Machine Learning on Causality

Report Internship Physics of Complex Systems

Davide Rossetti

Introduction

Causal relationships play a fundamental role in understanding the world around us. The ability to identify and understand cause-effect relationships is critical to making informed decisions, predicting outcomes, and developing effective strategies. However, deciphering causal relationships from observational data is a difficult task, as correlations alone may not provide definitive evidence of causality. In recent years, the field of machine learning (ML) has emerged as a powerful tool for causal analysis, offering new opportunities for uncovering hidden causal mechanisms and better understanding complex systems. ML algorithms can detect patterns and dependencies in data, enabling the discovery of causal links between variables. By leveraging sophisticated models and optimization techniques, ML approaches provide a data-driven and automated way to infer causal relationships. Causal analysis can be viewed from two different angles: Intervention causality and Observation causality. Interventional causality focuses on examining the causal effects of interventions or treatments. It aims to answer questions such as "What is the impact of a particular intervention on a particular outcome of interest?" Observational causality, on the other hand, is concerned with inferring causality from observational data where interventions or treatments are not explicitly controlled for. The goal of this work is to explore various ML techniques and methods to address the challenging task of causal analysis. The integration of machine learning techniques into causal analysis offers exciting opportunities for uncovering and understanding causal relationships from complex datasets. By harnessing the power of ML algorithms, researchers and practitioners can expand our knowledge of cause and effect, enabling more accurate predictions, better decision-making, and improved strategies in a wide range of domains.

Contents

Chapter I

Linear Response Theory and FDR

I

Linear Response Theory

The theory of Linear Response, in statistical mechanics, is a powerful tool for describing the evolution of a system away from or towards equilibrium when perturbed by the application of an external field. The main goal of response theory is to explain how the system responds to external conditions. When the original theory of the phenomenon is deformed by an external source, it is generally necessary to find another model. Instead, we can make progress with linear response analysis if we can describe the source as a small perturbation of the original system. For small perturbations, the changes in the properties of the system that couple to the external field are proportional to it. This constant of proportionality is called the Linear Response Function (or the After-Effect Function) and provides important information about the system [1]. By looking at the observable in a statistical way, there are many protocols for analysing the above problem. The linear response of the system can be studied in a static or dynamic description by going beyond the statistical mechanics of equilibrium and considering the system's time evolution.

Static Linear Response and FD Relation

In statistical equilibrium mechanics, the average value of thermodynamics observable and the magnitude of their fluctuations around their equilibrium values can be predicted. We consider a classical system described by a Hamiltonian $H_0(x)$, where x represents one of the possible microscopic configurations of the system consisting of N degrees of freedom. The thermal equilibrium of the system is given by the probability density function:

$$f_0(x) = \frac{1}{Z} e^{-H_0(x)/k_B T} \tag{1}$$

$$Z = \sum_x e^{-H_0(x)/k_B T} \tag{2}$$

A possible observable quantity (for example, in a paramagnetic system) is $M(x)$, the magnetic or dipole moment, whose macroscopic quantity is obtained at equilibrium by the thermal average:

$$\langle M\rangle_{eq} = \sum_x M(x) f_0(x) = \sum_x \frac{M(x)}{Z} e^{-H_0(x)/k_B T} \tag{3}$$

If we apply to the system a small field or, more generally, a small perturbation F and wait a sufficiently long time, the system reaches a new thermal equilibrium described by the new perturbed Hamiltonian: $H(x,F) = H_0(x) - FM(x)$ (we assume that $M(x)$ is the conjugate variable) and by the new probability density function:

$$f(x,F) = \frac{1}{Q(F)} e^{-\beta\left[H_0(x) - FM(x)\right]} \tag{4}$$

$$Q(F) = \sum_x e^{-\beta\left[H_0(x) - FM(x)\right]} \tag{5}$$

In classical statistical physics, the perturbed system can be extended to include the unperturbed system by using the expansion of the exponential factor. By a first-order linear response, we obtain:

$$e^{\beta FM(x)} = 1 + \beta FM(x) + O(F^2) \tag{6}$$

$$Q(F) = \sum_x e^{-\beta\left[H_0(x)\right]} \left(1 + \beta FM(x) + O(F^2)\right) \tag{7}$$

$$f(x,F) = \frac{1}{Q(F)} e^{-\beta\left[H_0(x)\right]} \left(1 + \beta FM(x) + O(F^2)\right) \tag{8}$$

and by exploiting the results of the equilibrium distribution:

$$f(x,F) = \left(1 - \beta\langle M\rangle F + \beta FM(x)\right) f_0(x) \tag{9}$$

These equations contain the equilibrium value of the average $\langle M\rangle$. If we now want to consider the average of any other observable B in the perturbed system, we derive that:

$$\langle B\rangle_F = \langle B\rangle_0 + \beta F\left(\langle MB\rangle_0 - \langle M\rangle_0\langle B\rangle_0\right) \tag{10}$$

where the indices F and o stand for the average performed in the disturbed and undisturbed systems, respectively. It can also be written as:

$$\langle B\rangle_F = \langle B\rangle_{\mathrm{o}} + F\chi_{MB} \tag{11}$$

where the coefficient χ_{BM} represents the susceptibility:

$$\chi_{MB} = \beta\left(\langle MB\rangle_{\mathrm{o}} - \langle M\rangle_{\mathrm{o}}\langle B\rangle_{\mathrm{o}}\right) = \beta\langle MB\rangle_{\mathrm{o}c} \tag{12}$$

The corresponding average change of a generic observable B, induced by the perturbation, is:

$$\langle \Delta B\rangle_F = \langle B\rangle_F - \langle B\rangle_{\mathrm{o}} = F\chi_{MB} \tag{13}$$

The relevant instance of the Fluctuation Response Theorem is the relationship that exists between the correlation function and susceptibility. This is a Fluctuation Dissipation Relation, which in statistical equilibrium mechanics relates the correlation functions to macroscopically measurable quantities such as specific heat, susceptibility and compressibility [2].

Dynamic Linear Response and FD Relation

To describe the non-equilibrium response of a system, we need to analyse its dynamic evolution. The non-equilibrium situation could be the result of the application of a field at a certain point in time or the relaxation dynamics after which this field was switched off (relaxation dynamics). To analyse this situation, we consider a system in phase space characterised by the Hamiltonian $H\left(p,q,t\right) = H_{\mathrm{o}}\left(p,q\right) - F\left(t\right)A\left(p,q\right)$, where H_{o} is the time-independent part and A is a general observable of the system.

$$H\left(p,q\right) = \begin{cases} H_{\mathrm{o}}\left(q,p\right) & \text{per } t < \mathrm{o} \\ H_{\mathrm{o}}\left(q,p\right) - A\left(q,p\right) & \text{per } t \geq \mathrm{o} \end{cases} \tag{14}$$

The system, when described by H_0, is in thermal equilibrium and its statistical properties are described by $f_{eq}(p,q)$. The evolution of the system is given by Hamilton's equations:

$$\frac{dq_i}{dt} = \frac{\partial H}{\partial p_i}, \tag{15}$$

$$\frac{dp_i}{dt} = -\frac{\partial H}{\partial q_i}, \tag{16}$$

In the phase space, the probability distribution evolves according to the Liouville Equation [3]:

$$\frac{\partial f(p,q,t)}{\partial t} + i\left[\hat{L_0} + \hat{L_{ext}}(t)\right] f(p,q,t) = 0 \tag{17}$$

$\hat{L_0}$ and $\hat{L_{ext}}$ are the Liouville operators associated with the unperturbed Hamiltonian and the perturbed Hamiltonian, respectively. By exploiting the linear approximation, we obtain an approximate solution:

$$f(p,q,t) \approx f_{eq}(p,q) - i\int_0^t dt' e^{-i\left(t-t'\right)\hat{L_0}} \hat{L_{ext}}\left(t'\right) f_{eq}(p,q) \tag{18}$$

where $f_{eq}(p,q) = f(p,q,t=0)$.
This equation allows us to calculate the ensemble average of any function $B(p,q)$, in linear order in $F(t)$:

$$\langle B(t)\rangle = \int dpdq\ B(p,q)\, f(p,q,t)\,. \tag{19}$$

In this way, we can compute the difference with respect to the equilibrium value:

$$\langle \Delta B(t)\rangle = -i\int dpdq\ B(p,q)\int_0^t dt' e^{-i\left(t-t'\right)\hat{L_0}} \hat{L_{ext}}\left(t'\right) f_{eq}(p,q) \tag{20}$$

Following the calculation in [3], we obtain the result:

$$\langle \Delta B(t)\rangle = \int_0^t dt' \chi\left(t-t'\right) F\left(t'\right) \tag{21}$$

where $\chi\left(t,t'\right)$ is the susceptibility or response function, (it is assumed for simplicity that it depends on the difference of its arguments) with the properties:

(a) Stationarity of unperturbed system (if the perturbation $f\left(t\right)$ at time t gives an output signal $s\left(t\right)$ then a perturbation $f\left(t_1\right)$ will give an output signal $s\left(t_1\right)$).

(b) Causality (the system responds only after an external perturbation F has been switched on).

