise and existence of outliers.

In order to solve the problems of noise and outliers when computing the persistent homology of real-world data, the use of more robust simplicial complex constructions is crucial. However, the computational cost of these constructions is always problematic when dealing with large data sets.

An approach to solve the problem of the high computational cost of persistent homology calculations, is to subsample the data. With this idea in mind, P. Bubenik [8] introduced in 2012 the concepts of *persistence landscape*, *p-landscape distance* and *mean persistence landscape*.

A persistence landscape is a function, $\lambda : \mathbb{N} \times \mathbb{R} \to \overline{\mathbb{R}}$, , where $\overline{\mathbb{R}} = [-\infty, \infty]$, created by taking a data sample and transforming its persistence diagram information into an additive function.

Definition 4.2.1 (Persistence landscape [8])**.** *Being b the birth and d the death of a persistence module, define $t = \frac{b+d}{2}$ and $m = \frac{d-b}{2}$. Then, a persistence landscape can also be defined as a sequence of functions $\lambda_k : \mathbb{R} \to \overline{\mathbb{R}}$, where $\lambda_k(t) = \lambda(k, t) = \sup\{m \geq 0 \mid \beta^{t-m,t+m} \geq k\}$ and $\beta^{t-m,t+m} = \dim(im(M(t - m \leq t + m)))$.*

Once a persistence landscape is calculated, it is interesting to distinguish it from another persistence landscape. This difference is measured by the p-landscape distance.

Definition 4.2.2 (p-Landscape distance [8]). *Given two persistence diagrams D and D', the p-landscape distance between them is given by*

$$\Lambda_p(D, D') = \|\lambda - \lambda'\|_p \tag{4.1}$$

Two important results are that persistence landscapes are stable with respect to the ∞-norm, and that the ∞-landscape norm is bounded by the bottleneck distance.

These definitions are useful to deal with the problem of analyzing a large data set, X. As mentioned before, this problem is approached by taking subsamples of X. By adding the persistence landscapes associated to different data samples, one may compute the mean persistence landscape. This then provides a topological statistical summary of the metric measure space.

Definition 4.2.3 (Mean persistence landscape [8]). *Consider $X_1, ..., X_n$ i.i.d. distributed copies of a data set X, and their associated persistence landscapes $\Lambda^1, ..., \Lambda^n$. Then the mean landscape is given by $\overline{\Lambda}^n(\omega) = \overline{\lambda}^n$, where*

$$\overline{\lambda}^n(k, t) = \frac{1}{n} \sum_{i=1}^{n} \lambda^i(k, t) \tag{4.2}$$

By the strong law of large numbers and the central limit theorem on Banach spaces, persistence landscapes convergence can be proven. Additionally, it is possible to define confidence intervals for the statistical approximation of persistence landscapes.

In 2014, Chazal et al. [18] added on the methodology of P. Bubenik for approximating the persistence landscape of a metric measure space. They found that a mean persistence landscape is stable to perturbations of the measure in the Wasserstein metric.

All their results consider that every data sample lies in a metric space $(\mathbb{M}, \rho)$ and is drawn *i.i.d.* from an unknown measure μ supported in a compact subset of $X_\mu \subseteq \mathbb{M}$. For any $a, b > 0$, $x \in X_\mu$ and $r > 0$, then $\mu(B(x, r)) \leq \min(ar^b, 1)$.

In the same year, Fasy et al. [48] provided a way to separate the noise in a data set. They defined confidence sets for different topological quantities. Their main idea was that the points within a certain band around the diagonal in a persistence diagram constitute noise.

Definition 4.2.4 (Confidence sets [48]). *Consider $S = \{X_1, ..., X_n\}$ a sample from a distribution P concentrated on or near $\mathbb{M} \in \mathbb{R}^d \subset \mathbb{R}^D$. Let $\mathcal{P}$ be the persistence diagram defined by the lower level sets $\{x : d_{\mathbb{M}}(x) \leq \varepsilon\}$, and $\hat{\mathcal{P}}$ be the persistence diagram of $\{x : d_{S_n}(x) \leq \varepsilon\}$. Given $\alpha \in (0, 1)$, define $c_n = c_n(X_1, ..., X_n)$ by*

$$\lim_{n \to \infty} \sup \mathbb{P}(W_\infty(\hat{\mathcal{P}}, \mathcal{P}) > c_n) \leq \alpha \tag{4.3}$$

Then, the confidence set C_n is a subset of all persistence diagrams whose distance to $\hat{\mathcal{P}}$ is almost c_n. This means

$$C_n = \{\tilde{\mathcal{P}} : W_\infty(\hat{\mathcal{P}}, \mathcal{P}) > c_n\} \tag{4.4}$$

For their theoretical calculations while subsampling, given a data point cloud with n data points, Fasy et al. [48] take $N = \binom{n}{b}$ subsamples with $b_n = O\big(\frac{n}{\log n}\big)$ data points each, where $b_n \to \infty$. These theoretical number of samples and their length, are sometimes costly to compute. However, it is sufficient to take a large number of samples.

However, there is another contribution within the same article of Fasy et al. [48] which is more interesting for us. They introduce a density-based method which is very insensitive to outliers and noise. Therefore, this approach seems to be more adequate when dealing with real-world data.

The main idea behind such density-based approach is to construct a density estimator out of the data. Then, taking a filtration of the upper level sets of such density estimator, it is possible to define a persistence diagram. For this approach, they consider the standard kernel density estimator (KDE).

Let p_h be the density of the probability measure $P_h = P \star \mathbb{K}_h$, a smooth version of P, where $\mathbb{K}_h(A) = h^{-D}\mathbb{K}(h^{-1}A)$ and $\mathbb{K}(A) = \int_A K(t)dt$, for K a smooth symmetrical kernel. Then an estimator of p_h is given by the KDE, $\hat{p}_h(x)$, where $\mathbb{E}(\hat{p}_h(x)) = p_h(x)$.

Definition 4.2.5 (Kernel density estimation approach [48]). *Let $X_1, ..., X_n$ be a sample from a distribution P concentrated on or near $\mathbb{M} \subset \mathbb{R}^D$. Then the kernel density estimator of p_h is given by*

$$\hat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^D} K\left(\frac{\|x - X_i\|_2}{h}\right) \tag{4.5}$$

4.3 *A new method for the identification of transitions between dynamical regimes in real-world data*

In this section we introduce a new method for the identification of transitions between different dynamical regimes in a system. This means, identifying the emergence of more attractors or a strong deformation of the attractors in the state space of a system.

Our method is summarized as follows. Let $\mathbf{U} = \{\mathbf{u}_0, \mathbf{u}_1, ..., \mathbf{u}_N\}$ be a time series data describing a non-stationary process and $\mathbf{X}$ the state space reconstructed from $\mathbf{U}$. Let $\mathbf{X}$ be the state space already reconstructed from $\mathbf{U}$. We proceed to divide X it into W windows of equal length. The time lapse of a window should be enough to capture a dynamical regime of the system. We then compute the topological summary for each window (mean h-persistence in the (r, c)-space, see Eq. 4.1) and compare these summaries between all windows using the so called total overlap (see Eqs. 4.2 and. 4.3).

This way, we can identify the topological differences between windows of measurements, which indicate the dynamical transitions in the time series data in a scale of time related to the length of the windows of measurements.

4.3.1 Estimating the mean h-persistence: a summary of the topological information of a time series

Let us start by assuming that the state space of a time series has already been reconstructed (in Section 4.4 we introduce a new method for state space reconstruct based on persistent homology), and denote this by $\mathbf{X}$. Then, as mentioned above, we divide X into W windows of length $M < N$: $\{X_0, ..., X_W\}$.

Every window should to include enough state space vectors to capture the dynamics of a given configuration of the state space. This means, enough state space vectors to sample all

the attractors existent at a certain moment of a possibly non-stationary process. However, to satisfy this condition, M may be large and we may face high computational costs for estimating the associated persistent homology of every window of state space vectors.

To solve this problem, we consider the statistical approach of Fasy et al. [48] to estimate a mean persistence diagram, containing the birth and death points, b and d respectively, of every h-homology group. However, we use variables $r = \frac{d+b}{2}$ and $c = \frac{d-b}{2}$, instead of b and d, as done by P. Bubenik [8]. Thus, we estimate a mean (r, c)-persistence diagram for every window of state space vectors.

Let $\mathcal{P}_h(\mathbf{X}_w)$ denote the mean persistence for the h-th homology group of the w-th window of state space vectors X_w. Then, to estimate $\mathcal{P}_h(\mathbf{X}_w)$ we take S random data samples of length $L \ll M$ from $\mathbf{X}_w$: $\{\mathbf{Y}_w^0, ..., \mathbf{Y}_w^S\}$. The computational details of how to obtain $\mathcal{P}_h(\mathbf{X}_w)$ are contained in Algorithm 4.

Let us denote by $\mathcal{P}_h(\mathbf{Y}_w^s)$ the h-persistence of the s-th sample of the w-th window of state space vectors. This is estimated using the Kernel density approach (see Section 4.2, Eq. 4.5), as the probability density function of the centers and radii in the persistence diagram associated to the s-th sample. For this estimation we use the multi-dimensional kernel density estimator (KDE) from the BEAST library [39].

This way, the mean persistence for the h-th homology group of the w-th window of state space vectors is then given by

$$\mathcal{P}_h(\mathbf{X}_w) = \frac{1}{S} \sum_{s=0}^{S} \mathcal{P}_h(\mathbf{Y}_w^s) \tag{4.1}$$

Algorithm 4 - Persistence statistics

1. Take a window of state space vectors of length M, $\mathbf{X}_w$.

2. Compute the h-persistence for S random data samples from $\mathbf{X}_w$:

for $s = 0$ **to** S **do**
- ▷ Take a random data sample of length $L \ll M$ from $\mathbf{X}_w$, $\mathbf{Y}_w^s$.
- ▷ Compute the persistence diagram of $\mathbf{Y}_w^s$, $\mathrm{D}(\mathbf{Y}_w^s)$.
- ▷ Obtain the maximum homology group populated in $\mathrm{D}(\mathbf{Y}_w^s)$, H.
- ▷ Obtain the maximum filtration value of $\mathrm{D}(\mathbf{Y}_w^s)$, f_w^s.

  for $h = 0$ **to** H **do**
  - ▷ Obtain birth and death points in $\mathrm{D}_h(\mathbf{Y}_w^s)$, (b, d).
  - ▷ Compute radii, $r = (d + b)/2$, and centers $c = (d - b)/2$.
  - ▷ Compute a two-dimensional KDE over the (r, c)-space to estimate the h-persistence of $\mathbf{Y}_w^s$, $\mathcal{P}_h(\mathbf{Y}_w^s)$.

  end for

end for

3. Compute the mean h-persistence for $\mathbf{X}_w$, $\mathcal{P}_h(\mathbf{X}_w)$, according to Eq. 4.1, for $h = [0, H]$.

