
Università degli Studi di Padova

Dipartimento di Fisica *Galileo Galilei*

Corso di Laurea Magistrale in Fisica

# DETECTING CAUSALITY IN COMPLEX SYSTEM

*Relatore:*
Dott. Samir Suweiss

*Laureando:*
Jacopo Schiavon

September 2017

# ABSTRACT

In this work, recent and classical results on causality detection and predicability of a complex system have been reviewed critically. In the first part, I have extensively studied the Convergent Cross Mapping method [1] and I tested it in cases of particular interest. I have also confronted this approach with the classical Granger framework for causality [2]. A study with a simulated Lotka-Volterra model has shown certain limits of this method, and I have obtained counterintuitive results.

In the second part of the work, I have made a description of the Analog method [3] to perform prediction on complex systems. I have also studied the theorem on which this approach is rooted, the *Takens' theorem*. Moreover, we have investigated the main limitation of this approach, known as the *curse of dimensionality*: when the system increases in dimensionality the number of data needed to obtain sufficiently accurate predictions scales exponentially with dimension. Additionally, finding the effective dimension of a complex system is still an open problem. I have presented methods to estimate the effective dimension of the attractor of dynamical systems known as Grassberger-Procaccia algorithm and his extensions [4]. Finally, I have tested all these data driven machineries to a well-studied dynamical system, *i.e.* the Lorenz system, finding that the theoretical results are confirmed and the data needed scales as $N \sim \epsilon^{d}$.

# CONTENTS

# ACRONYMS

**CCM** Convergent Cross-Mapping

**COM** Center of Mass

**DoF** Degrees of Freedom

**GC** Granger Causality

**LL** Local Linear

**LV** Lotka-Volterra

**MLE** Maximal Lyapunov Exponent

**PAI** Pairwise Asymmetric Inference

**PCC** Pearson Correlation Coefficient

**SDE** Stochastic Differential Equation

**TE** Transfer Entropy

# INTRODUCTION

The aim of this work is to study how to infer information on a complex dynamical system only from data and without explicitly knowing the underlying dynamics nor the relevant parameters of the systems.

When trying to predict the future of a system, one has to choose one of two main approach to the problem: the first is to understand the laws that govern the evolution of the system (if any) thus creating theoretical models that can explain and reproduce the observed system behaviour - described by data - and can also make predictions on its evolution, known as *generative models* [5, 6]. The alternative approach is more inductive and it attempts to predict the system evolution by inferring statistical models directly from the data, in this case called *discriminative models* [5, 6]. In this second approach one does not care of causes, but it only exploits statistical correlations to make predictions.

The scientific method used extensively in physics, from Galileo onward, has always been built upon the first kind of strategy. Observations, experiments and mathematical theoretical frameworks are used to to understand the fundamental and universal laws that govern the inanimate matter world. They are thus translated in models with parameters, and data are used to estimate them and to make new predictions. Finally experiments are done to falsify these models, closing the never closed loop of a scientific discovery. This has been, for example, the case of the Standard Model, with the recent discover of Higgs Boson [7]: the Standard Model have been developed in the second half of '900, and Higgs has published his paper in 1964 [8], a whole 50 years before the actual experimental confirm of the particle (made in 2012 at LHC [9, 10]).

In fact, physicists have always been skeptical about the second - data based - inductive approach. Indeed, Physics emerged from the awareness that mathematics could be used as a language to reason about and describe natural world [11]. The goal of physics has been to isolate the causation phenomena described by mathematical models that are able to generate synthetic data prior to any observations. In general, physicists do not think of these models being generative because what else could they be? But it is a choice nonetheless. Surely, the development of theoretical physics has been allowed by the possibility to access to the most outstanding experiments in science that thanks to the relentless efforts of experimental physicists have reached a precision incomparable with other branches of science. In fact, when physicists confront problems in biology, economics and social sciences, the available data are of a poor quality and with a very complex structure. And in most of these cases we are very far from identifying first principle laws that govern the dynamics of these systems.

Conversely, many successful applications of artificial intelligence use models essentially as black boxes: they map a set of known inputs and outputs (training data) by determining a set of parameters that give good performance when

generalized to pairs of input and output (test data) not used in the training. In recent years, therefore, the always increasing amount of data available to scientist is leading some very deep question about the approach that should be used when studying complex systems.

It is of great interest, then, the possibility of infer at least some properties of the system without the needs to create the full model and fit it to data, or at least to understand when this task is at least applicable. A very important and widespread example of this approach are neural networks: by using a set of techniques that are best thought of as a supervised learning approach, treating the experimental/observational data as a direct input, which is then used to iteratively improve the prediction power of the statistical model.

## OUTLINE OF THE WORK

The approach I will present in this work is a compromise between the ability to predict and the desire to gain some insight about the structure and the processes driving the system dynamics. For example, one would like to understand if there are variables more "important" than others, or if some of them have a fundamental role in predicting others. In this thesis, I will thus focus on two distinct problems: the detection of causal relationship and the determination of effective variables' number (i.e. the dimensionality of the system), giving us crucial information for the understanding of the analyzed system.

Accordingly the thesis is divided in two main parts. The first one is about the concept of causality: what it is and methods to measure and detect it. Here (Chapter 1) I have firstly made a survey of the methods known up to now, focusing in one of the newer and more promising one, which is Convergent Cross-Mapping (CCM). Then (in Chapter 2) I have studied a toy Lotka-Volterra (LV)-like model exploring what happens to the causality predicted by CCM method variating some of its parameters.

The second part, instead, focus on methods to estimate the number of effective variables, in order to give an hint about predictability. I devote a chapter (Chapter 3 on page 29) for a theoretical review of the tools used to estimate the dimensionality of a system (through an approach introduced by Grassberger and Procaccia [4]) and to make prediction by means of the historical record of the system (the so called Analog method [3]), and another one (Chapter 4 on page 43) where I apply this approach on a complex system.

# 1

# GENERAL DEFINITION OF CAUSALITY, GRANGER CAUSALITY AND CONVERGENT CROSS-MAPPING METHOD

In this chapter I will first introduce the concept of causality in some of its many variants, then focusing on the description of CCM [1] method and its generalizations [12].

## THE CONCEPT OF CAUSALITY

While *causality* is a concept that is widespread in common language, it did not had a consistent formal definition for time series until half of the '900, and even now there are many variant of the same concept.

The problem is that, while we have a grasp of what we mean by saying that a certain phenomena *cause* another, putting these words in a more formal mathematical language is not an easy tasks. When dealing with events that occurs in time, probably the first relevant problem is to understand how one time series can be related to the others. This problem was partially solved by Galton and Pearson with the introduction of *correlation* [13].

Even if correlation is a good way to measure similarity between two time series, to imply causation from correlation is a logical fallacia, as was clear even to Pearson itself (see [14] for a detailed discussion of this problem from the point of view of Pearson and his student Yule). This problem was finally overcame by Granger in 1969 [2], with a big perspective shift: the causal relation must come from the ability to use *informations* encoded in one series to make statements about the other. In this sense, Granger idea was the first step toward an information theory framework, even though the tools used in his definition are purely statistical.

A possible and intuitive definition for a phenomenon to cause another one is that the former (*cause*) should pass some kind of information to the latter (*effect*), thus creating a directed interaction between the two. Another way of thinking this concept is that from the (*cause*) we should be able to predict in some way the (*effect*). Finally, changing perspective, one can think that the *effect*

should help the prediction, given that it must in some way encode informations about the *cause*. These three naively presented approach to causality are the most used ones, defining the three main categories of techniques currently used to make causal inference: *Transfer Entropy*, *Granger Causality* and *Convergent Cross Mapping*.

## GRANGER CAUSALITY AND TRANSFER ENTROPY

I will begin this survey with Granger Causality (GC), introduced in 1969 [2] and based upon Wiener's work [15] about predictability. Informally, given $X$ and $Y$ two stationary stochastic variables, $Y$ is said to *Granger cause* $X$ if the ability to predict $X$ is improved by incorporating information about $Y$.

Rephrased, given two time series, under the GC framework, one is considered the cause and one the effect if informations encoded in the former are helpful in order to predict the latter. This is consistent with our experience, but we want to formalize this concept.

In order to give a formal definition we have to first introduce some notation. Let denote $X_t$ as a stationary stochastic process and $\overline{X}_t$ the set of *past* values of $X_t$, which is to say the set $\{X_{t-j}\}$ with $j \in \{1, 2, \dots, \infty\}$. We define also $U_t$ to be all the information in universe accumulated until time $t$ and $U_t - Y_t$ all this information *apart* from a specified series $Y_t$. Moreover, denoted the prediction of $A_t$ using the set $B_t$ with $P_t(A|B)$, we define $\sigma^2(A|B)$ to be the variance of the series of predictive error $\epsilon_t(A|B) = A_t - P_t(A|B)$, $\sigma^2(A|B) = \langle \epsilon_t(A|B)^2 \rangle - \langle \epsilon_t(A|B) \rangle^2$.

Then we can write the following definition:

*Granger definition of* causality

1.1 DEFINITION (GRANGER CAUSALITY). Let $X_t$ and $Y_t$ be two stationary stochastic processes. If and only if $\sigma^2(X|\overline{U}) < \sigma^2(X|\overline{U - Y})$, then we say that $Y_t$ *Granger cause* $X_t$ (which we will indicate as $Y_t \xrightarrow{\text{Grg}} X_t$).

This definition means that we improve our ability of predicting $X$ (measured by the variance of the series of predictive errors $\sigma^2$) using all the information in the universe with respect to the prediction obtained excluding information about $Y$.

This very same definition can be extended naturally to the concept of *feedback*, which means that $X_t \xrightarrow{\text{Grg}} Y_t$ and at the same time $Y_t \xrightarrow{\text{Grg}} X_t$:

*Granger definition of* feedback

1.2 DEFINITION (GRANGER FEEDBACK). We say that feedbak between $X_t$ and $Y_t$ is occurring ($X_t \xleftrightarrow{\text{Grg}} Y_t$) if:

- $\sigma^2(X|\overline{U}) < \sigma^2(X|\overline{U - Y})$ and
- $\sigma^2(Y|\overline{U}) < \sigma^2(Y|\overline{U - X})$.

The last concept that is interesting for our scope is the *causality lag*, which represent the time that occurs for a series $Y$ to cause $X$. Technically, it is defined as:

1.3 DEFINITION (GRANGER CAUSALITY LAG). If $Y_t \xrightarrow{\text{Grg}} X_t$, we define the *causality lag* $m$ to be the least value $k$ such that $\sigma^2(X|U - Y(k)) < \sigma^2(X|U - Y(k+1))$, which means that knowing the values $Y_{t-j}$ for $j = 0, 1, \ldots, m-1$ does not improve our ability to predict $X_t$.

*Granger definition of* causality lag

This definitions are probably the most used for causality in a wide variety of fields from economy and finance to demography, social science and biomedical applications (for some interesting, even if far from the scope of this work, examples see [16–18]).

### TRANSFER ENTROPY

Transfer Entropy (TE) is an information theoretic measure that quantify the overlap of information content of two systems and, more importantly, the dynamics of information transport, thus exploiting eventual relation of causality. It has been introduced by Schreiber [19] in 2000 and ever since has been used in a wide range of applications (see [20] for an example of application of transfer entropy to the study of Ising model).

The definition starts from Shannon entropy, which defines the *quantity of information* (bits) needed to optimally encode independent draws of a discrete random variable I, which follow a probability distribution function $p(I)$ (the possible states available to I are $i \in \mathcal{I}$):

$$H_I = -\sum_{i \in \mathcal{I}} p(i) \log p(i), \tag{1.1}$$

where the base of the logarithm determines the units used for bits (for example 2 in the usual bit). From this definition, we understand that we can optimally encode a signal if we know the correct probability distribution $p(i)$. If we use instead $q(i)$, we will use a wrong number of bit quantified by Kullback entropy:

$$K_I = \sum_i p(i) \log \frac{p(i)}{q(i)}.$$

In other words the Kullback entropy gives us a measure of our misunderstanding of the variable that we want to study: we think that its probability distribution is $q(i)$, and we thus need $H_I|_q$ bits to encode the possible states. In reality, though, the probability distribution is $p(I)$, and the Kullback entropy measures exactly the difference in entropy between the correct probability distribution and our wrong guessed one.

The definition of *Mutual Information* follows from this one, and represents the number of bit that we wrongly use when thinking that two phenomena are independent while they are correlated instead:

$$M_{IJ} = \sum_{ij} p(i,j) \log \frac{p(i,j)}{p(i)p(j)}. \tag{1.2}$$

This quantity does not take into account the dynamics of information, simply telling that the two systems are related, and is obviously symmetric for the exchange of I and J.

In order to consider dynamics of informations, which means some sort of directed flow, we should generalize a similar concept to the study of entropy's rates. This extension can be made by observing deviation from the generalized Markov property of the process[1]: in absence of information flow (and thus of causal relation) from J to I the state of J has no influence on transition probabilities of I. The difference in entropy between the model in which the two processes are unrelated and the one in which they indeed are can be, again, measured with a Kullback-like entropy, which we call *Transfer Entropy*:

$$T_{J\to I} = \sum_{ij} p\left(i_{n+1}|i_n^{(k)}, j_n^{(l)}\right) \log \frac{p\left(i_{n+1}|i_n^{(k)}, j_n^{(l)}\right)}{p\left(i_{n+1} \mid i_n^{(k)}\right)}, \tag{1.3}$$

i.e. given our null model that there is no causal relation between the two variables, the TE tell us how much we are wrong, in the sense that it quantify the quantity of extra information that we have to employ in order to describe the two variables as unrelated, without considering that a dipendence exist instead.

We highlight that this concept can be extended to continuous state systems with almost no effort by taking a limit of a coarse graining procedure, as explained in [19].

The practical definition of causality is thus:

*Transfer Entropy definition of causality*

1.4 Definition (TE causality). Let $X_t$ and $Y_t$ two stochastic processes and let $T_{X\to Y}$ and $T_{Y\to X}$ be the TE between the two as expressed from equation (1.3). Then we say that $X_t$ cause $Y_t$ if the information flow has a net value in the direction $X \to Y$, which means that $T_{X\to Y} \geqslant T_{Y\to X}$.

In 2009, Barnett *et Al.* [21] proved that, for Gaussian variables, TE and GC are completely equivalent, thus unifying the two fields in this sense and giving a solid information theoretical ground to GC, confirming our intuition that Granger framework was a first step into moving from a simple statistical correlation to a more information theoretical definition of causality.

### Why do we need to overcome Granger framework?

GC framework is based upon linear regression, and thus is not suited to takle nonlinear dynamical problems.

In fact, even if it is often overlooked in many applications, the GC framework heavily employs separability of variables: information content of one time series needs to be separated from the universe information set in order to determine if that variable is (or is not) causative in the model. While this request might looks trivial, and indeed it is in most applications, the nonlinear coupled dynamical systems are an important exception.

1 We say that a pair of processes obey a generalized Markov property if:

$$P\left(i_{n+1}|i_n^{(k)}\right) = P\left(i_{n+1}|i_n^{(k)}, j_n^{(l)}\right)$$

In order to understand this subtle implication, we should first understand exactly what is meant by *separability* by looking to an example in which this is indeed the case: a pair of coupled stationary Markov processes[2]. In this context, coupling means that transition probabilities for the process $(X_t)$ which is caused depends on the state of the process $(Y_t)$ which is the cause (thanks to the Markov property we can avoid to write all the precedings states):

$$P(X_{t_i} \to X_{t_{i+1}}) = P(X_{t_{i+1}} \mid X_{t_i}, Y_{t_i})$$

It is then clear that ignoring informations about $Y_t$ removes all the information about causality: depending on the exact functional form of $P(X_{t_{i+1}} \mid X_{t_i}, Y_{t_i})$ the loss can be small or quite relevant, and this gives us (by means of GC or TE approaches) a quantification of the causality relation.

In a linear dynamical system, the idea is almost the same: state $X_t$ depends only on the state $X_{t-1}$ and, if there is a coupling with another variable, on the state $Y_{t-1}$. Thus, removing $Y_t$ from our information set modify our ability of predicting $X_t$, and GC or TE frameworks assure us that we can employ this difference to detect the causality relation.

As we will understand better and in a formal way (by means of the powerful Takens' theorem) in the following section, let just analyze a simple case that can help in the visualization of the issue that appears in the case of nonlinear coupled dynamical systems. Consider, for example, this coupled logistic system:

$$\begin{cases} X(t+1) = A_x\, X(t)\,[1 - X(t) - \beta_x\, Y(t)] \\ Y(t+1) = A_y\, Y(t)\,[1 - Y(t) - \beta_y\, X(t)]\,. \end{cases} \tag{1.4}$$

In this system, we can rearrange the equations to use the values of $Y(t)$ and $Y(t+1)$ to express $X(t)$ and vice versa, which means that we exploit the information about one time series to estimate the other one. This gives:

$$\begin{cases} \beta_x\, Y(t) = 1 - X(t) - \dfrac{X(t+1)}{A_x\, X(t)} \\[2ex] \beta_y\, X(t) = 1 - Y(t) - \dfrac{Y(t+1)}{A_y\, Y(t)}. \end{cases} \tag{1.5}$$

In these equations, though, $X_t$ depends on $Y_t$ and its future state, but if we reinsert them back into equation (1.4), we obtain:

$$\begin{cases} X(t) = \frac{A_x}{\beta_y}\left[(1 - \beta_x Y_{t-1})\, g(Y) - \frac{1}{\beta_y} g(Y)^2\right] \\ Y(t) = \frac{A_y}{\beta_x}\left[(1 - \beta_y X_{t-1})\, g(X) - \frac{1}{\beta_x} g(X)^2\right], \end{cases} \tag{1.6}$$

2 A stationary stochastic Markov process is defined as a sequence of events $X_{t_i}$, where $t_i$ are ordered increasingly, for which the transition between a state and another (which is, the evolution from $X_{t_i}$ and $X_{t_{i+1}}$) is probabilistic (stochastic process), depends only on the last state (Markov property) and the process itself has a constant temporal average (stationarity). See [22] for an introduction to stochastic processes.

where

$$g(i) = \left(1 - i_{t-1} - \frac{i_t}{A_i i_{t-1}}\right).$$

These equations can be written as:

$$\begin{cases} X(t) = \mathcal{F}(Y_t, Y_{t-1}) \\ Y(t) = \mathcal{G}(X_t, X_{t-1}), \end{cases} \tag{1.7}$$

which means that we can reinsert them inside equation (1.4) in order to obtain a decoupled system:

$$\begin{cases} X(t+1) = A_x\, X(t)\,[1 - X(t) - \beta_x\, \mathcal{G}(X_t, X_{t-1})] \\ Y(t+1) = A_y\, Y(t)\,[1 - Y(t) - \beta_y\, \mathcal{F}(Y_t, Y_{t-1})]\,. \end{cases} \tag{1.8}$$

This apparently simple rearrangement of variables yields a profound consequence on GC framework: in fact, if we look at the causality relation between X and Y (represented in this case by the coefficient $\beta_y$), we see that GC cannot be applied anymore: we have effectively removed X from the dynamic of Y in the second of the two equations, but we have not lost any ability in predicting Y. The two systems leads to the same solutions, and then (if using GC or TE definitions) we would have incorrectly assumed that the two series are independent.