This is the general expression of the Fluctuation Dissipation Theorem, which is the relationship between the spontaneous fluctuations and the response to time-dependent external fields of a physical observable.
By considering similar calculations, we obtain the Kubo formula, which relates the fluctuations of the observed quantity B at time t to the related equilibrium-time correlation function of B and M (also known as the Kubo function $C_{MB}\left(t\right)$).

$$\langle \Delta B\left(t\right)\rangle = \beta F\langle B\left(t\right)M\left(0\right)\rangle_{oc} = \beta F C_{MB}\left(t\right) \tag{22}$$

Now we can establish a relationship between the two expressions and thus obtain the classical Kubo expression [1]:

$$\chi_{MB}\left(t\right) = -\beta\Theta\left(t\right)\dot{C}_{MB}\left(t\right) \tag{23}$$

Thanks to the temporal translational invariance at equilibrium, we can write:

$$\dot{C}_{MB}\left(t\right) = \langle \dot{B}\left(t\right)M\left(0\right)\rangle_{oc} = -\langle B\left(t\right)\dot{M}\left(0\right)\rangle_{oc} \tag{24}$$

Using the obtained results together, we can write the response function as follows:

$$\chi_{MB}\left(t\right) = -\beta\Theta\left(t\right)\langle \dot{B}\left(t\right)M\left(0\right)\rangle_{oc} \tag{25}$$

Properties of the Response Function

- **Translational Invariance**: If we assume that the response function is invariant under time translations we can write it as

$$\chi_{MB}\left(t\right) = -\beta\Theta\left(t\right)\langle \dot{B}\left(t\right)M\left(0\right)\rangle_{oc} \tag{26}$$

It is useful to perform the Fourier transform to work in frequency space. The Fourier transform for a generic function $f(t)$ is defined as:

$$\hat{f}(w) = \int_{-\infty}^{+\infty} dt f(t) e^{iwt} \tag{27}$$

$$f(t) = \int_{-\infty}^{+\infty} \frac{dw}{2\pi} \hat{f}(w) e^{-iwt} \tag{28}$$

If we consider the fluctuation relation for a generic observable B:

$$\langle \Delta B_i(t) \rangle = \int_0^t dt' \chi_{ij}\left(t - t'\right) \phi_j\left(t'\right) \tag{29}$$

Next, we take the Fourier transform:

$$\langle \delta B_i(w) \rangle = \chi_{ij}(w) \phi_j(w) \tag{30}$$

It can be observed that the response in frequency space is "local": if we perturb something at frequency w, the system responds at frequency w. For more advanced analysis, we need to go into the realm of non-linear response. If we consider the Fourier transform of the response function, we can introduce the following notation for the real and imaginary parts:

$$\chi(w) = Re(\chi(w)) + iIm(\chi(w)) = \chi'(w) + i\chi''(w) \tag{31}$$

Imaginary Part: This part is due to the part of the response function that is not invariant under time reversal (hence its Fourier counterpart is an odd function). It is called the spectral function and represents the dissipative part of the response function:

$$\chi''(w) = \frac{-i}{2} \int_{-\infty}^{+\infty} dt\ e^{iwt} \left[\chi(t) - \chi(-t)\right] \tag{32}$$

$$\chi''(-w) = -\chi''(w) \tag{33}$$

Real Part: It is called the reactive part of the response function and is an even function of w. The arrow of time plays no role here in relation to what happened in the case of the imaginary part.

$$\chi'(w) = \frac{1}{2}\int_{-\infty}^{+\infty} dt\ e^{iwt}\left[\chi(t) + \chi(-t)\right] \tag{34}$$

$$\chi'(-w) = \chi'(w) \tag{35}$$

- **Causality**: We cannot have an influence on the past. This means that any response function must satisfy:

$$\chi(t) = 0 \text{ for all } t < 0 \tag{36}$$

This requirement of causality in the frequency domain means that $\chi(w)$ is analytic for $Im(w) > 0$. This means that there is a relationship between the real part χ' and the imaginary part χ'': the Kramers-Kronig Relationship.

II

Onsager Regression

Onsager's regression hypothesis is a special case of the Fluctuation-Dissipation Theorem and states the following: "If a system is out of equilibrium at time t_0, it is impossible to know whether this state out of equilibrium is the result of an external disturbance or a spontaneous fluctuation. The relaxation of the system back to equilibrium will be the same in both cases (provided the initial deviation from equilibrium is small enough)" [4]. In his statement, Onsager considers irreversible processes in which the relaxation of a macroscopic variable $A(t)$ is observed when a small perturbation ΔH is turned off at time 0 (it can be considered to be applied from time $-\infty$ to time 0).
Considering the results of the previous equation, we can write (if we consider $M = B = A$):

$$\langle \Delta A(t)\rangle = \beta F \langle A(0)\, A(t)\rangle_{0c} = \beta F \langle \delta A(0)\, \delta A(t)\rangle_0 \tag{37}$$

Physically, the variable $A(t)$ fluctuates in time with spontaneous microscopic fluctuations around its average value:

$$\langle \Delta A(0) \rangle = \beta F \langle \delta A(0)\, \delta A(0) \rangle_0 \tag{38}$$

The regression hypothesis can thus be derived from the ratio between the two equations:

$$\frac{\langle \delta A(t) \rangle}{\langle \delta A(0) \rangle} \Big|_{out-eq} = \frac{\langle \delta A(t)\, \delta A(0) \rangle}{\langle \delta A(0)\, \delta A(0) \rangle} \Big|_{eq} \tag{39}$$

If we analyse the equilibrium correlation fluctuations between $\delta A(t)$ and an instantaneous fluctuation at time zero $C_{AA}(t) = \langle \delta A(0)\, \delta A(t) \rangle_0$, we obtain the following limits:

$$\lim_{t \to 0} C_{AA}(t) = \langle \delta A(0)^2 \rangle_0 \tag{40}$$

$$\lim_{t \to \infty} C_{AA}(t) = \langle \delta A(0) \rangle_0 \langle \delta A(t) \rangle_0 \tag{41}$$

At small times fluctuations are correlated, while at large times they are uncorrelated. Therefore, the decay of correlations is expressed as the regression of spontaneous fluctuations. In general, this decay is exponential:

$$C_{AA}(t) \approx C_{AA}(t)\, e^{-\frac{|t|}{\tau_A}} \tag{42}$$

where τ_A is the relaxation time of the observed quantity A. Thus, if a system can be brought out of equilibrium by a small perturbation F or by a spontaneous thermal fluctuation δA, its return to equilibrium will be characterised by equilibrium fluctuations.

Chapter II

Different types of Causality

The philosophical concept of causality refers to the set of all particular "cause-effect" relationships and this representation is a much debated argument in mathematics and physics. The recognition and clarification of cause-effect relationships between variables, events or objects are the fundamental questions of most natural and social sciences [5]. In most disciplines, we need to quantify the strength of a possible causal relation in order to explain past events, control present situations and predict future consequences [3]. There are two different ways of obtaining information about causal relations: the Interventional approach and the Observational approach. Both provide information about the presence or absence of cause and effect, but they differ in how this information is expressed.

I

Observational Causality

It is based on the observation of the autonomous behaviour of a system. Granger causality (GC) and Transfer Entropy (TE) are mainly based on this type of approach.

Granger Causality

GC can be quantified and measured computationally and makes two statements about causality: 1) the cause occurs before the effect and 2) the cause contains information about the effect that is unique and does not occur in any other variable. Consequently, the causal variable can help to anticipate the effect variable after other data have been used first [6]. This is a regression-based interpretation of past data that compares the statistical uncertainties of two predictions. Considering the time series of two events $x_j(t)$ and $x_i(t)$, the first prediction characterises the Restricted Model [6], where only the past

history of x_j is included to predict future values of itself:

$$x_j(t) = \sum_{k=1}^{T} \left[B_{j,k} x_j(t-k) + \epsilon_j(t) \right] \tag{43}$$

where B is the autoregressive coefficient and ϵ_j is the prediction error. The second prediction characterises the Full Model [6], in which the past of x_i is also included in the model for predicting future values of x_j:

$$x_j(t) = \sum_{k=1}^{T} \left[A_{ji,k} x_i(t-k) + A_{jj,k} x_j(t-k) + \epsilon_{j|i}(t) \right] \tag{44}$$

where, as before, the A's are the autoregressive coefficients and $\epsilon_{j|i}$ is the prediction error on x_j associated with the knowledge of variable x_i. If we register an improvement in prediction, i.e. a reduction in the variance of prediction errors, we say that x_i is a Granger cause of x_j. GC is thus quantified as:

$$F_{x_i \to x_j} = \ln \frac{Var\left(\epsilon_j\right)}{Var\left(\epsilon_{j|i}\right)} \tag{45}$$

So if this value is positive, we have an improvement in predictive accuracy. GC can also be seen as a statistical hypothesis test.

