

Building causation links in stochastic nonlinear systems from data

Original

Building causation links in stochastic nonlinear systems from data / Chibbaro, S., Furtlehner, C., Marchetta, T., Pantea, A., Rossetti, D.. - In: PHYSICAL REVIEW. E. - ISSN 2470-0045. - 113:(2026). [10.1103/kdyd-tc6n]

Availability:

This version is available at: 11583/3009064 since: 2026-09-26T18:30:33Z

Publisher:

APS

Published

DOI:10.1103/kdyd-tc6n

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

APS postprint/Author's Accepted Manuscript e postprint versione editoriale/Version of Record

(Article begins on next page)

Building causation links in stochastic nonlinear systems from data

Sergio Chibbaro

Université Paris-Saclay, CNRS, UMR 9015, LISN, F - 91405 Orsay cedex, France

Cyril Furtlehner

INRIA-Saclay, Université Paris-Saclay, LISN, F - 91405 Orsay cedex, France

Théo Marchetta

INRIA-Saclay, Université Paris-Saclay, LISN, F - 91405 Orsay cedex, France

Andrei-Tiberiu Pantea

Université Paris-Saclay, CNRS, UMR 9015, LISN, F - 91405 Orsay cedex, France

Davide Rossetti

*Politecnico di Torino, Corso Duca degli Abruzzi 24, I - 10129, Torino, Italy and
Université Paris-Saclay, UMR 9015, LISN, F - 91405 Orsay cedex, France*

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, offering new opportunities for uncovering hidden causal mechanisms and better understanding complex systems. In this work, we address the issue of detecting the intrinsic causal links of a large class of complex systems in the framework of the response theory in physics. We develop some theoretical ideas put forward by [1], and technically we use state-of-the-art ML techniques to build up models from data. We consider both linear stochastic and non-linear systems. Finally, we compute the asymptotic efficiency of the linear response based causal predictor in a case of large scale Markov process network of linear interactions.

I. INTRODUCTION

The general definition of causality may be slippery and has been a subject of scholar study since ever [3–5]. However, the recognition and clarification of some kind of cause-effect relationship between variables, events or objects are fundamental questions of natural and social sciences [6–8]. The motivation to analyze such a problem is very high, since in most practical cases we need to quantify the strength of a possible causal relation in order to explain past events, control present situations and predict future consequences [9–13]. One clear instance of the question is the interplay between carbon emissions and temperature, or to understand the possible relation between the contact with some substance et a following development of a disease. Causal analysis is a distinguished branch of statistics [7, 14], with a possible impact on machine learning [15].

The crucial point is to define properly the causality from a physical standpoint. In classical physics this idea is often associated with a principle of legal determinism, in which an antecedent uniquely determine the following state of the system, an idea rooted in classical mechanics [16, 17]. However, in this framework the notion of causality leads to considerable epistemological difficulties, and the notions of cause and effect are in fact not invoked. Instead, in a recent work [1] it has been shown that there is a conceptual parallelism between the statistical causal analysis and long-time response in physics [9]. There is still a crucial difference. In physics, the intervention on a system may cause some response only after some time, whereas statistical analysis of causality assumes immediate effects. In this sense, statistical analysis should be easier to work but generally not of physical relevance. The possibility to use the response theory to establish causal links is particularly interesting for those cases for which laboratory experiments are not possible, for instance in geophysical problems. As a matter of fact, in many realistic phenomena one is confronted to a time-series of few observables [2, 18].

On a general basis, we may formulate the problem as follows: a time series $\{x_{it}, x_{2t}, \dots, x_{dt}\}$ of d variables represent an observable system $\mathbf{x}_t$, and one wishes to determine unambiguously whether the behavior of x_i has been influenced by x_j during the dynamics, without knowing the underlying evolution laws. A first natural way to look at this problem is to look at correlations between variables $C_{ij}(t) = \langle x_{it}x_{jt} \rangle$, since a causal link is expected to manifest as a non-zero value for it for some $t > 0$, at least following intuition. Yet, it is well known that correlation does not imply causation, for instance, both x_i and x_j may be influenced by common-causal variables [7, 19].

More reliable tools to point out causal effects between two variables are the Granger causality (GC) test [20, 21], and transfer entropy (TE) approaches. GC allows to determine whether the knowledge of the past history of x_j increases the ability to predict future values of x_i . [22, 23]. TE is based on the definition of a degree of information exchange from x_j to x_i , which quantifies the loss of information about x_i that one experiences if $\{x_{jt}\}$ is ignored [24, 25]. However, from a physical point of view, these definitions may be not quite satisfactory. In physics, two variables are usually considered to have a mutual cause-effect relationship if an external action on one of them results in a change of the observed value of the second [1, 26]. As explained shortly, GE and TE provide a weaker information, whether, and to what extent, the knowledge of a certain variable is useful to the actual determination of future values of another. This kind of causality is called “observational”, while the stronger physical-oriented one is termed “interventional”. In a different framework, the latter is also favored in computer science by Pearl [7].

In some recent works [27, 28], the link between the interventional causal link and the physical response has been investigated, considering also applications of the utmost importance. However, some issues may hinder the possibility of using such approach, both from a theoretical and a practical point of view. This approach appears to be perfectly suited for linear Markov systems. It may fail in the case of nonlinear and/or non-Markovian systems, which however represent most of the relevant biological and physical ones. In particular, computing directly the response of a system implies the possibility to apply an infinitesimal perturbation to the system and observe the possible following statistical drift. The experiment must be repeated many times to have statistical convergence. Yet, in actual problems either we have a system we cannot act upon, either we have just a time-series and such experiments cannot be performed. Solely in the case of linear systems, covariance matrices are enough to reconstruct the entire response. Furthermore, from a more practical point of view, the data available are not necessarily enough to have a robust assessment of the response function, even in the linear case.

In this work, we address the issue of detecting causality in general nonlinear Markov or chaotic systems, from the physical perspective. The goal is to leverage the capability of machine learning algorithms to build up models from data. Machine learning (ML) has been notably developed in the field of computer vision [29], and it has brought new perspectives in the data assimilation and analysis of complex physical systems [30–33].

To the best of our knowledge, we present here the first study of the physical response problem in terms of a machine learning approach. In particular, we leverage this framework to establish causal links among dynamical variables. In this way, we go beyond previous studies and we generalize to very large Markov system and to a full nonlinear dynamics. We investigate whether it is possible to infer causal links purely from data, when these data are generated by complex noisy/chaotic systems. We consider both the case in which no a priori of the structure of the system is given, and when some physical information is used to design the data-driven approach. More specifically, we leverage advance ML technics to estimate the response function in various regimes linear and non-linear with or without cofounders; we set the basis to analyze the large scale behavior, by providing explicit formulas for a generic linear Markov process in the asymptotic limit, using the Random Matrix formalism. Since neural networks are very expressive nonlinear models, we are able to go beyond the linear response approximation considered in previous studies and to get at least qualitative estimates of the causal links, even in the general fully nonlinear cases.

The paper is organized as follows: In Section. II we discuss the two main different types of approaches to causal inference, namely observational and interventional, and how the latter can be virtually done via generalized response theory. In Section. III we recall the basics of generalized response theory, both for linear and non-linear settings. In Section. IV we revisit respectively the linear and non-linear stochastic examples studied in [27] with ML approaches. In Section. V we consider a setting with chaos by tackling the Lorentz model also ML techniques. Finally in Section. VI we analyze theoretically the efficiency of causal indicator based on linear response theory for large scale causal graphs with help of random matrix theory(RMT).

II. DEFINITION OF CAUSALITY

As anticipated in the introduction, a consistent and relevant definition of a causality relation may be slippery, and it is crucial to well define the terms of the problem. To obtain information about causal relations there are basically two broad types of approaches: the Interventional and the Observational one. Both provide some information about the presence or absence of cause and effect, but they differ in how this information is expressed. We quickly explain both approaches in the language of statistical physics, following the same line depicted in a recent work [27].

For the observational setting, two widely used methods to extract information are the Granger causality (GC) [20, 34, 35] for linear causal relations and the Transfer Entropy (TE) [18, 24], which generalizes the former to non-linear situations.

GC can be quantified and measured computationally, ususally in the context of linear regressive models. It 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 [36]. This is a regression-based interpretation of past data that compares the statistical uncertainties of two predictions. For example, let us consider the time series of two events $x_j(t)$ and $x_i(t)$, we can use a model where only the past history of x_j is included to predict future values of itself:

$$x_j(t) = \sum_{k=1}^T B_{j,k} x_j(t-k) + \epsilon_j(t) \quad (1)$$

where B is the autoregressive coefficient and ϵ_j is the prediction error. A second prediction may be obtained using a model, in which the past of x_i is also included in the model for predicting future values of x_j :

$$x_j(t) = \sum_{k=1}^T [A_{ji,k} x_i(t-k) + A_{jj,k} x_j(t-k) + \epsilon_{j|i}(t)] \quad (2)$$

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 . G-Causality may be quantified as:

$$F_{x_i \rightarrow x_j} = \log \frac{\text{Var}(\epsilon_j)}{\text{Var}(\epsilon_{j|i})} \quad (3)$$

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

TE derives causality from an information theory interpretation. It is a specific version of a Kullback-Leibler entropy for conditional probabilities. It is a non-parametric statistic that measures the amount of information exchanged between x and y . More specifically, it indicates the amount of uncertainty reduced in future values of y given knowledge of both the past values of x (and given past values of y) with respect to the knowledge of the past y alone [6, 18].

As affirmed by Granger himself, it is generally agreed that G-causality does not capture all aspects of causality, but enough to be worth considering in an empirical test [35]. Same may be said of Transfer of entropy [37]. A key observation regarding all these observational methods is their complete reliance on selecting the right variables. Causal factors excluded from the regression model cannot be reflected in the results. Consequently, Granger causality or Transfer Entropy should not be interpreted as direct representations of physical causal mechanisms [35]. In this sense, this type of causation is not necessarily satisfactory from a physical point of view and for this reason, is discarded in some situations in favour of a more ‘‘interventional’’ approach.

Interventional causality is more easily linked to the physical cause-effect relationship. The state of a variable is changed to manipulate the system and see if there is a reaction. Since we are in a statistical framework, it states that given a time series characterized by the vector $\mathbf{x}_t$ consisting of n entries, a perturbation of the variable x_j at time 0, $x_{j0} \rightarrow x_{j0} + \delta x_{j0}$, on average, causes a change in another variable x_{kt} with $t > 0$. In causal statistics [7, 39], interventional causality relies on causal Bayesian networks and directed acyclic graphs (DAG), and aims to understand the representation of the causal organisation and the response to external or spontaneous changes in variables. The main technical tool is the probability conditioned on an intervention, that is a change of some variable. This formalism permits to determine for a known graph whether one has sufficient data to correctly estimate the conditional probability, and in this case indicate what the conditional probability model would be for a modified model. The Pearl's approach may therefore be used to assess if enough data are available to answer a formalized question.

In [1], it was suggested that conceptually the Bayesian belief network used in causal statistics may be replaced in physical problems by a dynamical probabilistic model, which should contain the same dependencies. In this framework, the probability conditioned on a given intervention is related to the long-time response of the system to it. As a consequence, there is a link between the causality intended as a time-ordered relationship where a change in a variable determines a change on another variable at successive times. This idea can be mathematically formalized saying that there is a causal relation if, given a smooth function $\mathcal{F}(x)$, the relationship holds:

$$\frac{\overline{\delta \mathcal{F}(x_{kt})}}{\delta x_{j0}} \neq 0 \quad (4)$$

which means that a perturbation of the variable x_{j0} at the time 0 leads to a non-zero average variation $\mathcal{F}(x_{kt})$ (carried out over many realisations of the experiment) with respect to its unperturbed evolution [27]. It is possible to appreciate that the formulation given by Eq. (4) of the causality naturally leads to the statistical response of the system [9, 40, 41].

In this work we focus on this notion of interventional causality in the statistical physics framework.

III. RESPONSE THEORY FRAMEWORK

Starting from the general definition Eq. (4), a connection can be made with the response theory, when the system admits an invariant distribution and considering that the variation δx_{j0} is small enough.

We recall the main ideas and results of response theory. We focus in particular on the choice $\mathcal{F}(x_{kt}) = \delta x_{kt}$. Complements can be found in appendix, while a more detailed exposition can be found in Ref. [9]. Consider a Markov process $\mathbf{x}_t = (x_{1t}, \dots, x_{dt})$ of dimension d whose invariant p.d.f. ρ is smooth and nonvanishing. Given a (small) perturbation $\delta \mathbf{x}_0 = (\delta x_{10}, \dots, \delta x_{d0})$ at time $t = 0$, its effects at time t is obtained by measuring the difference between the vector $\mathbf{x}_t$ in the original dynamics and in the perturbed one, on average. Formally we compute

$$\overline{\delta x_{kt}} = \langle x_{kt} \rangle_p - \langle x_{kt} \rangle, \quad (5)$$

where $\langle \cdot \rangle_p$ and $\langle \cdot \rangle$ indicate the average over many realizations of the perturbed and of the original dynamics, respectively.

As shown in appendix A, it is possible to derive a closed formula for the response of a dynamical system valid under rather general hypotheses [42]; in particular, in its derivation no assumption of detailed balance is used, meaning that the formula also holds for out-of-equilibrium systems in stationary states. The result is the following, for $\mathbf{x}_t$ a stationary process characterized by an invariant probability density function $\rho(\mathbf{x})$, we can write the following relation:

$$\mathcal{R}_{j \rightarrow k}(t) \equiv \lim_{\delta x_{j0} \rightarrow 0} \frac{\overline{\delta x_{kt}}}{\delta x_{j0}} = - \langle x_{kt} \frac{\partial \ln \rho(\mathbf{x})}{\partial x_j} \Big|_{\mathbf{x}_0} \rangle \quad (6)$$

where $\mathcal{R}(t)$ represents the linear response matrix of the system under consideration at time t , and the average is taken over the joint stationary probability distribution function $p(\mathbf{x}_0, \mathbf{x}_t)$.

If we consider now the special case of a linear dynamics, whose dynamics is ruled by a discrete-time, stochastic linear evolution

$$\mathbf{x}_{t+1} = A\mathbf{x}_t + B\boldsymbol{\eta}_t \quad (7)$$

where A and B are $d \times d$ matrices and $\boldsymbol{\eta}_t$ is a t -dependent vector of delta-correlated random variables with zero mean, the relation (6) becomes simply

$$\mathcal{R}_{i \rightarrow j}(t) = [A^t]_{ji} = [C_t C_0^{-1}]_{ji}, \quad (8)$$

valid for linear Markov systems at discrete times, where we have introduced the covariance matrix $C_t = \langle \mathbf{x}_t \mathbf{x}_0^T \rangle$. This is a key result of linear response theory [9]. Therefore, for a linear system the knowledge of the covariance matrix fully characterizes the response of the system and thus the causality links.