As a simple real world example, consider the obvious causal relation between weather and the number of umbrella used in a given day. If we record the number of umbrellas used for quite a long period of time, we might use this kind of historical information to predict the number of umbrellas in the following day. The key point is that, the number of umbrella already contains information about the weather, and so considering explicitly also this time series does not improve our ability of prediction, and then GC would exclude this causality relation (incorrectly).

This reasoning, as we will see in the following section, will be generalized to almost all nonlinear dynamical system, with the exception of strongly coupled ones, in which the relative influence is so strong that the two series are effectively only one.

## CONVERGENT CROSS-MAPPING

Given what we have just seen, an approach to study the causal relations for nonlinear dynamical time series is needed. One method has been proposed by Sugihara in 2012 [1] and involves a change of the point of view: in this approach, $X_t \overset{\text{CCM}}{\longrightarrow} Y_t$ if is $Y_t$ that help us in predicting $X_t$, instead of the opposite perspective, typical of GC framework.

Rephrased, the CCM approach test for causality between X and Y by measuring the extent to which the historical record of Y values can reliably estimate states of X (and vice versa).

In order to understand why this can happen, we will need to introduce Takens' theorem, the main result of this chapter.

### TAKENS' THEOREM AND CONVERGENT CROSS-MAPPING

From a theoretical point of view, two time series are causally linked if they are generated from the coupled dynamical equations describing the evolution of the system under analysis.

For many dynamical system, there exist a subset of the phase space (called attractor) that is the smallest that has the following properties:

- is forward invariant: if a point is inside the attractor, it will forever be inside;
- there exist a neighbor of this set (called basin of attraction) and the point of this neighbors will eventually fall inside the attractor in the future.

For almost all physical problems, the time series observed in nature that are generated from the same dynamical system (of dimension $d$) shares a common attractor, which means that they are different component of a $d$ dimensional vector that evolves inside the attractor set. If the dynamical system is differentiable, then the attractor has a differential manifold structure and, moreover, it is a topological manifold if it is embedded in $\mathbb{R}^d$ with the euclidean topology. In figure 1.1 is represented an example of attractor, referred to the famous Lorenz system [23] defined in (1.12). It is clear that all the trajectory that leaves a point inside the classical butterfly shape stays inside the same set, revolving around the double center forever.

Takens' reconstruction theorem [24], of which a proof is given in Appendix A on page 49, gives us a powerful instrument to examine such attractors:

1.5 THEOREM (TAKENS' THEOREM FOR DYNAMICAL SYSTEMS).
*Let $M$ be the compact manifold represented by the attractor of the system (with dimension $d$). Let $(\Phi_\tau, X)$ be:*

*Takens' theorem for dynamical systems*

- *$\Phi_\tau \colon M \to M$ a smooth (at least $C^2$) diffeomorphism: the flux of the dynamics system after a time delay $\tau$ (we remember that $\Phi_\tau^2 = \Phi_{2\tau}$),*
- *$X \colon M \to \mathbb{R}$ a smooth (at least $C^2$) function, the projection of the attractor over one of the variable of the system (which is to say, the time series of this variable).*

*Then, it is a generic property that the $(2d+1)$-fold observation map ($p$ is a point in $M$):*

$$\begin{aligned} H[\Phi, X] \colon M &\to \mathbb{R}^{2d+1} \\ p &\mapsto (X(t), X(t-\tau), \dots, X(t-2d\tau)) \end{aligned}$$

*is an embedding (*i. e. *$H$ is a bijection between $M$ and its image $M_X$ after $H$, and both $H$ and $H^{-1}$ are continuous). By generic we mean that the set of pairs $(\Phi, X)$ for which this property is true is open and dense in the set of all pair $(\varphi, h)$ with $\varphi \colon M \to M$ a $C^2$ diffeomorphism and $h \colon M \to \mathbb{R}$ a $C^2$ function.*

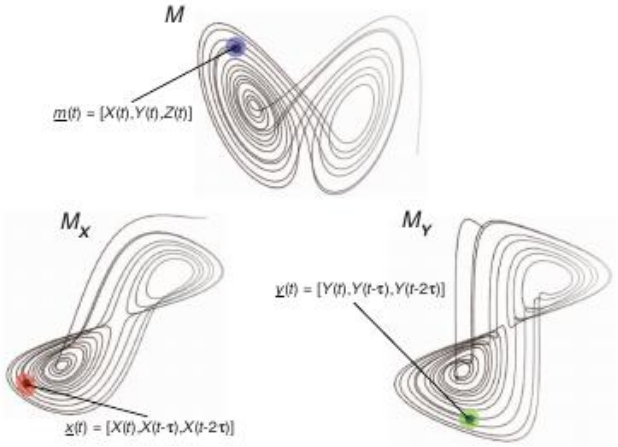



**Figure 1.1:** Schematic representation of the *shadow manifolds* obtained thanks to Takens' reconstruction theorem.

What this theorem tells us is that, under certain mild hypotesis, we can find an exact (from a topological point of view) copy of the behaviour of the system using only one of the variables and his past history. This explains why in the previous section we could manage to rearrange the coupled logistic system to express each variable only in term of itself and his past value.

Moreover, this allows us to make prediction: if we can reconstruct the dynamic thanks to one variable, and our reconstruction is topologically equivalent, we can use the reconstructed attractor to predict the value of the other variable.

Given a point $p \in M$ of the attractor, we choose the variable $X$ to obtain its projection in one coordinate and then, by means of Takens' theorem, we create the diffeomorphic copy of the attractor $M_x$, in which lives the copy $p_x \in M_x$ of the original point $p$. Takens' theorem assure us that the topologies over the two manifolds are homeomorphic, and thus the neighbors $\mathfrak{U}_p$ and $\mathfrak{U}_{p_x}$ are differentially the same, which means that to each point $q \in \mathfrak{U}_p$ corresponds one to one a point $q_i \in \mathfrak{U}_{p_x}$.

### CONVERGENT CROSS-MAPPING

We can then create a constructive method to obtain an estimate of one projection from another one.

Let's call the two series we are interested in $X(t)$ and $Y(t)$, the attractor manifold $M$ of dimension $d$, $E = 2d + 1$ the reconstruction dimension and $\tau$ the time lag we are interested in (this will usually correspond to the lag between two points in the time series, or a fraction of the period of the solution. More about this choice will be told in Chapter 3).

*Algorithm for Convergent Cross-Mapping*

1.6 ALGORITHM (CONVERGENT CROSS-MAPPING).

- Choose a projection direction $X$.
- Create the shadow manifold $M_x$ by defining (for each $t \geqslant E - 1$) the points $\mathbf{X}(t) = (X(t), X(t-\tau), \ldots, X(t-\tau(E-1)))$.

- For each point $Y(\tilde{t})$, choose the corresponding $\mathbf{X}(\tilde{t})$ and look for its $E+1$ nearest neighborhood.
- Calling $t_k$ the time index of the kth-nearest neighborhood, calculate some weights $w_k$ based on the distances between $\mathbf{X}(t_k)$ and $\mathbf{X}(\tilde{t})$.
- Finally calculate $\hat{Y}(\tilde{t})$ as a weighted sum

$$\hat{Y}(\tilde{t}) = \sum_{k \leqslant E+1} w_k Y(t_k).$$

As one can see, this algorithm provide an estimated time series $\hat{Y}(\tilde{t})$ (excluding the first $E-1$ points for which we cannot create the shadow projection) that can be evaluated against the real $Y(\tilde{t})$, for example by calculating Pearson Correlation Coefficient (PCC) $\mathrm{Pcc}(\hat{Y}, Y)$[3]. We will call the correlation between $\hat{Y}$ given $X$ and $Y$ as $\mathrm{Pcc}(X,Y)$, and usually we will use the parameter $C(X,Y)$ defined as:

$$C(X,Y) = [\mathrm{Pcc}(X,Y)]^2 \tag{1.9}$$

In order to calculate weights $w_k$ one should find a function that makes the nearest points more important: we want to give more importance to the nearest point, because they have a closer history. The function used by Sugihara is

$$w_k = \frac{1}{W} \max\left(\exp\left[-\frac{\|\mathbf{x}_k - \mathbf{x}_t\|}{d_{min}}\right], w_{min}\right) \tag{1.10}$$

with $W = \sum_k w_k$, $\mathbf{X}_k$ the kth-nearest neighborhood, $d_{min} = \|\mathbf{X}_{\tilde{t}} - \mathbf{X}_0\|$ the distance between $\mathbf{x}_{\tilde{t}}$ and its nearest neighborhood and $w_{min} = 10^{-6}$ a security threshold for computational issues. If $d_{min} = 0$, it means that the nearest point is actually superimposed to $\mathbf{X}_{\tilde{t}}$, and then the sum is reduced to just that point (assigning all other weights to 0)

### PROPERTIES OF CONVERGENT CROSS-MAPPING

CCM has many important property that can help us with our task of identifying causality. The first is the one that gives part of the name to the method:

*Property of convergence of Cross-Mapping*

PROPERTY 1.7 (CONVERGENCE): *Calling* $L$ *the lenght of the time series used to Cross-Map, if the hypothesis of Takens' theorem are satisfied,* $\mathrm{Pcc}[L]$ *is a function that converge to 1 as* $L$ *approaches infinite, formally:*

$$\lim_{L\to\infty} \mathrm{Pcc}[L] = 1$$

[3] PCC is defined for a population as

$$\mathrm{Pcc}(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}.$$

For a sample $x_i$ and $y_i$ of size $n$ this expression be rewritten in a more computationally convenient way as:

$$\mathrm{Pcc}(X,Y) = \frac{n\sum_i x_i y_i - \sum_i x_i \sum_i y_i}{\sqrt{n\sum_i x_i^2 - (\sum_i x_i)^2}\sqrt{n\sum_i y_i^2 - (\sum_i y_i)^2}}$$

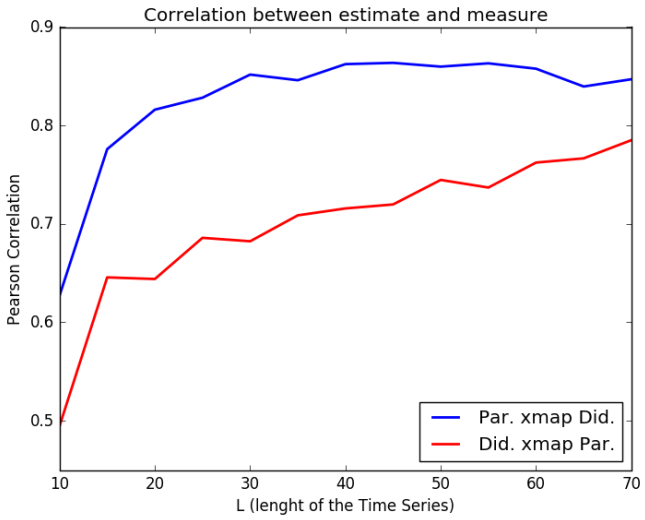



**Figure 1.2:** The prediction ability for the cross-mapping between two time series in a prey-predator system, versus the length of the time series used to create the shadow manifold and make prediction. For each length an estimated series has been calculated, and the PCC with the corresponding true series has been calculated. The solid line represent the fit from Equation 1.11. The data comes from [25] and are freely available at `https://robjhyndman.com/tsdldata/data/veilleux.dat`.

This property, exemplified in Figure 1.2, ensure us that, given a sufficient observation time, each series will predict the others in the same attractor with perfect ability.

In order to test the quality of a prediction, then, one can explore how fast the estimated series converge to the measured one, for example by evaluating a fit of the data with an exponential law:

$$\mathrm{Pcc}(L) = \rho_0 - \alpha e^{-\gamma L} \tag{1.11}$$

where $\rho_0$ should be as near to 1 as possible while $\gamma$ is the speed of convergence.

The fact that $\rho_0$ is not always 1 is not unexpected nor a problem: the presence of noise (both due to the model and the measurement) hinder slightly this method, because as seen from Theorem 1.5 on page 9 we need the time flux of the dynamics system to be $C^2$, which cannot be said for a Stochastic Differential Equation (SDE) in general[4]. Anyway, if the noise is sufficiently small, this approach works in the same way, but the prediction will not converge exactly to the measured one.

Moreover, as I will discuss in Section 1.3 on page 15 and show in Chapter 2 on page 17, the injection of a small quantity of noise will help the detection of causality relations.

Another important property which deserves to be mentioned is another direct consequence of the Taken's theorem:

*Property of the choice of projection*

PROPERTY 1.8 (CHOICE OF PROJECTION): *The quality of the prediction depends on the projection chosen. Moreover, in certain cases, exists projections that cannot predict other variables.*

4 For some particular class of SDE an extension of Takens' theorem has even been proven, see [26]

This property comes naturally when we think about the structure of the shadow manifold described in Theorem 1.5: we just need that the function $X\colon M \to \mathbb{R}$ to be "bad" in some sense, and the entire projection will be hindered. For example, if the function maps two point that in the attractor are far away very close to each other (as is the case with the Z axis in figure 1.3 on the next page) the only way to distinguish them will be with their past history, and the prediction will be worse. In the worst case, the projection will make two or more entire portions of the attractor indistinguishable, becoming completely useless for predictions.

An example of this behaviour is shown in Figure 1.3, where the Lorenz attractor [23] is depicted, as obtained from the three dimensional system:

$$\begin{cases} \dot{X}_t = \sigma\,(Y_t - X_t) \\ \dot{Y}_t = X_t\,(\rho - Z_t) - Y_t \\ \dot{Z}_t = X_t Y_t - \beta\, Z_t. \end{cases} \tag{1.12}$$

with $\sigma = 10$, $\beta = 8/3$ and $\rho = 28$.

In this case, projecting over Z-axes does not allow a good reconstruction, due to the fact that the projection itself does not represent correctly the attractor.

### CONVERGENT CROSS-MAPPING AS A METHOD TO DETECT CAUSALITY

The key concept from Sugihara [1] is that, given two time series from the same attractor, one can say that there is a causal relation (in Sugihara sense, or CCM sense) if at least one of them can predict the other. If both of them give good prediction, causality flow from the one that *is predicted* better to the other, but in this case it is better to talk about a feedback relation, in which both variable influence each other.

We define a parameter $\Delta(X, Y)$ in order to have a clearer definition of causality: given two time series $X_t$ and $Y_t$ and their Cross-Mapped correlation squared (as defined in equation (1.9)) $C(X, Y)$ and $C(Y, X)$, we define

$$\Delta(X, Y) = C(Y, X) - C(X, Y).$$

This allow us to make the following statement:

*CCM causality definition (Sugihara causality)*

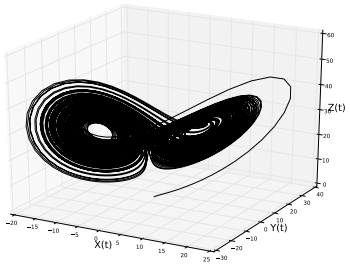
1.9 DEFINITION (CCM CAUSALITY). Given two time series $X_t$ and $Y_t$ from the same attractor, we say that:

$$X \xrightarrow{\text{CCM}} Y \qquad \text{iff} \qquad \Delta(X, Y) > 0,$$

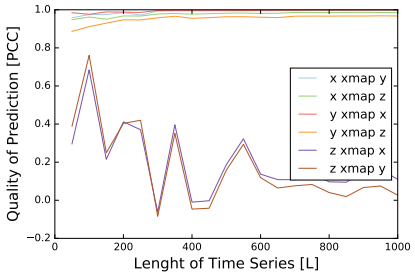
or, in words, X *CCM cause* Y if the ability of Y of predicting X is better than the ability of X of predicting Y (and so $C(Y, X)$ is bigger than $C(X, Y)$).

This definition give us an operative method to inquire causality: by creating shadow manifold from all the variables and estimating the other series, we can obtain a picture of causality relations in the system.

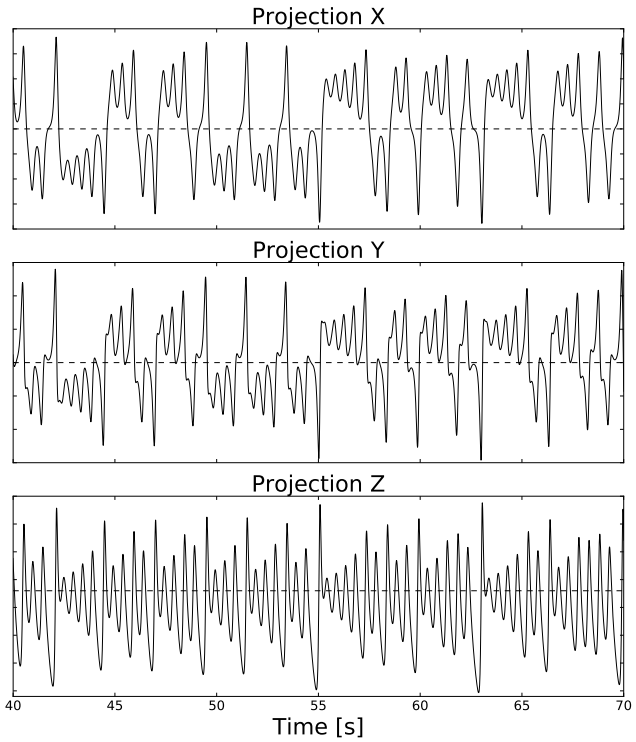
The peculiarity of this method is that it works in a different setting with respect to Granger framework, expanding the range of problems that can be tackled, and as such has been used in a wide range of fields [27, 28].