Transfer Entropy

TE derives causality from an information theory interpretation. It is a non-parametric statistic that measures the amount of information exchanged from y to x. More specifically, it indicates the amount of uncertainty reduced in future values of y given knowledge of the past values of x (given past values of y) [5]. The TE from x to y is defined as:

$$T_{y|x} = \sum f\left(y_{i+H}, \boldsymbol{y}_i^{(K)}, \boldsymbol{x}_i^{(L)}\right) \cdot \ln \frac{f\left(y_{i+H} | \boldsymbol{y}_i^{(K)}, \boldsymbol{x}_i^{(L)}\right)}{f\left(y_{i+H}, \boldsymbol{y}_i^{(K)}\right)} \tag{46}$$

where K and L are the numbers of time series for each embedded vector (x and y), f is the joint probability density function and H is the prediction horizon. The subscripts of x and y represent the first time series to be considered. The causality measure of this procedure [6] is then calculated by considering the difference between the influence of y given x and the influence of x given y:

$$T_{x\to y} = T_{y|x} - T_{x|y} \tag{47}$$

It can also be written as a function of Shannon entropy $H(x)$ and mutual information:

$$T_{x\to y} = H\left(y_t|y_{t-1:t-L}\right) - H\left(y_t|y_{t-1:t-L}, x_{t-1:t-L}\right) \tag{48}$$

$$T_{x\to y} = I\left(y_t; x_{t-1:t-L}|y_{t-1:t-L}\right) \tag{49}$$

The uncertainty in these process measurements is a weak point because TE and GC values are calculated even if there is no causal relationship. Consequently, there is also the reverse problem that the actual causal relationships are not captured. This type of causation is not satisfactory from a physical point of view and for this reason, is discarded in some situations in favour of a more interventional approach.

II

Interventional Causality

It is based on the physical cause-effect relationship. The state of a variable is changed to manipulate the system and see if there is a reaction. It states that given a time series characterised by the vector $\boldsymbol{x}_t$ consisting of n entries, a perturbation of the variable $x^{(j)}$ at time 0, $x_0^{(j)} \to x_0^{(j)} + \delta x_0^{(j)}$, on average, causes a change in another variable $x_t^{(k)}$ with $t > 0$. Mathematically speaking, there is a causal relation if, given a smooth function $F(x)$, the relationship holds:

$$\overline{\frac{\delta F\left(x_t^{(k)}\right)}{\delta x_0^{(j)}}} \neq 0 \tag{50}$$

which means that a perturbation of the variable $x^{(j)}$ at the time o leads to a non-zero average variation $F\left(x_t^{(k)}\right)$ (carried out over many realisations of the experiment) with respect to its unperturbed evolution [7].

Pearl Approach

It relies on causal Bayesian networks and directed acyclic graphs (DAG) to understand the representation of the causal organisation and the response to external or spontaneous changes in variables [8]. According to Pearl's idea, causal models tell us how the probabilities describing the variables would change as a result of an external disturbance. An important distinction is between the action $do\,(x)$, which sets the random variable X to a certain value x, and the observation $X = x$. To test whether a variable x_k has a causal influence on another variable x_j, the marginal distribution of x_j is calculated under the action $do\,(X_k = x_k)$ for all values x_k of X_k. By analysing the sensitivity of the distribution to different x_k, the causal relationship can be established [9]. In this approach, causal relationships are explained using deterministic functional equations, and the concept of probability is incorporated by assuming that some variables within the equations cannot be observed [8].

Linear Response Approach

The connection to linear response theory is made when the system admits an invariant distribution and the variation $\delta x_{\mathrm{o}}^{(j)}$ is small enough. In this case, the response variables are related to the spontaneous fluctuation correlations in the equilibrium dynamics by the Fluctuation-Dissipation Theorem. There is a strong and strict connection between responses and correlations: if $\boldsymbol{x}_t$ is a stationary process characterised by an invariant probability density function $f_s\,(\boldsymbol{x})$, the following relation can be proved under fairly general conditions:

$$R_t^{kj} \equiv \lim_{\delta x_{\mathrm{o}}^j \to \mathrm{o}} \frac{\overline{\delta x_t^k}}{\delta x_{\mathrm{o}}^j} = -\langle x_t^k \frac{\partial \ln f_s\,(\boldsymbol{x})}{\partial x^j} \Big|_{\boldsymbol{x}_{\mathrm{o}}} \rangle \tag{51}$$

where R_t represents the linear response matrix of the system under consideration at time t, and the average is calculated over the twice joint probability distribution function $f_s\,(\boldsymbol{x}_t, \boldsymbol{x}_{\mathrm{o}})$. The distribution characteristics are derived from the data if they are not known [7].

Chapter III

Linear Stochastic System

I

Linear Response and Correlations

Determining causal relationships in a system cannot be done by analysing correlations alone. To better understand the difference between correlations and causal connections, we can draw on linear response theory. We consider, as in the article [7], a three-dimensional vector $\boldsymbol{x} = (x, y, z)$ whose time evolution is given by the system:

$$x_{t+1} = ax_t + \epsilon y_t + b\eta_t^{(x)}, \tag{52}$$

$$y_{t+1} = ax_t + ay_t + b\eta_t^{(y)}, \tag{53}$$

$$z_{t+1} = ax_t + az_t + b\eta_t^{(z)}. \tag{54}$$

This system represents linear stochastic Markov dynamics at discrete times, where η's are independent Gaussian processes with zero mean and unitary variance. The parameters a, b and ϵ are constant. To simulate this system, we start from a generic random vector and evolve it over a sufficiently long period of time to approach equilibrium and reach a stationary state. To understand whether there is a causal relationship between two generic variables, we use an interventional approach. We need to analyse whether the perturbation of one variable at time 0 (for example δy_0) implies the average variation (different from 0) of another variable at time $t > 0$ (for example δz_t). In order to obtain accurate results, the analysis was performed on 10^5 trajectories, since there are different noise realisations and the variation of the variables must be small enough, as required by linear response theory. By analysing the average variation at each step for each variable, we get the response matrix. In this case, we want to find out whether y has a causal relationship with z. So we change the initial value of y by 1% and observe the average variation of z in the following steps. The analysis was carried out for different values of ϵ in order to obtain a quantitative description of the causal relationships and over many trajectories for avoiding the effects of the disorder. The parameters of a and b

are fixed at 0.5 and 1, respectively. By performing the numerical experiment and then graphing the outcomes, we obtain a series of diagrams. From these diagrams, we can deduce that causal relationships cannot be derived from the analysis of correlation functions alone. The two plots in Figure 1 give indeed different information.

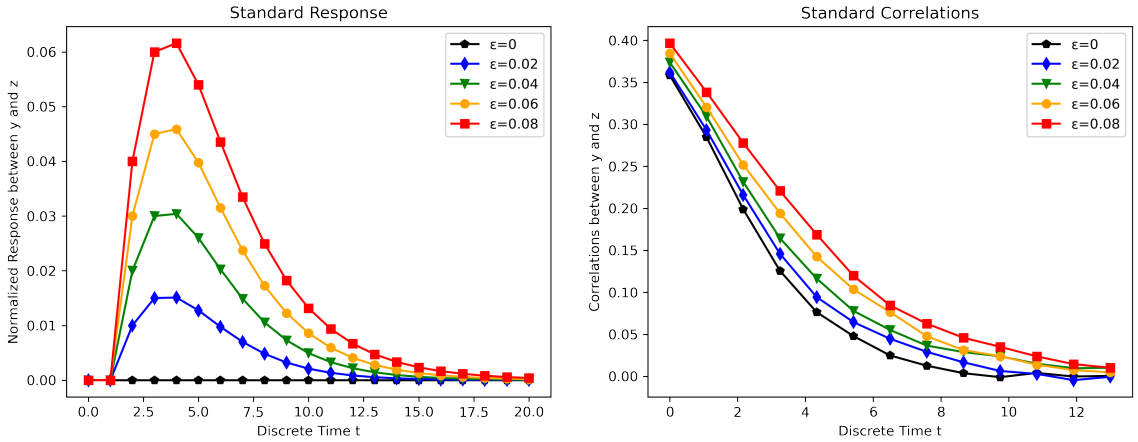

Figure 1 *Comparison between response and spurious correlations. The response has been rescaled with respect to the initial perturbation. The initial covariances were divided by the variance of y and z (calculated using a long series of data) to obtain the correlations.*

Since in this case, the system is linear, discrete-time and Markovian, we can derive the response function by simple operations on the covariance matrix [7] or by simple exponentiation of the propagator matrix A of the dynamics:

$$A = \begin{bmatrix} a & \epsilon & 0 \\ a & a & 0 \\ a & 0 & a \end{bmatrix} \qquad R_t = A^t = C_t C_0^{-1} \tag{55}$$

where R_t is the response matrix at time t, while C_t and C_0 represent the covariance matrix at time t and time 0 respectively. To verify the previous results in Figure 1, we calculate the response matrix for different times using the last formula (see Figure 2). The result is completely equal to the precedent one. This means that the two treatments are equivalent in this problem setting.