In general, it is not possible to have access to the joint probability distribution function $p(\mathbf{x}_0, \mathbf{x}_t)$, while the computation based on the covariance matrix is valid only for linear systems. For these reasons, in this work, we compute the response using another method, linked to the general definition (6). That is more fundamental from a conceptual point of view [43, 44], since we use directly the definitions of transport coefficients: one perturbs the system with an external force or imposes driving boundary conditions and observes the relaxation process. In our numerical simulations, we compute $\mathcal{R}_{i \rightarrow j}(t)$ by perturbing the variable x_i at time $t = t_0$ with a small perturbation of amplitude $\delta x_i(0)$ and then evaluating the separation $\delta x_j(t|t_0)$ between the two trajectories $\mathbf{x}(t)$ and $\mathbf{x}'(t)$ which are integrated up to a prescribed time $t_1 = t_0 + \Delta t$. At time $t = t_1$ the variable x_i of the reference trajectory is again perturbed with the same $\delta x_i(0)$, and a new sample $\delta \mathbf{x}(t|t_1)$ is computed and so forth. The procedure is repeated $N \gg 1$ times and the mean response is then evaluated:

$$\mathcal{R}_{i \rightarrow j}(\tau) = \frac{1}{N} \sum_{s=1}^N \frac{\delta x_j(t_s + \tau | t_s)}{\delta x_i(t_s)}. \quad (9)$$

where it is assumed that the delay $\Delta t = t_{s+1} - t_s$ between two consecutive examples is much larger than the auto-correlation time of the process.

IV. LOW DIMENSIONAL MARKOV SYSTEMS

Let us start with the 3-dimensional linear process example considered in [27], where Eq. (7) is specified by

$$A = \begin{pmatrix} a & \epsilon & 0 \\ a & a & 0 \\ a & 0 & a \end{pmatrix}$$

and $B = \sigma \mathbb{I}$ representing isotropic noise, with standard deviation σ . The parameters a , σ and ϵ are constant in time, with default valued $a = 0.5$ and $\sigma = 1$ in the following while the linear response is studied by varying ϵ . Thanks to the linearity of the system the interventional causal relationships relies on the sole auto-correlation of the signal as seen in Eq. (8). In this simple example the interventional causality of the system can be computed exactly [27] so that we can use this system as a benchmark to assess different data-driven techniques.

Here we focus on the possibility to find the same results directly from data, without prior knowledge of the model. We consider thus to have the time-series of a system, without knowledge of the underlying model, though we make the assumption that the system is linear, and use these data to empirically estimate the linear response (8) which actually corresponds to solving a Linear Regression [45]. There are two ways to do that: either we estimate the Markov matrix A (Markov regression) from the data and compute the linear response in (8) by taking powers of this estimate; either we directly compute the linear response (8) (linear regression) at finite time by estimating the auto-correlation of the process. As said, the linear response being equivalent to a linear regression, can be regularized by adding a penalization on the regression coefficients w_{ij} which actually coincide with response coefficient $\mathcal{R}_{j \rightarrow i}(t)$ for a given t . Hence we can combine these two inference methods with the Lasso Regularization, also known as L_1 regularization, in order to filter out small spurious regression coefficients and select only the relevant ones [30, 48]. This consists in adding a term of the form L1 Penalization = $\lambda \sum_{i,j=1}^n |w_{ij}|$ to the loss function, with λ the L_1 penalty. In our situation it should help to decide whether regression coefficient corresponding to ϵ is zero or not.

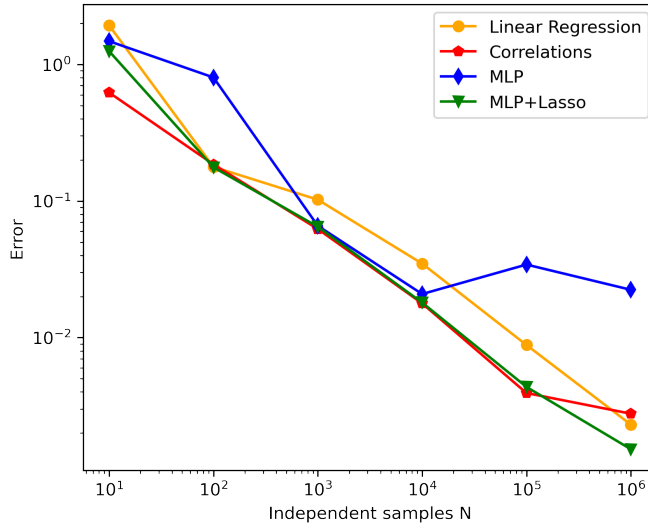


FIG. 1. Plot of the square error of the reconstruction of the matrix A for four different methods. The resulting error for each method was calculated for a different number of training samples.

If one is able to build the propagator A , then it is possible to find the causal links by using the Eq. (8). We estimate the global error for the different methods as the following: we compute the square root of the sum of the quadratic single error over each coefficient is given in Fig.1. All methods give similarly satisfactory results and converge as expected by Law of the large numbers as the number of independent samples increases. The L_1 regularization seems to reduce to fluctuations. So if the system is not explicitly known different machine learning techniques may still be used to reconstruct the Markov process and the response function.

Now, we turn to a non-linear generalization involving three interacting particles in one dimension also studied in [27]. 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 is determined by the following equations:

$$\begin{aligned}\dot{x} &= -U'(x) - (x - y) + d\eta^{(x)}, \\ \dot{y} &= -U'(y) - (y - x) - (y - z) + d\eta^{(y)}, \\ \dot{z} &= -U'(z) - (z - y) + d\eta^{(z)}.\end{aligned}\tag{10}$$

with the potential $U(x) = (1 - r)x^2 + rx^4$. The components of the noise $d\eta$ are iid centered normal variables with variance equal to the used time step. The parameter r tune the level of non-linearity of the dynamics. When $r = 0$ we recover a linear dynamics. For $r \geq 1$ the potential U takes on a pronounced double-well form.

Consistently with the study conducted in [27], we consider $r \in \{1, 2.4\}$ for our analysis. For simulations, we take a time step of 0.001 and a total time span of 5 (in arbitrary units). We used a stochastic Heun integrator to perform the

numerical simulations [49]. 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 undertake and iterate this particular system numerous times over many trajectories, assuming ergodicity of the dynamics of the system.

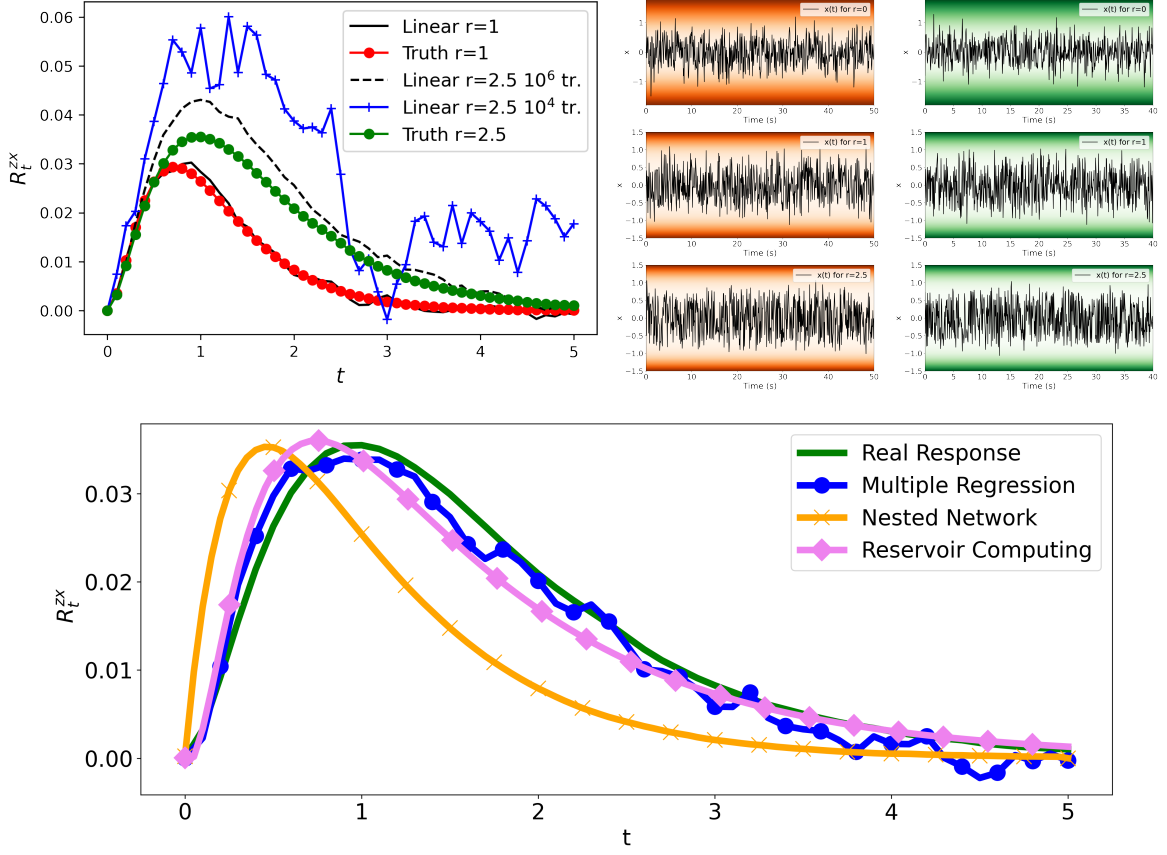


FIG. 2. (a) Plot of the response function $\mathcal{R}_{x \rightarrow z}$ in the non-linear dynamics for different values of r ; The response computed through definition (6) indicated as the “truth”, and the linear formula based on covariance matrix (8) are shown. All the plots have been obtained by averaging over 10^6 trajectories, but the blue-crossed line which has been obtained with 10^4 trajectories. (b)-(c): Visualisation of sample trajectories for different nonlinearity. The strength of the confining potential is qualitatively indicated by the colormap, with light areas corresponding to the minimum of the potential. The (b) orange figure represents the ground truth dynamics. The (c) green points are obtained with the Reservoir Computing, see section. [(d): Comparison between the reconstruction of the response. Multiple Regression (blue line), the threshold value of λ to $0.5 \cdot 10^{-3}$; Lasso L1 regularization was employed during the regression step, with a λ parameter set to $5 \cdot 10^{-6}$. The analysis was based on 10^5 trajectories. Rescaled Response computed from the NN model with 34 parameters (yellow line), the amplitude has been divided by 3; Response obtained with Reservoir Computing (Pink line). The total parameter count in the RC model was approximately 150. Ground truth response is given for comparison (red line).]

Fig. 2(a) shows a comparison between the response calculated by intervention measures and the response obtained through covariance in the linear approximation. When $r = 1$, an excellent agreement is found between the response computed via the general definition and its linear approximation. In the scenario where the non-linear contribution is higher, but not too large (as in the case of $r = 2.5$), the linearized response continues to provide meaningful insights into the causal relationships among the system variables. It is worth emphasising that many data must be used to avoid spurious correlations, about 10^6 trajectories in this case. For comparison, we show the results obtained with 10^4 samples. In this case, the error is much more important, and even qualitatively the response is much less clear. The set of equations (11) contains a nonlinear perturbation, which is still only local in the state variables, whereas the couplings remain linear. This is a rather general feature of physical stochastic systems, but it draws an important difference with fully non-linear chaotic dynamical systems. The reasonable success of applying the linear approximation to this nonlinear case stems from this structure. We consider now different approaches to specifically address the non-linear nature of the problem, in order to be able to cope with more general situations of

practical interest, where we are confronted just with time-series [18] without knowing the nature of the interaction. These approaches are purely data-driven and we want to critically assess whether they may be in some cases valuable alternatives. In particular, we shall consider approaches with increasing the level of physical knowledge available for the modelling.

a. Non-Linear Multiple Regression We first make use of a regression using a nonlinear basis containing the guesses state-vector of the system; Unlike simple linear regression, which assumes a linear relationship between variables, non-linear multiple regression allows the modelling of non-linear relationships between variables. In doing so we are back to a linear regression setting at the expense of a potentially much larger set of candidate features, depending on the prior knowledge on the problem. This relationship is represented by a function, which can take various forms, such as polynomial, exponential, logarithmic, or trigonometric functions, that is building upon a dictionary. Here we choose a dictionary of size 19 made of all 3-variables polynomials functions 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: $(n + d)!/(n! d!)$. For the same reason as in the linear case, we use a L_1 regularization to enforce sparsity of the solution in order to see if we are able to recover the exact form of the potential. The choice of the L_1 penalty λ is done in a standard way by cross-validation. As we shall see in Section V, once an optimal penalization is found, the choice of the dictionary is not key, and we would obtain similar results including a larger set of functions. Using this approach, we are to able to reconstruct the model from data with an excellent level of accuracy, despite the stochastic nature of the system. Let us now focus on the reconstruction of the response. As already mentioned, within the functional formulation the dynamics is linear and we are back to the linear response setting (8) without approximations but in a much larger space.

As for the linear case, we may regress the Markov operator and use iterations of this estimate operator to compute the generalized response. Hence, regression is performed between t and $t + 1$, and the linear response for arbitrary time horizon τ is obtained by taking powers of empirical covariances matrices, 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 [Some formula is needed here to understand what is exactly done]. Capturing the response requires summing over all 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 we cannot neglect the corresponding response generated by their influence. As can be seen from 2, this method lead to an accurate representation of the response curve, hence showing how the linear response theory can be adapted to nonlinear systems.