4.3.2 Identifying dynamical transitions

In an analogy to the use of linear correlation, we define the overlap between the mean h-persistence of different windows of state space vectors, $\mathbf{X}_{w_i}$ and $\mathbf{X}_{w_j}$, $OV_h(\mathbf{X}_{w_i}, \mathbf{X}_{w_j})$ by

$$OV_h(\mathbf{X}_{w_i}, \mathbf{X}_{w_j}) = \left|\min\left(\mathcal{P}_h(\mathbf{X}_{w_i}), \mathcal{P}_h(\mathbf{X}_{w_j})\right)\right| \tag{4.2}$$

This definition corresponds to the definition of Lee et al. [76] for the overlapping area between

two joint probability distribution functions, but adapted to the comparison of the mean h-persistence estimations.

By using persistent homology, we incorporate high-order topological information to the comparison between windows of measurements. This way we approach one of the general problems associated to the use of correlation measurements to compare data: their use of low-dimensional information of single attractors.

We define the total overlapping between these two windows of state space vectors by

$$OV(\mathbf{X}_{w_i}, \mathbf{X}_{w_j}) = \sum_{h=0}^{H} OV_h(\mathbf{X}_{w_i}, \mathbf{X}_{w_j}) \tag{4.3}$$

We say that those pairs of windows of state space vectors having larger overlap are associated to state space configurations with similar dynamics: similar geometry of their attractors, for example. The definition of boundaries for determining whether the overlap is low of high will not be covered for now, but left for future work.

Therefore, whenever two neighboring windows of state space vectors have low overlap, we say there is a dynamical transition.

4.4 Our criteria for an adequate state space reconstruction

As we mentioned in Section 2.3, there are various definitions of what a *good embedding* is. In general, an adequate reconstruction of the state space of a system from a time series should provide all the dynamical information of such system.

Now, considering the concepts of homotopy and homology reviewed in Sections 4.1.2 and 4.1.3, it turns inevitable to ask wether persistent homology can say anything about the embedding parameters necessary to produce an adequate state space reconstruction from a time series.

The definition of equivalence of embeddings of Cross and Gilmore [24] states that in an adequate state space reconstruction, there is a smooth map going from the reconstructed space to the true state space. This relates to the definition of homotopy, in which two continuous maps $f, g : \mathbb{X} \to \mathbb{Y}$ are homotopic if there is a continuous map $H : \mathbb{X} \times [0, 1] \to \mathbb{Y}$ such that $H(x, 0) = f(x)$ and $H(x, 1) = g(x) \, \forall x \in \mathbb{X}$ (see Def. 4.1.14). This way, we would expect a good embedding to be homotopy equivalent to the true state space.

In terms of persistent homology, this means that the persistence diagrams (or barcodes) of the true and the reconstructed state spaces should be the same or *almost* the same. Thus, the distance between them, measured with any of the distances given in Section 4.1.5, should be either zero or very small.

Additionally, the largest non-trivial homology class in the reconstructed space, should also be equal to the one of the true state space. This statement derives from the results of Niyogi, Smale and Weinberger [95], who have stated that, when identifying the homology of a submanifold of dimension d from a data sample, one would see the j-th Betti number be zero for all $j > d$.

However, when we analyze real-world data, the most common situation is not knowing what the true state space of the system underlying a time series is. In this case, the definition

of equivalence of embeddings is not enough. We suggest that an alternative approach could follow the definition of good state space reconstruction given by Palit et al. [97]. According to these authors, a good reconstruction not only has the smallest shape distortion of the attractor but also has a more dense reconstructed attractor, with less outliers.

Recently, Dey, Fan and Wang [29] have approached the problem of the state space reconstruction and provided a way to estimate the embedding dimension. Their topological methodology uses Vietoris-Rips complexes to detect the dimension of a hidden manifold from a data sample using local homology. Its computation depends on what they call the "reach", which means the minimum distance from any point in the sample to its medial axis. This methodology has the advantage of being less sensitive than other methods to the local distribution of points, since it does not require uniformity. However, since it depends on the computation of the medial axis transform (a shape descriptor), it is very sensitive to small perturbations. Our suggestion to alleviate the sensitivity to perturbations in this approach is to replace the medial axis by the "scale axis transform", introduced by Giesen et al [56], in the definition of reach of Dey, Fan and Wang. The scale axis transform is a scale adaptive skeletal shape representation based in the medial axis transform. The investigation of the use of the scale axis transform for the computation of the embedding dimension is an interesting topic for future research.

Another interesting topic for future investigation is the estimation of the embedding delay using k-order combinatorial Laplacians [91]. These Laplacians are the generalization of the graph Laplacian to simplicial complexes and relate to the existence of k-dimensional bottlenecks in a graph. Their spectrum indicate how far from having a non-trivial k-th homology class a simplicial complex is. To the best of our knowledge, the behavior of the k-order combinatorial Laplacians spectrum is not well studied yet.

In our new method for the reconstruction of the state space, introduced in Section 4.5, we consider that, in terms of persistent homology, a good state space reconstruction satisfies the following condition:

- When reconstructing the state space from a subsample of the data, its associated persistence diagram (or barcode) will be "similar" to the one obtained from another subsample.

By similar we mean that the main topological features should be recovered, within a confidence interval. And that the existing differences should be additional points close to the diagonal in the persistence diagram (or short intervals in the barcode) created from the sample.

This implies that the reconstructed attractor taking the full data set is dense. And that if a sample of the data is sufficiently long, it should be able to reconstruct the same attractor. This assumption is inspired by the criteria of Palit et al. [97] for an adequate state space reconstruction in terms of shape distortion (see Section 2.3).

4.5 *A new method for state space reconstruction – Based in persistent homology*

In Section 4.3 we introduced our new method for the identification of transitions between dynamical regimes in complex data. For this, we assumed having an adequate state space reconstruction from a time series. Such reconstruction could be performed with the method

presented in Section 3.4 or with classical methods, like false nearest neighbors and autocorrelation (see Section 2.1.2).

However, we want to make use of higher-order topological information for the reconstruction of the state space with the aim of uncovering dynamical features like different time-scales in the dynamics.

In this section we describe a new methodology for identifying adequate embedding parameters for the reconstruction of the state space of deterministic dynamical systems using persistent homology. This method uses the concepts defined in Section 4.3 and the ideas presented in Section 2.1.2 for the selection of adequate non-uniform embedding delays and embedding dimension. This methodology is described in Algorithm 5 and summarized as follows.

Given a time series $\mathbf{U}$, we start by creating a two-dimensional τ-delayed time series, $\mathbf{X}(\tau)$, where $\tau \in [0, \tau_{max}]$. And we take an initial window of state space vectors of length $M < N$, $\mathbf{X}_0(\tau)$.

As mentioned before, determining the size of M is a problem itself and will depend on the specific characteristics of the analyzed system, but it should be enough to capture a dynamical regime of the system.

Following the steps described in Section 4.3.1, we estimate a mean h-persistence diagram for the initial window $\mathbf{X}_0(\tau)$, using variables $r = \frac{d+b}{2}$ and $c = \frac{d-b}{2}$, instead of b and d (where b and d denote the birth and death points, respectively, of every h-homology group). We denote this h-persistence diagram by $\mathcal{P}_h(\mathbf{X}_0(\tau))$.

For the estimation of the mean h-persistence diagram for the initial window, we use S random data samples of length $L \ll M$ from $\mathbf{X}_0(\tau)$: $\{\mathbf{Y}_0^0(\tau), ..., \mathbf{Y}_0^S(\tau)\}$ (see Algorithm 4 for details).

Then, to obtain the embedding delays necessary for a non-uniform embedding, we follow the idea of selecting embedding delays in terms of local minima in linear autocorrelation, assuming that there is some similarity in their behavior[3].

This way, we select all embedding delays $\tau \in [1, \tau_{max}]$ that provide a local minima in the associated total overlapping (Eq. 4.3), $OV(\mathbf{X}_0(\tau), \mathbf{X}_0(\tau))$. These constitute the so-called set of candidate embedding delays, which we denote by $\mathbf{T}$.

The relation between the autocorrelation function and the total overlap is illustrated in Section 4.5.1.

Finally, we set the embedding dimension by using the embedding delays in $\mathbf{T}$ for the method of false nearest neighbors (see Section 2.1.2).

[3] As mentioned in Section 2.1.2, autocorrelation is one of the most common measurements used to determine the delays required for the reconstruction of a state space via the uniform embedding of a time series.

Algorithm 5 - STATE SPACE RECONSTRUCTION

1. Set maximum embedding delay to test, τ_{max}.

for $\tau = 1$ **to** τ_{max} **do**

▷ Create a two-dimensional τ-delayed time series, $\mathbf{X}(\tau)$.

▷ Compute mean persistence for the h-th homology group of $\mathbf{X}(\tau)$, $\mathcal{P}_h(\mathbf{X}_0(\tau))$. This is done according to Algorithm 4.

▷ Measure associated total overlapping, $OV(\mathbf{X}_0(\tau), \mathbf{X}_0(\tau))$, according to Eq. 4.3.

end for

2. Select the set of candidate embedding delays, $\mathbf{T}$, containing the embedding delays that provide local minima in total overlapping (Eq. 4.3).

3. Set embedding dimension using the embedding delays in the set of candidate embedding delays, $\mathbf{T}$, in the false nearest neighbors (FNN) method[a].

[a]To do a uniform embedding instead, use the first candidate embedding delay in $\mathbf{T}$ for the FNN method.

4.5.1 Total overlap *vs.* Autocorrelation function

To illustrate our assumption about the existence of a relation between the autocorrelation function (defined in Eq. 2.5) and the total overlap (defined in Eq. 4.3), let us use a second order autoregressive model, $AR(2)$.

An $AR(2)$ model is described by the following stochastic difference equation

$$x_{i+1} = \alpha_1 x_{i-\tau_1} + \alpha_2 x_{i-\tau_2} + \xi_i \tag{4.1}$$

When the roots of the characteristic equation given by $1 - \alpha_1 u - \alpha_2 u^2 = 0$ are complex conjugate, then x_τ is a stationary process.

For $\tau_1 = 3$, $\tau_2 = 5$, $\alpha_1 = 0.7$ and $\alpha_2 = 0.3$, ξ_i corresponds to white Gaussian noise. The time series corresponding to these parameters is shown in Fig. 4.5.