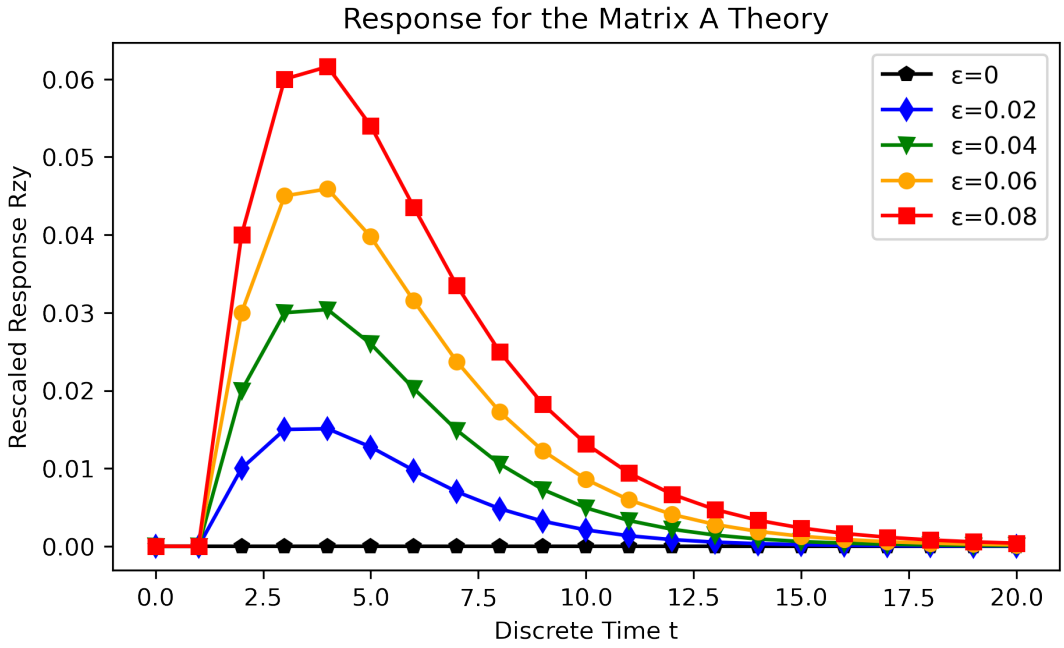

Figure 2 *Response matrix at time t calculated with the exponentiation t of the matrix A.*

II

Reconstruction of Propagation Matrix

A helpful approach to understanding the causal relationships within the examined system is to reconstruct the propagation matrix. There are several ways to reconstruct the propagation matrix of this linear system by assuming a time series of the data. In particular, we rely on independent samples: we thus wait, in each simulation, for the necessary time at each time series to obtain decorrelated samples.

The first method is to use Multiple Linear Regression, a statistical technique used to model the relationship between a dependent variable and multiple independent variables:

$$\begin{pmatrix} x_t \\ y_t \\ z_t \end{pmatrix} \xrightarrow{Model} \begin{pmatrix} x_{t+1} \\ y_{t+1} \\ z_{t+1} \end{pmatrix}$$

The second method relies on the calculation of the correlations across different variables of the system. More specifically, we turn our attention to the calculation of the mean value derived from the equations that capture the temporal evolution of the system.

$$V = \begin{bmatrix} E\left[x_{t+1}x_t\right] & E\left[x_{t+1}y_t\right] & E\left[x_{t+1}z_t\right] \\ E\left[y_{t+1}x_t\right] & E\left[y_{t+1}y_t\right] & E\left[y_{t+1}z_t\right] \\ E\left[z_{t+1}x_t\right] & E\left[z_{t+1}y_t\right] & E\left[z_{t+1}z_t\right] \end{bmatrix} \qquad C = \begin{bmatrix} E\left[x_tx_t\right] & E\left[x_ty_t\right] & E\left[x_tz_t\right] \\ E\left[y_ty_t\right] & E\left[y_ty_t\right] & E\left[y_tz_t\right] \\ E\left[z_tx_t\right] & E\left[z_ty_t\right] & E\left[z_tz_t\right] \end{bmatrix} \tag{56}$$

$$V = A \cdot C \quad \text{reversing:} \quad V \backslash C \ = A \tag{57}$$

The third method is to use a neural network to learn the coefficients. We start with the simple Multi-Layer Perceptron (MLP). A multilayer perceptron is a class of artificial neural networks characterised by a layered structure with multiple interconnected nodes or neurons. This neural network architecture consists of an input layer, one or more hidden layers and an output layer. Each layer consists of several neurons that are densely connected to the neurons of the neighbouring layers. In this case, it is characterised by an architecture with only two layers of three neurons each. The activation function is linear and the regularisation function is not used in this first step.

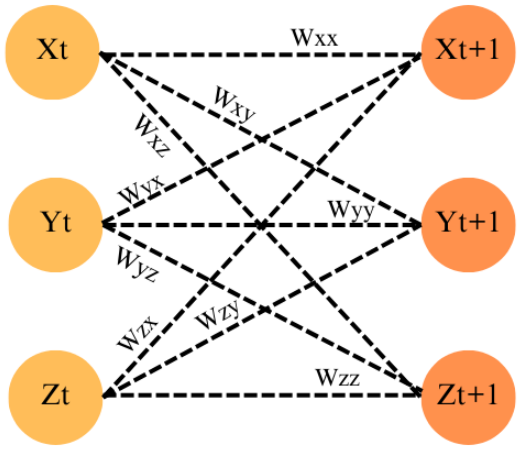

Figure 3 *Architecture of Multi-Layer Perceptron without hidden layers, three neurons in the input layer and three neurons in the output layer.*

By observing the errors of the coefficients of the propagation matrix obtained by the three previous methods, we will develop a better method to deal with the problem. We start considering the absolute error over each coefficient of the propagator matrix for a generic trajectory in Fig. 4

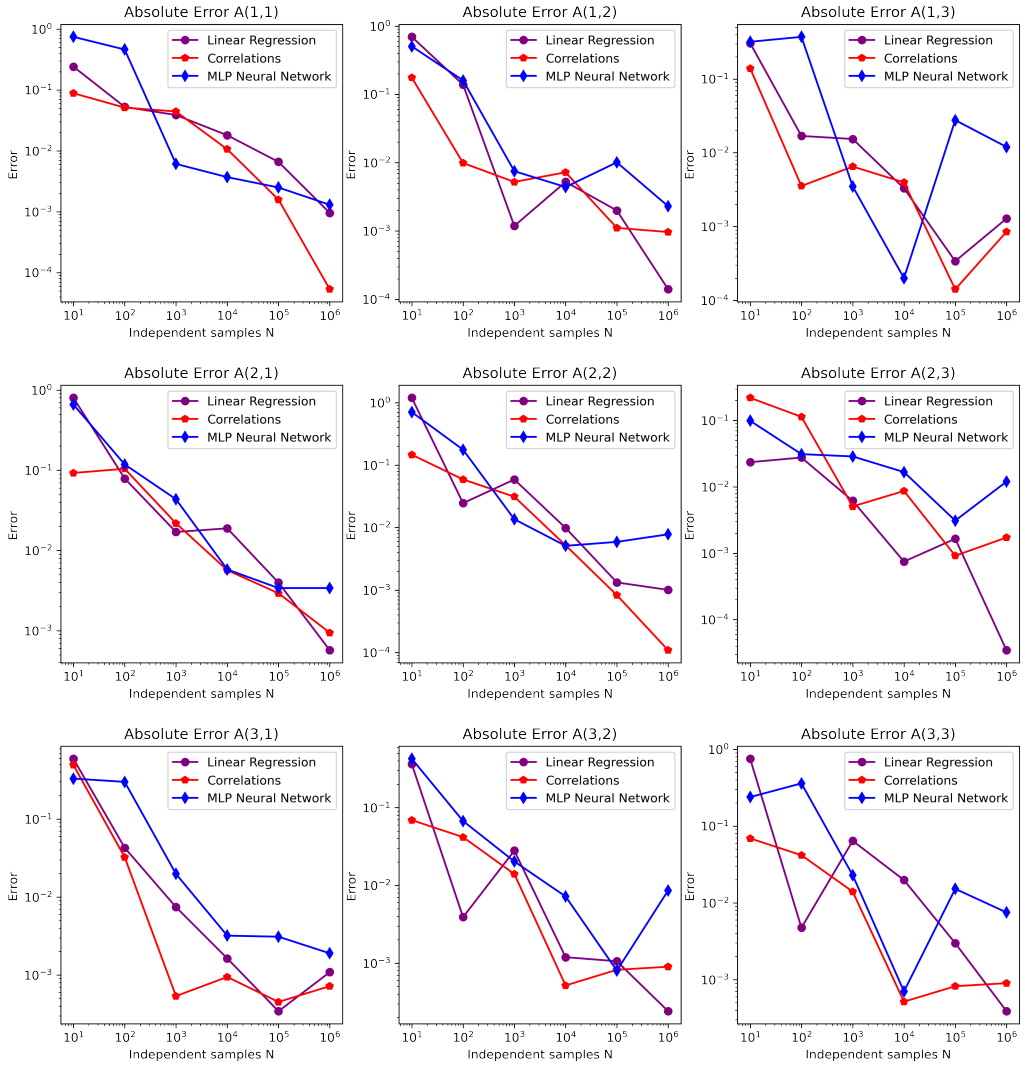

Figure 4 *Plot of the absolute errors obtained over each coefficient of the matrix A using the methods of multiple linear regressions, correlations and multi-layer perceptron as a function of different numbers of independent samples N.*

The general error for the three methods, obtained by the square root of the squared sum of the single error over each coefficient, is given in Fig. 5. The method of Linear Regression and the one of Correlations give excellent results and achieve better accuracy as the number of independent samples increases.