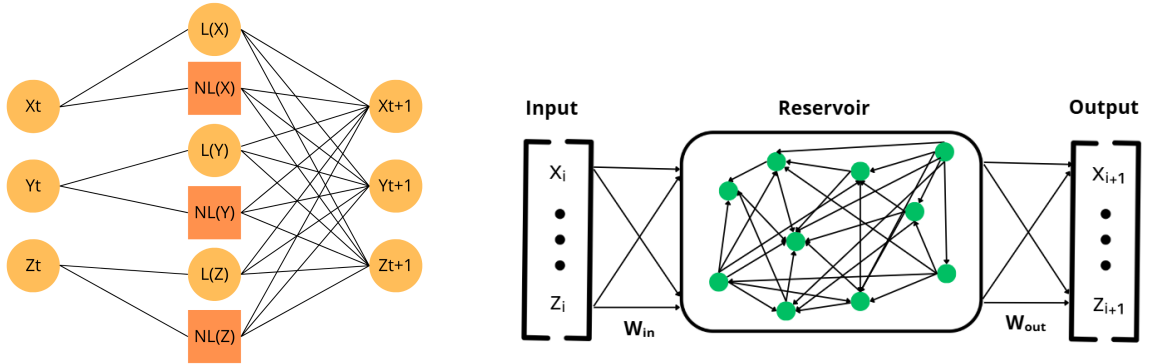


FIG. 3. Left: Architecture of the Nested Neural Network in which each node representing the variable is connected to a linear block and a non-linear block. The linear block is a single node, while the non-linear block is characterized by two labels with two neurons each. Everything is fully connected to the output label. The network is characterized by 34 parameters. Right: architecture of a reservoir computing. Only output weights are trained. Internal weights W are kept fixed.

b. Nested Neural Network Before going for a pure data-driven approach, we have also developed an original approach particularly designed to sort out causal relations among variables, given some a-priori on the form of the model. The approach is therefore data-driven but takes into account some physical considerations, leading to an efficient model in terms of need of data. That may be relevant when a pure data-driven approach becomes too greedy in data to correctly build a model of nonlinear systems [50]. The architecture proposed is sketched in Fig. 3. The model is formulated first by treating the linear and nonlinear part of the dynamics separately in terms of two distinct contributions. Again the focus is on predicting the future state of a system defined by three dynamic variables — x , y , and z — observed at a given time step t , and estimating their respective values at the following time step $t+1$. While in principle we do not make any assumptions on the amplitude of the nonlinearity, yet we consider implicitly that the nonlinear couplings are weak. The model processes the input through a dual-pathway design, combining both linear

transformations and nonlinear mappings via nested sub-networks. This hybrid approach leverages the interpretability of linear models along with the expressive power of neural networks to capture complex temporal dynamics. At each time step, the system receives a three-dimensional input vector composed of the current values of the variables x , y , and z . Each of these variables is then processed through two distinct computational channels: (i) **Linear Channel**: the first path applies a simple linear transformation to the input variable. This channel serves to retain the original signal in a scaled form, ensuring that straightforward linear trends or proportional changes are preserved throughout the network. Each variable is multiplied by a learnable coefficient, providing a direct and interpretable signal flow to the final output layer. (ii) **Nonlinear Channel (Nested Neural Network)**: In parallel, each variable is passed through an independent, nested neural network. These sub-networks comprise multiple hidden layers with nonlinear activation functions, enabling the modelling of complex and potentially nonlinear temporal relationships. Each nested network is dedicated to a single variable and is trained to learn higher-order representations of its temporal behaviour. The outputs from both the linear and nonlinear channels for each variable are then combined — typically through concatenation or summation — and forwarded to the final layer of the network. This final layer is composed of three neurons, each responsible for predicting the value of one of the original variables at the next time step. This dual-channel strategy offers several advantages. The linear pathway preserves a direct, interpretable signal, which can be important in domains where understanding the contribution of individual variables is necessary — such as causality and decision-making applications. Meanwhile, the nested neural networks allow the system to go beyond simple trend following and instead capture latent interactions, nonlinear dependencies, and dynamic behaviours that would be inaccessible to purely linear models. This architecture is modular. Each variable is handled through its own nested sub-network, allowing for isolated tuning, parallel processing, and potential reuse across tasks or models. The interest of this approach is its compatibility with explainability tools. The clear separation between linear and nonlinear components allows analyse the distinct contributions of each path to the final output, gaining insight into which aspects of the input are being handled linearly and which are subjected to more complex transformations.

The reference response to be compared with is computed using the standard recipe: the two systems are initialized with the same initial conditions and the value of x is shifted by 1% in the second system. In the next steps, we calculate the average variation of the variable z by evolving the system through the Nested Neural Network model. The causality can be determined by quantitatively considering the statistical weight of the neural network. To get complete insight into the performance of the model we need also to look at the response function in addition to whether the causal links are properly detected or not by the NN. Interestingly, even a small yet physically oriented model is able to provide decent results, as highlighted in Figure.2, where we show the results of the response obtained with a small network including only 34 parameters. It is remarkable that we have obtained correct qualitative results even with small training datasets.

c. Recurrent Network Model: Reservoir Computing Here we turn to a purely data-driven approach without any implicit assumption on the process. The idea is to learn a model of the process on which the interventional causal experiment can be explicitly performed in the form of (4). Many architectures/models can be considered for this kind of tasks. When one considers a dynamical system, a sensible approach is to use a recurrent neural network (RNN) [51]. A specific network which has been found particular successful for nonlinear dynamical systems is the Reservoir Computing (RC). Over the years, RC has attracted considerable attention due to its ability to solve problems in time series prediction, classification and control, providing reliable results with minimal parameter tuning. Despite its conceptual sophistication, the practical utility of RC has solidified its role as a central tool in computer modeling [52–55]. RC was introduced by models such as Echo State Networks (ESNs) and Liquid State Machines (LSMs) and simplifies the training process by using a randomly initialized, fixed recurrent layer, the so-called reservoir. Instead of training the entire network, as is the case with traditional RNNs, only the output layer is optimized, making RC extremely efficient and scalable. This paradigm emerged in response to the challenges of training RNNs, in particular the problems of vanishing gradients, high computational costs and the difficulty of capturing long-term dependencies with conventional gradient descent approaches. RC has been shown to be particularly effective when dealing with systems characterized by a completely Markovian nature, where transitions depend only on the current state and require no memory of previous time steps. In such contexts, traditional RNNs often lead to inefficient learning as they attempt to model unnecessary temporal dependencies that are irrelevant to the problem at hand. The core concept of RC is to project a non-linear system into a higher dimensional space in which its dynamics become approximately linear. This transformation allows the application of linear regression techniques, often augmented with Lasso L1 regularization, to manage sparsity and improve model generalization. Unlike standard RNNs, which require iterative and complex weighting updates, the RC reservoir remains fixed, significantly reducing computational complexity while preserving the expressive power of the system. This approach generalizes the fixed dictionary based regression by allowing the features to be learned. RC has proven to perform exceptionally well in a variety of applications, even when recurrent dynamics are disabled, as memory is not always required for certain problems. This black-box approach combines efficiency with versatility and is therefore particularly suitable for systems with minimal or uniform causal dependencies over time.

In the problem under consideration, a three-dimensional input space was mapped to a 50-dimensional feature space. Reducing the dimensionality to less than 50 dimensions led to a small shift of the response curve in Fig.2 to the left, but did not significantly change the overall results, and the agreement between the ground truth and the response predicted by the network is quite good. From a more qualitative point of view, we show in Fig. 2 the signals produced by the original stochastic equations (in orange) and by the RC, for different nonlinearity. It can be appreciated that trajectories appear very similar from a statistical point of view. This approach achieved highly accurate responses that were consistent with prior methodologies. These findings highlight the versatility and efficacy of RC in addressing a wide range of computational challenges.

Summarizing the results of this section, when the system is weakly non-linear insights on causal relationships may be found already with the linear approximation. Then for more strongly non-linear systems, a kind of physics-informed machine learning approach may be used based on a regression of nonlinear functions, allowing us to compute exactly the response in the linear framework. The main limitation comes from the size of the training set. When a large data set is given, the most efficient model is the RNN approach. Then in the situation of a limited dataset, a trade-off can be found with the predefined feature model by adjusting the size of the dictionary. Yet when data are too scarce, the nested model which has a very small number of parameters can still give a qualitative picture.

V. CAUSALITY IN FULLY NON-LINEAR SYSTEMS

We now consider the Lorenz '63 system, which exhibits a highly nonlinear and chaotic behaviour. This system's evolution equation reads

$$\begin{cases} \dot{x} = \sigma(y - x) \\ \dot{y} = rx - y - xz \\ \dot{z} = xy - bz \end{cases} \quad (11)$$

where σ represents the Prandtl number, r the Rayleigh number and $b > 0$ is a parameter linked to the ratio of the convective rolls. With $\sigma = 10$, $b = \frac{8}{3}$, $r = 28$ the system display fully chaotic behaviour [56, 57]. This special case fall into the general class of dynamical system characterized by d scalar variables $\mathbf{x} = (x_0, x_1, \dots, x_{d-1})$, given by the relation

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}), \quad (12)$$

for a vector valued function $\mathbf{f}: \mathbb{R}^d \rightarrow \mathbb{R}^d$.

As for the stochastic systems, we consider the system to be fluctuating around an invariant state and are interested in the study of a data-driven interventional approach for computing the response functions

$$\mathcal{R}_{j \rightarrow i}(\tau) = \frac{\overline{\delta x_i(t + \tau)}}{\delta x_j(t)}, \quad (13)$$

where $\bar{\cdot}$ represents the ensemble average and δx_i denotes the perturbation that x_i undergoes when x_j is perturbed with δx_j .

In order to do this, we assume having access to a long trajectory $(\mathbf{x}(t))_{t \in [0, T]}$, but $\mathbf{f}$ is unknown. Since we are not able to influence any of the variables of the system, we resort to a data-driven method for approximating the perturbed trajectories. First, we try to predict $\mathbf{f}$ based on the given $(\mathbf{x}(t))_{t \in [0, T]}$. To this purpose, we train an adequate Neural-Network with the accessible data. In order to approximate its response functions we have implemented two separate methods of approximating $\mathbf{f}$:

- (a) Multi Layer Perceptron (MLP) [58] made up of a 3 layers feedforward neural network with 50 neurons in each hidden layer and LeakyReLU(0.1) as an activation function. This model has been trained with up to 10^6 samples.
- (b) Sparse Identification of Nonlinear Dynamical systems (SINDy) [59] implemented using the PySINDy python package [60, 61], which is based on performing a nonlinear regression on a basis of functions. This model has been trained on 10^3 samples;

Then, as for the reservoir computing, we can perturb many trajectories of our surrogate model to numerically compute the response based on the simulated modified trajectories. Given that we have now a synthetic model, we can produce an arbitrarily large number of samples to avoid any statistical issues when computing the response function. [There should be as scaling in terms of N and the Lyapounov exponent to get the horizon of prediction, this is actually

discuss below. But there is a distinction, because for statistical quantities the error should behave like $1/\sqrt{N}$ while for deterministic one we should have something like $\lambda/\log(N)$ I presume?

We will compare the response results obtained via these approaches also with those of the linear approximation which consists in computing the response based on the covariance matrices.

Following the results obtained in the previous section, a simple data-driven approach like the MLP appears adequate. We expect the nonlinear chaotic systems to perform like a Markovian one [62], and therefore recurrent networks are not needed. Nonetheless, we have thoroughly checked that the results remain qualitative the same using a RNN, so that present results are to be considered general. For a more physics-oriented approach we have used SINDy, which is conceptually very near to the nonlinear regression we have used previously for stochastic processes. In SINDy, complex nonlinear dynamics are represented as a linear combination of simpler nonlinear functions. These functions (feature space) typically include terms inspired by the underlying governing equations, allowing the algorithm to incorporate relevant physical information [63]. In a nutshell, we may approximate a function as $\mathbf{f}(\mathbf{x}) \approx \sum_{k=1}^p \Xi_k(\mathbf{x})\alpha_k$, where the dictionary is $\Xi(\mathbf{x}) = [\mathbf{1} \ \mathbf{x} \ \mathbf{x}^2 \ \dots \ \mathbf{x}^k \ \dots \sin(\mathbf{x}) \ \exp(\mathbf{x}) \ \dots]$, and α_k is a sparse vector with only some nonzero terms. In our case, we have included in the dictionary polynomials[up to which degree?], trigonometric and exponential functions. Combining the training with the sparse penalisation[59], we obtain the coefficients in front of each feature. To this aim we minimize a Loss function $\alpha = \arg\min_{\alpha'} \|\dot{\mathbf{x}} - \Xi(\mathbf{x})\alpha'\|_2 + \lambda\|\alpha'\|_1$.

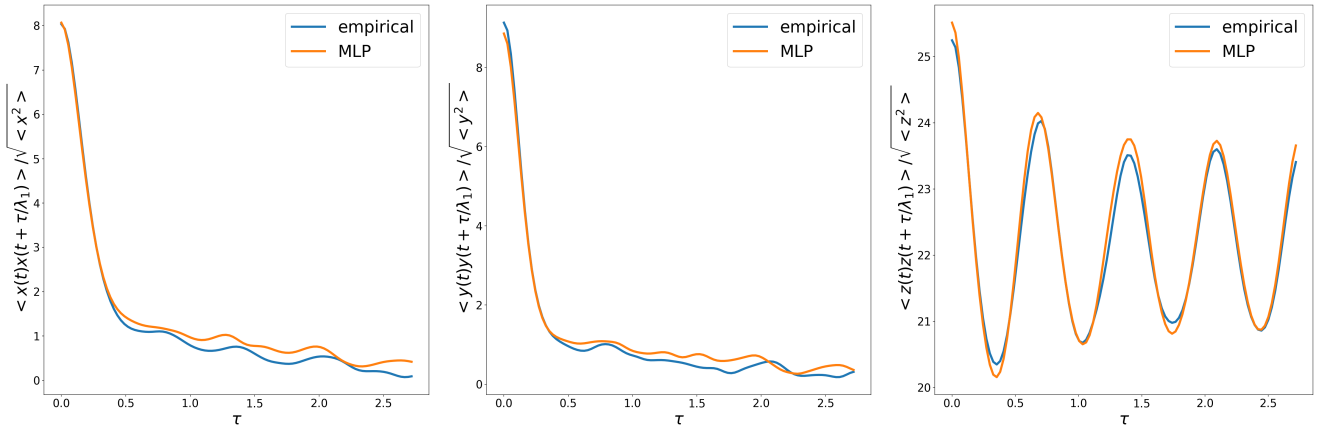


FIG. 4. Normalized auto-correlation function $\langle x_i(t)x_i(t+\tau) \rangle$ for the MLP model compared against that computed through direct integration of the Lorenz system Eq. (11) (blue line), and the SINDy model (green line). Lorenz and SINDy are indistinguishable. The time is expressed in terms of the Lyapunov time based on the maximum Lyapunov exponent.