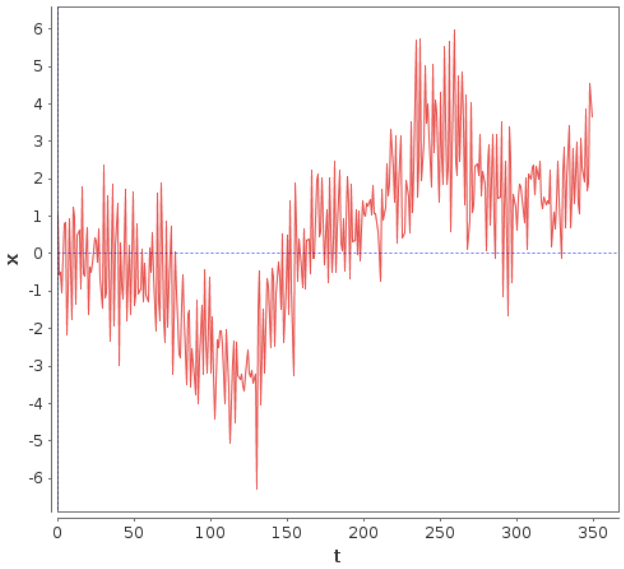
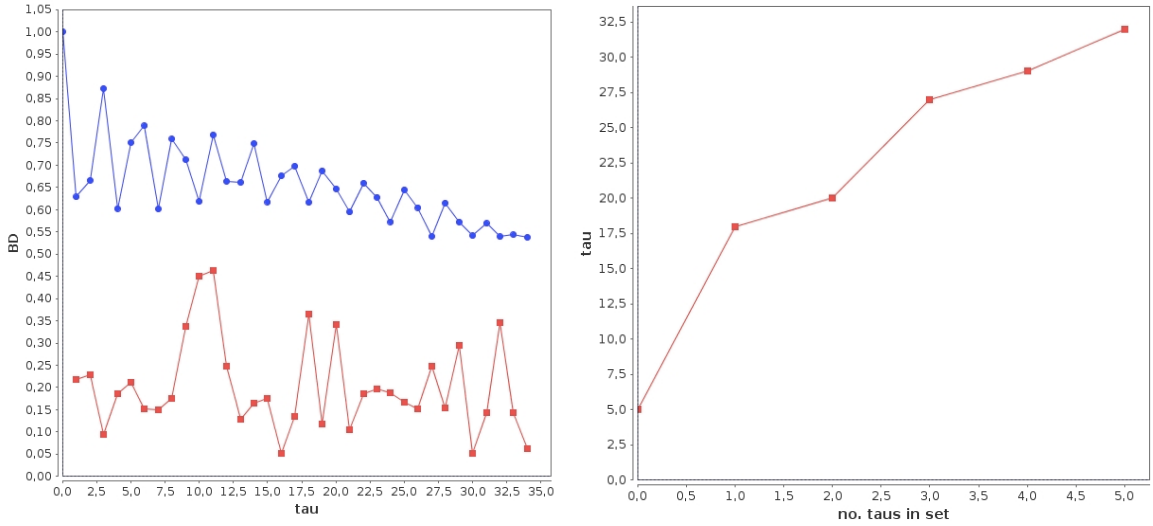


Figure 4.5: Time series for a second order autoregressive model, $AR(2)$, given by Eq. 4.1, using $\tau_1 = 3$, $\tau_2 = 5$, $\alpha_1 = 0.7$ and $\alpha_2 = 0.3$.

We calculate the autocorrelation and the total overlap over the time series shown in Fig. 4.5. These results are shown in Figure 4.6(a). This figure shows that the topological comparison

is better suited to identify the embedding parameters necessary to perform non-uniform state space reconstructions. This means, that it is better suited to analyze dynamical systems with more than one time scale in its dynamics.



(a) Autocorrelation (blue) *vs* Total overlap (red).

(b) Selected embedding delays.

Figure 4.6: (a) Plot that shows the relationship between the autocorrelation function and the total overlap between the persistence diagrams associated to the time series and their corresponding delayed time series. (b) Embedding delays that have persistence diagrams whose total overlap (see Eq. 4.3) from the non-delayed persistence diagram is a local minima.

In Fig. 4.6(b) we show the delays that, according to Figure 4.6(a), have local minima in total overlap. This plot shows the recovery of delay $\tau = 5$ and of other delays which are common multiples of $\tau = 3$.

4.6 Example 1: Two-dimensional double-well potential with varying depth in the potential wells

In this section we analyze a time series that shows a process in which the x-component of the barriers between the potential wells of a two-dimensional double-well potential increase and then decrease. The goal of analyzing this time series is to identify the increase in depth of the potential wells. This in turn, illustrates that the method introduced in Chapter 4 is a tool for the identification of dynamical transitions.

For all the computations of persistent homology we use JavaPlex [116]. And for all kernel estimations, we use a multi-variate kernel density estimator (KDE) from the BEAST library [39], using 30 intervals for the kernel approximation.

4.6.1 The system

An introduction to the standard one-dimensional double-well potential system was provided in Section 3.6. This time, we analyze a time series associated to a system describing the motion of a particle in a heat bath at temperature T, under the gradient of a two-dimensional double-well potential plus a random force (or Wiener process).

The starting configuration of the two-dimensional double-well potential is described by:

$$V(x, y, t) = \frac{1}{2}\big(x(t) - 1\big)^2 + \frac{1}{2}y(t)^2 \tag{4.1}$$

The configuration of this system changes over time following a process of W steps, during which the x-component of the barriers dividing the two potential wells increases. Let $\mathcal{T}$ denote the time step in such process.

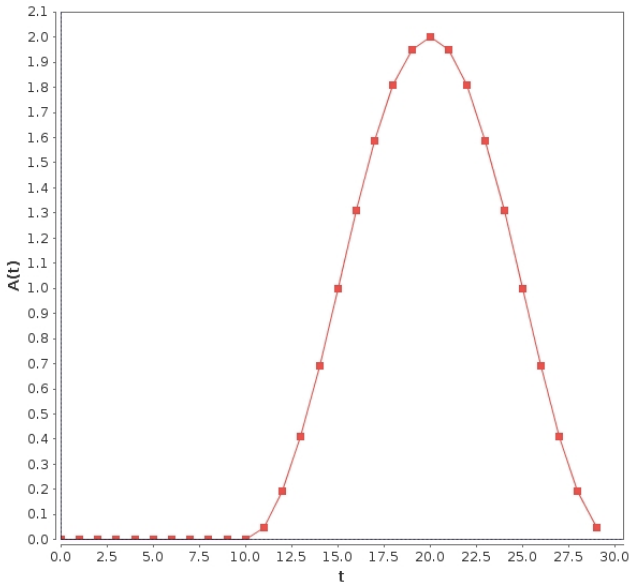


Figure 4.7: Plot of $A(\mathcal{T})$, the function defining the change in height of the barriers dividing the two potential wells of a two-dimensional potential. For the first ten steps of this process, $A(\mathcal{T}) = 0$. For the following twenty steps, it changes according to Eq. 4.2.

During the first $Z < W$ steps of the process, the configuration of the potential remains as in Eq. 4.1. However, during the remaining steps of the process, the height of the wall dividing these wells changes according to

$$V(x, y, \mathcal{T}, t) = \frac{1}{2}\big(x(t) - 1\big)^2 + \frac{1}{2}y(t)^2 \left(1 + A(\mathcal{T})e^{-\alpha\left(x(t)-x_0\right)^2}\right), \tag{4.2}$$

where $A(\mathcal{T})$ indicates the height of the wall dividing the potential wells and is given by:

$$A(\mathcal{T}) = 2\sin^2\left(\frac{\pi}{20}\mathcal{T}\right) \tag{4.3}$$

The variations in $A(\mathcal{T})$ for every step in the process are shown in Fig. 4.7. Consequently, we expect the total overlap to show a similar behavior to the change in potential barriers.

We construct a two-dimensional time series for a process with $W = 30$ steps, setting $Z = 10$, setting Eq. 4.2 to $\alpha = 50$. This produces the potential shown in Fig. 4.7.

For producing the two-dimensional time series, we proceed as in Section 3.6. This means, we integrate the double-well potential's Langevin dynamical equations using an Euler Maruyama integrator with lag time $\lambda = 0.001$, initial positions $(x_0, y_0) = (0, 1)$ and temperature $T = 400K$.

We create 5000 data points for every step in the process. And the time series for the entire process results from merging the time series simulated for every process step. This way, our full time series has 150000 data points.

4.6.2 Analysis results

Given the length of our time series, the calculation of its persistence homology is too costly in computational terms.

Therefore, we analyze it using the approach introduced in Section 4.3, which consists on dividing the time series into W different windows of measurements and estimating their mean (r, c)-persistence via subsampling.

We divide the time series into 30 windows of measurements; every window of measurements has 5000 data points. This way, every window corresponds to a different dynamical regime, well described by the two-dimensional time series it contains. By doing this, we are able to show empirically that the proposed method recovers the changes in dynamical regime.

This way, we can estimate a mean persistence diagram for every window of measurements, following the strategy described in Section 4.3 but substituting state space vectors by measurements data in it. We measure the differences in total overlap (see Eq. 4.3) between windows. Large differences in total overlap should correspond with transitions between dynamical regimes.

For the estimation of the mean (r, c)-persistence we use Algorithm 4, taking $S = 30$ random samples from every window of measurements. Each of these samples has $M = 700$ data points, corresponding to a 14% of the data points in a window. This is clearly a small number of samples and the results could be improved by taking a larger number of samples (see Section 4.2).

The persistence diagram is obtained by constructing a filtration of simplicial complexes and analyzing the topology of these simplicial complexes. The filtration we consider consists of 30 filtration intervals equally spaced. The simplicial complex construction we use is the witness complex with sequential *minmax* landmark selector, using L landmark points. The maximum homology group for which persistence is calculated is $H = 3$.

In order to stay within reasonable computational time when the number of data points used to compute a simplicial complex increases, we design the following heuristics to set the number of landmark points, L:

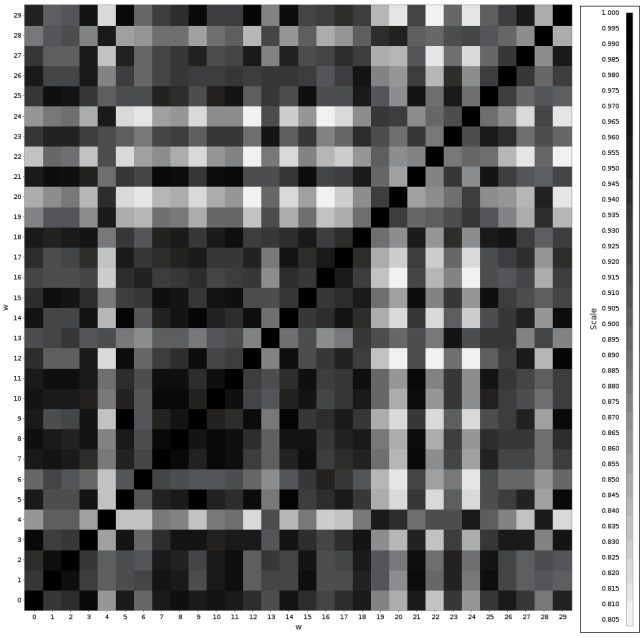
$$L = \frac{5M}{8}\left(e^{-M^{1/10}} - e^{-1} + 1\right) \tag{4.4}$$

In Fig. 4.8 we see the overlapping for homology groups 0 and 1. The homology group H_2 was not populated. And in Fig. 4.9 we see the total overlap between the 30 different windows of measurements.

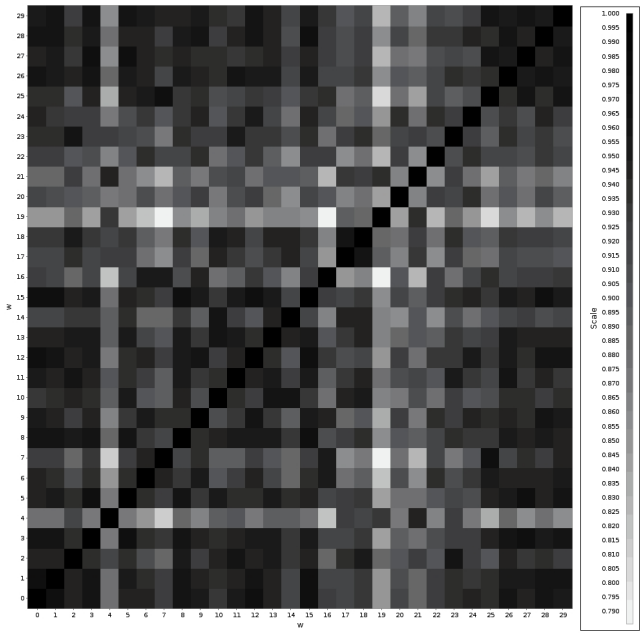
As seen in Fig. 4.9, total overlap is lower for windows of measurements 19 to 24 in relation to the other windows and the total overlap is high between all other windows. However, the total overlap is not high between windows of measurements 19 to 24.