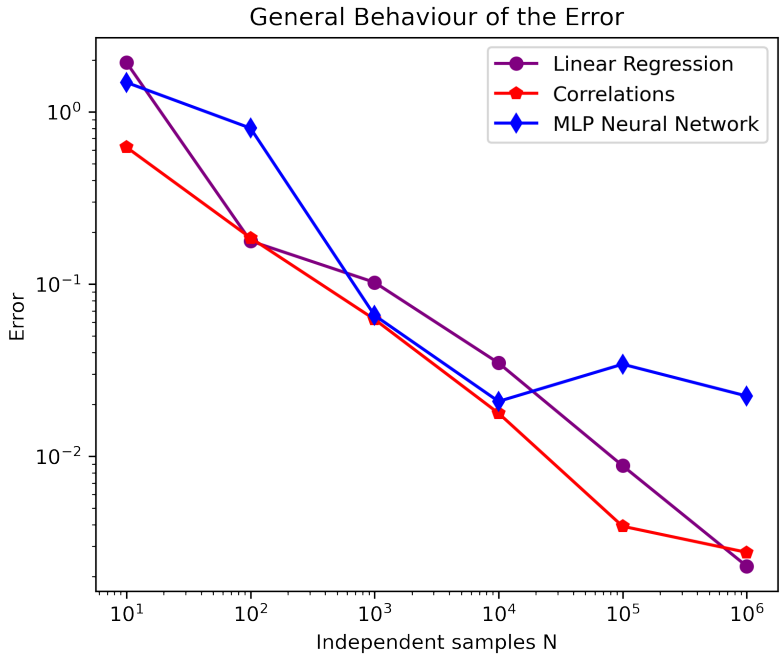

Figure 5 *Plot of the square error of the reconstruction of the matrix A for the three different methods. The resulting error for each method was calculated for a different number of training samples.*

On the contrary, we can see from the data that, if we increase the value of N for the MLP too much (in particular when we go over 10^4 independent samples), there is a possibility of overfitting the training data, leading to an increase in the prediction error. To mitigate this problem, we employ the Lasso Regularization function, also known as L1, to learn the parameters.

Lasso and Sparse Regression

L1 Regularization, also called Lasso Regularisation, is a technique used in machine learning and statistical modelling to introduce a penalty term that encourages the model to select a sparse set of features. It is usually used to prevent overfitting and to improve the generalization ability of the model. In L1 regularization, the penalty term added to the loss function is proportional to the absolute values of the coefficients of the model. Mathematically, it can be expressed as follows:

$$L1Penalization = \lambda \sum_{i=1}^{n} |w_i| \tag{58}$$

Where lambda (λ) is the regularization parameter that controls the strength of regularization, and $\sum_{i=1}^{n}|w_i|$ denotes the sum of the absolute values of the coefficients or weights of the model. By adding the L1 regularization term to the loss function, the optimization algorithm attempts to minimize both the loss and the size of the coefficients. As a result, L1 regularization forces many coefficients to become exactly zero, effectively performing feature selection by eliminating irrelevant or redundant features. Sparsity makes the model more interpretable by highlighting the most important features for prediction [10]. In this way, we can use the sparse matrix to distinguish the variables that have a causal effect on others and those that do not. In this way, we determine the optimal lambda parameter for the optimization task: we identify the range in which we can obtain good results and solve the problem by comparing different values of the parameter, as shown in Fig. 6.

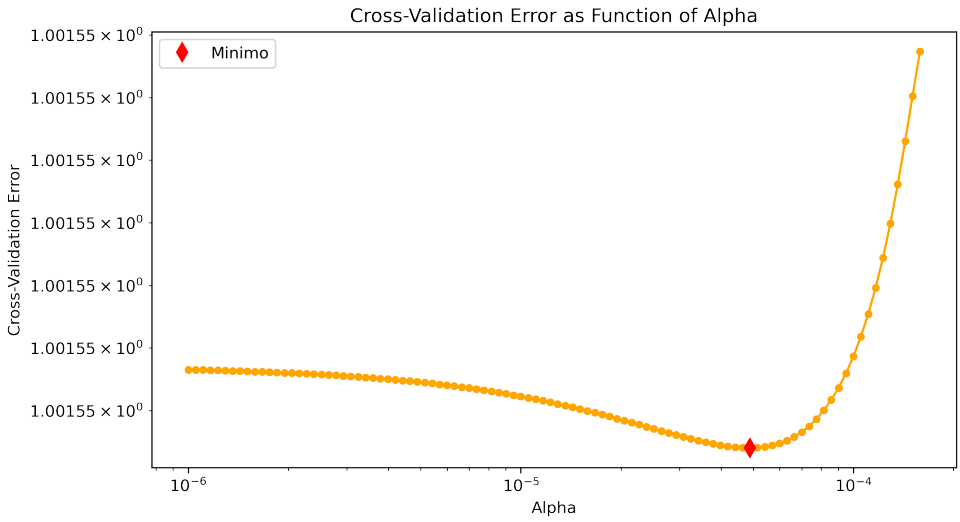

Figure 6 *Plot of the computational process employed to determine the optimal value of λ used in Lasso regularization for parameter calculation.*

Once we have identified the variables that play a role in the cause-effect relationships of the systems, we can apply the aforementioned methods to achieve higher accuracy in determining the parameters. In particular, we can analyze the problem a second time by assuming a situation where the coefficients that play no role in causality are set to 0. Alternatively, we can rely on the coefficients obtained through the optimisation process performed with the regularisation function. In this particular case, we have observed a significant improvement in the accuracy of the determination of the quantitative coefficient that captures the dynamics of the system. This remarkable improvement can be seen in the graphical representation in Fig. 7.

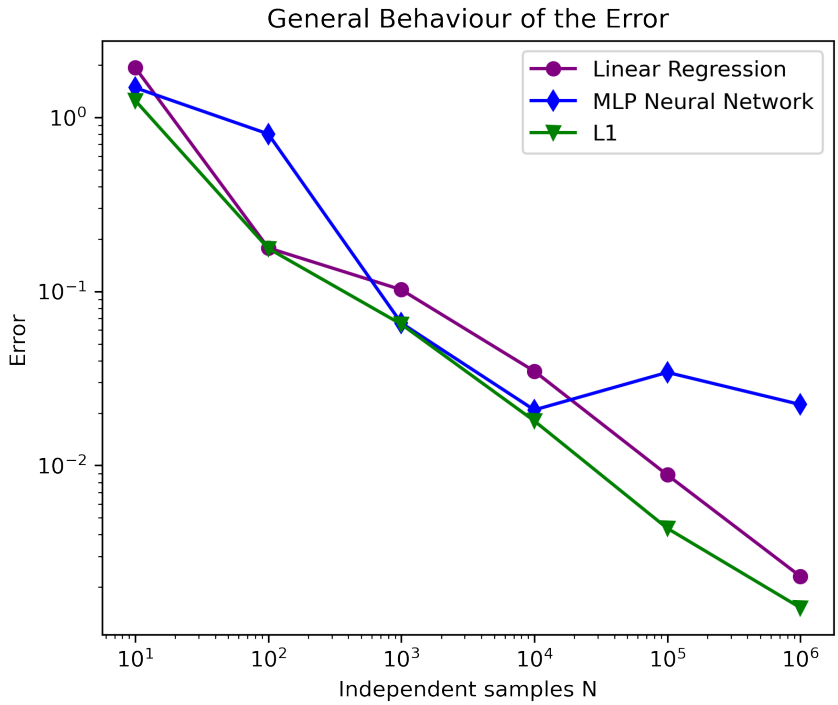

Figure 7 *Plot of the optimal error obtained by the Lasso regularization method.*

Of particular note is the significant improvement in accuracy seen in the contrast between the green and blue lines depicted in the Fig. 7, which is due to the implementation of L1 Regularization. This method entails a particular refinement in which the value 0 is deliberately and precisely assigned to those variables that do not exert a causal influence on the prediction of future values of other variables (as opposed to nominal or small values, as in the previous methods). This strategic move leads to the emergence of a sparse matrix, highlighting the importance and precision of the relevant features within the data. So, according to the results, among the available methods, the last approach is the one with which we can achieve the best results compared to the alternative techniques.

Chapter IV

Non-Linear System

Now, let's shift our focus to a non-linear system involving the consideration of three interacting particles in one dimension. These particles, described by variables x, y and z, are under the influence of a quartic potential, and we assume a dynamics characterized by overdamping. The evolution of the system state is determined by the following equations:

$$\dot{x} = -U'(x) - k(x - y) + b\eta^{(x)}, \tag{59}$$

$$\dot{y} = -U'(y) - k(y - x) - k(y - z) + b\eta^{(y)}, \tag{60}$$

$$\dot{z} = -U'(z) - k(z - y) + b\eta^{(z)}. \tag{61}$$

with the potential

$$U(x) = (1 - r)x^2 + rx^4 \tag{62}$$

In this context, k and b are considered fixed constants, while η represents Gaussian noise with 0 mean and variance equal to the used time step. The parameter r serves as an indicator of the degree of non-linearity within the dynamics. Specifically, when r is equal to 0, the external potential U has a harmonic nature. With values of r greater than 1, on the other hand, the potential U takes on a pronounced double-well form, as shown in Fig. 8.