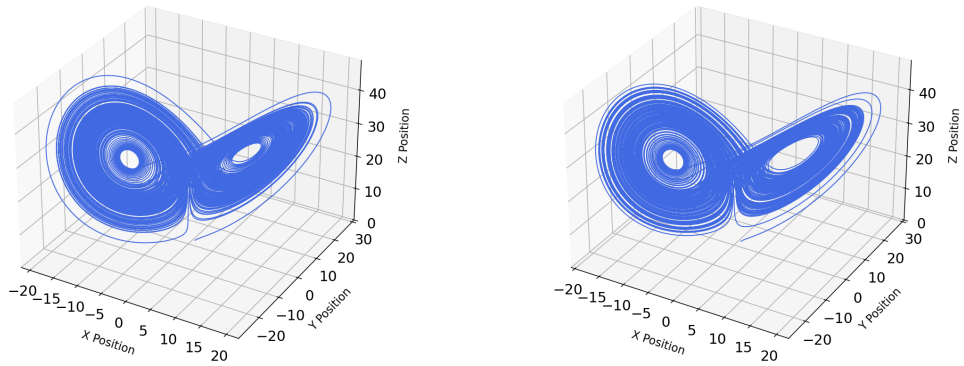


FIG. 5. Trajectory obtained through the numerical integration of the Lorenz system (Left) compared with that generated using an MLP NN (right).

First, we look at the quality of the reconstruction of the phase-space dynamics by the models. Through the sparse identification we are able to find the coefficient of the Lorenz model basically exactly. In Fig. 4, we show the autocorrelations for each variable and we can see that it is not possible to distinguish the True model from the SINDy one. Looking at the results obtained with the MLP, it is found that the agreement is excellent. If some small

difference may be observed, it is worth emphasising that all the observables are equivalent from a statistical point of view. Additional discussion concerning the correlation can be found in appendix. Pursuing the comparison of the dynamics, we show in Fig. 5 the typical attractor obtained using the true model or the MLP; It is evident that the classical butterfly Lorenz attractor is accurately reconstructed. It is worth noting that these particularly clean results have been obtained for the MLP thanks to a large data-set training; similar results have been obtained also with smaller chunks of data, with a little more of statistical noise. For the present proof-of-concept goal, and without loss of generality, we stick to the large dataset, to avoid possible issues related to the lack of training.

We consider now the analysis of the causality through the response. In the Lorenz model, there are different couplings between variables, such as linear and nonlinear. Moreover, there are self-linear interaction for all variables. It is interesting to analyse the results disentangling the various couplings. Figure 6 presents a comparison of the numerically obtained response functions for the linear interactions. First, It is not surprising that the response

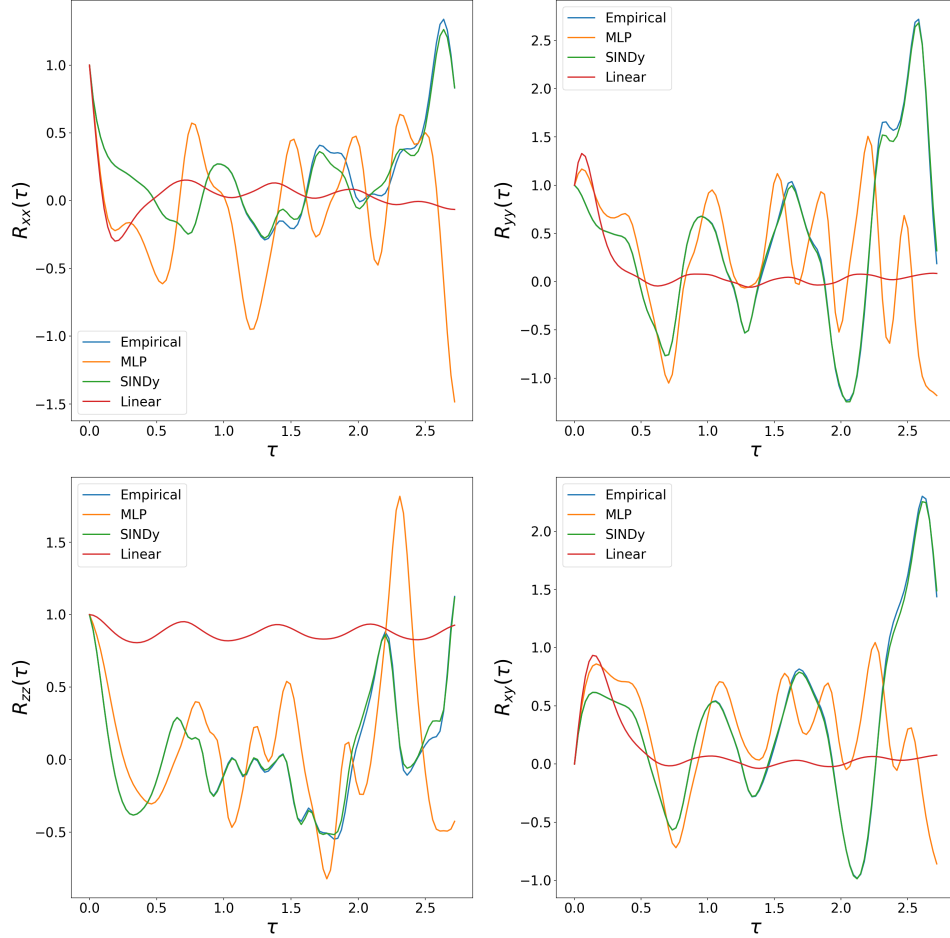


FIG. 6. The response functions $\mathcal{R}_{i \rightarrow j}(\tau)$ for $i, j = x, y, z$, obtained numerically with a sample size of 5000[Is it the size of the trajectory?], using the two methods presented above, the empirically obtained response functions using a Runge-Kutta method of order 4 for integrating the dynamical system (11) and the response functions predicted by the linear response theory. The unperturbed trajectory is computed using RK4. The time-scale is normalized by the Lyapunov time $1/\lambda \approx 500$ units, λ is the maximum Lyapunov exponent.

predicted for the linearized system performs poorly, due to the strongly nonlinear character of Lorenz '63. It is still interesting to remark that the linear approximation provides poor qualitative predictions even when considering just the linear part of interactions. When the model is fully nonlinear, the dynamics on the attractor cannot be trivially reduced to its linear part. As a consequence of the accurate reconstruction of the original model, SINDy performs outstandingly, the response functions obtained in this way overlap with the empirically obtained response for most of the time. Moreover, the MLP approach outperforms the pure linear model, and in particular shows good results in the case of the $\mathcal{R}_{y \rightarrow x}$. In both $\mathcal{R}_{x \rightarrow x}$ and $\mathcal{R}_{z \rightarrow z}$, the MLP is able to correctly detect the response, but it fails on the long run, in relation to the presence of chaos which impeaches a long horizon in presence of a finite, and yet small,

difference in the model. However, we can observe that the global qualitative and quantitative prediction is good, detecting the main frequencies of the response. Looking at the nonlinear couplings tells a slightly different story.

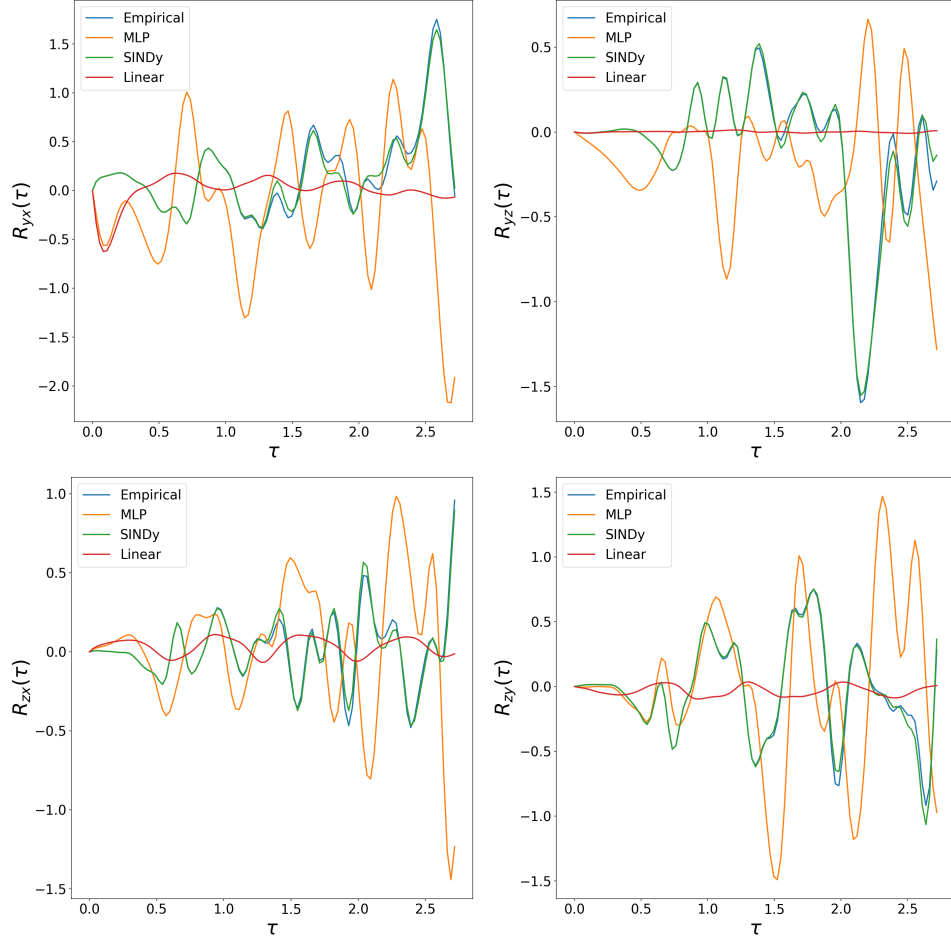


FIG. 7. The response functions $\mathcal{R}_{i \rightarrow j}(\tau)$ for $i, j = x, y, z$, computed as in the previous figure. We focus here on the non-linear couplings.

SINDy continues to reproduce perfectly the response. Yet, while the neural network is able to provide a better picture than the crude linear approximation, at least in terms of shape, the MLP is able to provide at best only a qualitative answer. It is interesting to underline that in fully nonlinear systems the accurate reconstruction of auto-correlations and of the dimension of the attractor is not enough to guarantee a good response of the system. We have found that some of the cross-correlations are only qualitatively predicted on the long-time by the MLP. That indicates that all the correlations play a non-trivial role in response, in line with what is found for the weakly non-linear Markov processes. In Appendix B, we show the cross-correlation.

We want now to build a specific set-up to get more insights on the role of chaos in the response dynamics. For such a purpose, we avoid to round the SINDy coefficients keeping a small but finite error, namely on the sixth digit. [I am not sure to understand this last sentence] Furthermore, for all methods, instead of perturbing the original data series, as we should do in real applications, we build a reference unperturbed trajectory for the three models available, and then we compute the response to a perturbation. In this way also the initial condition is slightly different and we emphasise the possible differences in the models. Few examples of the responses obtained in this new set-up are shown in Fig. 8. The linear approximation has not changed, and we show it just for reference. The MLP provides results very similar to those shown in Figs. 6-7. The MLP-based model has clearly some differences from the true Lorenz system is not sensitive to the way of producing the unperturbed trajectory does not affect much. Instead, the small error in the SINDy model is magnified in the present approach. The small error in the coefficients lead to generate a slightly different unperturbed trajectory and both errors jointly make an impact on the response. It is clear that in this case SINDy is not able to exactly reproduce the original. In particular, in all cases we notice a departure from the original Lorenz response around $t \sim 500$ which corresponds roughly to the Lyapunov time. It is also interesting

to note that in the linear or weakly non-linear couplings, panels (a)-(c), the results obtained by SINDy are not exact but provide a good description of the response. Yet, when considering a fully non-linear coupling as in panel (d), the model is not able anymore to give a good qualitative description of the response on the long-time limit.

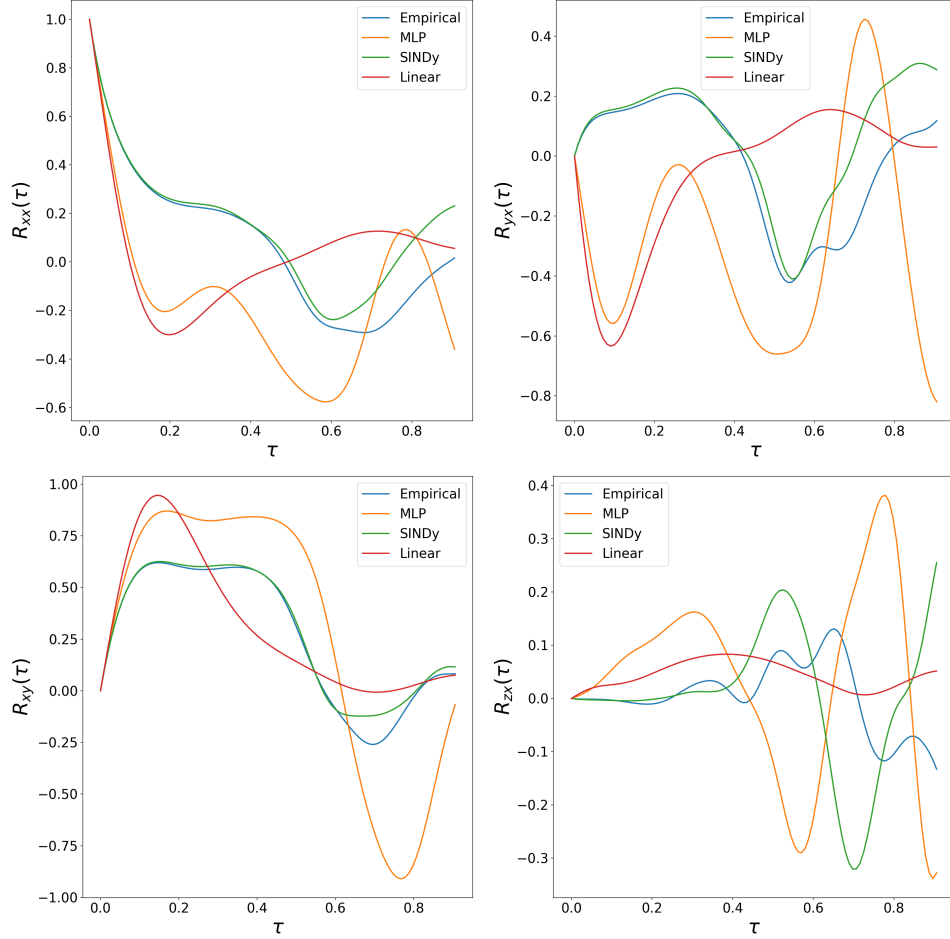


FIG. 8. The response functions $\mathcal{R}_{i \rightarrow j}(\tau)$ for $i, j = \overline{1, 3}$, computed as in the previous figure. We deal with a relatively short time-series here to emphasize the differences. As before, the time-scale is normalized by the Lyapunov time $1/\lambda \approx 500$ units, λ is the maximum Lyapunov exponent. The unperturbed trajectory is computed using the different methods presented. The error starts growing at a time of order λ .