**(a)** The attractor




**(b)** Correlations




**(c)** Projections along the three axes

**Figure 1.3:** The Lorenz attractor (equations (1.12)) and the correlation evaluated with increasing time series length. From figure 1.3b appears that variable Z is not a good predictor (its skill is way smaller than that of the others). The reason of this problem is bounded to the shape of the attractor (figure 1.3a) and the choice of projection. As appears clearly from the three projections (figure 1.3c), the Z axes cannot distinguish from the two centers around which the dynamics evolves. This means that, when seen from Z's perspective, it is exactly the same being in the first or in the second center, and so its predictions are inevitably unreliable.

## PROBLEM OF CONVERGENT CROSS-MAPPING AND EXTENSIONS

Unfortunately, though, also this method presents some problems that can hinder severely its applicability and that have been highlighted by McCracken *et Al.* in 2014 [12].

Mainly, the problem is that this method appears not to work in some range of parameters or for some particular systems, for which Taken's hypotesis does not apply. More precisely, McCracken discusses two cases in particular: the first difficulties appears when the system is strongly coupled. In this case, the strong coupling allows for both projections to be equally good at predicting the time series evolution and the method fails in detecting causality. In some cases, the presence of a strong coupling is evident from the data (for example if more than one variable has the same, strong periodicity, probably one of them is coupled strongly to the others and drives them) or is known from previous measurements or models. In those cases, probably Granger causality is a better test.

In the second case the method does not work depending on the shape of the attractor, and this is in generally difficult to say even a posteriori! In this case CCM may provide an unreliable result, but we do not have even a way to know if we are actually in this situation or not. In other word there is not a systematic way to know if CCM is working or not. This problem has been addressed in [12], where some systems are investigated within a certain range of parameters and CCM appears to change its prediction.

This issue is closely related to the attractor manifold and the chosen projection, and it is due to property 1.8 on page 12. In fact, if the attractor is shaped in such a way that one of the natural projection is not well suited for reconstruction (because it does not respect the hypotesis of Takens' theorem), then creating a reconstructed version of the attractor made with it will not be homeomorphic to the target one. This means that our prediction will become unreliable, but if we are not aware that the problem is in the hypotesis of the theorem, we might think that this is a genuine effect of causality. For example, if we calculate $\Delta(X, Z)$ or $\Delta(Y, Z)$ for Lorenz attractor (see Figure 1.3 on the preceding page) we find immediately that $Z$ is strongly driven from both $X$ and $Y$, which is definitely not true. Obviously, in this example one could see this effect just by watching the trajectories, but other systems can be much more subtle.

A possible simple solution to this problem could be to change the projection function, but of course this undermines the scope of the work, which is to understand the relation between variables and not between function of variables.

### PAIRWISE ASYMMETRIC INFERENCE

In order to solve this problem, in [12] a variant to CCM method, called Pairwise Asymmetric Inference (PAI), has been proposed.

Consider a $2d+2$ dimensional manifold created as the image of:

$$\begin{aligned} \hat{H}[\Phi, X, Y] \colon M &\to \mathbb{R}^{2d+1} \\ p &\mapsto (X(t), X(t-\tau), \dots, X(t-2d\tau), Y(t)) \end{aligned}$$

Estimating $X$ with this shadow manifold is like measuring the extent to which a single time step of $Y$ improves the ability to estimate $X$, similar to what Granger defines as causality.

Then, after calling $C(X, XY)$ the correlation squared between this estimation and the original time series $X$, one can define:

$$\hat{\Delta}(X, Y) = C(X, XY) - C(Y, YX),$$

and then, as done before, link this parameter to the causality direction: if $\hat{\Delta} > 0$ then the addition of a single time step of $Y$ improves the estimation of $X$ more than what adding a single step of $X$ does for $Y$, thus implying that $Y$ contains more information about $X$ than the other way around and so that $X$ causes $Y$. We can thus define the PAI causality as:

*PAI causality definition*

1.10 Definition (PAI causality). Given two time series $X_t$ and $Y_t$ from the same attractor, we say that:

$$X \xrightarrow{\text{PAI}} Y \qquad \text{iff} \qquad \hat{\Delta}(X, Y) > 0,$$

or, in the same way, $X$ *PAI cause* $Y$ if the addition of a single time step from time series $Y$ improves the self estimation of $X$ better than what a single time step of $X$ does when added to $Y$ (and so $C(X, XY)$ is bigger than $C(Y, YX)$).

The interesting thing about this alternative definition is that it shift back the perspective to something similar to Granger definition, but keeping the Takens' structure typical of the study of dynamical systems. In fact, we are evaluating if (and how much) the addition of one variable improves our ability to predict the other one.

# 2

# APPLICATIONS OF CONVERGENT CROSS-MAPPING ALGORITHM ON A MODEL ECOSYSTEM

In order to test the CCM method and its limits, we apply it on times series generated through a multi-species population dynamics model known as generalized Lotka-Volterra [29] (LV).

## DESCRIPTION OF THE SYSTEM

The model is described in equations (2.1), where $K$ is the carrying capacity for preys (whose population is indicated by $X$), $Y$ is the predators population and $E$ denotes state variable for the environment that is coupled with population dynamics. The equations are stochastic as white noise [22] is added to the system. Three distinct sources of noise have been considered:

1. An environmental noise, additive for $E$, whose strength is identified by its standard deviation $\sigma_E$;
2. Demographic (multiplicative white) noise in both predator and prey population dynamics with intensity $\sigma_X$ and $\sigma_Y$, respectively;
3. A measurement noise, for both $X$ and $Y$, mimicking uncertainties obtained by adding to the final series a white Gaussian noise with standard deviation $\sigma_{Meas}$.

The most general formulation of this system is thus:

$$\begin{cases} E(t) = \sin \omega t + \sigma_E dW \\ \dot{X}_t = a X_t \left(1 - \frac{X_t}{K}\right) - b X_t Y_t + \sigma_X X_t dW \\ \dot{Y}_t = (-c + \gamma E(t)) Y_t + d X_t Y_t + \sigma_Y Y_t dW \end{cases} \tag{2.1}$$

where $dW$ represents symoblically the infinitesimal increments of a Wiener random walk process $W_t = \int \xi_t dt$ with $\xi_t = \mathcal{N}(0, 1)$.

The parameters that I have fixed are:

$$a = 1.5 \qquad b = 0.5 \qquad c = 1.5 \qquad d = 1.5$$
$$\omega = 1$$
$$X_0 = 1 \qquad Y_0 = 1.$$

while coupling with environment $\gamma$, carrying capacity K and noises' standard deviations have been varied.

To integrate this SDE system I have implemented a stochastic Runge-Kutta second order method (also known as *improved Euler* or *Heun* method, see [30, 31] for an overview of numerical methods for SDE solutions), given by the update equation:

$$\begin{aligned} \bar{\mathbf{x}}_i &= \mathbf{x}_i + f(t_i, \mathbf{x}_i)\delta t + g(t_i, \mathbf{x}_i)dW \\ \mathbf{x}_{i+1} &= \mathbf{x}_i + \frac{\delta t}{2}\left[f(t_i, \mathbf{x}_i) + f(t_{i+1}, \bar{\mathbf{x}}_i)\right] + \frac{dW}{2}\left[g(t_i, \mathbf{x}_i) + g(t_{i+1}, \bar{\mathbf{x}}_i)\right] \end{aligned} \tag{2.2}$$

I have used a $\delta t$ of $1 \times 10^{-2}$ s, and integrated over a period of 150 s. Then I downsampled the resulting series with a factor 50 and excluded the first 50 seconds of the dynamics (to avoid transient regime), thus obtaining series of 200 points.

Before the application of CCM-method algorithm, I have standardized the time series, subtracting the mean value and dividing by their standard deviation. It is important to note that this procedure is done after the generation of the time series, and that change the meaning of X and Y in this context: in fact, the two variables loses their sense as population values (that should be always bigger or equal to zero). The reason is simply for computational and graphical convenience, because in this way the two series becomes comparable both from a numerical point of view (finite precision effect have less influence if both series have comparable excursion) and a visual one (in this way the two series can be compared in the same plot even if they originally are very different).

Finally we have applied the CCM method to infer causality relations among X, Y and E for different set of paramters. For each of the parameters combination, I have repeated this entire procedure 10 times and averaged the results of CCM.

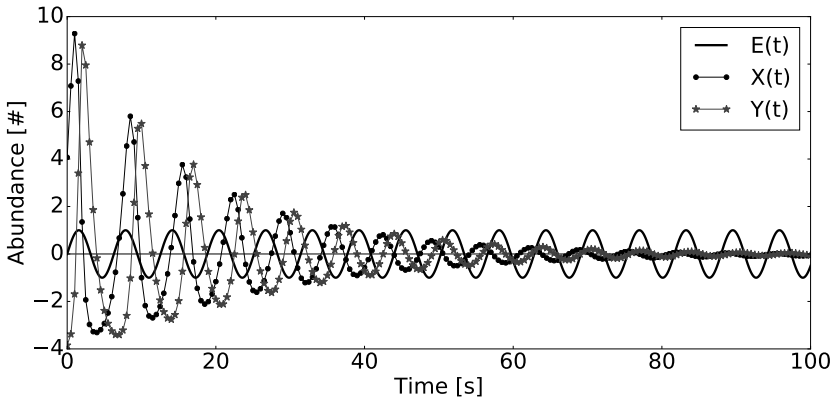
### EXPECTED RESULT FROM INTUITION

From the choice of parameters for the system that we have done, we expect the correlations directions to be as follows:

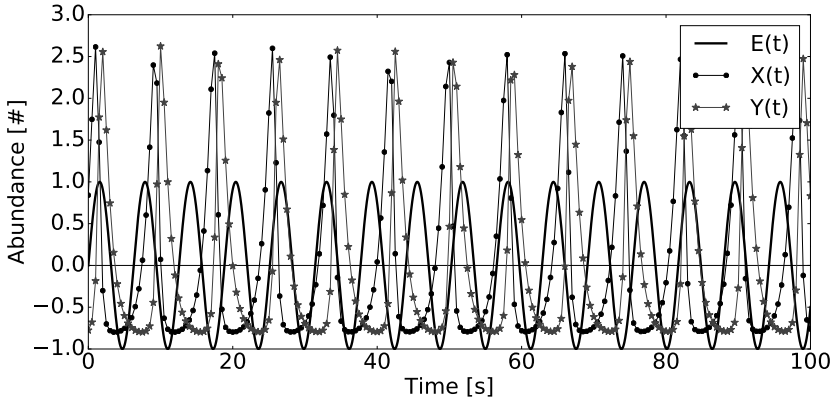
$$X \xrightarrow{\text{CCM}} Y \qquad \Longleftrightarrow \qquad \Delta_{XY} > 0 \tag{2.3}$$
$$E \xrightarrow{\text{CCM}} Y \qquad \Longleftrightarrow \qquad \Delta_{EY} > 0 \tag{2.4}$$
$$E \xrightarrow[\underline{Y}]{\text{CCM}} X \qquad \Longleftrightarrow \qquad \Delta_{EX} > 0 \tag{2.5}$$

where I have underlined that the environment cause X through Y.

Obviously, the condition in equation (2.3) is the trickiest one: there is a two-way relationship between X and Y and, according to Sugihara, the CCM method should be able to discern that $d > b$.




**(a)** $K = 5$




**(b)** $K = 500$

**Figure 2.1:** In figure are represented the first 100 s of the dynamics for the system (2.1) without environmental coupling. When $K$ is small, the series converges quickly to the equilibria solutions $(X_\infty, Y_\infty)$ (in this case, after the standardization, the equilibria are almost $(0, 0)$: reaching quickly the equilibrium means that the mean value for each series will be almost equal to $(X_\infty, Y_\infty)$ and thus the standardization removes it exactly), while for $K$ very big the dynamics is a dumping cycle (however, the dumping is small and the dynamics approach the equilibrium as slow as $1/K$, that we have to wait a long time interval to see that the series reach $X_\infty, Y_\infty$). $E(t)$ (the environment) is reported as a reference.
Points identified with circles and stars shows the downsampling operation, while the solid lines represents the entire solution.

Observing again system (2.1), we notice that when there is no coupling with environment, the system can be solved analytically by linearization of the equations around the equilibria $(X_\infty, Y_\infty)$, see Figure 2.1 on the preceding page. The only interesting equilibrium (in which neither of the variables goes extinct) is given by:

$$X_\infty = \frac{c}{d}$$
$$Y_\infty = -\frac{a(c - dk)}{bdk}.$$

The eigenvalues analysis of the linearized matrix shows that the two series in the phase space follow a spiral that falls to the equilibrium with a speed that depends (once fixed all the other parameters) on $1/K$. Thus, in the case of a small K the two series converge quickly to $(X_\infty, Y_\infty)$, while if K is much bigger that the other parameters, the equilibrium is reached after a long time.

Then, even if the second case is similar to an extended version of the transient regime of the first one, it allows us to explore what happens when there is a natural oscillation regime over the oscillations inducted by environment.

## SMALL CARRYING CAPACITY

With a small carrying capacity ($K = 5$, we are in the situation depicted in figure 2.1a on the previous page), the system quickly relaxes to $(0, 0)$ (after standardization). If we exclude also the next 50 points, thus beginning measurements after 100 s, the attractor is a point (a manifold of dimension $d = 0$) and then the hypothesis of Takens' theorem are not applicable anymore, leading to a failure of CCM method. This fact is solved with the addition of model noise or with a coupling with an external (driver) variable $E(t)$.

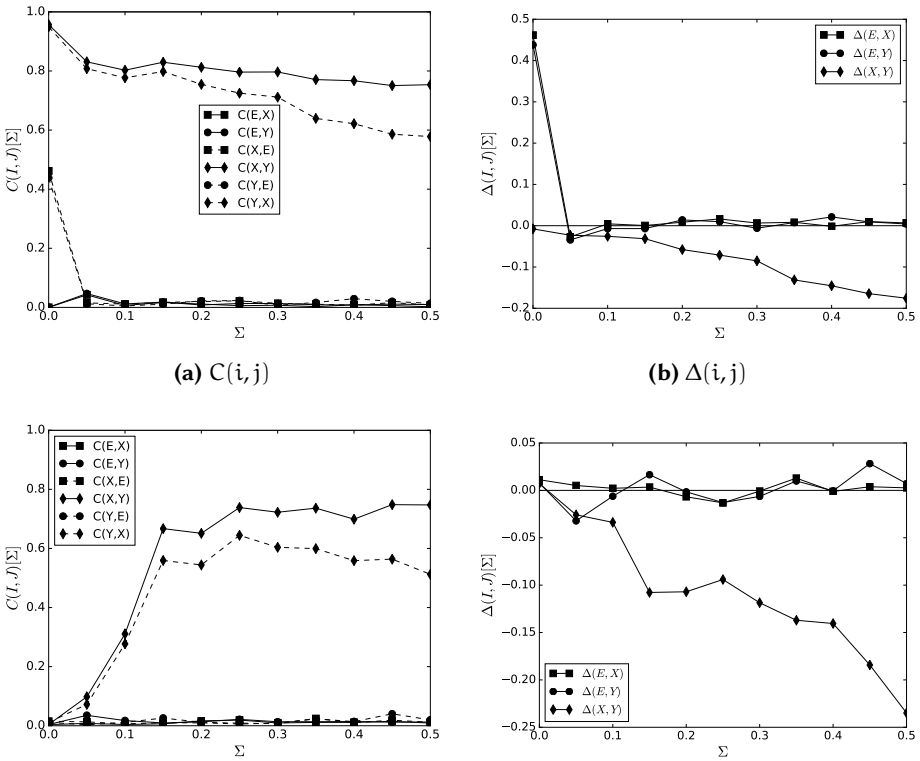
### EFFECT OF NOISE

The addition of model noise, most of the time, helps the detection of causality, even though it generally worsen the quality of prediction. In fact, thanks to the stochastic term, the solution deviates from $(X_\infty, Y_\infty)$, and the dynamics can be observed better. This is not the case for measurement noise, which does not modify the dynamics nor the attractor shape.

In all this section I will focus on the effect of both kind of noise: for demographic noise, I will vary $\Sigma = \sigma_x = \sigma_y = \sigma_E$ in the interval $[0, 0.5]$ with step of 0.05, while $\sigma_{Meas} \in \{0, 0.1\}$.

#### *Without environmental coupling ($\gamma = 0$)*

The result for this parameters setup are shown in Figure 2.2 on the facing page. When environment is completely decoupled from the dynamics of $X_t$ and $Y_t$, the importance of demographic noise is dominant. In fact, even a small quantity of noise makes us deviate from $(X_\infty, Y_\infty)$, allowing the possibility of prediction.




**(a)** $C(i,j)$ **(b)** $\Delta(i,j)$

**(c)** $C(i,j)$ **(d)** $\Delta(i,j)$

**Figure 2.2:** Results for a system uncoupled from environment and with a small carrying capacity ($K = 5$). In figures 2.2a and 2.2b the series have no measurement noise (which means $\sigma_{Meas} = 0$), while figures 2.2c and 2.2d refers to time series with $\sigma_{Meas} = 0.1$. All plots are represented for the variation over $\Sigma = \sigma_x = \sigma_y = \sigma_E$, which grows from 0 to 0.5.