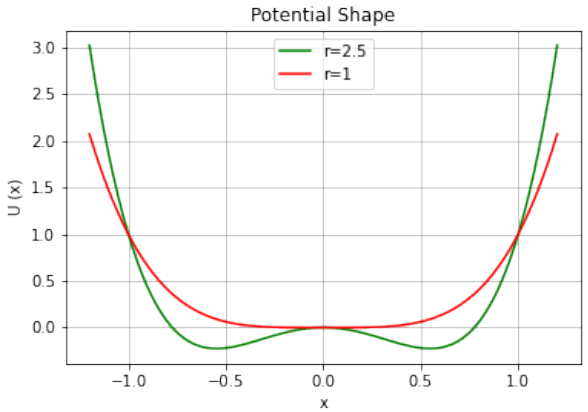

Figure 8 *Plot of potential shape for different values of r.*

Our main focus is to identify and establish the causal connections that exist among the variables involved. Additionally, we aim to evaluate the accuracy and validity of the formula $R_t = C_t C_0^{-1}$ used in the linear case in this particular non-linear system. Consistent with the study conducted in the article [7], we consider two different values of r for our analysis: 1 and 2.5. For each value, we proceed to measure the response obtained by the theoretical formula and the response obtained by multiplying the covariance matrix. As part of our analysis, we set the constants $b = 1$ and $k = 1$, assuming a time step of 0.001 seconds and a total time span of 5 seconds. To observe the response in the non-linear dynamics domain, we introduce a perturbation of 0.01 on variable x at time 0 and then examine the effects on variable z at future time steps. We used a stochastic Heun integrator to perform the numerical simulations. Starting from the initial problem:

$$\dot{\boldsymbol{x}} = f\left(t, \boldsymbol{x}\left(t\right)\right) + g\left(t, \boldsymbol{x}\left(t\right)\right) \qquad \boldsymbol{x}\left(t_0\right) = \boldsymbol{x}_0 \tag{63}$$

where $\boldsymbol{x}\left(t\right)$ represents the state of the system at time t, f is a deterministic function and g is the stochastic function. x_0 is the initial condition at time t_0: we obtain it by initialising the system in a random state and then bringing it to equilibrium by simulating the system for a sufficiently long period. The procedure for calculating the numerical solution is to first calculate the intermediate value $\tilde{\boldsymbol{x}}_{i+1}$ and then the final approximation $\boldsymbol{x}_{i+1}$ at the next integration point. It is described by the equations:

$$\tilde{\boldsymbol{x}}_{i+1} = \boldsymbol{x}_i + dt\left[f\left(t_i, \boldsymbol{x}_i\right) + g\left(t_i, \boldsymbol{x}_i\right)\right] \tag{64}$$

$$\boldsymbol{x}_{i+1} = \boldsymbol{x}_i + \frac{dt}{2}\left[f\left(t_i, \boldsymbol{x}_i\right) + f\left(t_{i+1}, \tilde{\boldsymbol{x}}_{i+1}\right)\right] + \frac{dt}{2}\left[g\left(t_i, \boldsymbol{x}_i\right) + g\left(t_{i+1}, \tilde{\boldsymbol{x}}_{i+1}\right)\right] \tag{65}$$

We undertake and iterate this particular system numerous times over several trajectories, assuming that the dynamics of the system under consideration correspond to ergodicity. Fig. 9 shows a comparison between the response calculated by intervention measures (the true response, represented by a red line) and the response obtained by correlation analysis (represented by a black line). This analysis is done for both values of the non-linear parameter. When $r = 1$, we get perfect results as in the linear case and can therefore use this approach in this very weak non-linear regime. In the scenario where the non-

linear contribution is sufficiently small (as in the case of $r = 2.5$), the linearized response (obtained by multiplying the covariance matrices) continues to provide meaningful insights into the causal relationships among the system variables.

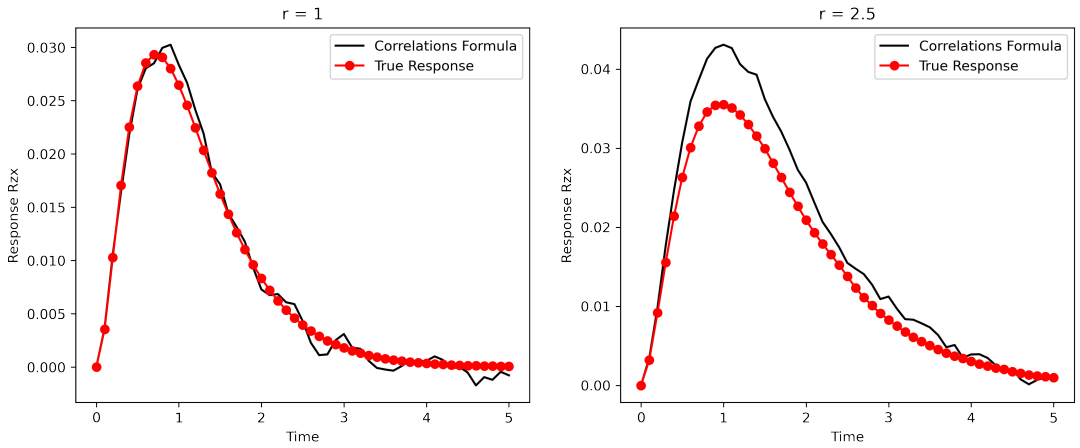

Figure 9 *Plot of response function R_{zx} in the non-linear dynamics for different values of r; each plot has been obtained by averaging over 10^6 trajectories.*

However, as the non-linear contribution increases, the task becomes more difficult because there is some uncertainty in the approximation of the true curve. Already for $r = 2.5$, we get some imperfections in the approximation of the curve, which means that a more precise approach to the problem is necessary.

I

Non-Linear Multiple Regression

Non-linear multiple regression is a statistical technique used to model complex relationships between multiple independent variables and a dependent variable. Unlike linear regression, which assumes a linear relationship between variables, non-linear multiple regression allows the modelling of non-linear relationships between variables. This relationship is represented by a function and can take various forms, such as polynomial, exponential, logarithmic, or trigonometric functions. The choice of the non-linear function depends on the underlying data and the problem at hand. This technique helps obtain a robust model whose predictions are reliable and in line with the trend that the data have followed in the past. Our analysis involved the use of a training data set consisting of approximately 10^6 independent trajectories. These trajectories were carefully generated with the non-linear parameter r set to 2.5.

Through the application of a non-linear regression approach, we formulated a polynomial representation and estimated the coefficients for each polynomial feature from degree 0 to degree 3. This gave a total of 20 coefficients, since in this case, we are dealing with three variables up to degree 3. In the general case, where we are dealing with n variables up to degree d, the number of coefficients is given by the formula:

$$\frac{(n+d)!}{n!\ d!} \tag{66}$$

1	2	3	4	5	6	7	8	9	10
const	x	y	z	x^2	xy	xz	y^2	yz	z^2
11	12	13	14	15	16	17	18	19	20
x^3	x^2y	x^2z	xy^2	xyz	xz^2	y^3	y^2z	yz^2	z^3

Threshold Method

With this methodology, we can derive discrete results that provide information about the causal relationships between the variables. By adjusting the threshold, we can obtain excellent results. However, we must remember that we are dealing with systems whose structures are not known a priori. Therefore, it is advisable to move to a more comprehensive and integrative approach. In this particular approach, we consider that the smaller coefficients contributing to the dynamics are of the order of 10^{-3}. Therefore, we set a threshold of $0.5 \cdot 10^{-3}$ power, which corresponds to half the time step (0.001 seconds), to eliminate coefficients that do not contribute to the dynamics and to obtain a sparse matrix. In this way, we can calculate the error by evaluating the sum of the squares of the differences between the coefficients of the obtained matrix and those of the original matrix, which describe the dynamics in this higher-dimensional space. This particular approach is suboptimal, and we have the opportunity to apply a novel method based on the results of the linear system.

Lasso Regularization-Threshold Method

Similar to our approach in the linear case, we used different values of the parameter lambda to perform this analysis by using the L1 regularization function, with the overall goal of obtaining a sparse matrix. We then created a graphical representation, visually shown in the Fig. 10. Following the method

discussed before, we set the threshold value to $0.5 \cdot 10^{-3}$, which is the same as the approach used in the previous practice. Using this modified approach, we found that the model exhibited a commendable level of accuracy in effectively identifying all causal factors associated with each variable within the expanded space.

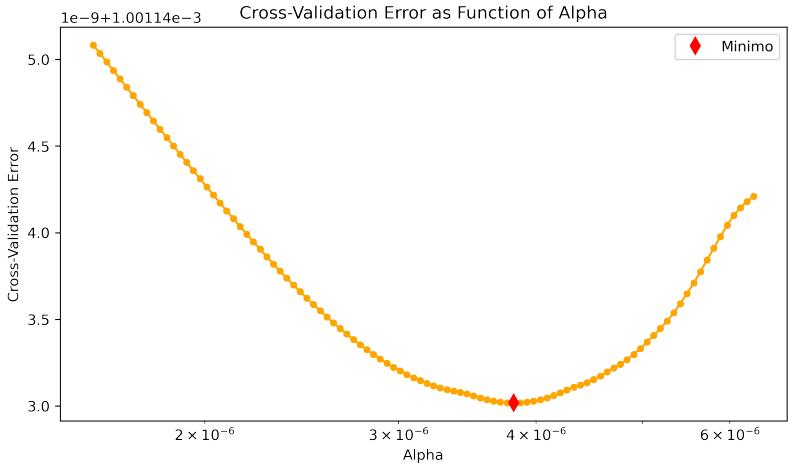

Figure 10 *Plot of the computational process employed to determine the optimal value of λ used in Lasso regularization for parameter calculation in the non-linear system.*

More precisely, the function establishes the value of zero for all variables that do not contribute to determining the future state of the system and, of consequence, the variables that play a central role in causality are emphasised. In terms of quantitative analysis, the model has an average error rate of $3.4 \cdot 10^{-4}$ per coefficient. This result can be considered favourable for a system of this nature, characterised by coefficients of such small magnitude.