These results lead us to ask about the role played by chaos in this context of causal inference. Van Kampen argued that the usual derivation of the Fluctuation-dissipation relation, based on a first-order truncation of the time-dependent perturbation theory for the evolution of probability density, is not fully justified [64]. In the presence of chaos the terms $\partial x_i(t)/\partial x_j(0)$ grow exponentially as $e^{\lambda t}$, where λ is the Lyapunov exponent, therefore it is not possible to use the linear expansion (??)[Missing equation] for a time larger than $(1/\lambda) \ln(L/|\delta \mathbf{x}(0)|)$, where L is the typical fluctuation size of the variable $\mathbf{x}$. Such an argument suggests that the linear response theory is valid only for extremely small and unphysical perturbations or for short times, an odd conclusion given the success of the linear theory for the computation of the transport coefficients. In fact, it has been demonstrated that the argument by Van Kampen is not relevant for the general application of the response theory [42], because the response deals with observables quantities, which are averages over ensembles of trajectories, and have a much more regular behaviour than the single trajectories. The effectiveness of the Linear Response theory lies in the “constructive role of chaos” [40]. However, in our case we compute the response through the definition (9). For such a problem, van Kampen’s objection is relevant, to some extent. In the presence of chaos, the two trajectories $\mathbf{x}(t)$ and $\mathbf{x}'(t)$ typically separate exponentially in time, therefore the mean response is the result of a delicate balance of terms which grow in time in different directions. A

naive estimate of the error in the computation of $\mathcal{R}_{i \rightarrow j}(\tau)$ suggests an increase in time as

$$e^{(N)}(\tau) = \left[\frac{1}{N} \sum_{k=1}^N \left(\frac{\delta x_j(t_k + \tau | t_k)}{\delta x_i(t_k)} \right)^2 - \left(\mathcal{R}_{i \rightarrow j}(\tau) \right)^2 \right]^{1/2} \sim \frac{e^{\frac{\lambda_2}{2} \tau}}{\sqrt{N}}, \quad (14)$$

where λ_2 is the generalized Lyapunov exponent of order 2 [66]:

$$\lambda_2 = \lim_{\tau \rightarrow \infty} \frac{1}{\tau} \ln \left\langle \left(\frac{|\delta \mathbf{x}(\tau)|}{|\delta \mathbf{x}(0)|} \right)^2 \right\rangle. \quad (15)$$

In the above equation $\delta \mathbf{x}(\tau)$ is assumed infinitesimal, i.e. evolving according the linearized dynamics (in mathematical terms, $\delta \mathbf{x}(\tau)$ is a tangent vector of the system) [67], and a lower bound for λ_2 is given by 2λ . Thus very large N seems to be necessary, in order to properly estimate this balance and to compute $\mathcal{R}_{i \rightarrow i}(\tau)[\mathcal{R}_{i \rightarrow j}(\tau) \text{ instead?}]$ for large τ .

Some remarks are in order. (i) In the case of linear quasi-gaussian systems, the Response is given in terms of time-correlations, in particular diagonal terms are proportional to auto-correlations. Here we show clearly that in the general nonlinear case a model built from data may reproduce very accurately the auto-correlations, without providing a satisfactory prediction of the response in all cases. (ii) Yet, to infer a model from data is possible, provided sufficient data for training are available. If no physical information is used, the model will give a decent representation of causality in terms of response for all linear or weakly nonlinear couplings. If one wants to obtain a fully predictive approach, it is needed to combine data and some prior physical modelling to infer the model. In this case, results are excellent. (iii) Finally, while it is not too difficult to learn a model able to accurately reproduce the phase-space dynamics, at least for moderately high-dimensional systems, the response turns out to be a very powerful but delicate statistical observable, which is subject to chaos. That hampers the straightforward generalization of the models inferred by data to new cases.[what is meant by new cases?]

VI. RESPONSE THEORY ANALYSIS FOR HIGH DIMENSIONAL LINEAR MARKOV SYSTEMS

Up to now, we have considered systems consisting of a few observables with different types of interactions, either linear, weakly non-linear or strongly non-linear leading to chaotic behavior. In this section we would like to consider the situation where the number of observables in interaction is extensive, and to which extend causal relations among them can be extracted. This is of interest for instance when facing networks of interactions with unknown structure. Here we will limit ourselves to linear interactions and investigate what the linear response theory can tell us about causal relations in this context, when the number of observations is proportional to the number of observables. The goal is to quantify the efficiency of the response theory based causal predictor in an asymptotic regime corresponding to the case of an extensive number of time series associated to sensors taking place on a large causal graph. We assume the underlying process to obey the linear dynamics given by (7) rewritten in a slightly different way in order to account relaxation for some of the process explicitly

$$\mathbf{x}_{t+1} = [(1 - \epsilon)\mathbb{I} + \epsilon A] \mathbf{x}_t + \sigma \boldsymbol{\eta}_t \quad (16)$$

controlled by a damping coefficient $0 < \epsilon \ll 1$, while A is a $d \times d$ incidence matrix, where $A_{ij} = 1$ represents an oriented link from j to i and σ the standard deviation of the noise, represented here by the d -dimensional normal variable $\eta_{it} = \mathcal{N}(0, 1)$, $i = 1, \dots, d$, decorrelated both in time and among the coordinates. An example of such network is shown on Figure 9. In order to make this concrete, we can picture this system as an hydraulic network, where only the nodes are observed and behave both as sources and sinks (the noise term), the oriented flow between nodes being determined by the incidence matrix, these being not observed. The data time series have a length of size T in terms of times steps $t = 0, \dots, T$ and we are interested by the so-called proportional asymptotic limit where both $d, T \rightarrow \infty$ with fixed ratio T/d . A itself encodes the topology of a sparse random oriented graph of mean connectivity $\bar{c}$. The goal is to study causal predictors $\mathcal{R}_t$ based on linear response theory.

A. Causal Response functions

In this case the linear response coincides with the linear regression of $\mathbf{x}_{t+\tau}$ from $\mathbf{x}_t$. For sake of simplicity we will limit the analysis to the ridge regularized regression, i.e. with L_2 penalty. We will refer to this as the linear ridge response (LRR) causal indicator. Sparse regression techniques, either based on L_1 or other prior like Gauss-Bernoulli, can be analyzed along similar lines as the compressed sensing problem [68, 69] and will be discussed in a separate

work [70] extending the present formalism to the sparse causal regression context. By convenience, for reasons which will be clearer later on, we denote the ridge penalty by α^{-1} . In order to express the linear response let us introduce some notations. We assume that we are able to extract a set of N independent samples $\{\mathbf{x}_{t(s)}, s = 1, \dots, N\}$, the proportional scaling limit being characterized by the ratio $\rho \stackrel{\text{def}}{=} \frac{N}{d}$. In case we want to use all the available samples, these cannot then be considered as independent, but in principle some corrections could be performed on ρ , based on the level of correlation between successive samples, so that the following analysis remains valid upon replacing ρ by some effective one, but we leave aside for sake of simplicity. We are interested in the empirical estimate of the response function corresponding to the time delay t based on these data. Let us define the following correlation matrices (assuming $\mathbf{x}_t$ to be centered)

$$C_t^{(N)} \stackrel{\text{def}}{=} \frac{1}{N} \sum_{s=1}^N \mathbf{x}_{t(s)+t} \mathbf{x}_{t(s)}^\top$$

and the resolvent

$$G^{(N)} \stackrel{\text{def}}{=} (\mathbb{I} + \alpha C_0^{(N)})^{-1},$$

then the empirical L_2 regularized linear response reads

$$\mathcal{R}^{(N)}(t) = \alpha C_t^{(N)} G^{(N)}. \quad (17)$$

This is an $d \times d$ matrix, and we denote its elements by

$$\mathcal{R}_{i \rightarrow j}^{(N)}(t) \stackrel{\text{def}}{=} \mathbf{e}_i^\top \mathcal{R}^{(N)}(t) \mathbf{e}_j$$

with $\{\mathbf{e}_i, i = 1, \dots, d\}$ the canonical vector basis of $\mathbb{R}^d$ and where the arrow indicates the causal order. Based on this we define the squared causal response (SCR) from node i to j

$$\begin{aligned} \chi_{i \rightarrow j}^{(N)}(t) &\stackrel{\text{def}}{=} \left\langle |\mathbf{e}_j^\top \mathcal{R}^{(N)}(t) \mathbf{e}_i|^2 \right\rangle_{\boldsymbol{\eta}} \\ &= \alpha^2 \left\langle \text{Tr} \left[G^{(N)} \mathbf{e}_i \mathbf{e}_i^\top G^{(N)} C_t^{(N)} \mathbf{e}_j \mathbf{e}_j^\top C_t^{(N)} \right] \right\rangle_{\boldsymbol{\eta}} \end{aligned}$$

where $\langle \rangle_{\boldsymbol{\eta}}$ denotes an averaging over the noise. We denote by $\chi_{i \rightarrow j}(t)$ its corresponding asymptotic limit in the proportional scaling. Given some non-negative threshold h and some inverse temperature β , we are interested by the LRR causal predictor which we define as

$$p_{i \rightarrow j}(t) \stackrel{\text{def}}{=} \sigma \left[\beta (\chi_{i \rightarrow j}(t) - h) \right], \quad (18)$$

where $\sigma(x) \stackrel{\text{def}}{=} 1/(1 + e^{-x})$ is the sigmoid function. In order to obtain this limit we first need the covariance of $\mathbf{x}_t$ at coinciding time in the population limit defined as

$$C \stackrel{\text{def}}{=} \lim_{N \rightarrow \infty} C_0^{(N)}.$$

From the definition (16) of the underlying process C satisfies a discrete-time Lyapounov equation

$$C = M C M^\top + \sigma^2 \mathbb{I} = \sigma^2 \sum_{\tau=0}^{\infty} M^\tau M^{\tau\top}$$

where $M = (1 - \epsilon) \mathbb{I} + \epsilon A$. We assume that the spectral radius of M is smaller than 1 for C to be well defined. In addition we have

$$C_t^{(N)} = M^t C_0^{(N)} + \frac{\sigma}{N} \sum_{s=1}^N \sum_{\tau=0}^{t-1} M^{t-\tau-1} \boldsymbol{\eta}_{t(s)+\tau} \mathbf{x}_{t(s)}^\top. \quad (19)$$

Using this we obtain for the SCR

$$\chi_{i \rightarrow j}^{(N)}(t) = \alpha^2 \text{Tr} \left[G^{(N)} \mathbf{e}_i \mathbf{e}_i^\top G^{(N)} \left(C_0^{(N)\top} M^{t\top} \mathbf{e}_j \mathbf{e}_j^\top M^t C_0^{(N)} + \frac{\sigma^2}{N} \sum_{\tau=0}^{t-1} \mathbf{e}_j^\top M^{t-\tau-1} M^{t-\tau-1\top} \mathbf{e}_j C_0^{(N)} \right) \right]$$

Thanks to the identity $\alpha G^{(N)} C_0^{(N)} = \mathbb{I} - G^{(N)}$ we finally obtain an expression which involve only the resolvent:

$$\begin{aligned} \chi_{i \rightarrow j}^{(N)}(t) &= \text{Tr} \left[(\mathbb{I} - G^{(N)}) \mathbf{e}_i \mathbf{e}_i^\top (\mathbb{I} - G^{(N)}) M^t{}^\top \mathbf{e}_j \mathbf{e}_j^\top M^t \right] \\ &+ \frac{\alpha \sigma^2}{N} \text{Tr} \left[G^{(N)} (\mathbb{I} - G^{(N)}) \mathbf{e}_i \mathbf{e}_i^\top \right] \times \sum_{\tau=0}^{t-1} \mathbf{e}_j^\top M^{t-\tau-1} M^{t-\tau-1}{}^\top \mathbf{e}_j \end{aligned} \quad (20)$$

where the first term which is $\mathcal{O}(1)$ corresponds to a bias independent of the noise while the second one which is $\mathcal{O}(t/N)$ is a variance term resulting from averaging over the noise. Hence we expect the second term to become meaningful for a regime where t/N is $\mathcal{O}(1)$, since averaging over the noise leads us to get rid of stochastic contributions which are typically of order $1/\sqrt{N}$. Note, that the average of the noise is in principle problematic, since it appears explicitly in (19) but is also present implicitly in the $\mathbf{x}_{t(s)}$ i.e. in $C_0^{(N)}$ leading to additional correlations that we do not take into account. The reason these are actually negligible relies actually on the hypothesis that $t(s+1) - t(s)$ is sufficiently large to insure independence between $\mathbf{x}_{t(s)}$ and $\mathbf{x}_{t(s+1)}$.