Even when our results indicate changes in the configuration of the state space between the windows of measurements, there is no full match between the indicated overlap and the change in potential barriers.

Some explanations to why these results do not have a higher quality may be found in the implementation of our method. For example, by increasing the number of samples S taken for the calculation of the mean persistence, could provide some improvement.



(a) Overlapping for homology group 0.



(b) Overlapping for homology group 1.

Figure 4.8: Overlapping for homology groups 0 and 1 measured for the 30 windows of measurements taken from the time series describing a two-dimensional double-well potential whose potential barriers change as shown in Fig. 4.7.

However, there are other possible explanations to the quality of our results. One immediate observation is that the dynamics of our system are stochastic and therefore the state space vectors do not for trajectories the way they would for a deterministic dynamical system. In this case, the state space vectors will more likely create objects with a certain fractal structure.

In such case, considering other variables, different to r and c, could provide some improve-

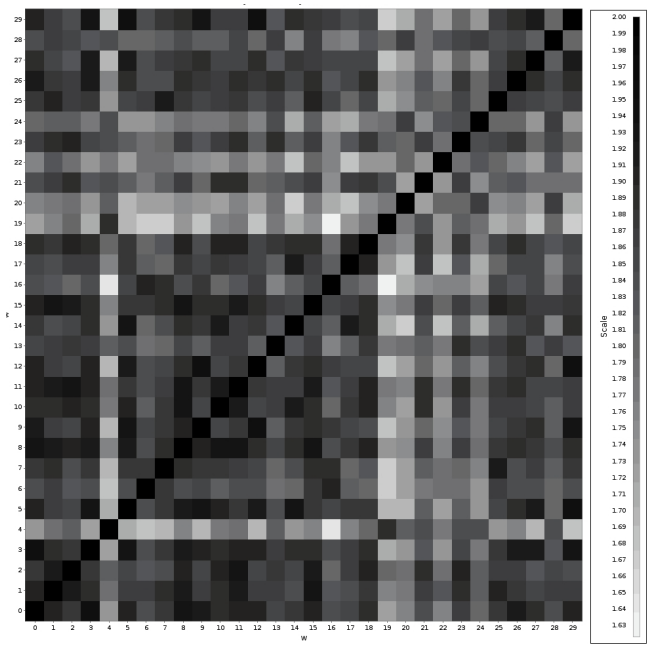


Figure 4.9: Total overlap between the 30 windows of measurements taken from the time series describing a two-dimensional double-well potential whose potential barriers change as shown in Fig. 4.7. Accordingly, we expect the total overlap to show a similar behavior.

ments in our results. For example, we could use the variables introduced by R. MacPherson and B. Schweinhart [80] to estimate fractal dimensions using persistence homology. The variables they introduce, called ratio and aspect, can be used to estimate the amount of empty space around an object. Therefore, these could potentially identify the increase or decrease in depth of a potential well. In Appendix B we provide a detailed explanation on the estimation of fractal dimensions using persistence homology.

4.7 Example 2: Logistic map with changing parameters

The time series we analyze in this section contains the dynamics of a logistic map where the parameter that determines whether the system shows non-chaotic or chaotic behavior varies in time. This illustrates the possibility of using the method described in Section 4.3 to identify transitions between dynamical regimes in a time series associated to a non-stationary deterministic system.

4.7.1 The system

A logistic map is a one-dimensional non-linear map given by Eq.4.1 where the value of parameter a determines whether the system shows non-chaotic or chaotic behavior.

$$x_{i-1} = ax_i(1 - x_i), \tag{4.1}$$

We generate a time series by merging 20 different logistic map time series, each computed with a different parameter a in Eq. 4.1. The set of parameters taken for these simulations is $A = \{3.2, 3.3, 3.4494, 3.4495, 3.45, 3.48, 3.51, 3.53, 3.5440, 3.5441, 3.55, 3.56995, 3.7, 3.75, 3.79, 3.82843, 3.8285, 3.83, 3.85, 3.9\}$.

In order to produce the time series associated to the initial parameter, $a = 3.2$, we take a random initial value between zero and one, x_0, and simulate 5000 data points. Then, the final value of this time series is used as initial value for the simulation of a time series with the second parameter $a = 3.3$. This procedure continues until we have simulated the time series for all the parameters in A. Fig. 4.10 shows the resulting time series.

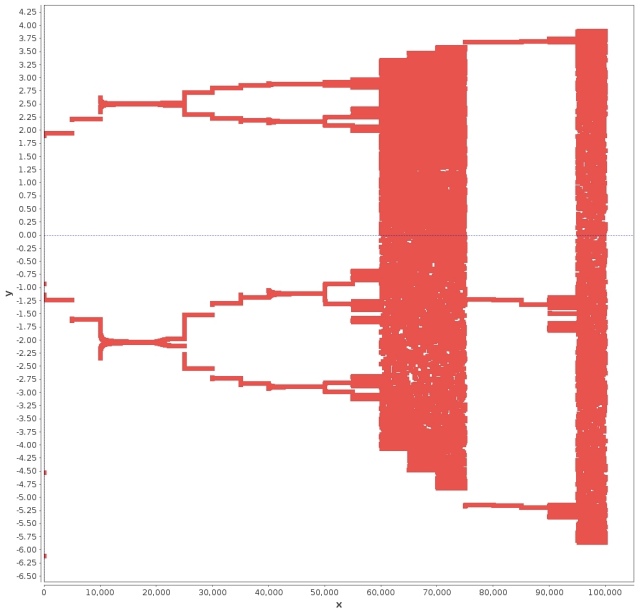


Figure 4.10: Time series describing the change in dynamics of a logistic map according to changes in parameter a (see Eq. 4.1). This time series with 100000 data points is obtained by gluing together 20 different time series generated by taking each of the values in set A (see text). Every time series associated to a parameter in A has 5000 data points.

As can be seen in the time series, for $a = 3.2$ and $a = 3.3$ the time series oscillates between two attractors.

The next 8 values, until $a = 3.56995$, are supposed to belong to a region of parameters that is supposed to produce time series oscillating between four attractors.

However, for $a = 3.4495$, a =3.45 and $a = 3.48$, we observe that two of these attractors are almost indistinguishable from two others, so we would expect the dynamics of this range in the time series to be more similar to the dynamics associated with the first two values of a.

For $a = 3.55$ we observe time series oscillating between eight attractors and for $a = 3.56996$ we observe a larger number of attractors.

For $a = 3.7$, $a = 3.75$ and $a = 3.79$, the system shows chaotic behavior. For $a = 3.82843$, $a = 3.8285$, $a = 3.83$ and $a = 3.85$, we again observe non-chaotic behavior. But chaotic behavior reappears for $a = 3.9$.

According to this, we expect three different degrees of overlapping. First, a group with large overlapping, containing the state space vectors windows 0 to 4, 5 to 9, and 16 to 18. A second group, containing windows 11, 12, and 16 to 19, should show a medium degree of overlapping with the windows of the first group. And finally, a third group containing windows 13 to 15 and 19, which should be reconstructed with the segments of time series showing chaotic behavior. This last group should show different measurements of overlapping with the first two groups and probably a larger overlapping with the windows in it.

4.7.2 Analysis results

To start the analysis of this system, we first need to reconstruct the state space from the time series. For this, we first divide the time series into 20 windows of measurements. Then we use the method described in Algorithm 5 on the first window of measurements.

The implementation of Algorithm 5 involves the estimation of mean (r, c)-persistence as in Algorithm 4.

For these calculations we take $S = 50$ random samples from the first window of measurements, with 700 data points each. Then we construct witness complexes using a *minmax* landmark selector with L landmark points, where L is set as in Eq. 4.4. And we estimate up to the $H = 3$ homology group.

This results in the estimation of embedding delays $\tau_0 = 5$ and $\tau_1 = 8$ for a non-uniform embedding. The state space is then reconstructed using these embedding parameters on the entire time series.

Once the state space has been reconstructed, we proceed to identify differences in the dynamics of the time series. For this, we divide the reconstructed state space vectors into 20 windows and measure the overlapping between windows.

The overlap results for homology groups 0 and 1 are shown in Fig. 4.11. The homology group H_2 was not populated but the plot of overlapping for H_1 is a nice surprise.

The plot of overlapping for H_1 plot is particularly interesting because it shows a clear distinction in behavior for the windows of state space vectors reconstructed with the segments of time series showing chaotic behavior (or the expected third group of overlapping behavior mentioned above). For these windows of measurements, the first homology group is populated and the total overlap between these windows is large.

The total overlap results are shown in Fig. 4.12. This plot does not show the existence of the three expected overlapping groups but of two, distinguishing between the windows of state space reconstructed with segments of time series showing chaotic behavior and those showing non-chaotic behavior.

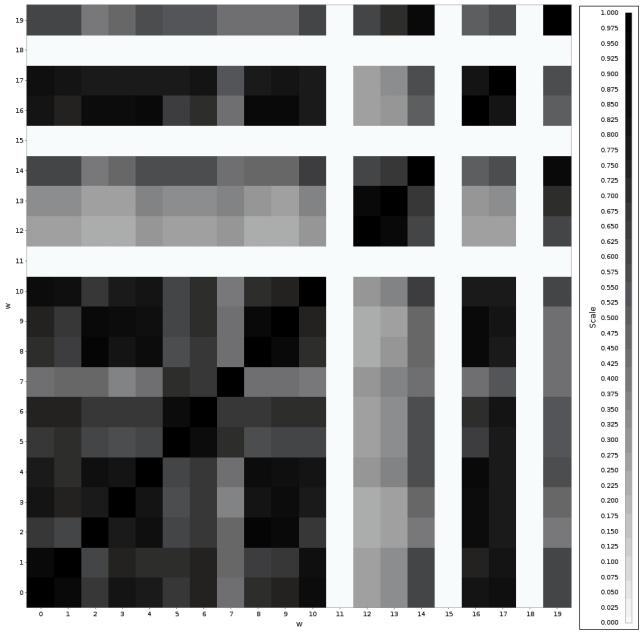
In general, our results suggest that the methods presented in Sections 4.3 and 4.5 allow the identification of transitions between dynamical regimes in deterministic systems.

Additionally, these results give us information about the reasons why we obtained results with low quality in the example analyzed in Section 4.6.