Linear Response Connection

By persisting in this particular context in a high-dimensional space, we can achieve another important result. In particular, we can compute the covariance matrices at different time intervals: At the moment, we derive matrices with dimensions 20x20, since the system is to be characterised by 20 different features corresponding to polynomial features up to degree 3. Employing the same method we used for the linear system, we calculate the response using the given formula. To avoid a singular matrix, it is important to eliminate the first row and column that correspond to the constant polynomial feature. In this way, we obtain matrices with reduced dimensions that measure 19x19. By

multiplication of covariance matrices, we obtain response matrices at time t, where we can selectively focus only on the first three rows, which refer to the variables x, y and z and ultimately define the state of the system. In this case, capturing the response requires summing the variables that depend on x, thus contributing to the dynamics of the system. In particular, the polynomial features x and x^3 play a central role and force us to aggregate the corresponding response generated by their influence. As can be seen from Fig. 11, our careful efforts have led to an accurate representation of the response curve, which, thanks to our extended dimensional space, restores the deep connection to the theory of statistical physics.

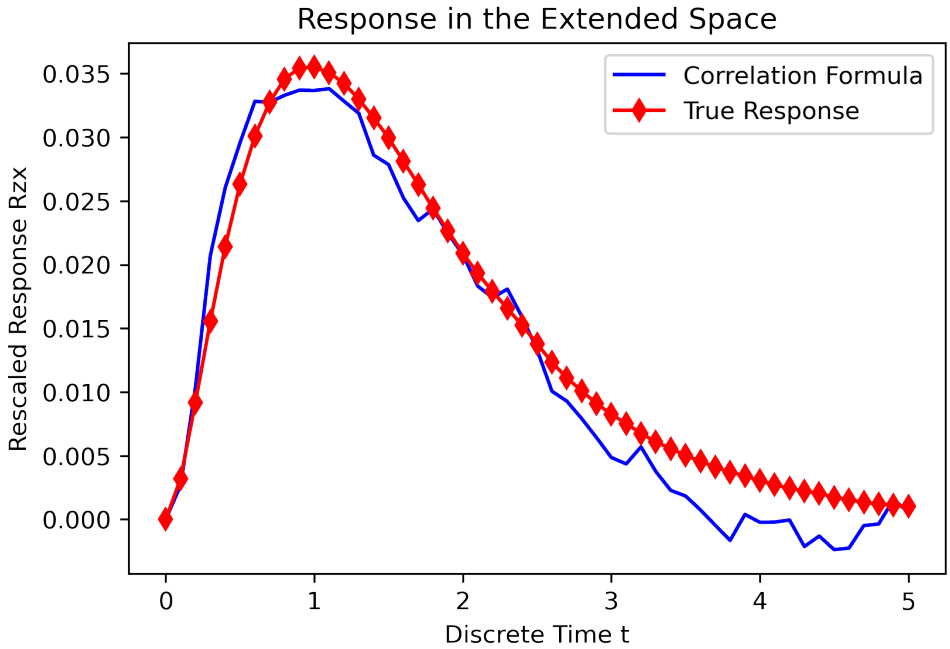

Figure 11 *Comparison between the reconstruction of the rescaled response in the extended space through the use of statistical physics formula (blue line) and real response (red line)*

This diagram has been carefully constructed by an extensive simulation involving an amount of more than $2.5 \cdot 10^6$ individual trajectories. It is imperative to capture such a large amount of data when working in the extended spatial domain, as it is essential to compensate for the inherent complexity and subtleties that arise to achieve the same level of precision and statistical convergence that mirrors the results that can be obtained in the conventional spatial framework.

II

Machine Learning and Statistical Physics Method

In a broader context, when confronted with a generic non-linear system, we have obtained remarkable results in determining the response function through an extensive study of correlations and ML techniques. This methodology involves the use of non-linear multiple regression to determine which variables within the expanded parameter space exert an influence on other variables. Then, the response is derived by performing matrix multiplications on the covariance matrices within this expanded parameter space. Through this process, we systematically derive the response of a variable at each successive time step, that follows the change in another variable, even in scenarios where the dynamics of the system are governed by a non-linear evolutionary law. In this way, we can access the system and examine how it responds through a numerical experiment based on its variables and relationships.
However, a limitation inherent in this approach becomes apparent when confronted with systems characterized by hidden variables in their dynamics. In such cases, this method becomes inaccurate and insufficient to decipher the influence of a single variable in the face of change, necessitating the search for the existence of alternative methods.

Low Variance Case

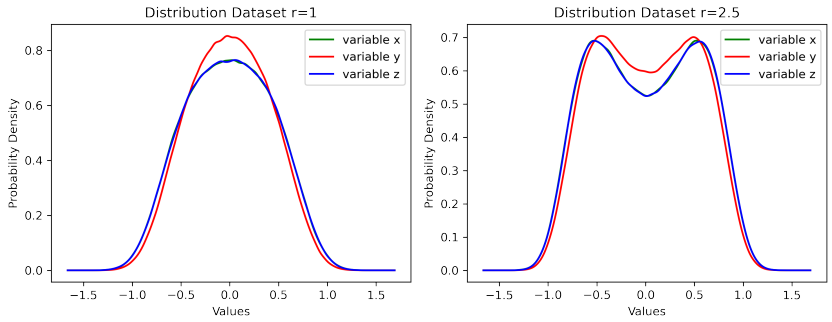

Figure 12 *Probability distribution of the variable x, y and z computed for different values of non-linear parameter r*

When we consider a system characterised by a bounded variance and values, in particular in the range from 0 to 1, analogous to the system studied before whose probability distribution is shown in Fig. 12, we can leverage a polynomial approximation strategy based on the principles of McLaurin's De-

velopment theory of evolution to elucidate each non-linear interrelationship between the individual variables. Specifically, within this operational regime, it is noteworthy that any non-linear relation or function can be carefully described in terms of polynomial terms. In this context, conjecture about the specific functional form underlying the data generation process is unnecessary, since the comprehensive summation of polynomial terms effectively includes the underlying mechanisms.

To ensure a high degree of prediction accuracy (which depends each time on the particular problem we are analyzing), we perform an analysis in which the power of each variable is examined up to the seventh degree, by assuming that there are no mixed terms. In the context of a system with three variables, this endeavour culminates in the derivation of a 22x22 matrix, accounting for the constant variable as well. Subsequently, employing the Machine Learning (ML) and Statistical Physics (SP) Methodology, we can reconstruct the system's evolution laws and thus extract the system's response from the collected data.

1	2	3	4	5	6	7	8	9	10	11
$const$	x	x^2	x^3	x^4	x^5	x^6	x^7	y	y^2	y^3
12	13	14	15	16	17	18	19	20	21	22
y^4	y^5	y^6	y^7	z	z^2	z^3	z^4	z^5	z^6	z^7

III

Recurrent Neural Network

Recurrent Neural Networks (RNNs) are a special architecture in the field of artificial neural networks. They are widely used in the fields of machine learning and natural language processing (NLP) and enable the modelling of sequential data. Such data can span different domains, from text sequences to time series. RNNs are designed (as we can see in Fig. 13) to process data with sequential structures where preserving the relationship in the information is of significant importance. RNNs offer important advantages over their counterparts, feed-forward neural networks:

Sequential Data Processing: RNNs perform well on tasks involving sequential data. They are able to process input data of varying lengths and capture dependencies over time.
Recurrent Connections: RNNs use recurrent linkages, a mechanism in which the output of a previous step serves as the input for the subsequent step. This mechanism allows RNNs to maintain a hidden state that facilitates the capture and storage of information from previous steps in the sequence, enabling memory and reuse.
Elaboration New Results: When generating new results, RNNs incorporate both current and historical data, taking into account the entire temporal evolution of the data.

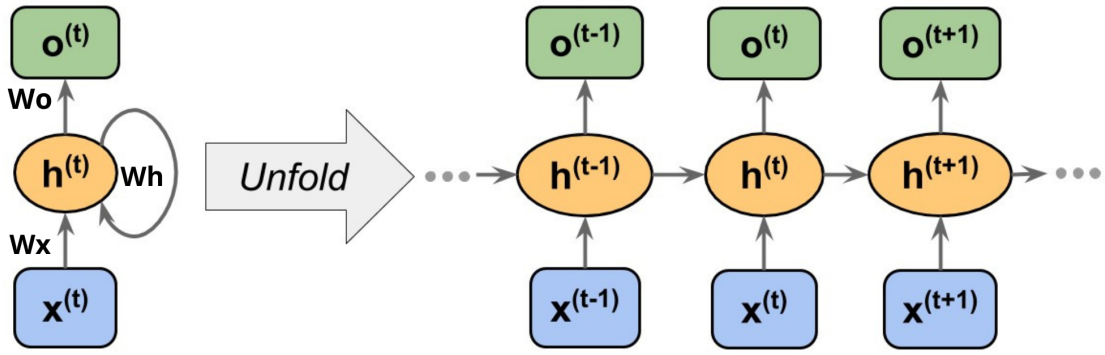

Figure 13 *Structure of a general one-layer RNN, where x corresponds to the input, h to the hidden state and o to the output. The weighting matrices W are the same at each time step.*

Nevertheless, there are some disadvantages associated with the use of this particular class of neural networks. The computational requirements resulting from the recurrent nature of RNNs lead to relatively long processing times. In addition, training RNNs tends to be more complicated and resource-intensive compared to feed-forward network models. Next, not all activation functions can be used to process longer sequences of information effectively. There are two important problems associated with RNNs: Gradient Exploding and Gradient Vanishing. Gradient Exploding occurs when gradients become excessively large during backpropagation, resulting in unstable training. Conversely, Gradient Vanishing occurs when gradients decrease to the point where the network has difficulty capturing long-term dependencies contained in sequential data. To overcome these challenges, advanced RNN variants such as Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRUs) networks have been introduced. LSTMs and GRUs incorporate gating mechanisms designed to mitigate problems associated with gradients by allowing the network to selectively retain or discard information from previous time steps.