B. Asymptotic efficiency of LRR causal predictor

In order to characterize the efficiency of the LRR causal predictor we need to introduce additional quantities. Assuming that the fluctuations of the linear response are Gaussian we need to characterize its bias and variance, conditionally to whether the causal link from $i \rightarrow j$ exists or not. Let us define by $\mathbf{f}_{\rightarrow i}(t)$ the vector of *true* underlying response coefficients associated to output node i and time delay t , corresponding to what would be observed in absence of noise

$$\mathbf{f}_{\rightarrow i}(t) \stackrel{\text{def}}{=} \mathbf{e}_i^\top M^t.$$

Considering $\{\mathcal{R}_{j \rightarrow i}^{(N)}(t), j = 1, \dots, d\}$ as a set of independent random variables, we assume these to be Gaussian conditionally to $f_{j \rightarrow i}(t)$, i.e. of the form

$$\mathcal{R}_{j \rightarrow i}^{(N)}(t) = \mu_{j \rightarrow i}^{(N)}(t) + \sigma_{\rightarrow i}^{(N)}(t) z_{j \rightarrow i}$$

where $z_{j \rightarrow i} \sim \mathcal{N}(0, 1)$ are iid. $\mu_{j \rightarrow i}^{(N)}(t)$ and $\sigma_{\rightarrow i}^{(N)2}(t)$ are respectively the bias and variance of $\mathcal{R}_{j \rightarrow i}^{(N)}(t)$ given by

$$\begin{aligned} \mu_{j \rightarrow i}^{(N)}(t) &= \frac{f_{j \rightarrow i}(t)}{\|\mathbf{f}_{\rightarrow i}(t)\|^2} \left\langle \mathbf{f}_{\rightarrow i}^\top(t) \mathcal{R}_{\rightarrow i}^{(N)}(t) \right\rangle_{\boldsymbol{\eta}}, \\ \sigma_{\rightarrow i}^{(N)2}(t) &= \left\langle \|\mathcal{R}_{\rightarrow i}^{(N)}(t)\|^2 \right\rangle_{\boldsymbol{\eta}} - \frac{\left\langle \mathbf{f}_{\rightarrow i}^\top(t) \mathcal{R}_{\rightarrow i}^{(N)}(t) \right\rangle_{\boldsymbol{\eta}}^2}{\|\mathbf{f}_{\rightarrow i}(t)\|^2}. \end{aligned}$$

$f_{j \rightarrow i} = \mathbf{e}_i^\top M^t \mathbf{e}_j$ is considered to be given but unknown and the other quantities like the overlap between $\mathbf{f}_{\rightarrow i}$ and $\mathcal{R}_{\rightarrow i}^{(N)}(t)$ take simple form in term of the resolvent:

$$\|\mathbf{f}_{\rightarrow i}(t)\|^2 = \mathbf{e}_i^\top M^t M^t{}^\top \mathbf{e}_i, \quad (21)$$

$$\left\langle \mathbf{f}_{\rightarrow i}^\top(t) \mathcal{R}_{\rightarrow i}^{(N)}(t) \right\rangle_{\boldsymbol{\eta}} = \text{Tr}[(\mathbb{I} - G^{(N)}) M^t{}^\top \mathbf{e}_i \mathbf{e}_i^\top M^t], \quad (22)$$

$$\left\langle \|\mathcal{R}_{\rightarrow i}^{(N)}(t)\|^2 \right\rangle_{\boldsymbol{\eta}} = \text{Tr}[(\mathbb{I} - G^{(N)})^2 M^t{}^\top \mathbf{e}_i \mathbf{e}_i^\top M^t] + \frac{\alpha \sigma^2}{N} \text{Tr}[G^{(N)} (\mathbb{I} - G^{(N)})] \times \sum_{\tau=0}^{t-1} \mathbf{e}_i^\top M^{t-\tau-1} M^{t-\tau-1}{}^\top \mathbf{e}_i. \quad (23)$$

As illustrated on Figure 9, each coefficient $\mathcal{R}_{i \rightarrow j}$ maintains its random nature in the asymptotic limit, such that the asymptotic estimation (C3) of the squared response averaged over the noise cannot be leveraged directly to analyze the asymptotic behavior of the LRR causal predictor. Instead, both $\mu_{j \rightarrow i}$ and $\sigma_{\rightarrow i}^2$ become deterministic in the asymptotic regime and can be fully determined analytically thanks to standard random matrix formulas [71, 72] upon solving RMT equations as stated in Appendix C. Then the performance of the causal predictor (18) being a simple function of these bias and variance is fully characterized. Equation (C4) given in Appendix C along with (21) fully determine $\mu_{\rightarrow i}(t)$ and $\sigma_{\rightarrow i}(t)$. In turn partitioning the set of links $\{i, j\}(t) = \bar{\mathcal{I}}(t) + \mathcal{I}(t)$ corresponding respectively

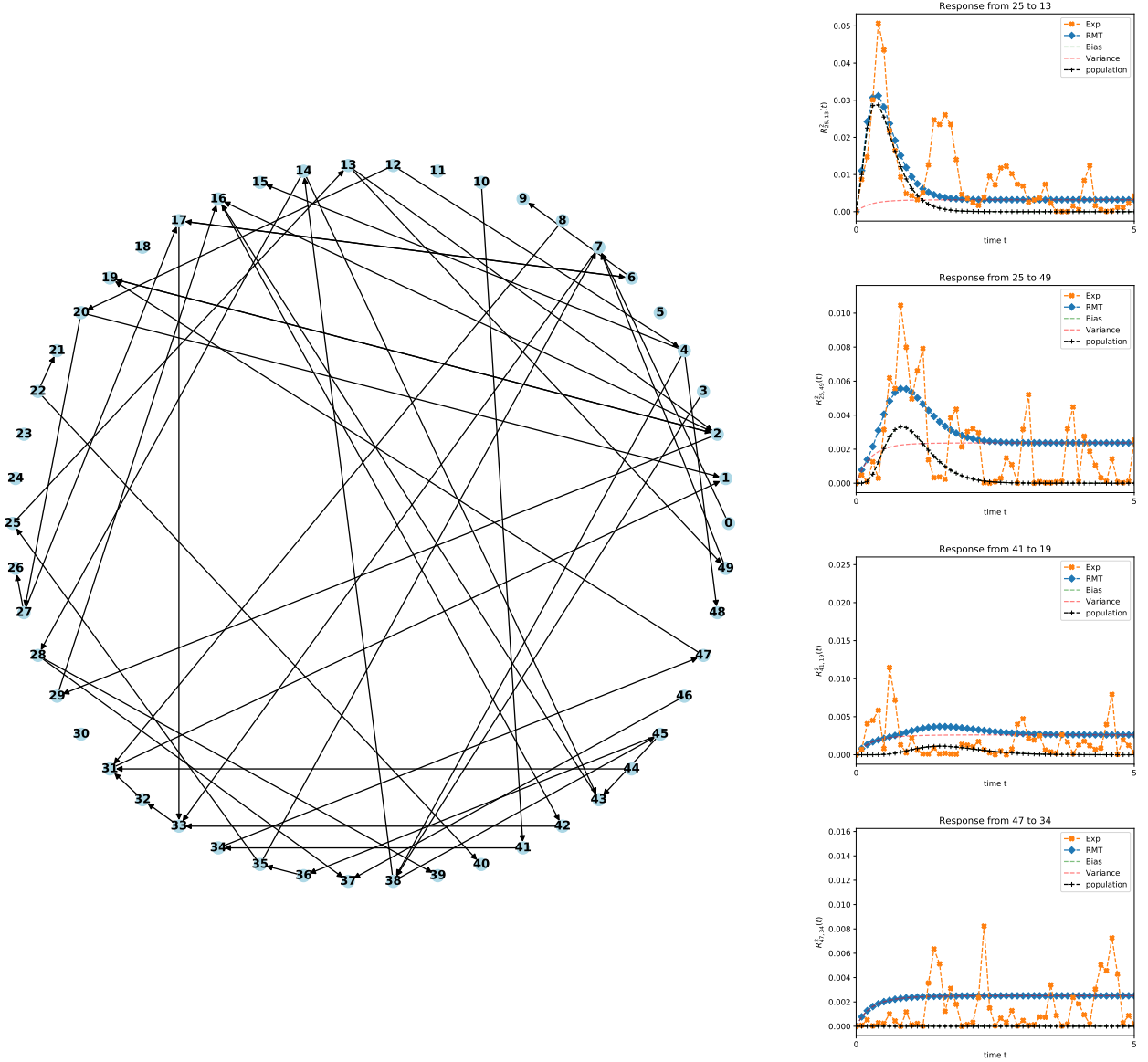


FIG. 9. (left) Random oriented graph of size $d = 50$ and output mean connectivity $\bar{c} = 1.2$. (Right) Comparison between response functions for resp. first, second third NN causal links and non-causal links for the last one. The black curves corresponds to the population responses, i.e. in the limit $N \rightarrow \infty$ at fixed d , the yellow are the empirical one, while the response predicted by RMT is in blue.

to the set of zero and non-zero entries in M^t leads to the following expressions for the expectation of true positive and false positive rates of causal links given by the LRR causal predictor:

$$\overline{TP}(h, t) = \left\langle \mathbb{E}_{\eta \sim \mathcal{N}(0,1)} \left[\sigma \left(\beta (|\mu_{i \rightarrow j}(t) + \sigma_{\rightarrow j}(t) \eta|^2) - h \right) \right] \right\rangle_{(i,j) \in \mathcal{I}(t)}$$

$$\overline{FP}(h, t) = \left\langle \mathbb{E}_{\eta \sim \mathcal{N}(0,1)} \left[\sigma \left(\beta (|\mu_{i \rightarrow j}(t) + \sigma_{\rightarrow j}(t) \eta|^2) - h \right) \right] \right\rangle_{(i,j) \in \bar{\mathcal{I}}(t)}$$

where h is an arbitrary threshold, and the empirical average over edges, i.e. implicitly on $f_{i \rightarrow j}(t)$ is denoted by $\langle \cdot \rangle$. A standard way to characterize the behavior of this causal predictor behaves is to display the Receiver Operating Characteristic (ROC) curve, i.e. by plotting number of true positive as a function of number of false positive when varying h in (18). The comparison between experiments and asymptotic results is shown on Figure 10, denoting a rapid onset of the asymptotic regime despite the high sparsity of the signal for small time response functions while for larger time the noise dominates rapidly rendering the LR causal predictor inefficient.

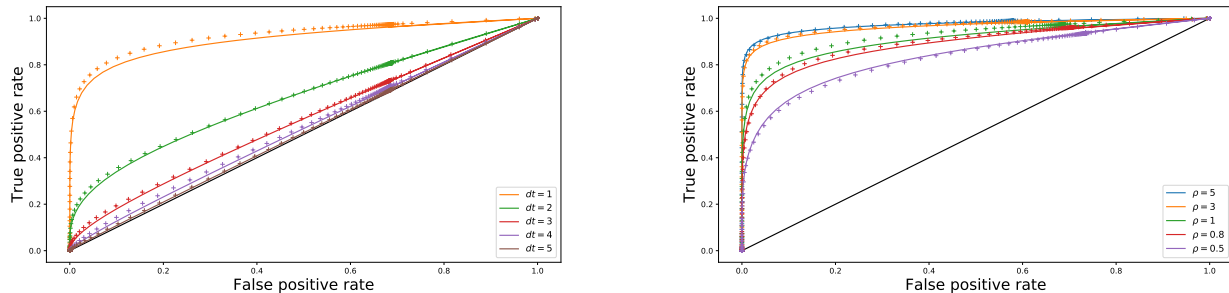


FIG. 10. ROC curve of the LRR causal predictor obtained for a network of size $d = 100$ mean connectivity $\bar{c} = 1.7$, inverse ridge penalty $\alpha = 10$, when varying respectively the time delay for the response on the left with $\rho = 2.5$ and the aspect ratio ρ with $dt = 1$ on the right panel.

VII. CONCLUSIONS

In this work, we have investigated in some generic examples to which extent and with what accuracy the response theory combined with ML techniques can be used to infer from data the causality links between different components of a physical system. The approach is grounded in the statistical physics framework, which allows us to the linear response of the system as a proxy for doing interventional causality experiments. This can be done both for linear and nonlinear systems, provided that we work with a functional representation of the dynamical system in the latter case.

For linear systems, conceptually the problem is solved in terms of the covariance matrix, which fully determined the response function. As expected, we have shown that in this case, for linear stochastic Markov processes, it is possible to efficiently extract the response from data with different techniques. The use of a sparse regularized neural network turned out to be particularly efficient, with a very small number of parameters needed.

In the nonlinear case, the response cannot be calculated from the covariance, but in principle we should access to the whole invariant measure. We have studied both a stochastic system perturbed by a nonlinear potential, and the Lorenz63 system, paradigm of the fully chaotic systems. The difference between the two, is first that one is stochastic while the other is not and in the first the non-linearity does not couple the variables in contrary to the Lorenz63 model. Although both systems are nonlinear, we have found a different behavior. [I don't get this point, it seems similar techniques are used in both cases with similar accuracy, thanks to working in the functional representation, potentially of much larger dimension but where the dynamics is linear. What about Lorenz? The remaining of the conclusion should be clarified]

The nonlinear Markov system possesses an underlying linear structure, which makes it possible to capture the system's response in a qualitative sense, even through a linear approximation. In this setting, various data-driven approaches prove effective.

On one hand, if one can physically identify a suitable functional basis for the system's state vector, the response can be reconstructed with near-exact accuracy. On the other hand, in a purely data-driven framework, powerful recurrent architectures—such as reservoir computing—can still provide a reasonably good agreement.

The situation is different in the chaotic case. Here, the linear approximation fails entirely and offers no insight into the actual response or causality. A properly parameterized neural network can successfully reconstruct the phase dynamics in detail, but it still only delivers a qualitative account of the full response. By contrast, sparse identification makes it possible to infer the correct model directly from data, starting from a large dictionary of candidate functions, as long as the relevant terms are included. The resulting model is nearly indistinguishable from the true one.

These findings point to a key conclusion: the response of a system is a particularly delicate statistical observable—far more challenging to characterize accurately than other two-point statistics such as auto-correlations. This calls for special care when analyzing responses in complex systems. Nevertheless, we have shown that a physics-guided approach can indeed yield highly accurate results.

Finally, we have also considered the case in which the system is simple enough (linear) but the number of degrees of freedom is so high that it imposes a lack of observability. In this case, it is possible in principle to access to the response via covariance, but we have not enough data to compute it accurately. That may be relevant for network applications. Leveraging the asymptotic limit we have dealt theoretically with this case in the Random Matrix theory framework. We have shown that it is possible to obtain a rather accurate description of the response in most of the cases, adding some form of regularization.

Concluding, we have shown how it is possible to obtain the response, and thus the causality links, from data in linear and nonlinear systems. For nonlinear systems the problem is arduous, and a purely data-driven approach is bound to fail quantitatively. We hope that this work will pave the way to apply these methods to relevant applications, notably in geophysics.

VIII. ACKNOWLEDGEMENTS

The project has received the funding by French ANR grant SPEED (ANR-20-CE23-0025-01) and the ANR grant SCALP (ANR-24-CE23-1320).