Different markers identifies couples of time series (Diamonds indicate the couple $(X,Y)$, squares $(E,X)$ and circles $(E,Y)$), while solid and dashed lines identify the two possible directions of the coupling: the solid line corresponds to $i \xrightarrow{\text{Predict}} j$, the dashed one with the same markers corresponds to $j \xrightarrow{\text{Predict}} i$.

On the other hand, with no noise, because of the null dimensionality of the attractor space, the CCM fails.

From Figure 2.2a and 2.2c one can see the expected results: correlation between series $X$ and $Y$ are moderately high (even if they decrease slightly with increasing noise). As said before, the presence of a small demographic noise, when the prediction ability is severely hampered by measurement noise, help definitely to predict something.

Interestingly, in Figure 2.2a, the first point (which has $\Sigma = 0 = \sigma_{\text{Meas}}$) shows a big prediction skill, even if there should be almost no dynamics at all. The point is, from $t = 50\,s$ to approximately $t = 100\,s$, the time series are not yet completely fixed in $(X_\infty, Y_\infty)$ and they oscillate a little bit. In fact, all the prediction skill is based upon the first few points. This can be seen clearly if we make CCM run only with points $X(t), Y(t)$ with $t > 100\,s$: the method does not work[1].

The other interesting feature is that $\Delta(X, Y) \leqslant 0$, in contrast with what we expect from condition (2.3), and the difference even increase by increasing $\Sigma$ or $\sigma_{\text{Meas}}$. This shows that CCM method can gives prediction which are against intuition.

*With strong environmental coupling ($\gamma = 2$)*

When we couple the system with environment by increasing the value of $\gamma$, we find some interesting results, shown in Figure 2.3.

First of all, now $\Delta(E, X)$ and $\Delta(E, Y)$ are significantly bigger than zero, and also show a certain hierarchy that is similar to what expected (since $X$ is caused by $E$ only through $Y$, we expect this causal relation to be weaker, or similarly that $E$ and $X$ are worst in predicting each other).

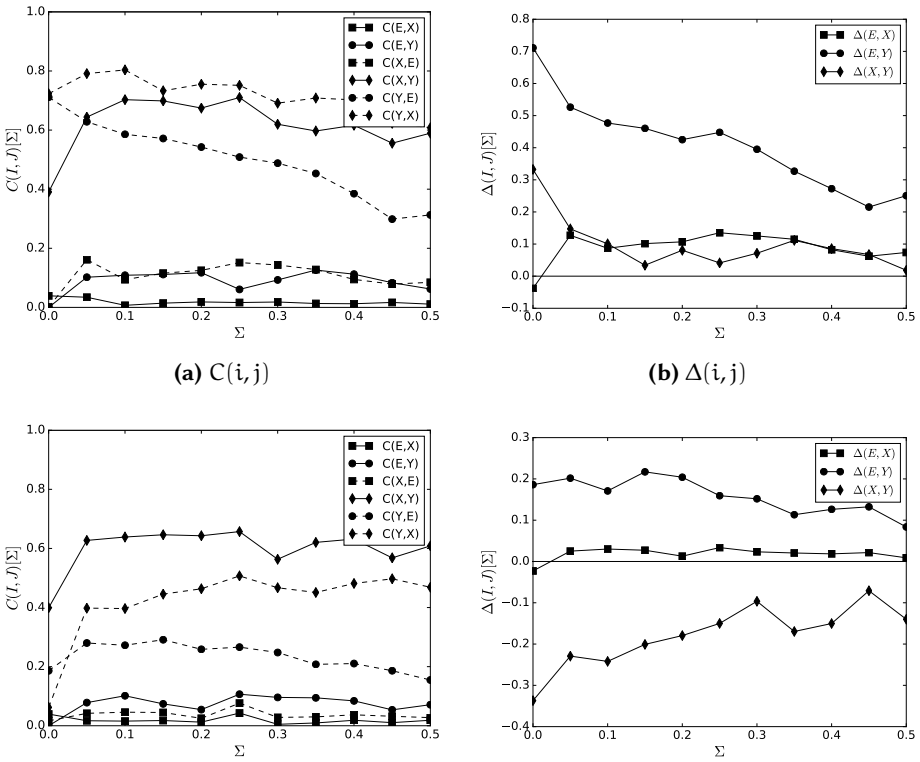
Second, now for $\sigma_{\text{Meas}} = 0$ we find $\Delta(X, Y) \geqslant 0$, which is more in accordance with our expectations as expressed in equation (2.3).

### EFFECT OF COUPLING WITH ENVIRONMENT

Let us consider now the effect of coupling with the external driver, variating $\gamma$ between 0 (no coupling at all) and 2 (quite big coupling), as shown in Figure 2.4.

The most interesting thing that appears from the plot of the correlation coefficient of figures 2.4a, 2.4c is that increasing the coupling strength leads to a degradation of the prediction of the causality relation between the environment and the system (for $\sigma_{\text{Meas}} \geqslant 0.1$. Even more interestingly, the sign of $\Delta(X, Y)$ changes (as clearly visible in figure 2.4b) while increasing $\gamma$.

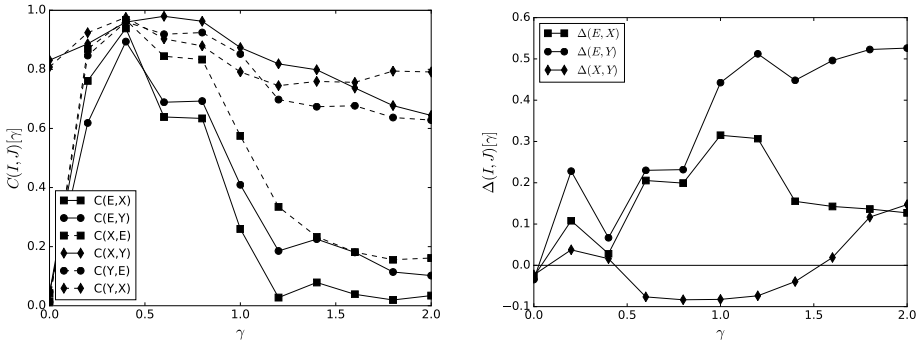
1 The problem does not come from CCM itself: predicting the evolution of something constant is, in fact, quite trivial. The point is that it is a meaningless problem and, moreover, PCC is not well defined for two constant process.



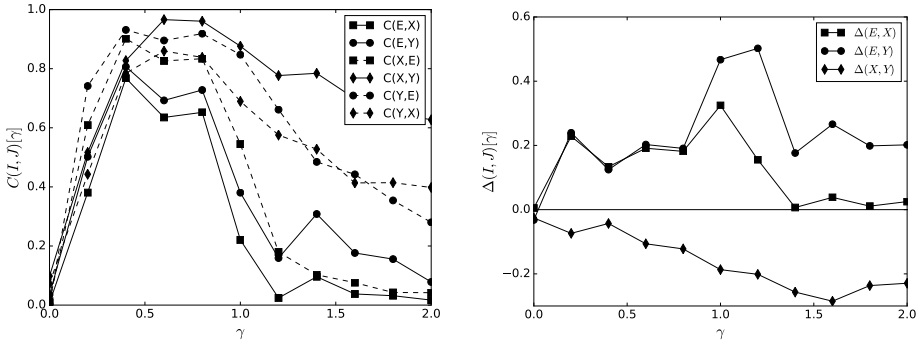

**(a)** $C(i,j)$ **(b)** $\Delta(i,j)$

**(c)** $C(i,j)$ **(d)** $\Delta(i,j)$

**Figure 2.3:** Results for a system strongly coupled with environment (coupling constant $\gamma = 2$) and with a small carrying capacity ($K = 5$). In figures 2.3a and 2.3b the series have no measurement noise (which means $\sigma_{\text{Meas}} = 0$), while figures 2.3c and 2.3d refers to time series with $\sigma_{\text{Meas}} = 0.1$. All plots are represented for the variation over $\Sigma = \sigma_x = \sigma_y = \sigma_E$, which grows from 0 to 0.5.
Different markers identify pair of time series (Diamonds indicate the couple $(X,Y)$, squares $(E,X)$ and circles $(E,Y)$), while solid and dashed lines identify the two possible directions of the coupling: the solid line corresponds to $i \xrightarrow{\text{Predict}} j$, the dashed one with the same markers corresponds to $j \xrightarrow{\text{Predict}} i$.




**(a)** $C(i,j)$ **(b)** $\Delta(i,j)$



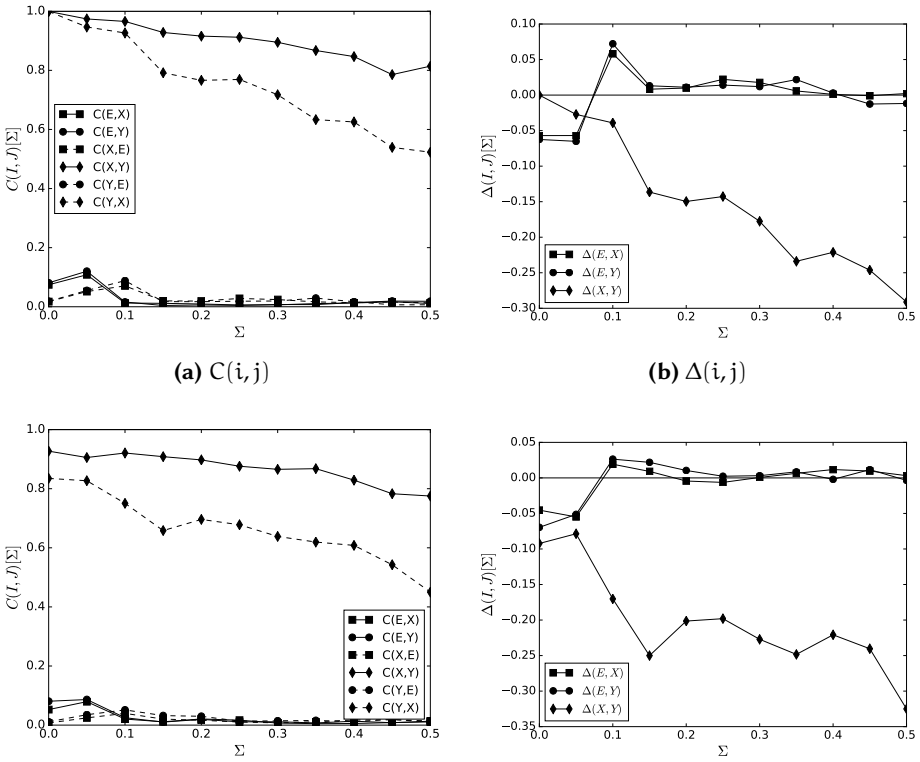

**(c)** $C(i,j)$ **(d)** $\Delta(i,j)$

**Figure 2.4:** Results for a system with a small quantity of noise ($\Sigma = 0.05$) and a small carrying capacity ($K = 5$). In figures 2.4a and 2.4b the series have no measurement noise (which means $\sigma_{Meas} = 0$), while figures 2.4c and 2.4d refers to time series with $\sigma_{Meas} = 0.1$. All plots are represented for the variation over $\gamma$, which grows from 0 to 2.
Different markers identify pair of time series (Diamonds indicate the couple $(X,Y)$, squares $(E,X)$ and circles $(E,Y)$), while solid and dashed lines identify the two possible directions of coupling: the solid line corresponds to $i \xrightarrow{\text{Predict}} j$, the dashed one with the same markers corresponds to $j \xrightarrow{\text{Predict}} i$.

## BIG CARRYING CAPACITY

By increasing the carrying capacity, as we have seen in Figure 2.1 on page 19, we change consistently dynamics to the system.

In fact, by taking $K \gg 1$ (e.g. $K = 500$) we see that the ratio between K and all the other parameters is of the order $10^2$, which means that the quadratic term in X can be almost neglected. This transform the system in a sort of classic LV, with all of its well known dynamics. For example, the slightly dumped oscillations can be considered a neutral limit cycle around a center equilibrium[2].




**(a)** $C(i,j)$ **(b)** $\Delta(i,j)$

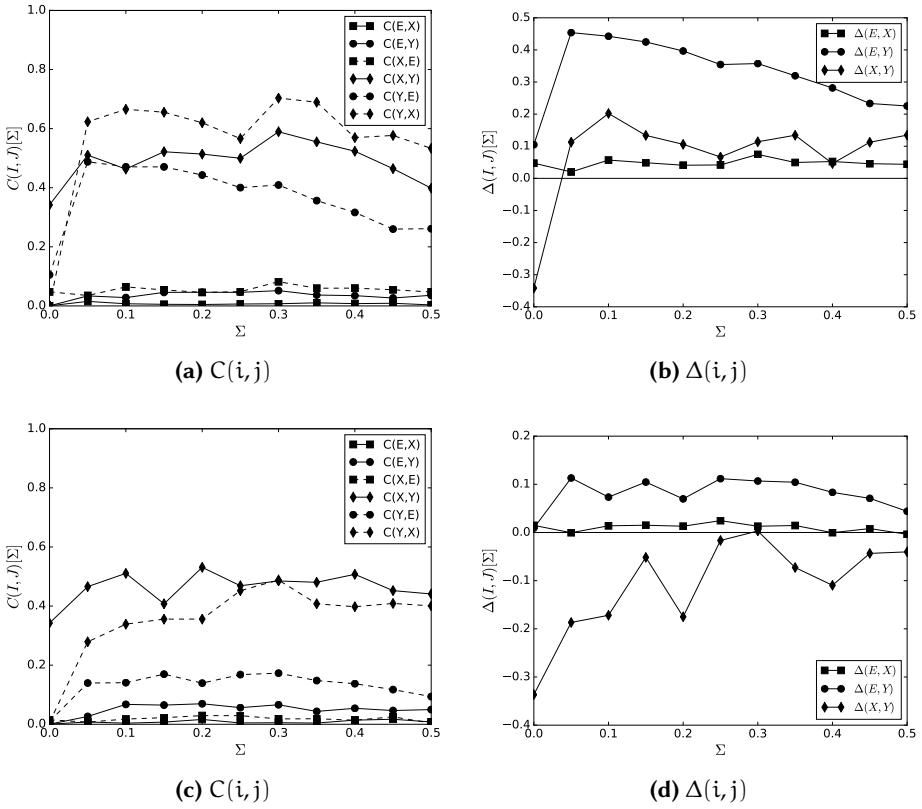
**(c)** $C(i,j)$ **(d)** $\Delta(i,j)$

**Figure 2.5:** Results for a system uncoupled from environment and with a big carrying capacity ($K = 500$). In figures 2.5a and 2.5b the series have no measurement noise (which means $\sigma_{Meas} = 0$), while figures 2.5c and 2.5d refer to time series with $\sigma_{Meas} = 0.1$. All plots are represented for the variation over $\Sigma = \sigma_x = \sigma_y = \sigma_E$, which grows from 0 to 0.5.
Different markers identify pair of time series (Diamonds indicate the couple $(X, Y)$, squares $(E, X)$ and circles $(E, Y)$), while solid and dashed lines identify the two possible directions of the coupling: the solid line corresponds to $i \xrightarrow{\text{Predict}} j$, the dashed one with the same markers corresponds to $j \xrightarrow{\text{Predict}} i$.

2 We call neutral limit cycle a cycle that is neither stable nor unstable, because it surrounded by other limit cycle for a big portion of the phase space. This is a behavior typical of LV equations [32].

### EFFECT OF NOISE

The presence of a finite (even if big) K allows us to add a stochastic noise to the system, making it different from a classic LV model in which the cycles are not attractively stable and a stochastic term makes the solutions diverge. Results in this case are reported in figures 2.5 and 2.6 for both cases of no and small coupling. Interestingly, they show very similar features with respect to the case $K \sim 1$, namely the fact that according to CCM the causality between X and Y changes direction as the coupling with environment increase.



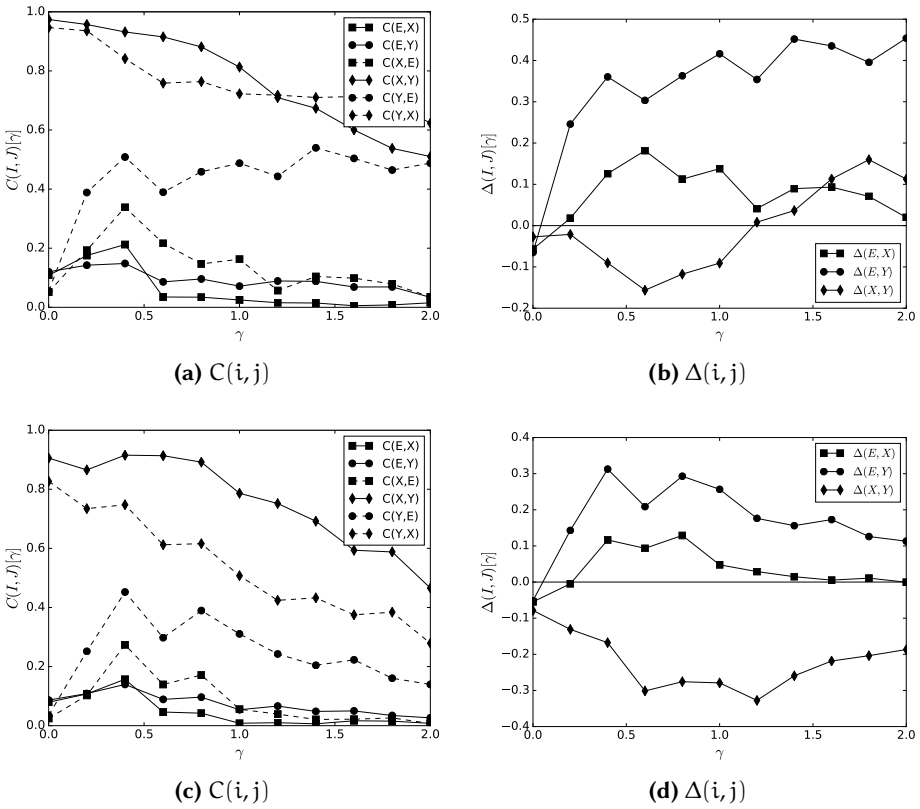

**(a)** $C(i,j)$ **(b)** $\Delta(i,j)$

**(c)** $C(i,j)$ **(d)** $\Delta(i,j)$

**Figure 2.6:** Results for a system strongly coupled with environment (coupling constant $\gamma = 2$) and with a big carrying capacity ($K = 500$). In figures 2.6a and 2.6b the series have no measurement noise (which means $\sigma_{Meas} = 0$), while figures 2.6c and 2.6d refers to time series with $\sigma_{Meas} = 0.1$. All plots are represented for the variation over $\Sigma = \sigma_x = \sigma_y = \sigma_E$, which grows from 0 to 0.1.
Different markers identify pair of series (Diamonds indicate the couple $(X,Y)$, squares $(E,X)$ and circles $(E,Y)$), while solid and dashed lines identify the two possible directions of the coupling: if a solid line corresponds to $i \xrightarrow{\text{Predict}} j$, the dashed one with the same markers corresponds to $j \xrightarrow{\text{Predict}} i$.