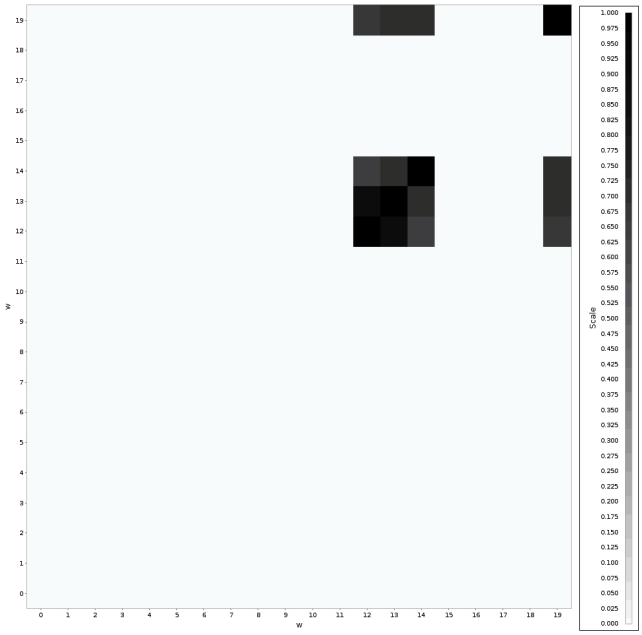
The main differences between that example and the one analyzed in this section lie in the dynamical properties of the time series. In the first example there is no variation in the number of attractors, but in the shape of those. Additionally, the equations describing the dynamics of the system analyzed in the first example are stochastic, whereas the dynamics of the current example are deterministic.

In terms of implementation, the only difference between these two examples was an increase in the number of random samples used to estimate the persistence of every window of state space vectors. However, this increase in number of samples is not drastic. We are still in the same order of magnitude.

Thus, we suggest that the quality in results provided by our method of analysis does not reside in the implementation but in the dynamical properties of the system analyzed. Our



(a) Overlapping for homology group 0.



(b) Overlapping for homology group 1.

Figure 4.11: Overlapping for homology groups 0 and 1 measured for the 20 windows into which the reconstructed state space vectors from the time series describing a logistic map with changing dynamics are divided. The state space vectors are reconstructed from the time series using embedding delays $\tau_0 = 5$ and $\tau_1 = 8$ in a non-uniform embedding.

method might be adequate for the identification of differences in dynamical regimes but not when the transitions imply small deformations of the attractors.

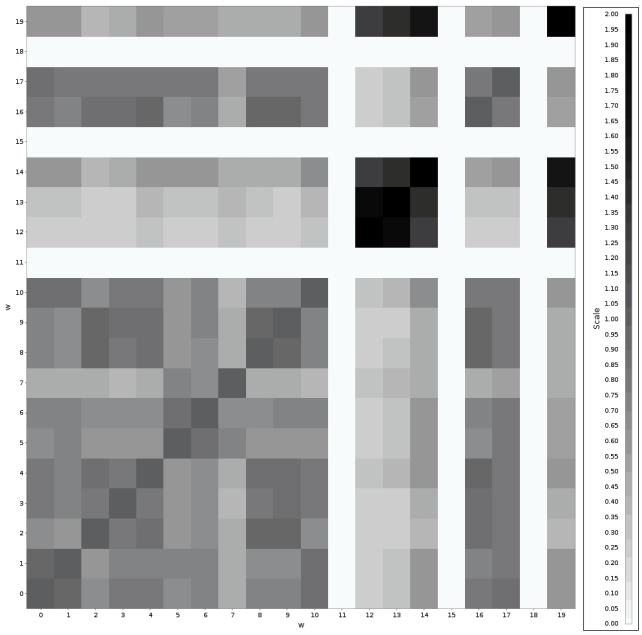


Figure 4.12: Total overlap between the 20 windows into which the reconstructed state space vectors from the time series describing a logistic map with changing dynamics are divided. The state space vectors are reconstructed from the time series using embedding delays $\tau_0 = 5$ and $\tau_1 = 8$ in a non-uniform embedding.

4.8 Final remarks

In this Chapter, we introduced our new method for the identification of transitions between dynamical regimes in real-world time series via persistent homology. We analyzed two examples with it.

In the first example, consisting in a time series describing a double-well potential with varying depth in its potential wells, we divide the reconstructed state space vectors into a given number of windows and measure the total overlap (Eq. 4.3) between the mean h-persistence of different windows (Eq. 4.1).

Even when our results indicate changes in the configuration of the state space between, our results did not show a full match between the changes in total overlap and the changes in potential barriers.

In the second example, the time series corresponds to the merge of thirty logistic map time series produced with different parameters.

The results of this analysis suggests that our new method for the identification of transitions in dynamical regime is adequate for analyzing systems with this type of transitions.

In terms of implementation, the only difference between the analysis of these two examples was an increase in the number of random samples used to estimate the mean h-persistence of different windows, but the number of samples did not increase dramatically.

Thus, we suggest that our method might be better suited for the identification of transitions between dynamical regimes in systems where the transitions imply large deformations of the attractors.

In the future, in order to better understand the type of transitions between dynamical regimes that we can identify with our method, we intend to study the differences in results when

increasing the number of samples in the estimation of the mean h-persistence significantly

5. Conclusions

In this thesis we introduce two new methodologies for the identification of different dynamical features in a system from the analysis of a real-world time series. We focus in the dynamical features corresponding to the different dynamical (metastable) states in a system and the transitions between dynamical regimes in a system.

For the identification of dynamical features in a system from a time series, we first need to reconstruct the state space, which is the space containing all the states of a system. The state space, together with an evolution operator, entirely describe a dynamical system. Analyzing the geometry of the state space, one is able to identify the different attractors of a system.

Having an adequate reconstruction of the state space is fundamental for obtaining robust results from the analysis of dynamical features. Many criteria to identify an adequate state space reconstruction rely on the comparison of the reconstructed space with the true state space. However, when analyzing real data, the true state space of a system is almost never known. Few criteria have been developed for the case where the true state space is unknown. Among these we find the criteria of Uzal et al. [119], where an adequate state space is identified in terms of the probability of predicting future states.

In this thesis, we consider delay-coordinate maps for the reconstruction of the state space. Delay-coordinate maps are commonly used to reconstruct the state space for deterministic dynamical systems, but its use has also been extended to some types of stochastic systems. In Section 2.1.1 we specifically talk about delay-coordinate maps. These maps depend on two embedding parameters: the time delay and the embedding dimension. Thus, our general criteria for the identification of an adequate state space reconstruction consists on measuring the gain or loss of geometrical information when varying the embedding parameters for a delay-coordinate map. Chapter 2 is entirely dedicated to the general criteria considered for obtaining an adequate state space reconstruction.

Once the state space of a system has been adequately reconstructed, the identification of dynamical features can be performed in many ways. We decided to approach this problem from a geometrical and topological perspective.

Initially, we took a recurrence analysis approach; explained in Section 3.1. A recurrence is the return of a state space trajectory, given sufficiently long time, to a state arbitrarily close to a former state. The regions in state space to which a system recurs the most can be associated to the existence of attractors, which indicate different dynamical states in the system.

Additionally, we introduce the concept of metastability to emphasize the possibility of a system having different and well distinguished time scales in its dynamics. This concept, to the best of our knowledge, had not been previously introduced to the analysis of recurrences,

despite being a common characteristic of physical phenomena. A discussion on metastability can be found in Section 3.2.3.

This way, we introduced a new method for the identification of different metastable states in a real-world time series, described in Section 3. It consists on the construction of a recurrence network (see Section 3.1.2) using the reconstructed state space vectors of a system and a parameter of proximity. From the analysis of this network with a fuzzy module finding algorithm specifically developed for the identification of metastability, we identify metastable modules which correspond to metastable states in the system.

We also introduce a new method for the reconstruction of the state space from a time series, based on the analysis of filtrations of recurrence networks constructed from the state space vectors of different state space reconstructions via delay-coordinate maps with different embedding parameters. This method is described in Section 3.4, and the specific criteria necessary for an adequate state space reconstruction within this approach are described in Section 3.2.

We illustrate the performance of our new method with two different complex time series. The first time series describes the motion of a particle in a heat bath with temperature $T = 100K$, under the gradient of a double-well potential and a random force. The second time series describes the changing molecular configurations of a molecule of trialanine at low temperature, $T = 300K$.

The results of these analysis suggest that our method is suitable for the identification of metastable states in real-world time series data. And that this is robust (see Section 3.6.4) to the addition of noise with an amplitude of up to 6% the amplitude of the original time series, and to the removal of up to 7% of the data measurements.

However, this method has some restrictions, imposed to by the use of recurrence networks and a module finding algorithm. This way, it is only suitable for the analysis of time series coming from dynamical systems with clearly distinguishable attractors, having one clearly dominant time scale in the dynamics and being entirely described by low-order topological information.

In order to overcome these drawbacks and analyze complex time series whose reconstructed state spaces have high-order topologies, we took a different approach for the identification of dynamical features in a system: persistent homology. This is an algebraic topological approach which consists on the analysis of high-order topological features that persist in a set of coverings with different fixed radius (proximity parameter) ϵ, this means, in a filtration depending on ϵ. Persistent homology is introduced in Section 4.1.

Our method is described in Section 4.3. In this, from the state space adequately reconstructed from a time series, we select a given number of windows of state space vectors of equal length. The time lapse of a window should be enough to capture a dynamical regime of the system. We then compute the topological summary for each window, called the mean h-persistence in the (r, c)-space (see Eq. 4.1) and compare these summaries between all windows using the so called total overlap (see Eqs. 4.2 and. 4.3). The differences in overlap between windows of measurements indicate dynamical transitions in the time series.

As done with the previous approach, we also introduce a method for the state space reconstruction from a time series, now based in the persistent homology approach. This method is described in Section 4.5, and the specific criteria necessary for an adequate state space reconstruction within this approach are described in Section 4.4.

We illustrate the performance of our new method with two different complex time series. The first time series describes a process in which the x-component of the barriers between the potential wells of a two-dimensional double-well potential increase and then decrease. The second time series describes the dynamics of a logistic map where the parameter that determines whether the system shows non-chaotic or chaotic behavior varies in time.

From the results of these analysis, we suggest that our method is able to identify dynamical transitions in deterministic systems where there is a change in the number of attractors, or where the dynamical transitions do not correspond to small deformations in the attractors.

However, we have some suggestions for future work to overcome this problems. First, for the identification of changes in the shape of attractors of deterministic systems, it may be possible to get good results by using other variables, like the ones mentioned in Appendix B. For the ever more challenging analysis of stochastic dynamical systems, it might be necessary to try a completely new approach and introduce state space reconstructions that implement the embedding theorems presented in Section 2.2. This way, the possibilities to expand this research topic are vast and exciting.

Other topics for future research include improvements in the state space reconstruction using the suggestions mentioned in Section 4.4. Or the improvement in the computation of persistent homology for large data sets and higher-order topological features. The application of our method based in persistent homology to real time series is also an exciting project for the future.

A. The adjusted rand index (ARI)

There are various measures to quantify the distance between either crisp or fuzzy partitions. A crisp partition, or hard partition, is one in which every object is assigned to only one cluster in a binary way. A fuzzy partition, or soft partition, is one in which every object is assigned to various clusters with different weight values.

In general, the distance between partitions can be measured in three ways: based on counting pairs of elements, on summations of overlaps, or on differences in mutual information or entropy−like the normalized mutual information (NMI) or the Jaccard index−. Depending on the way the partitions we analyze are created, one may use different types of measures.