-
- [1] Erik Aurell and Gino Del Ferraro. Causal analysis, correlation-response, and dynamic cavity. In *Journal of Physics: Conference Series*, volume 699, page 012002. IOP Publishing, 2016.
 - [2] C.K. Assaad, E. Devijver, and E. Gaussier. Survey and evaluation of causal discovery methods for time series. *Journal of Artificial Intelligence Research*, 73:767–819, 2022.
 - [3] Aristotle. *The complete works of Aristotle*. Princeton University Press Princeton, 1984.
 - [4] David Hume. *A treatise of human nature*. Clarendon Press, 1896.
 - [5] Bertrand Russell. On the notion of cause. In *Proceedings of the Aristotelian society*, volume 13, pages 1–26. JSTOR, 1912.
 - [6] 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.
 - [7] Judea Pearl. *Causality*. Cambridge university press, 2009.
 - [8] David D Zhang, Harry F Lee, Cong Wang, Baosheng Li, Qing Pei, Jane Zhang, and Yulun An. The causality analysis of climate change and large-scale human crisis. *Proceedings of the National Academy of Sciences*, 108(42):17296–17301, 2011.
 - [9] 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.
 - [10] Hong-Li Zeng, Erik Aurell, Mikko Alava, and Hamed Mahmoudi. Network inference using asynchronously updated kinetic ising model. *Physical Review E*, 83(4):041135, 2011.
 - [11] Rudolf Friedrich, Joachim Peinke, Muhammad Sahimi, and M Reza Rahimi Tabar. Approaching complexity by stochastic methods: From biological systems to turbulence. *Physics Reports*, 506(5):87–162, 2011.
 - [12] Marco Baldovin, Fabio Cecconi, Massimo Cencini, Andrea Puglisi, and Angelo Vulpiani. The role of data in model building and prediction: a survey through examples. *Entropy*, 20(10):807, 2018.
 - [13] Federica Ferretti, Victor Chardes, Thierry Mora, Aleksandra M Walczak, and Irene Giardina. Building general langevin models from discrete datasets. *Physical Review X*, 10(3):031018, 2020.
 - [14] Jonathan S Yedidia, William T Freeman, Yair Weiss, et al. Understanding belief propagation and its generalizations. *Exploring artificial intelligence in the new millennium*, 8(236–239):0018–9448, 2003.
 - [15] Sifan Wang, Shyam Sankaran, and Paris Perdikaris. Respecting causality is all you need for training physics-informed neural networks. *arXiv preprint arXiv:2203.07404*, 2022.
 - [16] Alexandre Kojève. *L'idée du déterminisme dans la physique classique et dans la physique moderne*. FeniXX, 1990.
 - [17] Jack Cohen and Ian Stewart. *The collapse of chaos: Discovering simplicity in a complex world*. Penguin UK, 2000.
 - [18] Holger Kantz and Thomas Schreiber. *Nonlinear time series analysis*. Cambridge university press, 2003.
 - [19] Herbert A Simon. Spurious correlation: A causal interpretation. *Journal of the American statistical Association*, 49(267):467–479, 1954.
 - [20] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.
 - [21] Anil Seth. Granger causality. *Scholarpedia*, 2(7):1667, 2007.
 - [22] Alex J Cadotte, Thomas B DeMarse, Ping He, and Mingzhou Ding. Causal measures of structure and plasticity in simulated and living neural networks. *PloS one*, 3(10):e3355, 2008.
 - [23] Adam B Barrett, Lionel Barnett, and Anil K Seth. Multivariate granger causality and generalized variance. *Physical review E*, 81(4):041907, 2010.
 - [24] Thomas Schreiber. Measuring information transfer. *Physical review letters*, 85(2):461, 2000.
 - [25] Terry Bossomaier, Lionel Barnett, Michael Harré, Joseph T Lizier, Terry Bossomaier, Lionel Barnett, Michael Harré, and Joseph T Lizier. *Transfer entropy*. Springer, 2016.
 - [26] Lionel Barnett, Adam B Barrett, and Anil K Seth. Granger causality and transfer entropy are equivalent for gaussian variables. *Physical review letters*, 103(23):238701, 2009.
 - [27] Marco Baldovin, Fabio Cecconi, and Angelo Vulpiani. Understanding causation via correlations and linear response theory. *Physical Review Research*, 2(4):043436, 2020.
 - [28] Marco Baldovin, Fabio Cecconi, Antonello Provenzale, and Angelo Vulpiani. Extracting causation from millennial-scale climate fluctuations in the last 800 kyr. *Scientific Reports*, 12(1):15320, 2022.
 - [29] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

- [30] Steven L Brunton and J Nathan Kutz. *Data-driven science and engineering: Machine learning, dynamical systems, and control*. Cambridge University Press, 2022.
- [31] Steven L Brunton, Bernd R Noack, and Petros Koumoutsakos. Machine learning for fluid mechanics. *Annual review of fluid mechanics*, 52:477–508, 2020.
- [32] Michele Buzzicotti, Fabio Bonaccorso, P Clark Di Leoni, and Luca Biferale. Reconstruction of turbulent data with deep generative models for semantic inpainting from turb-rot database. *Physical Review Fluids*, 6(5):050503, 2021.
- [33] Julien Brajard, Alberto Carrassi, Marc Bocquet, and Laurent Bertino. Combining data assimilation and machine learning to emulate a dynamical model from sparse and noisy observations: A case study with the lorenz 96 model. *Journal of computational science*, 44:101171, 2020.
- [34] Norbert Wiener. Nonlinear prediction and dynamics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 3: Contributions to Astronomy and Physics*, volume 3, pages 247–253. University of California Press, 1956.
- [35] A. Seth. Granger causality. *Scholarpedia*, 2(7):1667, 2007. revision #127333.
- [36] 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.
- [37] Ryan G James, Nix Barnett, and James P Crutchfield. Information flows? a critique of transfer entropies. *Physical review letters*, 116(23):238701, 2016.
- [38] Ernesto Pereda, Rodrigo Quian Quiroga, and Joydeep Bhattacharya. Nonlinear multivariate analysis of neurophysiological signals. *Progress in neurobiology*, 77(1-2):1–37, 2005.
- [39] Pearl Judea. An introduction to causal inference. *The International Journal of Biostatistics*, 6(2):1–62, 2010.
- [40] Ryogo Kubo, Morikazu Toda, and Natsuki Hashitsume. *Statistical physics II: nonequilibrium statistical mechanics*, volume 31. Springer Science & Business Media, 2012.
- [41] Roberto Livi and Paolo Politi. *Nonequilibrium statistical physics: a modern perspective*. Cambridge University Press, 2017.
- [42] Massimo Falcioni, Stefano Isola, and Angelo Vulpiani. Correlation functions and relaxation properties in chaotic dynamics and statistical mechanics. *Physics Letters A*, 144(6-7):341–346, 1990.
- [43] Giovanni Ciccotti and Gianni Jacucci. Direct computation of dynamical response by molecular dynamics: The mobility of a charged lennard-jones particle. *Physical Review Letters*, 35(12):789, 1975.
- [44] Denis J Evans and Gary P Morriss. *Statistical mechanics of nonequilibrium liquids*. ANU Press, 2007.
- [45] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Springer, 2006.
- [46] F Rosenblatt. Principles of neurodynamics: Perceptrons and the theory of brain mechanisms, 1961.
- [47] Bernhard Mehlig. *Machine learning with neural networks: an introduction for scientists and engineers*. Cambridge University Press, 2021.
- [48] 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.
- [49] Peter Eris Kloeden, Eckhard Platen, and Henri Schurz. *Numerical solution of SDE through computer experiments*. Springer Science & Business Media, 2012.
- [50] Michele Alessandro Bucci, Onofrio Semeraro, Alexandre Allauzen, Sergio Chibbaro, and Lionel Mathelin. Curriculum learning for data-driven modeling of dynamical systems. *The European Physical Journal E*, 46(3):12, 2023.
- [51] Pantelis R Vlachas, Jaideep Pathak, Brian R Hunt, Themistoklis P Sapsis, Michelle Girvan, Edward Ott, and Petros Koumoutsakos. Backpropagation algorithms and reservoir computing in recurrent neural networks for the forecasting of complex spatiotemporal dynamics. *Neural Networks*, 126:191–217, 2020.
- [52] Benjamin Schrauwen, David Verstraeten, and Jan Van Campenhout. An overview of reservoir computing: theory, applications and implementations. In *Proceedings of the 15th european symposium on artificial neural networks. p. 471-482 2007*, pages 471–482, 2007.
- [53] Jaideep Pathak, Brian Hunt, Michelle Girvan, Zhixin Lu, and Edward Ott. Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach. *Physical review letters*, 120(2):024102, 2018.
- [54] L Storm, Kristian Gustavsson, and Bernhard Mehlig. Constraints on parameter choices for successful time-series prediction with echo-state networks. *Machine Learning: Science and Technology*, 3(4):045021, 2022.
- [55] L Storm, Hampus Linander, J Bec, Kristian Gustavsson, and Bernhard Mehlig. Finite-time lyapunov exponents of deep neural networks. *Physical Review Letters*, 132(5):057301, 2024.
- [56] Edward N Lorenz. Deterministic nonperiodic flow. *Journal of atmospheric sciences*, 20(2):130–141, 1963.
- [57] Edward Ott. *Chaos in dynamical systems*. Cambridge university press, 2002.
- [58] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.
- [59] Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, March 2016.
- [60] Brian de Silva, Kathleen Champion, Markus Quade, Jean-Christophe Loiseau, J. Kutz, and Steven Brunton. Pysindy: A python package for the sparse identification of nonlinear dynamical systems from data. *Journal of Open Source Software*, 5(49):2104, 2020.
- [61] Alan A. Kaptanoglu, Brian M. de Silva, Urban Fasel, Kadierdan Kaheman, Andy J. Goldschmidt, Jared Callahan, Charles B. Delahunt, Zachary G. Nicolaou, Kathleen Champion, Jean-Christophe Loiseau, J. Nathan Kutz, and Steven L. Brunton. Pysindy: A comprehensive python package for robust sparse system identification. *Journal of Open Source*

Software, 7(69):3994, 2022.

- [62] Angelo Vulpiani. *Chaos: from simple models to complex systems*, volume 17. World Scientific, 2010.
- [63] Jean-Christophe Loiseau and Steven L Brunton. Constrained sparse galerkin regression. *Journal of Fluid Mechanics*, 838:42–67, 2018.
- [64] Nicolaas G Van Kampen. The case against linear response theory. *Physica Norvegica*, 5(3-4):279–284, 1971.
- [65] Rep Kubo. The fluctuation-dissipation theorem. *Reports on progress in physics*, 29(1):255, 1966.
- [66] Patrizia Castiglione, Massimo Falcioni, Annick Lesne, and Angelo Vulpiani. *Chaos and coarse graining in statistical mechanics*. Cambridge University Press, 2008.
- [67] Guido Boffetta, Massimo Cencini, Massimo Falcioni, and Angelo Vulpiani. Predictability: a way to characterize complexity. *Physics reports*, 356(6):367–474, 2002.
- [68] F. Krzakala, M. Mézard, F. Sausset, Y.F. Sun, and L. Zdeborová. Statistical-physics-based reconstruction in compressed sensing. *Physical Review X*, 2(2):021005, 2012.
- [69] S. Ganguli and H. Sompolinsky. Statistical mechanics of compressed sensing. *Physical review letters*, 104(18):188701, 2010.
- [70] S. Chibbaro and C. Furtlehner. Typical behavior of sparse causal regression indicator on large scale causal graph. In preparation.
- [71] V. A. Marchenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.
- [72] O. Ledoit and S. Péché. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1):233–264, 2011.
- [73] C. Furtlehner. Free dynamics of feature learning processes. *J.Stat.Phys.*, 190(3):51, 2023.
- [74] The study of an “impulsive” perturbation is not a severe limitation, for instance in the linear regime the (differential) linear response describes the effect of a generic perturbation.
- [75] Giovanni Paladin and Angelo Vulpiani. Anomalous scaling laws in multifractal objects. *Physics Reports*, 156(4):147–225, 1987.
- [76] David Ruelle. General linear response formula in statistical mechanics, and the fluctuation-dissipation theorem far from equilibrium. *Physics Letters A*, 245(3-4):220–224, 1998.

Appendix A: Response theory

We report a derivation of a general FDR, based on previous works [9, 28] Consider a dynamical system $\mathbf{x}(0) \rightarrow \mathbf{x}(t) = S^t \mathbf{x}(0)$ with states $\mathbf{x}$ belonging to a N -dimensional vector space. We consider the possibility that the time evolution could be described by stochastic differential equations. We assume that the system is mixing and that the invariant probability distribution $\rho(\mathbf{x})$ enjoys some regularity property, while no assumption is made on N . The purpose is to express the average response of a generic observable A to a perturbation, in terms of suitable correlation functions, computed according to the invariant measure of the unperturbed system. The first step is to study the behavior of a single component of $\mathbf{x}$, say x_i , when the system, described by $\rho(\mathbf{x})$, is subjected to an initial (non-random) perturbation $\mathbf{x}(0) \rightarrow \mathbf{x}(0) + \Delta \mathbf{x}_0$ [74]. This instantaneous kick modifies the density of the system into $\rho'(\mathbf{x})$, which is related to the invariant distribution by $\rho'(\mathbf{x}) = \rho(\mathbf{x} - \Delta \mathbf{x}_0)$. We introduce the probability of transition from $\mathbf{x}_0$ at time 0 to $\mathbf{x}$ at time t , $W(\mathbf{x}_0, 0 \rightarrow \mathbf{x}, t)$. For a deterministic system, with evolution law $\mathbf{x}(t) = S^t \mathbf{x}(0)$, the probability of transition reduces to $W(\mathbf{x}_0, 0 \rightarrow \mathbf{x}, t) = \delta(\mathbf{x} - S^t \mathbf{x}_0)$, where $\delta(\cdot)$ is the Dirac delta function. Then we can write an expression for the mean value of the variable x_i , computed with the density of the perturbed system:

$$\langle x_i(t) \rangle' = \int \int x_i \rho'(\mathbf{x}_0) W(\mathbf{x}_0, 0 \rightarrow \mathbf{x}, t) d\mathbf{x} d\mathbf{x}_0 . \quad (\text{A1})$$

The mean value of x_i during the unperturbed evolution can be written in a similar way:

$$\langle x_i(t) \rangle = \int \int x_i \rho(\mathbf{x}_0) W(\mathbf{x}_0, 0 \rightarrow \mathbf{x}, t) d\mathbf{x} d\mathbf{x}_0 . \quad (\text{A2})$$

Therefore, defining $\overline{\delta x_i} = \langle x_i \rangle' - \langle x_i \rangle$, we have:

$$\begin{aligned} \overline{\delta x_i}(t) &= \int \int x_i \frac{\rho(\mathbf{x}_0 - \Delta \mathbf{x}_0) - \rho(\mathbf{x}_0)}{\rho(\mathbf{x}_0)} \rho(\mathbf{x}_0) W(\mathbf{x}_0, 0 \rightarrow \mathbf{x}, t) d\mathbf{x} d\mathbf{x}_0 \\ &= \langle x_i(t) F(\mathbf{x}_0, \Delta \mathbf{x}_0) \rangle \end{aligned} \quad (\text{A3})$$

where

$$F(\mathbf{x}_0, \Delta \mathbf{x}_0) = \left[\frac{\rho(\mathbf{x}_0 - \Delta \mathbf{x}_0) - \rho(\mathbf{x}_0)}{\rho(\mathbf{x}_0)} \right] . \quad (\text{A4})$$

Note that the system is assumed to be mixing, so that the decay to zero of the time-correlation functions prevents any departure from equilibrium.