### EFFECT OF COUPLING WITH ENVIRONMENT

Again, when studying the variation of prediction skills with different couplings, we find that the sign of $\Delta(X,Y)$, and thus our identification of the causality




**(a)** $C(i,j)$ **(b)** $\Delta(i,j)$

**(c)** $C(i,j)$ **(d)** $\Delta(i,j)$

**Figure 2.7:** Results for a system with a small quantity of noise ($\Sigma = 0.05$) and a big carrying capacity ($K = 500$). In figures 2.4a and 2.4b the series have no measurement noise (which means $\sigma_{\text{Meas}} = 0$), while figures 2.4c and 2.4d refers to time series with $\sigma_{\text{Meas}} = 0.1$. All plots are represented for the variation over $\gamma$, which grows from 0 to 2.
Different markers identifies couples of series (Diamonds indicate the couple $(X, Y)$, squares $(E, X)$ and circles $(E, Y)$), while solid and dashed lines identify the two possible directions of the coupling: the solid line corresponds to $i \overset{\text{Predict}}{\longrightarrow} j$, the dashed one with the same markers corresponds to $j \overset{\text{Predict}}{\longrightarrow} i$.

direction, depends on the variation of $\gamma$. In fact, as can be seen by figure 2.7, when $\gamma$ changes from $\gamma \leqslant 1$ to $\gamma > 1$ the sign of $\Delta(X, Y)$ changes.

Again, the other interesting thing is that prediction ability decrease with the increasing of $\gamma$, while the relation $E \overset{\text{CCM}}{\longrightarrow} X$ almost cannot be detected.

## SUMMARY

The key of this example is that, by changing a parameter not directly linked to the coupling $X \leftrightarrow Y$, one obtains opposite prediction from CCM method. Then, one can be in a difficult position when trying to make detection of causality, depending on the exact dynamics of the system.

In fact, as we have seen, results obtained from CCM method applied to system (2.1) show some peculiarity that we already expressed in section 1.3: CCM effectiveness in detecting causal relationship does depend on the system parameters,

and thus on the type of the population dynamics of the system in the ecosystems and its coupling with the environment, which in principle should be the thing that one would want to determine with this method.

| | | $K = 5$ | $K = 500$ |
|---|---|---|---|
| $\gamma = 0$ | $\sigma_{\text{Meas}} = 0$ | $\Delta(X, Y) < 0$ | $\Delta(X, Y) < 0$ |
| | $\sigma_{\text{Meas}} = 0.1$ | $\Delta(X, Y) < 0$ | $\Delta(X, Y) < 0$ |
| $\gamma = 2$ | $\sigma_{\text{Meas}} = 0$ | $\Delta(X, Y) > 0$ | $\Delta(X, Y) > 0$ |
| | $\sigma_{\text{Meas}} = 0.1$ | $\Delta(X, Y) < 0$ | $\Delta(X, Y) < 0$ |

**Table 2.1:** Summary of results obtained within this section for $\Delta(X, Y)$. The underlined row shows that with a certain set of parameters the results obtained by CCM method are unreliable.

# 3

# THE METHOD OF ANALOGS AND THE CURSE OF DIMENSIONALITY: IS PREDICTION POSSIBLE?

In most cases and for many complex systems (finance, brain activity, earthquakes, etc...) we do not know the equations describing their evolution, or even if we know them in theory, they cannot be computed exactly even with the most powerful of the super computer (e.g. whether forecast). However, if the system dynamics is regulated by deterministic (even if unknown) laws, and if we know, through data, the past history of the system for enough time, can we predict its future evolution?

In this chapter we present the *method of analog*, first introduced by Lorenz in 1969, and that tries to answer to this fundamental question. Two main extensions of this method exist, and will be also presented. The concept on which the method is based is simple, and relies in the deterministic approach that *From the same antecedent follows the same consequent* [3]: it is thus sufficient to find a similar enough antecedent to make prediction about the consequent. This fundamental approach has been criticized already in XIX century by Maxwell stating [33]: *It is a metaphysical doctrine that from the same antecedents follow the same consequents. [...] But it is not of much use in a world like this, in which the same antecedents never again concur, and nothing ever happens twice. [...] The physical axiom which has a somewhat similar aspect is that from like antecedents follow like consequents.* In fact, Maxwell genius foresaw what now is well known from chaos theory: the ubiquitous presence of irregular evolutions due to deterministic chaos.

Nevertheless, we are now in the era of Big Data and there is a growing optimism that what could not have been achieved in the past, it is now possible and effective [34]. As Chris Anderson, former editor-in-chief at Wired magazine, stated in a provocative way: *This is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear. Out with every theory of human behaviour, from linguistics to sociology. Forget taxonomy, ontology, and psychology. Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves.* So, it is possible, if we have enough data, to effectively predict the dynamics of complex system? Can the

method of analogs be successfully applied on a different range of disciplines opening up new perspective in the predictive analytics?

Following a recent work of a group of Italian physicists [3] we will show that unfortunately the optimism must be limited: in fact there is an intrinsic limitation of the method that lies in the curse of dimensionality: if the system dimensionality is very large (and typically is), then the applicability of the method is spoiled, even using the largest data set available today.

## METHOD OF ANALOGS

The concept of the method of analogs, as said, is very straightforward in its intuitive formulation and can be expressed as *If a system behaves in a certain way, it will do so again*, in agreement with a strict deterministic view of the world.

In order to present the method of analogs, as introduced by Lorenz [35, 36] in its mathematical formulation, I will begin with a few notation.

Assume that $\mathbf{x}_t$ describes the time series of a particular state of a complex system (e.g. the expression activity of different genes at time $t$). Suppose that we have collected $N$ samples $\mathbf{x}_k$ with $k = 1, \dots, N$ and that we want to forecast what state the system will assume in the future, that is to say that we want to predict the value of $\mathbf{x}_{N+\tau}$.

The original idea is to search for a state among $\mathbf{x}_1, \dots, \mathbf{x}_{N-1}$, let us call it $\mathbf{x}_k$, that is the most similar to $\mathbf{x}_N$, and then use its consequents as proxies for the evolution of $\mathbf{x}_N$. Formally, calling $\epsilon$-analog the nearest point to $\mathbf{x}_N$ for which holds $\|\mathbf{x}_k - \mathbf{x}_N\| \leqslant \epsilon$, the prediction after $\tau$ time units is then:

$$\hat{\mathbf{x}}_{N+\tau} = \mathbf{x}_{k+\tau}.$$

The first, simple, generalization is the so called Center of Mass (COM) prediction, where we consider for our estimate all the $\epsilon$-analogs of $\mathbf{x}_N$, which means all the $n$ points $\mathbf{x}_{k_i}$, $i = 1, \dots n$, for which holds $\|\mathbf{x}_N - \mathbf{x}_{k_i}\| \leqslant \epsilon$. From this set of points, we determine the evolution of its center of mass, which is the weighted average of the analogs:

$$\hat{\mathbf{x}}_{N+\tau} = \sum_i \mathbf{W}_i \mathbf{x}_{k_i+\tau},$$

where $\mathbf{W}$ is a suitable weight matrix, that in the simplest version of the method is taken to be simply the $n$ dimensional identity matrix or a more refined expression such as the one in equation (1.10).

A simple sketch to illustrate this variant of the method is presented in Figure 3.1 on the facing page. Intuitively, from this picture, one can understand that the quality of prediction will depends strongly on $\epsilon$ and the number of $\epsilon$-analogs.

Crucially, an analytical approximation of the uncertainty of the prediction can be done.

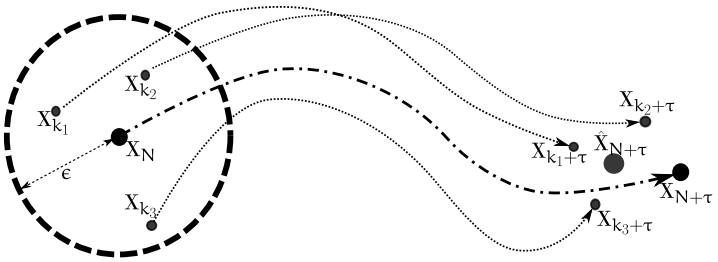



**Figure 3.1:** Sketch of the method of analogs, where $\epsilon$ is the neighborhood dimension, $\mathbf{X}_{k_i}$ is the $i$th $\epsilon$-analog and $\hat{\mathbf{X}}_{N+\tau}$ is the estimate of future point.

In order to calculate the accuracy of prediction $\|\hat{\mathbf{x}}_{N+\tau} - \mathbf{x}_{N+\tau}\|$, one has to estimate $\mathbf{x}_{N+\tau}$. Assuming that at least one good analog has been found, then the $\epsilon$-analog $\mathbf{x}_k$ can be assumed as a proxy of the present state $\mathbf{x}_N$, with an uncertainty $\delta_0$ set apart, i.e. $\mathbf{x}_k = \mathbf{x}_N + \delta_0$, where $\delta_0 \leqslant \epsilon$.

Then, we can express the expansion of this uncertainty over time with the Maximal Lyapunov Exponent (MLE)[1], using the same approach of the discussions about sensitivity to initial conditions typical of chaotic dynamical systems:

$$\delta_\tau \simeq \delta_0 e^{\lambda\tau}, \tag{3.1}$$

therefore, by defining $\Delta$ as the maximal tolerance about the error on prediction, we find that our estimated state will be $\Delta$-accurate up to a time:

$$\tilde{\tau}(\delta_0, \Delta) \approx \frac{1}{\lambda} \ln \frac{\Delta}{\delta_0}, \tag{3.2}$$

i.e. the time of prediction scales logarithmically with $\Delta/\delta_0$.

We thus have $\mathbf{x}_{N+\tau} = \mathbf{x}_{k+\tau} + \delta_\tau$ and substituting Eq. (3.1) the relative error of the prediction is approximated by

$$\frac{\|\hat{\mathbf{x}}_{N+\tau} - \mathbf{x}_{N+\tau}\|}{\hat{\mathbf{x}}_{N+\tau}} \approx \frac{\|\mathbf{x}_{k+\tau}(1 + \delta_0 e^{\lambda\tau})\|}{\mathbf{x}_{k+\tau}} \propto \delta_0 e^{\lambda\tau}, \tag{3.3}$$

and because the MLE is fixed by the dynamic of the system, the only way to improve the prediction is to minimize $\delta_0$.

The other (less natural) extension to Lorenz's version of the method of analogs is called Local Linear (LL) prediction. As with all the methods in the analog method framework, also for LL predictor one finds the $k$ $\epsilon$-analogs of the point $\mathbf{x}_N$. Then, instead of taking the (weighted) average of the successors of the analogs to estimate the successor of $\mathbf{x}_N$, the predictor is obtained by fitting a linear map $\mathcal{L}$ to the set of $\epsilon$-analogs $\mathbf{x}_i$ such that it minimize the difference between $\mathcal{L}(\mathbf{x}_i)$ and $\mathbf{x}_{i+\tau}$, and then applying this map to the point $\mathbf{x}_N$.

1 We recall that the Lyapunov exponents describe the behavior of vectors in the tangent space of the phase space, and thus characterize the rate of separation of two points infinitesimally close in the phase space. The importance of the MLE follows directly from the fact that this rate is exponential $\delta_t \sim e^{\lambda t}$ and thus (if it is positive) the relevance of the bigger exponent will obliterate all the other. For a more accurate introduction to Lyapunov exponents see [37].

Even if this procedure is more refined than the COM approach, it is less intuitive and presents the same problem (namely the search for nearest neighbors) of the simpler methods, and as such will not be discussed further.

## TAKENS' EMBEDDING THEOREM AND ITS APPLICATION TO THE METHOD OF ANALOGS

In the above presented explanation of the method of analogs, we have considered $\mathbf{x}$ as a faithful representation of a given state of the system, and as such it is considered to be a point of the d-dimensional phase space ($\mathbf{x} \in \mathbb{R}^d$) measured with arbitrary precision.

Obviously, this is far from the reality of experiments, where measurements have some noise determined by the experimental setup. Moreover, we can often measure only one or a few scalar variables linked to the real state of the system through some unknown projection function. In other words we actually do not know the phase space of the system.

Nevertheless, we can use Takens' embedding theorem 1.5 on page 9 or its extensions[2] in order to reconstruct the phase space and then apply the method of analog on this reconstructed space to make predictions.

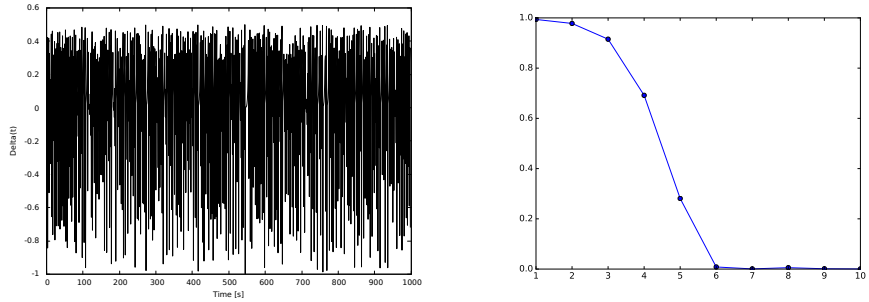
In Figure 3.2 an example of this approach is shown, applied to a first-difference time series obtained from the *tent map* [40]:

$$x_{t+1} = \begin{cases} \mu x_t & x_t < \frac{1}{2} \\ \mu(1 - x_t) & x_t \geqslant \frac{1}{2}. \end{cases} \tag{3.4}$$

In particular, we have studied the case $\mu = 2$, which is interesting because the generated time series has an autocorrelation function that goes to zero so rapidly that the series is indistinguishable (with this approach) from a pure white random sequence. In this example, the embedding dimension used is $E = 3$, which is more than enough to display the complete dynamics of the attractor (which has an attractor dimension $0.97 \pm 0.03$ [41], determined with the Grassberger approach, see section 3.3), and a time lag $\tau = 1$. As Figure 3.2b shows, the prediction ability steeply decreases after few time steps in the future, a symptom of the chaotic nonlinear dynamics (and the Lyapunov divergence of orbits, equation (3.1)).

The reconstruction of the phase space is one of the main procedure used when dealing with chaotic dynamical series, also because often just a simple visual inspection of the data in the embedding space allows some understanding of the system. This approach, though, does not overcome nor mitigate the problems that will be examined in the later sections.

2 For example, if the attractor has a fractal dimension, Takens' theorem cannot be applied but its extension by Sauer *et Al.* [38] called *Fractal Delay Embedding Prevalence Theorem* still holds. Other notable extensions are proved in [26, 39] and generalize the theorem to some stochastic equations.




**(a)** $\Delta_t$ **(b)** $\mathrm{Pcc}(\Delta_t, \hat{\Delta}_t)$ vs $T_P$

**Figure 3.2:** In figure 3.2a are shown 1000 points of a time series obtained as $\Delta_t = x_{t+1} - x_t$ where $x_t$ is generated from the tent map (3.4). The first 500 points of the series are used to construct the training attractor (with $E = 3$ and $\tau = 1$), and then prediction $T_p$ step in the future are made for each of the next 500 points (excluding the last $T_p$ points for which we don't have a "measured" version $T_p$ step in the future). In figure 3.2b is shown the PCC between the predicted points and the real ones as a function of the number of step in the future we want to predict.

### DELAY RECONSTRUCTION AND THE PROBLEM OF FINDING A GOOD EMBEDDING

In this section, I will explain some details of Takens approach to the method of analogs following [42, § 3, 9].

Suppose that the only measure that we have access to is a scalar which is a function of the (unknown) state vector $\mathbf{x}$:

$$u_n = h(\mathbf{x}(n\,\delta t)) + \eta_n$$

with $h\colon M \to \mathbb{R}$ some unknown scalar projection and $\eta_n$ some white random measurement noise. We can then define the vector:

$$\mathbf{u}_n = \left(u_n, u_{n-\tau}, u_{n-2\tau}, \ldots, u_{n-(m-1)\tau}\right) \tag{3.5}$$

where $\tau$ is the *lag* and $m$ is the *embedding dimension*. In order to fulfill the delay embedding theorem requirements, the choice of $m$ must be done in accordance to a strict criterion, while the choice of $\tau$ is more free, even if some particular values are better than others.

#### *Choice of the embedding dimension*

The choice of $m$ is fixed by Takens' theorem and its extensions, and must be such that:

$$m > 2\,D_M$$

where $D_M$ is the dimension of the subset $M$, called attractor, of the (unknown) phase space in which the orbits of the system are bounded. If $D_M$ is non integer, $M$ is said to be a fractal attractor and this condition says that $m$ must

be strictly larger than twice $D_M$, while if $D_M$ is an integer we get back the familiar Takens' claim that $m = 2D_M + 1$.

The problem is that most of the time we have no idea of the attractor dimension, and we may even do not know if an attractor exist at all (that is to say, we don't know if the system is regulated by a (non)linear deterministic dynamics or if it is driven solely by a stochastic process).