The adjusted rand index (ARI), as defined by Hubert and Arabie in 1985 [61], measures the agreement between two partitions. When the partitions are not similar at all it is equal to zero, and when the partitions are equivalent it is equal to one. This measure is based on counting pairs of objects classified simultaneously in the same cluster for two clustering partitions. It has a generalized hypergeometric distribution as null hypothesis [66], where the partitions are drawn randomly considering a fixed number of clusters and elements in each cluster, even when the number of clusters in both partitions might differ.

The use of the ARI for evaluating classification can be read in the work of Santos and Embrechts [103]. Their results show that this measure is meaningful even in the cases where the labeling of the partitions is switched. According to these authors, the ARI is defined as follows.

Let us imagine a set of objects $S = \{O_1, ..., O_N\}$. There are $\binom{N}{2}$ combinations of pairs of elements of set S. Set two partitions of S, $P = \{p_1, p_2, ..., p_A\}$ and $Q = \{q_1, q_2, ..., q_B\}$, such that $\cup_{a=1}^{A} p_a = \cup_{b=1}^{B} q_b = S$, $p_a \cap p_{a'} = \emptyset$ for any $a \neq a'$, and $q_b \cap q_{b'} = \emptyset$ for any $b \neq b'$. Let t_{ab} be the number of objects in S that were classified in the a-th subset of P and in the b-th subset of Q simultaneously. Then, $ARI = F_1/F_2$, where:

$$F_1 = \binom{n}{2} \sum_{a=1}^{A} \sum_{b=1}^{B} \binom{t_{ab}}{2} - \sum_{a=1}^{A} \binom{t_{a\cdot}}{2} \sum_{b=1}^{B} \binom{t_{\cdot b}}{2}$$

$$F_2 = \frac{1}{2}\binom{n}{2} \left[\sum_{a=1}^{A} \binom{t_{a\cdot}}{2} + \sum_{b=1}^{B} \binom{t_{\cdot b}}{2} \right] - \sum_{a=1}^{A} \binom{t_{a\cdot}}{2} \sum_{b=1}^{B} \binom{t_{\cdot b}}{2}$$

In 2013, Hueffner et al. [62] introduced a modification to the ARI of Arabie and Hubert in order to compare fuzzy partitions as obtained by the MSM clustering algorithm [104] (see Section 3.6.4).

B. Fractal dimension estimation with persistence homology

The importance of estimating the fractal dimension of an attractor was commented in Section 2.1.2. In short, this can be used, considering Whitney's theorems of Section 2.1, for estimating an embedding dimension which is adequate to reconstruct the state space from a time series.

Some recent investigations have shown the use of persistence homology to estimate the fractal dimension. For example, R. MacPherson and B. Schweinhart introduced in 2012 [80] two variables called the ratio, r, and the aspect, z, and used them to define a persistent homology dimension, d_{PH}.

$$r = (b + d)/2 \tag{B.1a}$$

$$z = \operatorname{arcsec}(\frac{d}{b}) \tag{B.1b}$$

Given a barcode where every interval has a birth point b and death point d, the ratio and aspect are given respectively by Eqs. B.1b.

$$d_{PH} = \frac{\ln w}{\ln \rho} \tag{B.2}$$

Considering two parameters: $w > 0$ and $\rho > 1$, such that for $n(r, z)$ denoting the number of pairs (r, z) in a barcode, w and ρ satisfy that $n(r, z) = w\, n(\rho r, z)$. This way, d_{PH} is given by Eq. B.2.

In 2014, Máté and Heermann [88] found that, for fractals which are not exactly self similar, such persistence homology dimension estimates how the empty space around an object defined by the data sample scales. This way, d_{PH} can be understood as the fractal dimension of such empty space.

The radius is easily understandable as the *life time* of a topological feature, but for understanding what the aspect indicates let us imagine that, while creating ε-balls around every state space vector, some n-dimensional holes will be created. Then, the aspect can be interpreted as "the angular opening at the edge of a gulf in the structure" [80]. Additionally, according to R. MacPherson and B. Schweinhart, the dependency of the distribution of persistence homology points on the aspect may show differences in the statistical shape of the state space vectors which are not related to the fractal dimension.

For these clear relations to the fractal dimension, is that we use the radius and the aspect variables in Chapter 4 to define the persistence homology.

Nevertheless, we consider important to mention a different approach to measure the fractal dimension of an object using persistence homology. This approach is the one introduced by Máté and Heermann in [88]. For exactly self-similar fractal objects, these authors extract the fractal dimension directly from the barcode. They base this on the fact that the missing parts or holes of a fractal, scale in the same way as its volume.

$$d_F = \frac{\ln[m]}{\ln[s]} \tag{B.3}$$

This way, considering a scale constriction factor s and the number of self-similar copies on the smaller scale (or multiplicity) m, a generic fractal dimension, d_F, is given as by Eq. B.3.

Because of self similarity, the death point of an interval and the number of intervals ending at that point, $n(d)$, change according to a power law when going from one scale to the other. Then, there is a constant c such that the following equation holds:

$$\ln[n(d)] = \alpha \ln[d] + c \tag{B.4}$$

Considering two different scales d_i and d_j, such that $d_i < d_j$ and $n(d_i) > n(d_j)$, then

$$\alpha = \frac{\ln[n(d_i)] - \ln[n(d_j)]}{\ln[d_i] - ln[d_j]} = -\frac{\ln[n(d_j)/n(d_i)]}{\ln[d_i/d_j]} \tag{B.5}$$

Now, in the case when the underlying dynamical process is random, it is likely that no two intervals in the barcode will have the same length. Then, if the death points concentrate around a set of values, it is possible to compute averages on every scale d_i. This way, an assignment function $a_i(\epsilon)$ can be created, such that $a_i(\epsilon) = 1$ only if an interval of length ϵ

$$\langle d \rangle_i = \frac{1}{n[\langle d \rangle_i]} \sum_{\beta \in B} d^{(\beta)} a_i(d^{(\beta)}) \tag{B.6}$$

$$n[\langle d \rangle_i] = \sum_{\beta \in B} a_i(d^{(\beta)}), \tag{B.7}$$

appears in scale d_i. With this function, one can define a characteristic scale $\langle d \rangle_i$ for a barcode B. This characteristic scale is given as in Eq. B.6, where $d^{(\beta)}$ denotes the endpoint of an interval β in B and $n[\langle d \rangle_i]$ denotes the number of intervals belonging to scale d_i and is given by Eq. B.7.

This way, expression B.4 can be rewritten as $\ln[n(\langle d \rangle)] = \alpha \ln[\langle d \rangle] + c$.

This estimation of the fractal dimension could be used to estimate an adequate embedding dimension for the reconstruction of the state space from a time series. However, we have found (although not reported here) that the α estimates are very sensitive to noise. Therefore, we have not followed this approach for the analysis of real-world time series.

Bibliography

[1] Abarbanel, H. D. I. *Analysis of Observed Chaotic Data*. Springer Verlag, Berlin, 1996.

[2] Baptista, M., et al. Kolmogorov-Sinai entropy from recurrence times. *Physics Letters A 374* (2010), 1135–1140. doi:10.1016/j.physleta.2009.12.057.

[3] Barreira, L. Poincaré recurrence: Old and new. *World Scientific* (2005), 415–422.

[4] Bendich, P., Galkovskyi, T., and Harer, J. Improving homology estimates with random walks. *Inverse Problems 124002* (2011), MR2854318.

[5] Boccaletti, S., Valladares, D., Pecora, L. M., Geffert, P., and Carroll, T. Reconstructing embedding spaces of coupled dynamical systems from multivariate data. *Phys. Rev. E 65* (2002), 035204. doi:10.1103/PhysRevE.65.035204.

[6] Bovier, A. Metastability: A potential theoretic approach. *International Congress of Mathematicians 3* (2006), 499–518. doi:10.1007/978-3-642-23811-6_18.

[7] Bovier, A. Metastability: From mean field models to SPDEs. *Probability in Complex Physical Systems, Springer Proceedings in Mathematics 11* (2012), 443–462. doi:10.1007/978-3-642-23811-6_18.

[8] Bubenik, P. Statistical topological data analysis using persistence landscapes. *arXiv:1207.6437* (June 2012).

[9] Carlsson, G. Topology and data. *Bulletin (New Series) of the American Mathematical Society 46*, 2 (2009), 255–308.

[10] Carlsson, G., and de Silva, V. Topological estimation using witness complexes. In *Symposium on Point-Based Graphics, ETH, Zurich, Switzerland* (2004).

[11] Carlsson, G., and Silva, V. D. Topological approximation by small simplicial complexes. Tech. rep., Mischaikow, and T. Wanner, 2003.

[12] Carlsson, G., and Zomorodian, A. The theory of multidimensional persistence. *Discrete and Computational Geometry 42*, 1 (2009), 71–93. doi:10.1007/s00454-009-9176-0.

[13] Casdagli, M., Eubank, S., Farmer, J. D., and Gibsoni, J. State space reconstruction in the presence of noise. *Physica D 51* (1991), 52–98.

[14] Cerri, A., di Fabio, B., Ferri, M., Frosini, P., and Landi, C. Betti numbers in multidimensional persistent homology are stable functions. *Mathematical Methods in the Applied Sciences 36*, 12 (2013), 1543–1557. doi:10.1002/mma.2704.

[15] Cerri, A., Fabio, B. D., Jabłoński, G., and Medri, F. Comparing shapes through multi-scale approximations of the matching distance. *Computer Vision and Image Understanding 121*, 0 (2014), 43–56. doi:10.1016/j.cviu.2013.11.004.

[16] Cerri, A., and Landi, C. The persistence space in multidimensional persistent homology. In *Discrete Geometry for Computer Imagery*, R. Gonzalez-Diaz, M.-J. Jimenez, and B. Medrano, Eds., vol. 7749 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2013, pp. 180–191. doi:10.1007/978-3-642-37067-0_16.

[17] Chazal, F., Clisse, M., Labruère, C., and Michel, B. Convergence rates for persistence diagram estimation in topological data analysiss. In *Proceedings of the 31st International Conference on Machine Learning, Beijing, China* (2014), vol. 32.

[18] Chazal, F., et al. Subsampling methods for persistent homology. *arXiv:1406.1901v1 [math.AT]* (June 2014), 16.

[19] Chen, Y., and Yang, H. Multiscale recurrence analysis of long-term nonlinear and nonstationary time series. *Chaos, Solitons and Fractals 45* (2012), 978–987. doi:10.1016/j.chaos.2012.03.0135.

[20] Cohen-Steiner, D., Edelsbrunner, H., and Harer, J. Stability of persistence diagrams. *Discrete and Computational Geometry 37*, 1 (2007), 103–120. doi:10.1007/s00454-006-1276-5.

[21] Cohen-Steiner, D., Edelsbrunner, H., Harer, J., and Mileyko, Y. Lipschitz functions have L_p-stable persistence. *Foundations of Computational Mathematics 10*, 2 (2010), 127–139.

[22] Collins, A., Zomorodian, A., Carlsson, G., and Guibas, L. J. A barcode descriptor for curve point cloud data. *Computers and Graphics 28* (2004), 881–894. doi:10.1016/j.cag.2004.08.015.

