


M2 International Physics of Complex Systems

Master's thesis under the supervision of
Cyril Furtlehner [1] and Sergio Chibbaro [2]

Causality in Machine Learning and the Large Dimensional Limit

Abstract

In this thesis, we investigate the incorporation of causality into machine learning models, particularly focusing on large dimensional data. The work starts with an exploration of linear regression within machine learning, highlighting its connections with linear response theory, a field of statistical physics. We derive a closed-form solution for the empirical response function, accounting for bias and variance. The study extends to high-dimensional data, addressing the complexities of identifying causal relationships and the curse of dimensionality. Numerical simulations show that observational causality can be inferred in high-dimensional contexts with linear dependencies among variables. This thesis bridges principles from statistical mechanics and machine learning, offering new perspectives on causal relationships in complex systems.

Théo MARCHETTA
2023-2024

LISN, team TAU
INRIA, SACLAY

15 june 2024

[1] cyril.furtlehner@inria.fr
[2] sergio.chibbaro@universite-paris-saclay.fr

Contents

1 Introduction

Nature is a complex causal network where every event can be seen as the end point of a causal chain. Thus, the causal relationships between different entities has always been of interest for the scientists. This is particularly studied in the context of Earth's climate system, an up to date exemple being *El Niño*. The origin of *El Niño* is a time-periodic perturbation of the atmospheric circulation. This phenomenon then causes an heating of the east coast of the pacific ocean, itself inducing some global movements of fishes and a perturbation of the climate of the area [1]. A better knowledge of the behavior of this climate phenomenon would be helpful to predict and limitate the socio-economical effects. For a physical system, causality is a principle at the basis of special relativity, which induces no faster-than-light flow of information. This has been a source of motivation to understand deeper our knowledge of the world due to numerous paradoxes that emerged, the most famous one being the EPR paradox, where locality and causality are closely related. This paradox has been solved in the early 80's by the 2022 Nobel Prize Alain Aspect [2].
Note also that two variables may be correlated even when no causal link is present. This is called a spurious relationship, the main reason being the presence of a third variable being causal of both of them. This misconception has been already identified in the old societies, which already warned the fallacy of the argument with the sentence "cum hoc sed non propter hoc" ("with this, therefore because of this"), to be distanced from "post hoc ergo propter hoc" ("after this, therefore because of this").

There are two ways to obtain information about causal relationships. The first one is called the *interventional causality*, where you modify one parameter of your system and see the evolution of the others. This approach has been built on a bayesian theory by the computer scientist Judea Pearl [3]. When the intervention on your system is small enough, one can use the linear response theory, which has first been developed in order to study out-of-equilibrium systems in statistical physics. For a nondeterministic system with two variables x_i and x_j, the response from x_j at time 0 to x_i at time t is defined as :

$$R_{j\to i}(t) = R_{ij}(t) \equiv \lim_{\delta x_{0,j}\to 0} \frac{\overline{\delta x_{t,i}}}{\delta x_{0,j}} \tag{1}$$

where the overline stands for an average over the probability distribution of the system we are looking at.
However, if you apply a small perturbation on your system, this one doesn't "know" if the perturbation is done by an external agent or by random fluctuations. Thus, it is possible for certain systems to compute the theoretical relaxation after the perturbation. This is Onsager's regression hypothesis [4].
In the case of non-linear systems, depending on the strength of the non-linearity, the linear response approximation may break down. In this case, one may use the tools from Machine Learning (ML) and try to build a specific neural network which is able to learn the true non-linear causal links between variables. This

idea has been explored by an ex master student in the lab [5]. However, new architectures may be needed in order for the problem to be called solved.

In a real-world setting, having an intervention on a specific system may be difficult or impossible due to technical reasons. Thus, one would like to infer the causality between variables just by having in possession a set of data. This *observational causality* is said to be harder to reach, because not perturbing the system gives you less information about it. It is however a much more flexible technique. Finding a general framework in order to consistently recover the causality between different parameters would be a lead forward as it would significantly help our comprehension of the world around us.
The objective of this master's thesis is to infer observationaly and analyse the causality between different variables interacting in a linear way. In particular, we will see how the noise, which may be intrinsic to our system or linked to hidden variables may impact the dynamics of learning of our model. Interestingly, if the number of parameters interacting together is sufficiently big, one can expect some sort of concentration of measures [6].
This would mean that, in certain limits, the theory of Random Matrices, which will be introduced thereafter may be useful in order to make conclusions on our model and to get additional insights.

2 Model

2.1 Vector autoregressive model

We will put ourself in the frame of graph theory. Let $G = (V, E)$ be a directed graph where V is the set of vertices. In the following, $|V| = D$ where D will be called the dimension of our system while E is the set of vertices. This graph can be totally described by an adjacency matrix A where the value A_{ij} represents the weight of the edge linking node i to node j. If $A_{ij} = 0$, no connexion is present between edges i and j.
We are interested in the system :