Of course, if we choose a very high $\tilde{m}$, bigger than the true $m$, we will obtain a correct embedding anyway, so one can be tempted to just set a very high $\tilde{m}$ and use that to make prediction. This approach, thus, leads to two non trivial problems: one is linked to the number of analog within radius $\epsilon$ (and will be addressed in the next section), while the other is more theoretical and linked to the MLE: the larger is $m$, the more far in the past will be the last component of $\mathbf{u}$, which will be at a lag $m\tau$. This implies that our algorithm for prediction is processing (with the same weight) information that we know are almost unrelated with the point of interest.

One of the most useful method to choose $m$ is the so called method of *False nearest neighbours* [43].

Let $\tilde{m}$ be the real but unknown embedding dimension, and let $m < \tilde{m}$ be the dimension in which, using the available data, we perform the embedding instead. The map that transform points from $\tilde{m}$ to $m$ is a projection in which some of the axes are eliminated (precisely $\tilde{m} - m$ axes are eliminated). Thus, points that are separated with a large distance along one axes that is deleted by the projection appears, in the $m$-dimensional space, as neighbors even if they are not so (in this sense the method is called of false neighbors).

If we call

$$d_i^m = \left\| \mathbf{u}_i^{(m)} - \mathbf{u}_{k(i)}^{(m)} \right\|_{\text{Max}}$$

the distance (in $m$ dimensions) between $\mathbf{u}_i$ and its nearest neighbors $k(i)$ within the maximum norm, we can express the fraction of false neighbors going from dimension $m$ to $m+1$ as:

$$\chi_{\text{fnn}}^{(m)}(r) = \frac{1}{N} \sum_{i=1}^{N} \Theta\left( \frac{d_i^{m+1}}{d_i^m} - r \right), \tag{3.6}$$

where $\Theta(\bullet)$ is the Heaviside step function that counts the couple for which the ratio between their distance in $m+1$ and $m$ dimension is bigger than a certain threshold $r$.

A visual explanation of this approach is represented in Figure **??**.

A correction to this method, proposed by Hegger and Kantz [44], to account for noise in the data (which keep the number of false neighbours steady because the embedding dimension of a random process is infinite) is to consider in Equation 3.6 only those points having a distance in $m$ dimension

$$d_i^m < \frac{\sigma}{r},$$

where we have defined $\sigma$ to be the standard deviation of data. This correction takes into account the simple but important fact that two points cannot be false

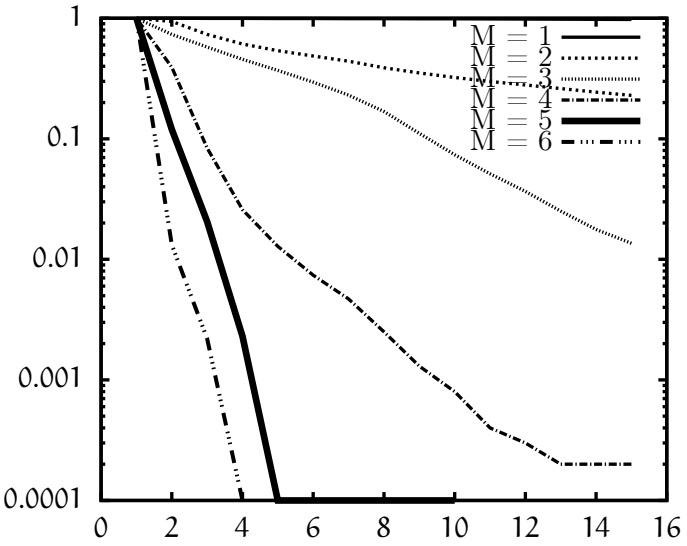



**Figure 3.3:** Method of the False Nearest Neighbors for the Lorenz attractor. Each line represents a different embedding dimension M. As can be seen, passing from dimension $M = 4$ to dimension $M = 5$ there is a sharp increase in the steepness of the curve describing the number of false neighbors as a function of the ratio of the distance in the two dimensions, suggesting that $M = 5$ is the correct embedding dimension. This is in agreement with the known correlation dimension of Lorenz attractor, which is slightly above two ($D_A \simeq 2.04$), leading to an embedding dimension $\tilde{M} > 2D_A \simeq 4$.

neighbors if already in $m$ dimension their distance is larger than the standard deviation times $1/r$ (on average they cannot be at a distance larger than $\sigma$).

*Choice of time lag*

The other free parameter in the delay embedding approach is the time lag $\tau$, which presents a further difficulty: in fact, even though theoretically this parameters does not have an importance in the phase space reconstruction (it does not appear in the formulations of Takens' theorems nor in its extension), from a practical point of view choosing the right $\tau$ has some important consequences in the ability to resolve the attractor and then to make effective prediction.

In fact, if $\tau$ is too short with respect to the characteristic time of the system dynamics (such as the main period, if it exists), successive coordinates of the delay vectors are strongly correlated, and all the points $\mathbf{u}_i$ tend to places themselves around the diagonal of $\mathbb{R}^m$.

Instead, if $\tau$ is chosen to be too large, the $\mathbf{u}_i$ points are almost completely uncorrelated and they fills a very large cloud of $\mathbb{R}^m$, becoming non informative.

There are two main statistical approaches to this problem, based on the autocorrelation function and on the mutual information, respectively.

The first one is a minimalist approach, and it is based upon the idea that to maximize the amount of information by extracting pair of points, they have to be uncorrelated. Then, the best choice $\tau_0$ would be the delay that gives the

first zero in the autocorrelation function[3] $R(\tau_0) = 0$ (the first zero because, albeit many roots of $R(\tau)$ exists, obviously as the distance in time increases the correlation between two points decrease, but they are also less informative). The problem of this approach is that the autocorrelation test only is reliable for *linear* dependence, and as such it is not best suited for strongly non-linear system.

The second approach [45], based on mutual information, consists in choosing the time delay that corresponds to the first minimum of the mutual information (1.2) between $u(t)$ and $u(t-\tau)$. This assure us that the information shared between $u(t)$ and $u(t-\tau)$ is minimal and thus the information gain of considering both of them is the biggest.

Obviously, both the autocorrelation function and the mutual information will tend to zero with $\tau \to \infty$, and the interesting zero (minimum) will be the first one.

The two approach are exemplified in Figure 3.4.

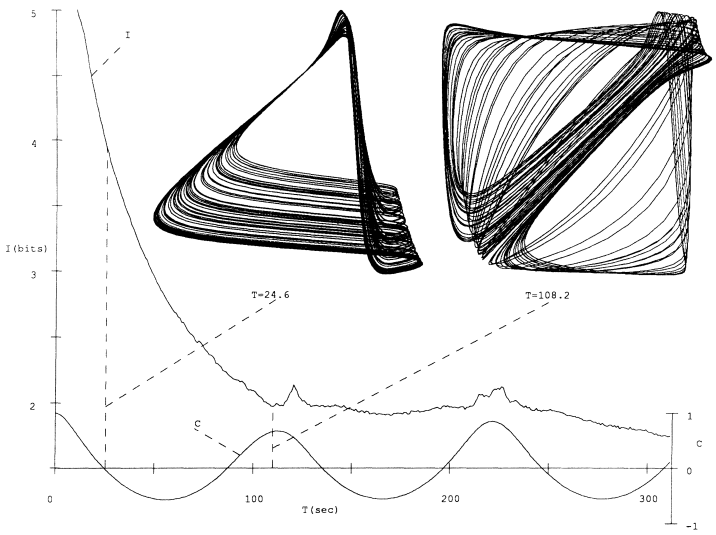



**Figure 3.4:** This figure is taken from [45]. It represent two different reconstruction of the Roux attractor, one using the autocorrelation function (labeled with c) criterion and the other one with the mutual information (labeled with I)criterion. The two chosen time lags are shown in a scale that corresponds to the caratheristic time of the system, which is the time of a *quasi*-periodic orbit.

A RULE OF THUMB to find out the right time lag if the signal has a strongly (quasi-)periodic component, is to choose $\tau$ to be between one half and one tenth of the period, often around one quarter. Infact, if the system has an embedding dimension between two and ten, this allows the spanning of most of the orbit with each point in the embedded space. Unfortunately, if there is not a strong

3 with autocorrelation function we mean the function

$$R(\tau) = \sum_t X(t) * X(t-\tau)$$

which is commonly used in signal processing.

periodic component or if the system is very large, this reasoning cannot be applied.

## THE CURSE OF DIMENSIONALITY: STATISTICAL DIFFICULTIES FOR GOOD ANALOGS AND KAC'S LEMMA

The accuracy estimation of the prediction given by (3.3) highlights how analogs that are distant $\delta_0$ from the state $\mathbf{x}_N$ will give exponentially distant far prediction from the "real" one as a function of the future time. This result reminds to the typical problem of sensitivity to initial conditions in dynamical systems with chaotic behaviour: the same dynamics with initial condition that are different of a little amount $\delta_0$ will give exponentially (with time) far trajectories [37].

However, if we look deeply in Eq. (3.3) things are different in the two cases. In fact, as shown in the work of Cecconi *et Al.* [3], the main issue in Eqs. (3.3) and (3.2) is finding good analogs, which means finding points $\epsilon$-near to the current state with a sufficiently small $\epsilon$ that allows $\tilde{\tau}$ to be large enough so that we can actual make future predictions within a maximal tolerance in the error of the prediction ($\Delta$).

However, as we will see in the next section, because of the curse of dimensionality the amount of data needed to obtain a good analogs increases exponentially with the system size.

### DIMENSIONALITY AND DEGREES OF FREEDOM

The theory of ergodic dynamical systems is founded upon the principle that long-term statistical properties of a system can be described with the time-independent probability distribution $\mu(\sigma)$ of finding the system in any specific region $\sigma$ of the phase space which, for a system with $d$ Degrees of Freedom (DoF), is a region $\sigma \subset \mathbb{R}^d$ of the phase space.

If the evolution is conservative (as in Hamiltonian mechanics), it conserves volumes in the phase space, then the probability $d\mu(\mathbf{x})$ of finding the system in a small region $dV$ around $\mathbf{x}$ is proportional to the measure of $dV$. Instead, in dissipative systems, volumes in the phase space are contracted (on average) until probability $d\mu(\mathbf{x})$ is concentrated in a subset $M \subset \mathbb{R}^d$ of dimension $D_M$ called attractor. $D_M$ is thus what we call the system effective dimension, *i. e.* the dimension of the subspace of the phase space that is really spanned by the system.

Formally, the dimension $D_M$ describes the small-scale behavior of $\mu_\epsilon$ of finding points $\mathbf{x} \in M$ inside the $d$-dimensional sphere $B^d_{\mathbf{y}}(\epsilon)$ of radius $\epsilon$ and centered around $\mathbf{y}$:

$$\mu\left(B^d_{\mathbf{y}}(\epsilon)\right) = \int_{B^d_{\mathbf{y}}(\epsilon)} d\mu(\mathbf{x}) \sim \epsilon^{D_M} \tag{3.7}$$

Thus, in dissipative systems, trajectories of the system are effectively described by a number $D_M < d$ of DoF, even though they are defined in a d-dimensional space. If $D_M$ is non integer, the attractor is said to be *fractal*, and this condition is typical for chaotic systems (consider for example [46])[4].

In order to obtain an estimate of $D_M$, we follow the approach of Grassberger and Procaccia [4], first defining the correlation sum as a function of the distance $\epsilon$:

$$C(\epsilon) \equiv \lim_{N\to\infty} \frac{1}{N^2} \sum_{i,j=1}^{N} \Theta\left(\epsilon - \left\|\mathbf{x}_i - \mathbf{x}_j\right\|\right). \tag{3.8}$$

This definition can be rewritten in term of the standard correlation function

$$c(\mathbf{r}) = \sum_{i,j}^{N} \left[\langle \mathbf{x}_i(0)\mathbf{x}_j(\mathbf{r})\rangle - \langle \mathbf{x}_i(0)\rangle \langle \mathbf{x}_j(\mathbf{r})\rangle\right]$$

as

$$C(\epsilon) = \int_0^{\epsilon} d\mathbf{r}\; c(\mathbf{r}).$$

The key point proved by Grassberger and Procaccia is that this function behaves as a power law of $\epsilon$ with an exponent $\nu$ when $\epsilon$ is sufficiently small:

$$C(\epsilon) \sim \epsilon^{\nu} \tag{3.9}$$

and that this exponent is an estimate of the system dimensionality, $\nu \sim D_M$. In figure 3.5 is represented the estimation of $\nu$ from the Correlation sum (3.8) using a simple counting algorithm. It is important to observe that the power law scaling holds only for small $\epsilon$, which will have consequences in the following discussion. In order to understand intuitively the Grassberger-Procaccia claim, we see that the definition of correlation sum (3.8) can be thought as the fraction of $\epsilon$-analogs of a certain point $\mathbf{x}_{\tilde{k}}$ averaged over all the points $\mathbf{x}_k$ of the time series.

Now we can see the analogy with ergodic theory: because (3.8) is an estimate of the averaged probability of finding the system in a sphere of radius $\epsilon$, by exploiting relations (3.7) and (3.9) we can derive:

$$\epsilon^{\nu} \sim C(\epsilon) \approx \langle \mu_\epsilon \rangle \sim \epsilon^{D_M} \qquad \Longrightarrow \qquad \nu \approx D_M \approx \mathcal{D}.$$

In the following, I will use $\mathcal{D}$ as a unified symbol for both $\nu$ and $D_M$, neglecting the small difference between the two.

The Grassberger-Procaccia exponent $\nu$, which is called *correlation dimension*, is computationally (at least in principle) easy to compute: it is sufficient to calculate the function in the limit (3.8) for some value of $\epsilon$ and then fit the power law (3.9).

Before explaining the reason of why this happens only in principle, I will introduce the relation between correlation dimension and recurrence time (and Poincaré recurrence theorem).

4 In general, attractor are inhomogeneous in $D_M$, and to different $\mathbf{Y}$ corresponds different $D_M$. In this thesis we will ignore this technicality and study only attractor with a single fractal dimension $D_M$. For a more detailed discussion, that goes beyond the scope of this work, about multifractal attractor see [47]

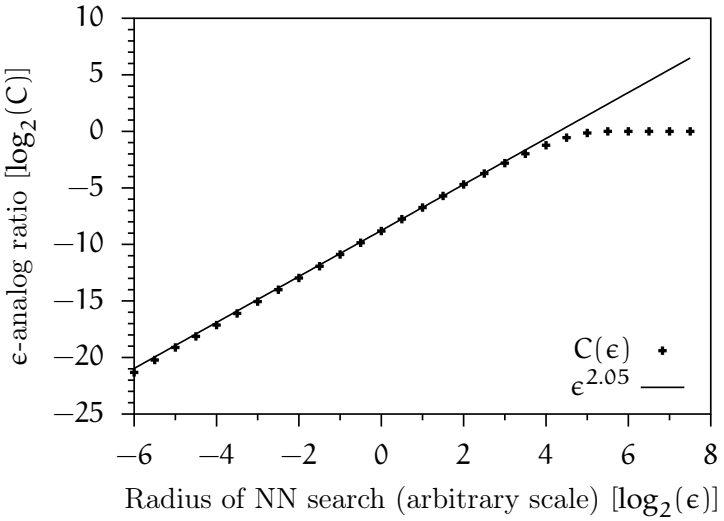



**Figure 3.5:** Estimate of the correlation dimension $\nu$ from the correlation sum for the Lorenz attractor.

### RELATION BETWEEN CORRELATION DIMENSION AND POINCARÉ RECURRENCE THEOREM

We recall the classical result of Poincaré recurrence theorem [48], phrased in a mathematical language that applies to our problem:

3.1 THEOREM (POINCARÉ RECURRENCE THEOREM).
*Lets call $\Phi_t$ the flow of a volume-preserving differential system with bounded orbit.* *Poincaré recurrence theorem*

*Then, for each open set $\mathfrak{U}$ of the phase space almost all the orbits $\Phi_t(x)$ with $x \in \mathfrak{U}$, intersect $\mathfrak{U}$ infinitely many times, which means that there exists infinitely many $\tau$ such that $\Phi_\tau(x) \in \mathfrak{U}$. Thus, the set of points $x \in \mathfrak{U}$ for which this property does not hold has zero measure.*

Even if the theorem stresses that $\Phi_t$ must be volume-preserving, one can extend this result even to dissipative systems, provided that $\mathfrak{U} \subset M$, which means that the only trajectories that return infinitely many times are the ones that start from the attractor.

Moreover, another straight extension is the one to ergodic systems, considering the measure on the bounded phase space as the invariant probability $\mu(\sigma)$, then the points that does not return have zero probability. In this framework the *recurrence time* $\tau_\sigma(\mathbf{x}_0)$ can be defined as the time that the trajectory that flows from a point $\mathbf{x}_0$ needs in order to return inside the set $\sigma$ that contains $\mathbf{x}_0$, which can be written as:

$$\tau_\sigma(\mathbf{x}_0) = \inf_k \{k \geqslant 1 \mid \mathbf{x}_k \in \sigma\}\, \delta t. \tag{3.10}$$

Averaging this time over all point in $\sigma$ one obtains the average recurrence time for points in $\sigma$:

$$\langle \tau_\sigma \rangle = \frac{1}{\mu(\sigma)} \int_\sigma d\mu(\mathbf{x})\ \tau_\sigma(\mathbf{x}) \tag{3.11}$$

Now, we recall the classical Kac's result [49] that states:

3.2 THEOREM (KAC'S LEMMA).

*Kac's lemma*

*For an ergodic system, given $\sigma \subset \mathcal{M}$ a non empty subset on the attractor $\mathcal{M}$, $\mu(\sigma)$ an invariant probability measure on $\mathcal{M}$ and $\tau_\sigma(\mathbf{x})$ defined as in equation (3.10), then:*

$$\int_\sigma d\mu(\mathbf{x})\ \tau_\sigma(\mathbf{x}) = 1,$$

*which means that, recalling equation (3.11), the average recurrence time is:*

$$\langle \tau_\sigma \rangle = \frac{1}{\mu(\sigma)}.$$

*In other words, the average recurrence time to a region $\sigma$ of the attractor is just the inverse of the probability of finding the system in a state inside $\sigma$.*

This result can be seen as one of the foundation principle of statistical mechanics [3]: in fact, for a systems in $\mathbb{R}^d$ (and thus with $d$ DoF) and with an accessible volume $L^d$ (which is the mean excursion of each component of $\mathbf{x}$), if the phase space volumes is preserved and we consider $\sigma$ as the hypercube of linear dimension $\epsilon$,then

$$\mu(\sigma) \sim \left(\frac{\epsilon}{L}\right)^d \qquad \Longrightarrow \qquad \langle \tau_\sigma \rangle \sim \left(\frac{L}{\epsilon}\right)^d.$$

This means that the recurrence time grows exponentially with $d$, which means that for system of macroscopic size (where $d$ is typical of the order $\sim 10^{23}$, for example, but this holds even in much smaller system of around $10^3$ particles) $\langle \tau_\sigma \rangle$ becomes enormous for any $\sigma$. This is the classical concept of *irreversibility* as stated by Boltzmann himself (even without knowing Kac's lemma) [50].