[23] Cooper, C., and Frieze, A. Component structure of the vacant set induced by a random walk on a random graph. *Random Structures and Algorithms 42*, 2 (2013), 135–158. doi:10.1002/rsa.20402.

[24] Cross, D. J., and Gilmore, R. Representation theory for strange attractors. *Phys. Rev. E 80* (Nov 2009), 056207. doi:10.1103/PhysRevE.80.056207.

[25] Dall, J., and Christensen, M. Random geometric graphs. *Phys. Rev. E 66* (Jul 2002), 016121. doi:10.1103/PhysRevE.66.016121.

[26] d'Amico, M., Frosini, P., and Landi, C. Using matching distance in size theory: A survey. *International Journal of Imaging Systems and Technology 16*, 5 (2006), 154–161. doi:10.1002/ima.20076.

[27] Dechert, W. D., and Gençay, R. The topological invariance of Lyapunov exponents in embedded dynamics. *Physica D 90* (1996), 40–55.

[28] Deuflhard, P., and Weber, M. Robust Perron cluster analysis in conformation dynamics. *Linear Algebra and its Applications 398* (2005), 161–184. doi:10.1016/j.laa.2004.10.026.

[29] Dey, T. K., Fan, F., and Wang, Y. Dimension detection with local homology. *arXiv:1405.3534 [cs.CG]* (2014), 18.

[30] Djurdjevac, N., Bruckner, S., Conrad, T. O. F., and Schuette, C. Random walks on complex modular networks. *Journal of Numerical Analysis Industrial and Applied Mathematics 6*, 1-2 (2012), 29–50.

[31] Donges, J. F. *Functional network macroscopes for probing past and present Earth system analysis: Complex hierarchical interactions, tipping points, and beyond.* PhD thesis, Humboldt University, 2012.

[32] Donges, J. F., et al. Identification of dynamical transitions in marine palaeoclimate records by recurrence network analysis. *Nonlinear Processes in Geophysics 18*, 5 (2011), 545–562. doi:10.5194/npg-18-545-2011.

[33] Donges, J. F., et al. Nonlinear detection of paleoclimate-variability transitions possibly related to human evolution. In *Proceedings of the National Academy of Sciences of the United States of America* (2011), vol. 108, pp. 20422–20427. doi:10.1073/pnas.1117052108.

[34] Donges, J. F., Heitzig, J., Donner, R. V., and Kurths, J. Analytical framework for recurrence network analysis of time series. *Phys. Rev. E 85* (Apr 2012), 046105. doi:10.1103/PhysRevE.85.046105.

[35] Donner, R. V., et al. Recurrence networks a novel paradigm for nonlinear time-series analysis. *New Journal of Physics 12* (2010), 033025. doi:10.1088/1367-2630/12/3/033025.

[36] Donner, R. V., et al. The geometry of chaotic dynamics - a complex network perspective. *Eur. Phys. J. B 84* (2011). doi:10.1140/epjb/e2011-10899-1.

[37] Donner, R. V., et al. Recurrence-based time series analysis by means of complex network methods. *International Journal of Bifurcation and Chaos 21*, 4 (2011), 1019–1046. doi:10.1142/S0218127411029021.

[38] Donner, R. V., Zou, Y., Donges, J. F., Marwan, N., and Kurths, J. Ambiguities in recurrence-based complex network representations of time series. *Phys. Rev. E 81* (Jan 2010), 015101. doi:10.1103/PhysRevE.81.015101.

[39] Drummond, A. J., Suchard, M. A., Xie, D., and Rambaut, A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution 29*, 8 (2012), 1969–1973. doi:10.1093/molbev/mss075.

[40] Dumas, J. G., Heckenbach, F., Saunders, B. D., and Welker, V. Computing simplicial homology based on efficient smith normal form algorithms. *Algebra, Geometry, and Software Systems MR2011758 (2004i:55009)* (2003), 177–207.

[41] Eckmann, J. P., Kamphorst, S. O., and Ruelle, D. Recurrence plots of dynamic-systems. *Europhysics Letters 4*, 9 (1987), 973–977.

[42] Eckmann, J. P., Kamphorst, S. O., Ruelle, D., and Ciliberto, S. Liapunov exponents from time series. *Phys. Rev. A 34* (1986), 4971.

[43] Eckmann, J. P., and Ruelle, D. Ergodic theory of chaos and strange attractors. *Reviews of Modern Physics 57*, 3 (1985), 617–656.

[44] Edelsbrunner, H., and Harer, J. L. *Computational Topology, An Introduction.* American Mathematical Society, National Science Foundation US, 2010.

[45] Edelsbrunner, H., Letscher, D., and Zomorodian, A. Topological persistence and simplification. *Discrete and Computational Geometry 28* (2002), 511–533.

[46] Emrani, S., Gentimis, T., and Krim, H. Persistent homology of delay embeddings and its application to wheeze detection. *Signal Processing Letters, IEEE 21*, 4 (2014), 459–463.

[47] Eroglu, D., Marwan, N., Prasad, S., and Kurths, J. Finding recurrence networks' threshold adaptively for a specific time series. *Nonlinear Processes in Geophysics 21*, 6 (2014), 1085–1092. doi:10.5194/npg-21-1085-2014.

[48] Fasy, B. T., et al. Confidence sets for persistence diagrams. *Tha Annals of Statistics 42* (2014), 2301–2339.

[49] Faure, P., and Korn, H. A new method to estimate the Kolmogorov entropy from recurrence plots: its application to neuronal signals. *Physica D 122* (1998), 265–279.

[50] Feldhoff, J. H., et al. Geometric detection of coupling directions by means of inter-system recurrence networks. *Physics Letters A 376* (2012), 3504–3513. doi:10.1016/j.physleta.2012.10.008.

[51] Fischer, A., et al. Identification of biomolecular conformations from incomplete torsion angle observations by hidden Markov models. *J. Comput. Chem. 28* (2006), 2453–2464. doi:10.1002/jcc.20692.

[52] Fraser, A. M., and Swinney, H. L. Independent coordinates for strange attractors from mutual information. *Phys. Rev. A 33* (Feb 1986), 1134–1140. doi:10.1103/PhysRevA.33.1134.

[53] Galka, A. *Topics in Nonlinear Time Series Analysis: With Implications for EEG Analysis*. Advanced series in nonlinear dynamics. World Scientific, 2000.

[54] Gao, Z., and Jin, N. Flow-pattern identification and nonlinear dynamics of gas-liquid two-phase flow in complex networks. *Phys. Rev. E 79* (2009), 066303. doi:10.1103/PhysRevE.79.066303.

[55] Ghrist, R. Barcodes: The persistent topology of data. *Bull. Amer. Math. Soc. 45*, 1 (2008), 61–75.

[56] Giesen, J., Miklos, B., Pauly, M., and Wormser, C. The scale axis transform. In *Proceedings of the Twenty-fifth Annual Symposium on Computational Geometry* (New York, NY, USA, 2009), SCG '09, ACM, pp. 106–115. doi:10.1145/1542362.1542388.

[57] Grassberger, P., and Procaccia, I. Measuring the strangeness of strange attractors. *Physica D 9*, 1-2 (1983), 189–208.

[58] Hasson, C. J. Influence of embedding parameters and noise in center of pressure recurrence quantification analysis. *Gait and Posture 27*, 3 (2008), 416–422.

[59] Hinton, G. E., and Roweis, S. T. Stochastic neighbor embedding. In *Advances in neural information processing systems* (2002), pp. 833–840.

[60] Hsing, T., and Rootzén, H. Extremes on trees. *Ann. Probab. 33*, 1 (2005), 413–444.

[61] Hubert, L., and Arabie, P. Comparing partitions. *Journal of Classification 2* (1985), 193–218.

[62] Hueffner, S., Kayser, B., and Conrad, T. O. F. Finding modules in networks with non-modular regions. In *Lecture Notes in Computer Science* (2013), vol. 7933, pp. 188–199. Proceedings of the 12th international symposium, SEA 2013.

[63] Hughes, N. P., and Tarassenko, L. Novel signal shape descriptors through wavelet transforms and dimensionality reduction. In *Proc. SPIE* (2003), vol. 5207, pp. 763–773. doi:10.1117/12.506045.

[64] Huke, J., and Broomhead, D. Embedding theorems for non-uniformly samples dynamical systems. *Nonlinearity 20*, 9 (2007), 2205. doi:10.1088/0951-7715/20/9/0111.

[65] Huke, J. P. Embedding nonlinear dynamical systms: A guide to takens' theorem. Tech. rep., Defence Research Agency, Malvern, UK, 1993.

[66] Hullermeier, E., Rifqi, M., Henzgen, S., and Senge, R. Comparing fuzzy partitions: A generalization of the rand index and related measures. *IEEE Trans. Fuzzy Syst. 20*, 3 (2012), 546–556. doi:10.1109/TFUZZ.2011.217930.

[67] Judd, K., and Mees, A. Embedding as a modeling problem. *Physica D 120* (1998), 273–286. doi:10.1016/S0167-2789(98)00089-X.

[68] Kantz, H., and Olbrich, E. Scalar observations from a class of high-dimensional chaotic systems: Limitations of the time delay embedding. *Chaos 7*, 3 (1997), 423–429. doi:10.1063/1.166215.

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

[70] Kong, Z., and Yeh, E. M. On the critical density for percolation in random geometric graphs. In *Proceedings of IEEE International Symposium on Information Theory, ISIT 2007* (2007), vol. 1-7, pp. 151–155. doi:10.1109/ISIT.2007.4557082.

[71] Kramers, H. A. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica 7* (1940), 284–304.

[72] Kraskov, A., Stögbauer, H., and Grassberger, P. Estimating mutual information. *Phys. Rev. E 69* (2004), 066138. doi:10.1103/PhysRevE.69.066138.

[73] Krishnan, A., Giuliani, A., Zbilut, J., and Tomitah, M. Implications from a network-based topological analysis of ubiquitin unfolding simulations. *PLoS ONE 3* (2008).

[74] Krishnan, A., Zbilut, J., Tomita, M., and Giuliani, A. Proteins as networks: usefulness of graph theory in protein science. *Curr. Prot. Peptide Sci. 9* (2008), 28–38.

[75] Lafon, S., and Lee, A. Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE Transactions on Pattern Analysis and Machine Intelligence 28*, 9 (2006), 1393–1403. doi:10.1109/TPAMI.2006.184.

[76] Lee, H., Kim, J., Waliser, D., Loikith, P., Mattmann, C., and McGinnis, S. Using joint probability distribution functions to evaluate simulations of precipitation, cloud fraction and insolation in the north america regional climate change assessment program (narccap). *Climate Dynamics 45*, 1-2 (2015), 309–323. doi:10.1007/s00382-014-2253-y.

[77] Lee, J., McGough, L., and Safdi, B. R. Rènyi entropy and geometry. *Physical Review D 89* (2014), 125016. doi:10.1103/PhysRevD.89.125016.

[78] Letellier, C. Estimating the Shannon entropy: Recurrence plots versus symbolic dynamics. *Physical Review Letters 96* (2006), 254102. doi:10.1103/PhysRevLett.96.254102.