LSTM (Long Short Term Memory)

Unlike conventional RNN, the hidden layers of LSTM consist of concrete memory cells, and the flow of data is controlled by computational units called "gates". Within each cell, three different types of gates can be activated: the first is responsible for resetting the state of the cell, the second updates the state and the third is specialised in changing the hidden units.

GRU (Gated Recurrent Unit)

GRU represents a variant that features a relatively simplified architecture compared to LSTMs and offers higher computational efficiency. Similar to LSTMs, GRUs do not contain separate memory cells, which reduces architectural complexity. GRUs consist of two gates: the reset gate and the update gate. The reset gate controls which information from the previous time step should be reset or forgotten. It takes into account the input from the previous hidden state and the current input and ultimately returns values between 0 and 1 for each element in the hidden state. This mechanism enables the network to recognise the relevance of information to the current time step. The update gate determines the extent to which the previous hidden state should be preserved and combined with the new candidate state. Similar to the reset gate, it considers the previous hidden state and the current input and generates values between 0 and 1 for each element in the hidden state.

Causality RNN Model

By using this particular neural network architecture, we have created an RNN that consists of a GRU layer followed by a linear layer. The inclusion of a linear layer adds flexibility to the structure of the model. After the GRU layer has processed sequential information, the linear layer can identify complicated relationships between the information output. This property proves particularly valuable when dealing with non-linear or complicated relationships. The neural network starts with an initialisation characterised by an input dimension of 3 and a hidden dimension of 3 within the GRU layer. A linear layer is then used to transform the output into a 3-dimensional representation. As a loss function, we chose the MSE (minimum squared error), which quantifies the discrepancy between the model's predictions and the target values, and an Adam optimiser to serve as a tool to update the model weights throughout the training process. During the simulation of the system, we observed the phenomenon (the same as the non-linear system) for an approximate duration of 10 minutes. During this period, we recorded the positions of each of the three particles at 1 millisecond intervals, keeping the non-linear parameter r at a fixed value of 2.5. This observation generates a dataset of $5 \cdot 10^5$ sequences, each consisting of 250 time steps. For the purpose of cross-validation, we reserved 90% of the data for training and 10% for testing. It is important to note that our dataset is structured as a 3D tensor $(5 \cdot 10^5, 250, 3)$. After these preliminary steps, we start training the RNN and then evaluate its predictive capabilities over 100 different trajectories. The RNN model is able to detect the causal link between variable x and variable z by examining the response function. It is worth noting that the RNN model is sensitive to the initial conditions, especially in terms of the magnitude obtained by averaging the responses over the 100 trajectories. However, despite this sensitivity, the RNN model reproduces the same qualitative response as the original system (simulated without noise in this case). This consistency in the qualitative output confirms the efficiency of the RNN in capturing the causal relationship between the two variables of interest. Nevertheless, as we can see in Fig. 14 two important differences emerge when comparing the RNN model with the original system:

- Quantitatively, the response produced by the RNN model differs from the results obtained in the original system;
- Temporally, the time interval in which the response is observed in the RNN model is much shorter and is 0.2 seconds, as opposed to the 3.5 seconds of the original system.

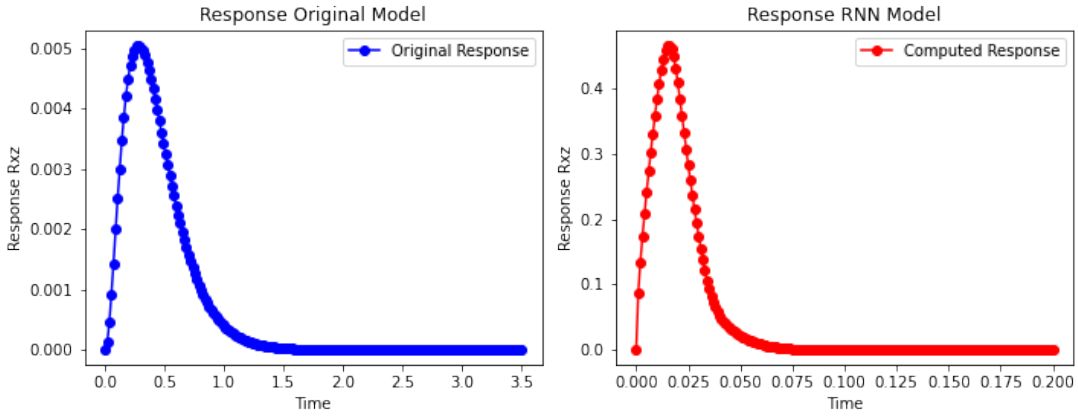

Figure 14 *Comparison between the original response (blue line) and the response computed by the trained RNN (red line)*

The system is rather complex and finding the causality relation is a good achievement, but it would be very useful to find a more accurate method or network to achieve better precision in the quantitative description of the response and in the temporal decorrelation of the initial intervention.

Chapter V

Conclusions

In conclusion, this study has explored the integration of Linear Response Theory with the application of machine learning techniques (ML) in the area of causality and offers insights into the potential and limitations associated with the use of ML algorithms to infer causal relationships. We first presented Linear Response Theory and the Fluctuation-Dissipation theorem to introduce the response function and explain the concept of causality. Our comprehensive analysis included a thorough examination and discussion of the two fundamental approaches to causality, namely interventional and observational. Subsequently, we turned our focus to a stochastic linear Markov model, recognizing that mere correlations are not sufficient to elucidate causal relationships. In this context, where the causal relationships are known, we demonstrated the effectiveness of the methods of ML in identifying causal factors and making accurate predictions based on observational data. In our analysis, we employed three different methods: Linear Regression, a Correlation mathematical approach, and a simple Multi-Layer Perceptron (MLP). It is worth noting that using the Multi-Layer Perceptron resulted in an increased error due to overfitting when we increased the number of independent samples used for the cross-validation task. To address this problem, we successfully solved it by applying Lasso Regularization, which led to very accurate results. Thus, we have effectively addressed the challenge of establishing the causal connection within the linear problem. Building upon the findings from the linear response analysis, we have extended our study to investigate the dynamics of a non-linear system. We recognize that many real-world systems exhibit nonlinear behaviour, and understanding their causal relationships is crucial. Initially, we sought to ascertain the extent to which the linear response analysis provides meaningful insights in the presence of nonlinearity. Subsequently, we turned our attention to addressing this challenge by harnessing the power of ML, by trying and analyzing new techniques. To confront this task, we have included non-linear regression models and RNN, to capture the intricate interactions and identify non-linear cause-effect relationships within the system. By applying these methodologies and integrating them with the principles of statistical physics, we have been able to find out the response for a non-linear system. In this direction, the work also discusses the limitations and challenges in non-linear systems associated with applying machine learn-

ing in causal analysis. The use of machine learning for causal analysis paves the way for future research directions to enhance effectiveness and reliability in establishing cause-effect relationships. Moving forward, our investigations will extend beyond linear and nonlinear analysis to incorporate hidden variables into the causal modelling framework. This extension is crucial as hidden variables can exert a substantial influence on the observed relationships between variables, offering a deeper comprehension of causal connections within complex systems.

References

[1] Ryogo Kubo. Statistical-mechanical theory of irreversible processes. i. general theory and simple applications to magnetic and conduction problems. *Journal of the Physical Society of Japan*, 12(6):570–586, 1957.

[2] Robert Zwanzig. *Nonequilibrium statistical mechanics*. Oxford university press, 2001.

[3] Umberto Marini Bettolo Marconi, Andrea Puglisi, Lamberto Rondoni, and Angelo Vulpiani. Fluctuation–dissipation: response theory in statistical physics. *Physics reports*, 461(4-6):111–195, 2008.

[4] Lars Onsager. Reciprocal relations in irreversible processes. i. *Physical review*, 37(4):405, 1931.

[5] Katerina Hlaváčková-Schindler, Milan Paluš, Martin Vejmelka, and Joydeep Bhattacharya. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441(1):1–46, 2007.

[6] Brian Lindner, Lidia Auret, Margret Bauer, and Jeanne WD Groenewald. Comparative analysis of granger causality and transfer entropy to present a decision flow for the application of oscillation diagnosis. *Journal of Process Control*, 79:72–84, 2019.

[7] Marco Baldovin, Fabio Cecconi, and Angelo Vulpiani. Understanding causation via correlations and linear response theory. *Physical Review Research*, 2(4):043436, 2020.

[8] Judea Pearl. *Causality*. Cambridge university press, 2009.

[9] Pearl Judea. An introduction to causal inference. *The International Journal of Biostatistics*, 6(2):1–62, 2010.

[10] Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre GR Day, Clint Richardson, Charles K Fisher, and David J Schwab. A high-bias, low-variance introduction to machine learning for physicists. *Physics reports*, 810:1–124, 2019.