However this results also means that the chance of finding good analogs decrease exponentially with the dimension of the system.

### CONSEQUENCES ON THE METHOD OF ANALOGS

Let's consider a time series $\mathbf{x}_1, \dots \mathbf{x}_k$ of points sampled at a time interval $\delta t$. We call $\mathcal{K}(\epsilon)$ the number of $\epsilon$-analog of the last point $\mathbf{x}_k$ that have been found in this series. Then, the average time between two such $\epsilon$-analog of $\mathbf{x}_k$ (which gives us an estimate of the time needed to return $\epsilon$-close to $\mathbf{x}_k$) is:

$$\bar{\tau}_k = \frac{(k-1)\delta t}{\mathcal{K}(\epsilon)}. \tag{3.12}$$

Consider the correlation sum defined in equation (3.8), then we can implicitly define $C_k(\epsilon)$

$$C(\epsilon) = \frac{1}{N}\sum_k^N C_k(\epsilon)$$

that represents the fraction of $\epsilon$-analog of the kth point of the series over the total data points N. Clearly, given this definition and equation (3.12)

$$C_k(\epsilon) = \frac{\mathcal{K}(\epsilon)}{k-1} = \frac{\delta t}{\overline{\tau}_k}. \tag{3.13}$$

In order to make a $\Delta$-accurate estimate in predicting the last point $\mathbf{x}_N$ through the past trajectoriy, we need at least $N \geqslant \tau_N$. Recalling the scaling law (3.9) and using the relation (3.13), we prove that the number of points needed scales as:

$$N \sim \epsilon^{-\mathcal{D}}, \tag{3.14}$$

which is inversely proportional to the accuracy $\epsilon$ (as expected) but also exponential in $\mathcal{D}$. This results highlights how for large systems (where the attractor dimension can larger than 10) prediction becomes almost impossible at a reasonable $\epsilon$.

Moreover, it states the precise limit of the Grassberger-Procaccia procedure to estimate the correlation dimension, given that the larger the correlation dimension the larger is the number of points that are needed to sample the attractor with a given accuracy $\epsilon$, also considering that the scaling (3.9) holds in the limit $\epsilon \to 0$.

Of course one of the main challenge is to estimate both $\mathcal{D}$ and $\epsilon$ for real system, when only a set of time series are known.

For example, in one of his work, Smith [51] proposed an estimae of $\epsilon$ in order to calculate the effective dimension $\mathcal{D}$ using the Grassberger-Procaccia algorithm. In particular, he found that the minimum number of point that needs to be measured from the attractor (following the scaling law in equation (3.14)) is

$$N \sim 42^{\mathcal{D}}.$$

This means that for a $\mathcal{D} \simeq 5$, N should be already of the order $\sim 10^8$ to obtain a consistent estimate of the system relevant parameters.

# 4

# THE CURSE OF DIMENSIONALITY, A NUMERICAL EXAMPLE

In this chapter I will provide some numerical example of the considerations made in chapter 3 on page 29, studying a simple system (the so called Lorenz attractor [23]).

## DESCRIPTION OF THE SYSTEM: LORENZ ATTRACTOR IN 3 DIMENSION

The first example is the classical three dimensional Lorenz attractor, that is described by

$$\begin{cases} \dot{X}_t = \sigma (Y_t - X_t) \\ \dot{Y}_t = X_t (\rho - Z_t) - Y_t \\ \dot{Z}_t = X_t Y_t - \beta Z_t. \end{cases} \tag{4.1}$$

with $\sigma = 10$, $\beta = 8/3$ and $\rho = 28$.

In order to simulate this system I have implemented the same second order Runge-Kutta method as in chapter 2 equation (2.2), using a time step $\delta t = 0.01$, which I then downsampled of a factor 50 in order to get a sampling frequency of 2 Hz.

For each time series, I have runned the algorithm ten times choosing random initial condition inside the attractor basin, then averaging over the results in order to avoid statistical coincidence.

## GRASSBERGER-PROCACCIA ALGORITHM FOR THE CORRELATION DIMENSION

The first algorithm that I have studied is the Grassberger-Procaccia algorithm to estimate the correlation dimension, which has been obtained by Grassberger and Procaccia using a time series of 15 000 points [4] and is equal:

$$\mathcal{D} = 2.05 \pm 0.01. \tag{4.2}$$

In Figure 4.1 the correlation dimension is estimated using time series of increasing length. It can be seen that the methods produces wrong results if the number of data is to small.

In fact, as shown in the previous chapter, the minimum number of points to get a reliable estimation of $\mathcal{D}$ is, according to Smith's work [51],

$$N_{min} = 42^{\mathcal{D}} \simeq 2125 \quad \text{points.} \tag{4.3}$$

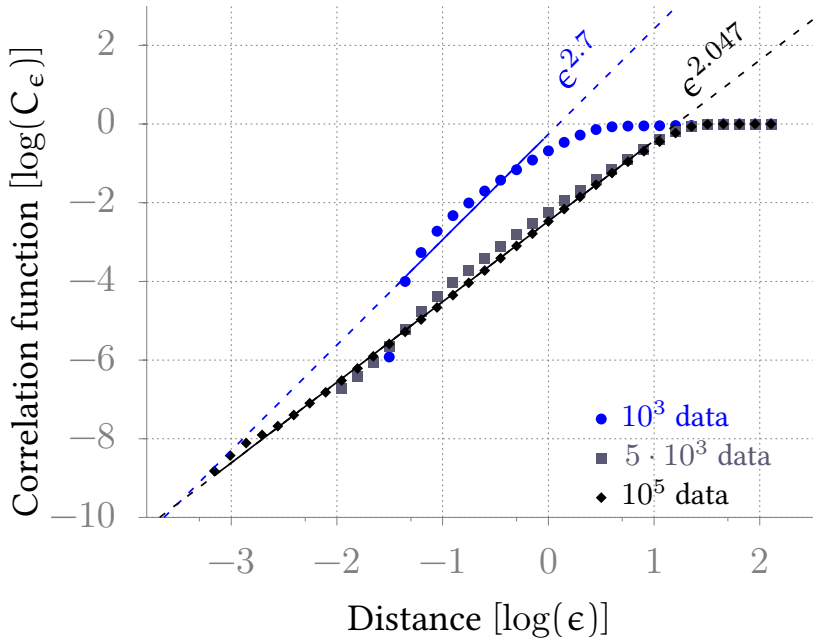



**Figure 4.1:** Correlation function as calculated by the Grassberger-Procaccia algorithm [4], represented in a log-log plot. The different colors corresponds to different length of the input time series, and the lines represent the best fit obtained for the time series with one thousand points and the one with one hundred thousands points. The straight lines represents the best fitting power law $C = a\epsilon^{\mathcal{D}}$ to two series of data (the shortest and the longest one). It can be noted that the points obtained with the series of 5000 points follows almost exactly the line with $\mathcal{D} = 2.05$.
The solid range of the lines represents the range of data used for the fit.

The best result for the correlation dimension obtained fitting a power law $C = a\epsilon^{\mathcal{D}}$ is (using the time series with 100 000 points)

$$\hat{\mathcal{D}}_{10^5} = 2.047 \pm 0.002 \tag{4.4}$$

which is perfectly compatible with the values determined by Grassberger and Procaccia, in equation (4.2).

However, also calculating the fit with 5000 points yields a result compatible with the expected one:

$$\hat{\mathcal{D}}_{5\cdot10^3} = 2.03 \pm 0.02,$$

while the fit with 1000 points gives a wrong result as $\hat{\mathcal{D}}_{10^3} = 2.67 \pm 0.1$.

Obviously, in order to make the fit, points with $\epsilon$ to large have been discarded, because the power law holds only in the limit $\epsilon \to 0$. A clear example is given by figure 4.1, in which only the points in range $[10^{-3}, 10^{1}]$ have been used to fit the powelaw with $10^5$ data points, and an even smaller range for the other series.

## CENTER OF MASS PREDICTION ALGORITHM

Then I have implemented a prediction algorithm based upon the method of analog (specifically, I used a COM algorithm using weights as given in equation (1.10)). In this case, the minimum number of points required to apply the method is estimated to scale as

$$N_{min} \sim \epsilon_r^{-\mathcal{D}},$$

where I have defined $\epsilon_r$ to be the relative goodness of the analog, *i.e.*

$$\epsilon_r = \frac{\mathbf{x}_N - \mathbf{x}_k}{\|\mathbf{x}_N\|}$$

and I have used the correlation dimension estimated in the previous section in equation (4.4).

This means that, if we consider a goodness $\epsilon = 0.1$, we expect to need $N_{min} \sim 10^2 = 100$ points to get sufficiently good $\epsilon$-analog. Remembering the sensitivity to initial condition expressed by equation (3.1), we expect the prediction to worsen if we tries to predict to far in the future. In [52] the MLE is estimated to be $\lambda_M = 0.906$, which means that after one second of evolution, the goodness of the analog decrease to approximately one half, and after two seconds it worsen by a factor 6. This effect should be observed by repeating the prediction at increasing time in the future.

In order to evaluate this result, I have simulated the system for different times, then taking the first half of the series as the training set (which represents the historical information about the system) and the second half as the test set. For each point of the test set I have made the prediction $\tau$ step into the future (excluding clearly the last $\tau$ points) using the first half as the set of the possibles $\epsilon$ analogs. In order to test the goodness of prediction, I have extended the definition of PCC in order to take into account the dimensionality of the system, and I have thus defined:

$$\rho = \frac{\sum_i (\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{y}_i - \overline{\mathbf{y}})}{\sqrt{(\mathbf{x}_i - \overline{\mathbf{x}})}\sqrt{(\mathbf{y}_i - \overline{\mathbf{y}})}}. \tag{4.5}$$

In figure 4.2 are reported the result of the correlation between the predicted and the measured valued (averaged over the length of the predicted series) as a function of the number of steps into the future made during prediction. The different length of the time series implies different initial goodness of the analog, and thus accuracy of the prediction. From the plot, the decreasing in accuracy,

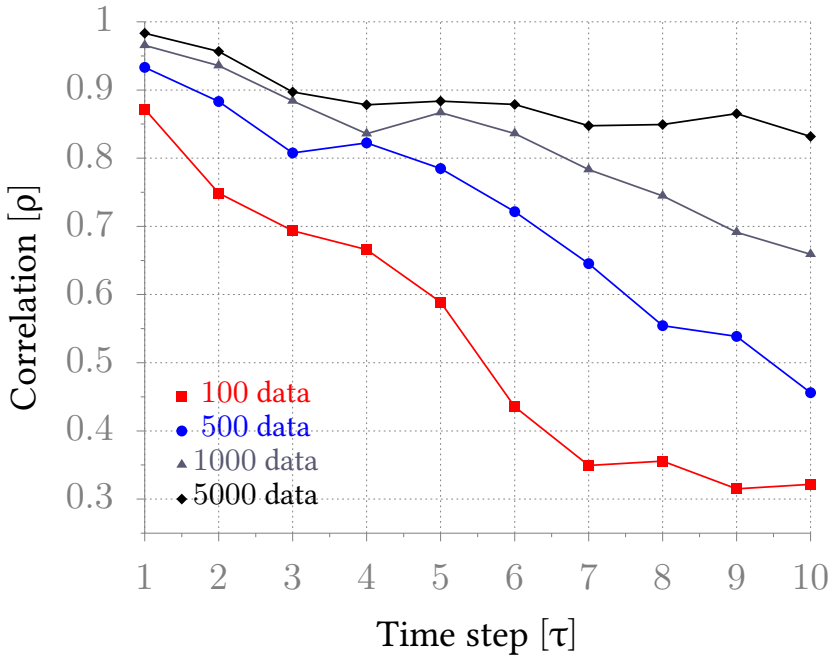



**Figure 4.2:** Correlation coefficient ρ as a function of the time step into the future τ. The correlation measure is defined in equation (4.5). It appears clear that the initial precision depends on the number of data, but also the possibility of mantaining a good precision into the future depends on this parameter (as already seen with the Lyapunov maximal exponents in (3.1)).

due to the MLE appears clearly, mostly for the plot with few points in the data set.

It is worth notinjg that the relatively small number (of the order of $10^3$) of data point used for this method scale exponentially with the dimensionality, as already seen in equation (3.14) in the previouos chapter.

# 5

# CONCLUSIONS

In this work I have reviewed some recent results on causality detection and predictability of a complex non linear system, for which we may not know the laws governing its dynamics, but we possibly have many ("big") data describing its behaviour in time. We have thus analyzed two main research question, namely: the determination of the causal laws that relate variables and the number of effective DoF of the system ("the system dimensionality"), together with the possibility of making prediction based on the recorded history of the system.

Alternative to the classical physics approach, based on building generative models unveiling causal relationships among variables, here I have presented a "data driven" inductive approach that tries to infer these information solely through the data describing the system's dynamics.

First of all, I have examined a number of definitions of causality (in chapter 1) that have been developed in the last years, and I have critically highlighted, both from a theoretical point of view and through specific examples, their limitation.

In particular I have presented the CCM algorithm, which is based on the powerful Takens' theorem of time delay embedding. I have shown how its success may depend on many different and often uncontrollable factors: the degree of non-linearity of the dynamics, the presence or not of oscillatory cycles, the presence or not of noise. In fact, in Chapter 2 I have studied a simple toy model with some free parameters that allowed me to generate several type of non linear times series with different couplings, and then tests if the output of the CCM algorithm recovers the known relationships. Results of the this experiment are summarised in Table 2.1 on page 28.

In the second part of the thesis, I have studied the problem of system dimensionality detection and predictability, which are strictly related. First, I have presented the method of the analog with its variants, explaining the deterministic foundation of the method, its merits and limitations. In particular I have discussed the so called "curse" of dimensionality, i.e. the intrinsic limitation of applicability of the method to system of high dimensionality, even if many data are available. To quantitive understand the minimum amount of data needed for effective predictability is thus strictly related to the capacity to infer the system dimension. I have presented two different approaches to this problem, one based upon the Grassberger-Procaccia algorithm and the other one founded on the Takens' theorem. Both of the methods, however, need the statistics of near-

est neighbors as input in order to determine the effective dimensionality of the system. I have then derived a scaling law for the minimum number of data required to make valid prediction (and at the same time to correctly estimate the effective dimension), showing that this number grows exponentially with the dimension of the system. This result is what is called the *curse of dimensionality*.

In Chapter 4 I have then shown a simple example of a system for which I have estimated the effective dimension and predicted the future behaviour, showing that the minimum number of points needed to make reliable estimate agrees with the theoretical expectations.

An interesting development of this work is the application of these methods to a real world examples, so to infer the possibility to apply a purely data driven approach to make predictions for practical applications (i.e. cancer detection). Clearly, a first step should be the study of a system for which a model already exists, in order to verify that technical difficulties (like noise or systematic errors in the dataset) does not hinder the estimations, and eventually studying some methods to overcome the obstacles.

In conclusion, I can summarize my work by saying that, even if physicists have always been skeptical about a data driven approach to the description of nature, this point of view is nonetheless worth of more credit, especially in the era of "big data", where we have for the first time an amount of data on different type of systems that can be really exploited in order to improve many societal problems from medicine to traffic. Of course this data driven approach must be made with extreme attention and sensibility, without over-claiming the generality of the results, and testing the specific conditions on which they hold.

On the other hand, I have shown that even for a relative simple, of low dimension, and controlled case of complex system (as the multi-species Lotka-Volterra or the Lorentz dynamics), the optimism of who says that *"this is a world where massive amounts of data and applied mathematics replace every other tool that might be brought to bear"* needs to be confronted with the reality of nature, which most of the time is more difficult than expected.

# A

# PROOF OF TAKENS' THEOREM.

I give here a proof of Takens' delay embedding theorem, firstly proved in the form we have used by Takens in 1981 [24], even if similar results were proven at the same time by other authors (see [53, 54]). Since then, many extension have been proved (see for example [26, 38, 39, 55]), and other author have provided alternative proofs [56].

A.1 Theorem (Takens' time delay reconstruction theorem).
*Let* $M$ *be a compact manifold of dimension* $d$. *Let* $(\varphi, h)$ *be:*

- $\varphi\colon M \to M$ *a smooth (at least* $C^2$*) diffeomorphism,*
- $h\colon M \to \mathbb{R}$ *a smooth function.*

*Then, it is true that the* $(2d+1)$*-fold observation map* $H[\varphi, h]\colon M \to \mathbb{R}^{2d+1}$ *defined by:*

$$x \mapsto (h(x), h(\varphi(x)), \dots, h(\varphi^{2d}(x))) \tag{A.1}$$

*is an immersion (*i. e. $H$ *is one-to-one between* $M$ *and its image after* $H$*, and with both* $H$ *and* $H^{-1}$ *differentiable).*

*Proof.* The proof of the theorem goes along several passages, that begins from a simple, particular case and generalize afterwards.