For an infinitesimal perturbation $\delta\mathbf{x}(0) = (\delta x_1(0) \cdots \delta x_N(0))$, the function in (A4) can be expanded to first order, if $\rho(\mathbf{x})$ is non-vanishing and differentiable, and one obtains:

$$\begin{aligned} \overline{\delta x_i}(t) &= - \sum_j \left\langle x_i(t) \frac{\partial \ln \rho(\mathbf{x})}{\partial x_j} \Big|_{t=0} \right\rangle \delta x_j(0) \\ &\equiv \sum_j \mathcal{R}_{j \rightarrow i}(t) \delta x_j(0) \end{aligned} \quad (\text{A5})$$

which defines the linear response

$$\mathcal{R}_{j \rightarrow i}(t) = - \left\langle x_i(t) \frac{\partial \ln \rho(\mathbf{x})}{\partial x_j} \Big|_{t=0} \right\rangle \quad (\text{A6})$$

of the variable x_i with respect to a perturbation of x_j . One can easily repeat the computation for a generic observable $A(\mathbf{x})$, obtaining:

$$\overline{\delta A}(t) = - \sum_j \left\langle A(\mathbf{x}(t)) \frac{\partial \ln \rho(\mathbf{x})}{\partial x_j} \Big|_{t=0} \right\rangle \delta x_j(0) . \quad (\text{A7})$$

In the above derivation of the FDR, we did not use any approximation on the evolution of $\delta\mathbf{x}(t)$. Starting with the exact expression (A3) for the response, only a linearization of the initial perturbed density is needed, and this implies only the smallness of the initial perturbation. From the evolution of the trajectories difference, one can define the maximum Lyapunov exponent λ , by considering the positive quantities $|\delta\mathbf{x}(t)|$, so that at small $|\delta\mathbf{x}(0)|$ and large enough t one can write

$$\left\langle \ln |\delta\mathbf{x}(t)| \right\rangle \simeq \ln |\delta\mathbf{x}(0)| + \lambda t . \quad (\text{A8})$$

Differently, in the derivation of the FDR, one deals with averages of quantities with sign, such as $\overline{\delta\mathbf{x}(t)}$. This apparently marginal difference is very important and underlies the possibility of deriving the FDR without incurring in van Kampen's objection [42, 64].

At this point one could object that in chaotic deterministic dissipative systems the above machinery cannot be applied, because the invariant measure is not smooth. It is worth emphasising that, even though in chaotic dissipative systems the invariant measure is singular [75], the previous derivation of the FDR is still valid if one considers perturbations along the expanding directions. In addition, one is often interested in some specific variables, so that a projection is performed, making irrelevant the singular character of the invariant measure [76]. In these cases, a general response function has two contributions, corresponding respectively to the expanding (unstable) and the contracting (stable) directions of the dynamics. The first contribution can be associated to some correlation function of the dynamics on the attractor (i.e. the unperturbed system). Nevertheless, a small amount of noise, always present in physical systems and numerical simulations, smoothens the $\rho(\mathbf{x})$ and the FDR can be derived. Then, the assumption on the smoothness of the invariant measure along the unstable manifold still allows to avoid subtle technical difficulties [76].

Appendix B: Lorenz system Cross-correlations

In this appendix, we provide some complementary results to those presented in section V. In Fig. 11, we show the cross-correlations between the different variables on a long-time horizon. We focus on the MLP architecture, since SINDy basically overlaps with the empirically simulated Lorenz system. While the correlation between x and y are quite well reproduced, differences are displayed in the others where z is present. These correlations are indeed much weaker, and the ratio noise-signal larger. In these cases, qualitative dynamics may be captured, but the difference is important even at short times. These results shows that the MLP is able to well predict dominant contributions, which leads to an excellent reproduction of the phase-space dynamics. Yet, sub-dominant correlations are not well captured, and these errors eventually are magnified by chaotic dynamics leading to the difference in the Response function.

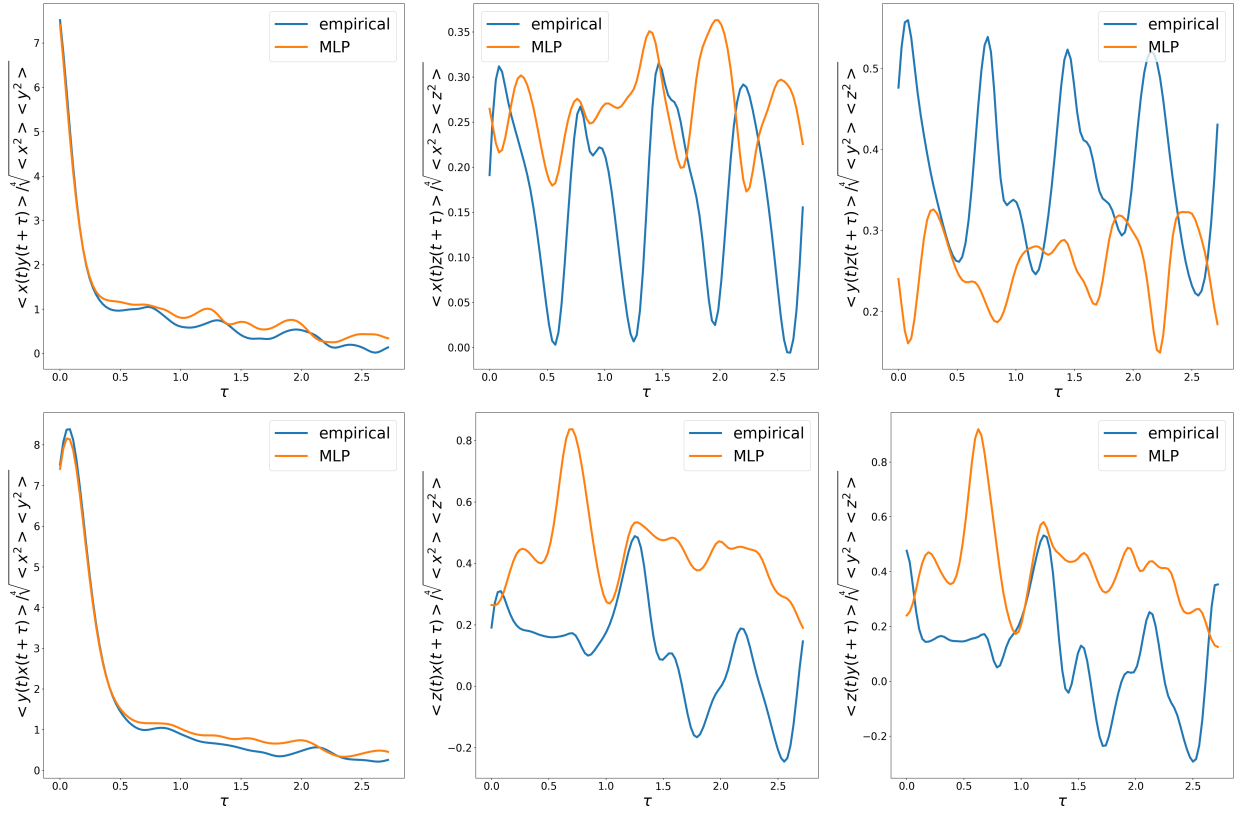


FIG. 11. Normalized auto-correlation function $\langle x_i(t)x_j(t+\tau) \rangle$ for the MLP model compared against that computed through direct integration of the Lorenz system Eq. (11) (blue line). The time is expressed in terms of the Lyapunov time based on the maximum Lyapunov exponent.

Appendix C: Asymptotic regime of the causal response with random matrix theory

We are interested to obtain the asymptotic behavior of the LRR causal indicator in the so-called proportional scaling limit, namely when the number of independent samples N goes to infinity, along with the number of features - i.e. the number of observed nodes d in our case - with fixed ratio $\rho = N/d$. This can be obtained by some random matrix formulas based on RMT [71, 72], which we simply state here. Let ν be the spectral distribution of C . The rescaled modulus $|\mathbf{x}_t|^2/d$ concentrates on a deterministic value in the large scale limit. This leads us to obtain a simpler form of the RMT equations than the general Marchenko-Pastur equations. Considering the quantity

$$\Gamma \stackrel{\text{def}}{=} \lim_{N, P \rightarrow \infty} \frac{\alpha}{N} \text{Tr}[G^{(N)}C],$$

we get the following self-consistent equations:

$$\Gamma = \frac{\alpha}{\rho} \int \nu(dx) \frac{x}{1 + \Lambda x} \quad (\text{C1})$$

$$\Lambda = \frac{\alpha}{1 + \Gamma} \quad (\text{C2})$$

These in turn define a resolvent $G = (\mathbb{I} + \Lambda C)^{-1}$, where Λ represents a self-energy when using the terminology of field theory. G represents a deterministic equivalent of $G^{(N)}$ [? ?], in the sense that for any sequence of $\mathbb{R}^d \times \mathbb{R}^d$ matrix M of Frobenius norm equal to one, $\text{Tr}[(G - G^{(N)})M]$ goes to zero typically like $\sqrt{\log(N)/N}$. In order to leverage this we need first to get rid of factors of the form $G^{(N)^2}$ or $G^{(N)}\mathbf{e}_i\mathbf{e}_i^\top G^{(N)}$ present in equation (20). For $G^{(N)^2}$ we make use of the following identity:

$$G^{(N)^2} = G^{(N)} + \alpha \frac{d}{d\alpha} G^{(N)}$$

This means that we need the derivative of G with respect to α :

$$\frac{d}{d\alpha} G = \frac{\Lambda'(\alpha)}{\Lambda} (G^2 - G)$$

$\Lambda'(\alpha)$ can be obtained from the self-consistent equation (C1,C2). We have

$$\Lambda'(\alpha) = \frac{\Lambda^2}{\alpha^2} \frac{1}{1-Q}$$

after introducing the quantity

$$Q = \frac{1}{\rho} \int \nu(dx) \frac{(\Lambda x)^2}{(1 + \Lambda x)^2}.$$

Note that $Q < 1$ for $\rho > 1$, while in the overparameterized regime ($\rho < 1$) it can tend to 1 when the L_2 regularization tends to zero, which results in a divergence of the test error in that case (see e.g. [73] for a thorough discussion and interpretation of these quantities regarding overfitting). For the second term, we use a similar trick by defining

$$G_\gamma^{(N)} = (\mathbb{I} + \gamma \mathbf{e}_i \mathbf{e}_i^\top + \alpha C^{(N)})^{-1},$$

as a generating function. As can be shown the corresponding deterministic equivalent takes the form

$$G_\gamma = (\mathbb{I} + \gamma \mathbf{e}_i \mathbf{e}_i^\top + \Lambda_i(\gamma) C)^{-1},$$

with Λ_i solution of the fixed point equations (C1,C2) now depends on γ . Expressing the fixed point equations formally at finite N to make the expansion with respect to γ meaningful, we now have

$$\Gamma(\gamma) = \frac{\alpha}{N} \text{Tr} \left[\frac{1}{\mathbb{I} + \gamma \mathbf{e}_i \mathbf{e}_i^\top + \Lambda_i(\gamma) C} C \right],$$

while the factor of interest now takes the desired form

$$G^{(N)} \mathbf{e}_i \mathbf{e}_i^\top G^{(N)} = \frac{d}{d\gamma} G_\gamma^{(N)} \Big|_{\gamma=0},$$

directly amenable to asymptotic analysis. Indeed the deterministic equivalent, which now depends on γ can be inserted to yield

$$\frac{d}{d\gamma} G_\gamma \Big|_{\gamma=0} = G(\mathbf{e}_i \mathbf{e}_i^\top + \Lambda'_i(0) C) G.$$

Using the fixed point equation we get

$$\Lambda'_i(0) = \frac{\Lambda}{N} \frac{\mathbf{e}_i^\top G(1-G) \mathbf{e}_i}{1-Q}.$$

Assembling everything we finally get the following asymptotic estimation of the response in the proportional limit (up to leading $\mathcal{O}(1/N)$ contribution)

$$\begin{aligned} \chi_{i \rightarrow j}(t) &= \text{Tr} \left[(\mathbb{I} - G) \mathbf{e}_i \mathbf{e}_i^\top (\mathbb{I} - G) M^t M^\top \mathbf{e}_j \mathbf{e}_j^\top M^t \right] \\ &+ \frac{1}{N} \frac{1}{1-Q} \mathbf{e}_i^\top G (\mathbb{I} - G) \mathbf{e}_i \left(\text{Tr} \left[G (\mathbb{I} - G) M^t M^\top \mathbf{e}_j \mathbf{e}_j^\top M^t \right] + \Lambda \sigma^2 \sum_{\tau=0}^{t-1} \mathbf{e}_j^\top M^{t-\tau-1} M^{t-\tau-1^\top} \mathbf{e}_j \right) \end{aligned} \quad (\text{C3})$$

The first two terms represent the deterministic (regularized) causal response obtained in absence of noise. They basically corresponds to the (regularized) weighted sum of round trips between i and j . The last term is a variance term resulting from the noise. Quantities like the overlap (22) and $\|\mathcal{R}_{\rightarrow i}^{(N)}(t)\|$ become deterministic asymptotically. We actually obtain

$$\begin{aligned} \mathbf{f}_{\rightarrow i}^\top \mathcal{R}_{\rightarrow i}(t) &= \mathbf{e}_i^\top M^t (\mathbb{I} - G) M^t M^\top \mathbf{e}_i, \\ \|\mathcal{R}_{\rightarrow i}^{(N)}(t)\|^2 &= \mathbf{e}_i^\top M^t (\mathbb{I} - G)^2 M^t M^\top \mathbf{e}_i + \Delta \left[\mathbf{e}_i^\top M^t G (\mathbb{I} - G) M^t M^\top \mathbf{e}_i + \Lambda \sigma^2 \sum_{\tau=0}^{t-1} \mathbf{e}_i^\top M^{t-\tau-1} M^{t-\tau-1^\top} \mathbf{e}_i \right] \end{aligned} \quad (\text{C4})$$

with

$$\Delta \stackrel{\text{def}}{=} \frac{1}{1-Q} \int \nu(dx) \frac{\Lambda x}{(1+\Lambda x)^2}.$$