[79] Letellier, C., Moroz, I. M., and Gilmore, R. Comparison of tests for embeddings. *Phys. Rev. E 78* (Aug 2008), 026203. doi:10.1103/PhysRevE.78.026203.

[80] MacPherson, R., and Schweinhart, B. Measuring shape with topology. *Journal of Mathematical Physics 53* (2012), 073516. doi:10.1063/1.4737391.

[81] Maila, M., and Shi, J. A random walks view of spectral segmentation. In *AI and STATISTICS (AISTATS) 2001* (2001).

[82] Marwan, N. *Encounters With Neighbours - Current Developments Of Concepts Based On Recurrence Plots And Their Applications.* PhD thesis, University of Potsdam, 2003.

[83] Marwan, N. A historical review of recurrence plots. *Eur. Phys. J. Special Topics 164* (2008). doi:10.1140/epjst/e2008-00829-1.

[84] Marwan, N., Romano, M. C., Thiel, M., and Kurths, J. Crossed recurrence plot based synchronization of time series. *Nonlinear Processes in Geophysics 9* (2002), 325–331.

[85] Marwan, N., Romano, M. C., Thiel, M., and Kurths, J. Recurrence plots for the analysis of complex systems. *Physics Reports 438* (2007), 237–329.

[86] Marwan, N., Schinkel, S., and Kurths, J. Recurrence plots 25 years later — gaining confidence in dynamical transitions. *EPL 101* (2013). doi:10.1209/0295-5075/101/20007.

[87] Marwan, N., Wessel, N., Meyerfeldt, U., Schirdewan, A., and Kurths, J. Recurrence-plot-based measures of complexity and their application to heart-rate-variability data. *Phys. Rev. E 66* (Aug 2002), 026702. doi:10.1103/PhysRevE.66.026702.

[88] Máté, G., and Heermann, D. W. Persistence intervals of fractals. *Physica A 405* (2014), 252–259. doi:10.1016/j.physa.2014.03.037.

[89] Metzner, P., Putzig, L., and Horenko, I. Analysis of persistent non-stationary time series and applications. *CAMCoS 7*, 2 (2012). doi:10.2140/camcos.2012.7.1753.

[90] Mindlin, G. B., and Gilmore, R. Topological analysis and synthesis of chaotic time series. *Physica D 58* (1992), 229–242.

[91] Mukherjee, S., and Steenbergen, J. Random walks on simplicial complexes and harmonics. *arXiv:1310.5099 [math.CO]* (2013).

[92] Munch, E. *Applications of Persistent Homology to Time Varying Systems.* PhD thesis, Duke University, 2013.

[93] Munz, M., and Biggin, P. C. Jgromacs: a Java package for analyzing protein simulations. *J. Chem. Inf. Model* (2012), 255–259.

[94] Nichkawde, C. Optimal state-space reconstruction using derivatives on projected manifold. *Phys. Rev. E 87* (Feb 2013), 022905. doi:10.1103/PhysRevE.87.022905.

[95] Niyogi, P., Smale, S., and Weinberger, S. Finding the homology of submanifolds with high confidence from random samples. *Discrete and Computational Geometry 39* (2008), 419–441. doi:10.1007/s00454-008-9053-2.

[96] Packard, N. H., Crutchfield, J. P., Farmer, J. D., and Shaw, R. S. Geometry from a time series. *Phys. Rev. Lett. 45* (Sep 1980), 712–716. doi:10.1103/PhysRevLett.45.712.

[97] Palit, S. K., Mukherjee, S., and Bhattacharya, D. A high dimensional delay selection for the reconstruction of proper phase space with cross auto-correlation. *Neurocomputing 113*, 0 (2013), 49–57. doi:10.1016/j.neucom.2013.01.034.

[98] Penrose, M. *Random Geometric Graphs*. Oxford University Press, Oxford, 2003.

[99] Poincaré, H. Sur le probléme des trois corps et les équations de la dynamique. *Acta Mathematica 13*, 1–270 (1890).

[100] Preis, R., et al. Dominant paths between almost invariant sets of dynamical systems. http://publications.mi.fu-berlin.de/64/.

[101] Richter, M., and Schreiber, T. Phase space embedding of electrocardiograms. *Phys. Rev. E 58* (Nov 1998), 6392–6398. doi:10.1103/PhysRevE.58.6392.

[102] Rulkov, N. F., Sushchik, M. M., Tsimring, L. S., and Abarbanel, H. D. I. Generalized synchronization of chaos in directionally coupled chaotic systems. *Phys. Rev. E 51* (Feb 1995), 980–994. doi:10.1103/PhysRevE.51.980.

[103] Santos, J. M., and Embrechts, M. On the use of the adjusted rand index as a metric for evaluating supervised classification. In *Artificial Neural Networks — ICANN 2009*, C. Alippi, M. Polycarpou, C. Panayiotou, and G. Ellinas, Eds., vol. 5769 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2009, pp. 175–184.

[104] Sarich, M., Djurdjevac, N., Brückner, S., Conrad, T., and Schütte, C. Modularity revisited: A novel dynamics-based concept for decomposing complex networks. *Journal of Computational Dynamics 1*, 1 (2014), 191–212. doi:10.3934/jcd.201.1.191.

[105] Sarich, M., and Schuette, C. Approximating selected non-dominant timescales by Markov state models. *Comm. Math. Sci. 10*, 3 (2012), 1001–1013.

[106] Sauer, T., Yorke, J. A., and Casdagli, M. Embedology. *J. Stat. Phys. 65*, 3-4 (1991), 579–616.

[107] Saul, L. K., Weinberger, K. Q., Ham, J. H., Sha, F., and Lee, D. D. Spectral methods for dimensionality reduction. In *Semisupervised Learning*. MIT Press, 2006, pp. 293–308.

[108] Schinkel, S., Dimigen, O., and Marwan, N. Selection of the recurrence threshold for signal detection. *The European Phys. J. 164* (2011), 45–53. Special Topics.

[109] Schutz, A. P., Zou, Y., Marwan, N., and Turvey, M. T. Local minima-based recurrence plots for continuous dynamical systems. *Int. J. Bifurcation and Chaos 21*, 4 (2011), 1065–1075. doi:10.1142/S0218127411029045.

[110] Seghouane, A. K., and Amari, S. I. The aic criterion and symmetrizing the kullback–leibler divergence. *Trans. Neur. Netw. 18*, 1 (Jan. 2007), 97–106. doi:10.1109/TNN.2006.882813.

[111] Simon, G., and Verleysen, M. High-dimensional delay selection for regression models with mutual information and distance-to-diagonal criteria. *Neurocomputing 70* (2007), 1265–1275.

[112] Stark, J. Delay embeddings for forced systems. i. deterministic forcing. *Journal of Nonlinear Science 9*, 3 (1999), 255–332. doi:10.1007/s00332990007.

[113] Stark, J., Broomhead, D., Davies, M., and Huke, J. Delay embeddings for forced systems. ii. stochastic forcing. *Journal of Nonlinear Science 13*, 6 (2003), 519–577. doi:10.1007/s00332-003-0534-4.

[114] Stark, J., Broomhead, D. S., Davies, M. E., and Huke, J. Takens embeddings theorems for forced and stochastic systems. *Nonlinear Anal. 30*, 9 (1997), 5303–5314. doi:10.1016/S0362-546X(96)00149-6.

[115] Takens, F. *Detecting strange attractor in turbulence*, vol. 898 of *Lecture Notes in Mathematics*. Springer Verlag, Berlin, 1981.

[116] Tausz, A., Vejdemo-Johansson, M., and Adams, H. JavaPlex: A research software package for persistent (co)homology. In *Proceedings of ICMS 2014* (2014), H. Hong and C. Yap, Eds., Lecture Notes in Computer Science 8592, pp. 129–136. Software available at http://appliedtopology.github.io/javaplex/.

[117] Teng, L., Li, H., Fu, X., Chen, W., and Shen, I.-F. Dimension reduction of microarray data based on local tangent space alignment. In *Proceedings of the Fourth IEEE International Conference on Cognitive Informatics* (Washington, DC, USA, 2005), ICCI '05, IEEE Computer Society, pp. 154–159.

[118] Thiel, M., Romano, M. C., Read, P. L., and Kurths, J. Estimation of dynamical invariants without embedding by recurrence plots. *CHAOS 14*, 2 (2004), 234–243.

[119] Uzal, L. C., Grinblat, G. L., and Verdes, P. F. Optimal reconstruction of dynamical systems: A noise amplification approach. *Phys. Rev. E 84* (Jul 2011), 016223. doi:10.1103/PhysRevE.84.016223.

[120] van der Maaten, L., Postma, E., and van den Herik, J. *Dimensionality Reduction: A Comparative Review*. Tilburg University, Netherlands, 2009.

[121] Vega, I., Schütte, C., and Conrad, T. Finding metastable states in real-world time series with recurrence networks. *Physica A: Statistical Mechanics and its Applications 445* (2016), 1–17. doi:10.1016/j.physa.2015.10.041.

[122] Vlachos, I., and Kugiumtzis, D. Non uniform state-space reconstruction and coupling detection. *Phys. Rev. E 82* (2010), 016207. doi:10.1103/PhysRevE.82.016207.

[123] Wang, J. *Geometric Structure of High-Dimensional Data and Dimensionality Reduction*. Springer Berlin Heidelberg, 2012.

[124] Whitney, H. Differentiable manifolds. *Ann. Math. 37* (1936), 645–680.

[125] Xu, X., Zhang, J., and Small, M. Superfamily phenomena and motifs of networks induced from time series. *PNAS 105*, 50 (2008), 19601–19605. doi:10.1073/pnas.0806082105.

[126] Yang, C., and Wu, C. Q. A robust method on estimation of Lyapunov exponents for a noisy time series. *Nonlinear Dyn. 64* (2011), 279–292. doi:10.1007/s11071-010-9860-x.

[127] Yang, C., Wu, C. Q., and Zhang, P. Estimation of Lyapunov exponents from a time series for n-dimensional state space using nonlinear mapping. *Nonlinear Dyn. 69* (2012), 1493–1507. doi:10.1007/s11071-012-0364-8.

[128] Yang, Y., and Yang, H. Complex network-based time series analysis. *Physica A 387* (2008), 1381–1386. doi:10.1016/j.physa.2007.10.055.

[129] Young, L. Entropy in dynamical systems. In *Entropy* (Princeton, New Jersey, USA, 2003), A. Greven, G. Keller, and G. Warnecke, Eds., Princeton Series in Applied Mathematics, Princeton Univ. Press, pp. 313–328.

[130] Zbilut, J. P., and jr. Webber, C. L. Embeddings and delays as derived from quantification of recurrence plots. *Physics Letters A 171* (1992), 199–203.

[131] Zbilut, J. P., Zaldivar-Comenges, J.-M., and Strozzi, F. Recurrence quantification based Liapunov exponents for monitoring divergence in experimental data. *Physics Letters A 297*, 3-4 (2002), 173–181. doi:10.1016/S0375-9601(02)00436-X.

[132] Zou, Y., et al. Identifying complex periodic windows in continuous-time dynamical systems using recurrence-based methods. *CHAOS 20* (2010), 043130. doi:10.1063/1.3523304.