- As a beginning, we assume that if $x$ is a point with period $k$ of $\varphi$ (which means that $\varphi^k(x) = x$) and $k \leqslant 2d+1$, all eigenvalues of $d\varphi^k_x$ are different and different from 1. We also assume that no different fixed point of $\varphi$ share the same image after $h$. For $H[\varphi, h]$ to be an immersion near a fixed point $x$, the co-vectors $dh_x, dh\varphi_x, \dots, dh\varphi^{2m}_x$ must span $T^*_x(M)$. This is true for each $h$ given that $\varphi$ satisfies the above mentioned condition at each fixed point.
- In the exact same way one proves that $H[\varphi, h]$ is generically an immersion and even an embedding when restricted to the periodic points with period $k \leqslant 2m+1$. So we may assume that for generic $(\overline{\varphi}, \overline{h})$ we have $H[\overline{\varphi}, \overline{h}]$, restricted to a compact neighborhood $V$ of the set of points with period $k \leqslant 2m+1$ is an embedding. Now this means also that for some $(\varphi, h) \in \mathfrak{U}$, which is a neighborhood of $(\overline{\varphi}, \overline{h})$, $H[\varphi, h]|_V$ is an embedding.

- From this, we show that exists some $(\varphi, h) \in \mathfrak{U}$, arbitrarily near to $(\overline{\varphi}, \overline{h})$ for which $H[\varphi, h]$ is an embedding.
  For any point $x \in M$, which is not a point of period $k \leqslant 2m+1$ for $\overline{\varphi}$, the co-vectors $d\overline{h}_x, d\overline{h}\varphi_x, \ldots, d\overline{h}\varphi_x^{2m} \in T_x^*M$ can be perturbed independently by perturbing y.
  Hence arbitrarily near $\overline{h}$ there is $\overline{\overline{h}}$ such that $(\overline{\varphi}, \overline{\overline{h}}) \in \mathfrak{U}$ and such that $H[\overline{\varphi}, \overline{\overline{h}}]$ is an immersion. Then there is a positive $\epsilon$ such that whenever $0 < \rho(x, x') < \epsilon$, $H[\overline{\varphi}, \overline{\overline{h}}](x) \neq H[\overline{\varphi}, \overline{\overline{h}}](x')$ (with $\rho$ some fixed metric on $M$). There is even a neighborhood $\mathfrak{U}' \subset \mathfrak{U}$ of $(\overline{\varphi}, \overline{\overline{h}})$ such that for any $(\varphi, h) \in \mathfrak{U}'$, $H[\varphi, h]$ is an immersion and $H[\varphi, h](x) \neq H[\varphi, h](x')$ whenever $x \neq x'$ and $\rho(x, x') \leqslant \epsilon$. From now on we also assume that each component of $V$ has diameter smaller than $\epsilon$.
- Finally, we have to show that in $\mathfrak{U}'$ we have a pair $(\varphi, h)$ with $H[\varphi, h]$ injective. For this we need a finite collection $\{U_i\}_{i=1}^N$ of open subsets of $M$, covering the closure of $M \setminus \{\bigcap_{j=0}^{2m} \varphi^j(V)\}$ and such that:
  1. for each $i = 1, \ldots, N$ and $k = 0, 1, \ldots, 2m$, diameter $\overline{\varphi}^k(U_i) < \epsilon$;
  2. for each $i, j = 1, \ldots, N$ and $k, l = 0, 1, \ldots, 2m$, $\overline{\varphi}^k(U_i) \cap U_j \neq \emptyset$ and $\overline{\varphi}^l(U_i) \cap U_j \neq \emptyset$ imply that $k = l$;
  3. for $\overline{\varphi}^j(x) \in M \setminus (\bigcup_i U_i)$, $j = 0, \ldots, 2m$, $x' \notin V$ and $\rho(x, x') > \epsilon$, no two points of the sequence $x, \overline{\varphi}(x), \ldots, \overline{\varphi}^{2m}(x), x', \overline{\varphi}(x'), \ldots, \overline{\varphi}^{2m}(x')$ belong to the same $U_i$.

  We take a partition of the unity $\{\lambda_i\}$ that correspond to the finite set of $U_i$ (that is, $\lambda_i$ is non negative, has support $\overline{U_i}$ and $\sum_i \lambda_i(x) = 1$ for all $x \in \overline{M \setminus V}$).
  Consider the map $\psi \colon M \times M \times \mathbb{R}^N \to \mathbb{R}^{2m+1} \times \mathbb{R}^{2m+1}$ which is defined in the following way

$$\psi(x, x', \epsilon_1, \ldots, \epsilon_N) = (H[\overline{\varphi}, \overline{\overline{h}}_\epsilon](x), H[\overline{\varphi}, \overline{\overline{h}}_\epsilon](x'))$$

  where $\epsilon$ stands for $(\epsilon_1, \ldots, \epsilon_N)$ and $\overline{\overline{h}}_\epsilon = \overline{\overline{h}} + \sum_i \epsilon_i \lambda_i$.
  We define $W \in M \times M$ as $W = \{(x, x') \in M \times M \mid \rho(x, x') \geqslant \epsilon$ and not both $x, x' \in \mathrm{inv}(V)\}$. $\psi$, restricted to a small neighborhood of $W \times 0$ is transverse with respect to the diagonal of $\mathbb{R}^{2m+1} \times \mathbb{R}^{2m+1}$. This follows immediately from all the conditions imposed on the covering with $U_i$. From this, we conclude that there are arbitrarily small $\rho \in {}^N$ such that $\psi(W \times \rho) \cap \Delta = \emptyset$. If also for such an $\rho$ holds that $(\overline{\varphi}, \overline{\overline{h}}_\epsilon) \in \mathfrak{U}$ then $H[\overline{\varphi}, \overline{\overline{h}}_\epsilon]$ is injective and hence an embedding.
- This proves that for a dense set of pairs $(\varphi, h)$, $H[\varphi, h]$ is an embedding. Since the set of all embeddings is open in the set of all mappings, there is an open and dense set of pairs $(O, Y)$ for which $H[\varphi, h]$ is an embedding.

This proves the theorem. □

# BIBLIOGRAPHY


[1] G. Sugihara, R. May, H. Ye, C. Hsieh, E. Deyle, M. Fogarty, and S. Munch. Detecting causality in complex ecosystems. *Science*, 338(6106):496–500, 2012. (Cited on pages iii, 3, 8, and 13.)

[2] C. W. J. Granger. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, 37(3):424–438, 1969. (Cited on pages iii, 3, and 4.)

[3] F. Cecconi, M. Cencini, M. Falcioni, and A. Vulpiani. Predicting the future from the past: An old problem from a modern perspective. *American Journal of Physics*, 80(11):1001–1008, 2012. (Cited on pages iii, 2, 29, 30, 37, and 40.)

[4] P. Grassberger and I. Procaccia. Characterization of Strange Attractors. *Phys. Rev. Lett.*, 50:346–349, Jan 1983. doi: 10.1103/PhysRevLett.50.346. (Cited on pages iii, 2, 38, 43, and 44.)

[5] Byers Jeff. The physics of data. *Nat Phys*, 13(8):718–719, aug 2017. (Cited on page 1.)

[6] Editorial. The thing about data. *Nat Phys*, 13(8):717–717, aug 2017. (Cited on page 1.)

[7] Sau Lan Wu. Brief history for the search and discovery of the Higgs particle — A personal perspective. *International Journal of Modern Physics A*, 29(27):1430062, 2014. doi: 10.1142/S0217751X14300622. (Cited on page 1.)

[8] P. W. Higgs. Broken symmetries and the masses of gauge bosons. *Phys. Rev. Lett.*, 13(16):508, 1964. (Cited on page 1.)

[9] Georges Aad, T Abajyan, B Abbott, J Abdallah, S Abdel Khalek, AA Abdelalim, O Abdinov, R Aben, B Abi, M Abolins, et al. Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. *Physics Letters B*, 716(1):1–29, 2012. (Cited on page 1.)

[10] Serguei Chatrchyan, Vardan Khachatryan, Albert M Sirunyan, Armen Tumasyan, Wolfgang Adam, Ernest Aguilo, T Bergauer, M Dragicevic, J Erö, C Fabjan, et al. Observation of a new boson at a mass of 125 GeV with the CMS experiment at the LHC. *Physics Letters B*, 716(1):30–61, 2012. (Cited on page 1.)

[11] Eugene P. Wigner. The unreasonable effectiveness of mathematics in the natural sciences. Richard courant lecture in mathematical sciences delivered at New York University, May 11, 1959. *Communications on Pure and Applied Mathematics*, 13(1):1–14, 1960. ISSN 1097-0312. (Cited on page 1.)

[12] J. M. McCracken and R. S. Weigel. Convergent cross-mapping and pairwise asymmetric inference. *Phys. Rev. E*, 90:062903, Dec 2014. doi: 10.1103/PhysRevE.90.062903. (Cited on pages 3 and 15.)

[13] Joseph Lee Rodgers and W. Alan Nicewander. Thirteen Ways to Look at the Correlation Coefficient. *The American Statistician*, 42(1):59–66, 1988. (Cited on page 3.)

[14] John Aldrich. Correlations Genuine and Spurious in Pearson and Yule. *Statist. Sci.*, 10(4):364–376, 11 1995. (Cited on page 3.)

[15] N. Wiener. The theory of prediction. *Modern mathematics for engineers*, 1: 125–139, 1956. (Cited on page 4.)

[16] C. Hiemstra and J. D. Jones. Testing for linear and nonlinear Granger causality in the stock price-volume relation. *The Journal of Finance*, 49(5): 1639–1664, 1994. (Cited on page 5.)

[17] A. Roebroeck, E. Formisano, and R. Goebel. Mapping directed influence over the brain using Granger causality and fMRI. *Neuroimage*, 25(1):230–242, 2005.

[18] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011. doi: 10.1016/j.jocs.2010.12.007. (Cited on page 5.)

[19] T. Schreiber. Measuring Information Transfer. *Phys. Rev. Lett.*, 85:461–464, Jul 2000. doi: 10.1103/PhysRevLett.85.461. (Cited on pages 5 and 6.)

[20] F. A. Razak and H. J. Jensen. Quantifying "causality" in complex systems: understanding transfer entropy. *PloS one*, 9(6):e99462, 2014. (Cited on page 5.)

[21] L. Barnett, A. B. Barrett, and A. K. Seth. Granger Causality and Transfer Entropy Are Equivalent for Gaussian Variables. *Phys. Rev. Lett.*, 103:238701, Dec 2009. doi: 10.1103/PhysRevLett.103.238701. (Cited on page 6.)

[22] C. Gardiner. *Stochastic Methods: A Handbook for the Natural and Social Sciences*. Springer Series in Synergetics. Springer Berlin Heidelberg, 2009. ISBN 9783540707127. (Cited on pages 7 and 17.)

[23] E. N. Lorenz. Deterministic Nonperiodic Flow. *Journal of the Atmospheric Sciences*, 20(2):130–141, 1963. doi: 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2. (Cited on pages 9, 13, and 43.)

[24] F. Takens. Detecting strange attractors in turbulence. *Lecture Notes in Mathematics, Berlin Springer Verlag*, 898:366, 1981. doi: 10.1007/BFb0091924. (Cited on pages 9 and 49.)

[25] C. Jost and S. P. Ellner. Testing for predator dependence in predator-prey dynamics: a non-parametric approach. *Proceedings of the Royal Society B: Biological Sciences*, 267(1453):1611–1620, 2000. doi: 10.1098/rspb.2000.1186. (Cited on page 12.)

[26] J. Stark, D.S. Broomhead, M.E. Davies, and J. P. Huke. Takens embedding theorems for forced and stochastic systems. *Nonlinear Analysis: Theory, Methods and Applications*, 30(8):5303–5314, 1997. doi: 10.1016/S0362-546X(96)00149-6. Proceedings of the Second World Congress of Nonlinear Analysts. (Cited on pages 12, 32, and 49.)

[27] J. C. McBride, X. Zhao, N. B. Munro, G. A. Jicha, F. A. Schmitt, R. J. Kryscio, C. D. Smith, and Y. Jiang. Sugihara causality analysis of scalp EEG for detection of early Alzheimer's disease. *NeuroImage: Clinical*, 7:258–265, 2015. (Cited on page 13.)

[28] A. E. BozorgMagham, S. Motesharrei, S. G. Penny, and E. Kalnay. Causality analysis: Identifying the leading element in a coupled dynamical system. *PloS one*, 10(6):e0131226, 2015. doi: 10.1371/journal.pone.0131226. (Cited on page 13.)

[29] Chengyi Tu, Jacopo Grilli, Friedrich Schuessler, and Samir Suweis. Collapse of resilience patterns in generalized Lotka-Volterra dynamics and beyond. *Phys. Rev. E*, 95:062307, Jun 2017. (Cited on page 17.)

[30] G. N. Milstein. *Numerical integration of stochastic differential equations*, volume 313. Springer Science & Business Media, 1994. (Cited on page 18.)

[31] K. Burrage, P. M. Burrage, and T. Tian. Numerical methods for strong solutions of stochastic differential equations: an overview. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 460(2041):373–402, 2004. (Cited on page 18.)

[32] Vito Volterra. Variations and fluctuations of the number of individuals in animal species living together. *ICES Journal of Marine Science*, 3(1):3–51, 1928. (Cited on page 25.)

[33] L Campbell and W. Garnett. *The Life of James Clerk Maxwell*. Macmillan and Company, 1884. (Cited on page 29.)

[34] R. Kitchin. *The data revolution: Big data, open data, data infrastructures and their consequences*. Sage, 2014. (Cited on page 29.)

[35] E. N. Lorenz. Three approaches to atmospheric predictability. *Bull. Amer. Meteor. Soc*, 50(3454):349, 1969. (Cited on page 30.)

[36] E. N. Lorenz. Atmospheric predictability as revealed by naturally occurring analogues. *Journal of the Atmospheric sciences*, 26(4):636–646, 1969. (Cited on page 30.)

[37] S. De Souza-Machado, R.W. Rollins, D.T. Jacobs, and J.L. Hartman. Studying chaotic systems using microcomputer simulations and Lyapunov exponents. *American Journal of Physics*, 58(4):321–329, 1990. (Cited on pages 31 and 37.)

[38] T. Sauer, J. A. Yorke, and M. Casdagli. Embedology. *Journal of Statistical Physics*, 65(3):579–616, 1991. doi: 10.1007/BF01053745. (Cited on pages 32 and 49.)

[39] J. C. Robinson. A topological delay embedding theorem for infinite-dimensional dynamical systems. *Nonlinearity*, 18(5):2135, 2005. (Cited on page 32.)

[40] P. Collet and J.P. Eckmann. *Iterated Maps on the Interval as Dynamical Systems*. Modern Birkhäuser Classics. Birkhäuser Boston, 2009. ISBN 9780817649265. (Cited on page 32.)

[41] J. C. Sprott and G. Rowlands. Improved correlation dimension calculation. *International Journal of Bifurcation and Chaos*, 11(07):1865–1880, 2001. (Cited on page 32.)

[42] H. Kantz and T. Schreiber. *Nonlinear time series analysis*. Cambridge university press, 2004. (Cited on page 33.)

[43] Matthew B. Kennel, Reggie Brown, and Henry D. I. Abarbanel. Determining embedding dimension for phase-space reconstruction using a geometrical construction. *Phys. Rev. A*, 45:3403–3411, Mar 1992. doi: 10.1103/PhysRevA.45.3403. (Cited on page 34.)

[44] Rainer Hegger and Holger Kantz. Improved false nearest neighbor method to detect determinism in time series data. *Phys. Rev. E*, 60(4):4970, 1999. doi: 10.1103/PhysRevE.60.4970. (Cited on page 34.)

[45] A. M. Fraser and H. L. Swinney. Independent coordinates for strange attractors from mutual information. *Phys. Rev. A*, 33(2):1134–1140, 1986. doi: 10.1103/PhysRevA.33.1134. (Cited on page 36.)

[46] O.E. Rössler. An equation for continuous chaos. *Physics Letters A*, 57(5): 397–398, 1976. doi: 10.1016/0375-9601(76)90101-8. (Cited on page 38.)

[47] G. Paladin and A. Vulpiani. Anomalous scaling laws in multifractal objects. *Physics Reports*, 156(4):147–225, 1987. (Cited on page 38.)

[48] H. Poincaré. Sur le problème des trois corps et les équations de la dynamique. *Acta mathematica*, 13(1):A3–A270, 1890. (Cited on page 39.)

[49] M. Kac. On the notion of recurrence in discrete stochastic processes. *Bulletin of the American Mathematical Society*, 53(10):1002–1010, 1947. (Cited on page 40.)

[50] C. Cercignani. *Ludwig Boltzmann: The Man who Trusted Atoms*. Oxford University Press, 1998. ISBN 9780198501541. (Cited on page 40.)

[51] L. A. Smith. Intrinsic limits on dimension calculations. *Physics Letters A*, 133(6):283–288, 1988. (Cited on pages 41 and 44.)

[52] J.C. Sprott. *Chaos and Time-series Analysis*. Oxford University Press, 2003. ISBN 9780198508403. (Cited on page 45.)

[53] D. Aeyels. Generic observability of differentiable systems. *SIAM Journal on Control and Optimization*, 19(5):595–603, 1981. (Cited on page 49.)

[54] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw. Geometry from a Time Series. *Phys. Rev. Lett.*, 45:712–716, Sep 1980. doi: 10.1103/PhysRevLett.45.712. (Cited on page 49.)

[55] E. R. Deyle and G. Sugihara. Generalized theorems for nonlinear state space reconstruction. *PLoS One*, 6(3):e18295, 2011. (Cited on page 49.)

[56] L. Noakes. The Takens embedding theorem. *International Journal of Bifurcation and Chaos*, 01(04):867–872, 1991. doi: 10.1142/S0218127491000634. (Cited on page 49.)