$$\boldsymbol{X_{t+1}} = A\boldsymbol{X_t} + \sigma\boldsymbol{\xi_t} \tag{2}$$

where $\boldsymbol{X_t}, \boldsymbol{\xi_t} \in \mathbb{R}^D$, $\boldsymbol{\xi_t} \sim \mathcal{N}(0, \mathbb{1})$ is a gaussian white noise, that is $\left\langle \boldsymbol{\xi_t}\boldsymbol{\xi_{t'}^\top} \right\rangle = \delta_{t,t'}\mathbb{1}_D$ and $\mathbb{1}_D$ is the identity matrix of size $D \times D$. Since no confusion will occur in the following, we will simply write it as $\mathbb{1}$. For the sake of convergence, the spectral radius of A need to be strictly smaller than 1. If one of the eigenvalues of the response matrix is equal to 1, then the system is said to be marginal. However, due to the noise term, no statistical convergence will be observed as the system would obey an usual D-dimensional random walk.
Eq (2) has the structure of a *vector autoregression*, that is a model evolving in time where every variable depends linearly on the others. More specially, it is said to be of order 1 because the state of the D-dimensional vector $\boldsymbol{X}$ at time t only depends on the state of $\boldsymbol{X}$ at time $t - 1$.
This type of equation has been largely studied in the litterature. Let us give

some usual properties that will be needed in the following of this thesis.

2.2 Some properties

Eq (2) is a discrete stochastic differential equation. In the limit of large time, this can be made continuous :

$$d\boldsymbol{X} = (A - \mathbb{1})\boldsymbol{X}dt + \sigma d\boldsymbol{W}(t) \tag{3}$$

where $d\boldsymbol{W}(t)$ is a Wiener process. We see that this model converges to an Ornstein-Uhlenbeck process, which eventually relaxes to its mean value in the case of a positive semi-definite matrix A. The full derivation of the probability of a specific state $\boldsymbol{X^*}$ as a function of time, $p(\boldsymbol{X} = \boldsymbol{X^*}, t)$ has been done in appendix A.

- Expected value.

$$\mathbb{E}[\boldsymbol{X_{t+1}}] = A\mathbb{E}[\boldsymbol{X_t}] + \mathbb{E}[\boldsymbol{\xi_t}] \tag{4}$$

Suppose that the process starts at a time $t_{start} = -\infty$. Then, we may expect that for a time $t > 0$, the process has attained a stationary state, that is all the statistical variables such as the expected value do not depend on time. This means that we have $\mathbb{E}[\boldsymbol{X_{t+1}}] = \mathbb{E}[\boldsymbol{X_t}] \equiv \mathbb{E}[\boldsymbol{X}]$ and, by using the property of the noise :

$$\mathbb{E}[\boldsymbol{X}] = 0 \tag{5}$$

- Correlation.

By multiplying Eq (2) by $\boldsymbol{X_{t+1}^\top}$, in the case of a stationary state, we end up with :

$$C_0 \equiv \mathbb{E}[\boldsymbol{X}\boldsymbol{X}^\top] = AC_0A^\top + \sigma^2\mathbb{1} \tag{6}$$

This equation is know as a discrete-time Lyapunov equation. The closed form solution can be written as an infinite sum :

$$C_0 = \sigma^2 \sum_{k=0}^{+\infty} A^k A^{k\top} \tag{7}$$

The subscript 0 is associated to the time delay between the two variables for which we compute the correlation time. C_0 is also called the *population matrix*.

-Autocorrelation.

The autocorrelation matrix can be expressed as a function of the population matrix. For a delay of 1, we have :

$$\begin{aligned} C_1 \equiv \mathbb{E}[\boldsymbol{X_{\tau+1}}\boldsymbol{X_\tau^\top}] &= \mathbb{E}[(A\boldsymbol{X_\tau} + \boldsymbol{\xi_\tau})(\boldsymbol{X_\tau})] \\ &= AC_0 \end{aligned} \tag{8}$$

This is because the noise sampled at time τ, $\boldsymbol{\xi_\tau}$ can only impact the variables at time at least $t+1$. Consequently, $\mathbb{E}[\boldsymbol{\xi_\tau} \boldsymbol{X_\tau}] = 0$. More generally, if one looks at the correlation at a time difference t :

$$C_t \equiv \mathbb{E}[\boldsymbol{X_{t+\tau}} \boldsymbol{X_\tau}^\top] = A^t C_0 \tag{9}$$

Finally, another useful relation for the following is :

$$\boldsymbol{X_t} = A^t \boldsymbol{X_0} + \sum_{\tau=0}^{t-1} A^{t-\tau-1} \boldsymbol{\xi_\tau} \tag{10}$$

which can be found by doing a simple recurrence, starting from Eq (2).

2.3 Linear regression in Machine Learning is related to linear response theory

Now, imagine that you have a finite set of measures during a time serie of length T, $\{\{\boldsymbol{X_t^s}\}, t \in [\![0; T-1]\!]; s \in [\![1; N]\!]\}$ and that you want to do a linear regression in order to fit your data. The objective is to infer the response matrix $\hat{R}(t)$ at each discrete time that fits the best your observations, that is:

$$\boldsymbol{X_t} = \hat{R}(t) \boldsymbol{X_0} \tag{11}$$

In the following, the notation $\hat{R}(t)$ is intended to be read as $\hat{R}^t$ for any matrix. This is to avoid any confusion with the transpose of a matrix.
We see that, in this case, the matrix $\hat{R}(t)$ can be interpreted as a response function defined in the linear response theory. To begin with, we have :

$$\hat{R}_{ij}(t) = \frac{\partial X_{t,i}}{\partial X_{0,j}} \tag{12}$$

Moreover, by taking an empirical average defined, for an observable O as $\langle O \rangle_N \equiv 1/N \sum_{s=1}^N O^s$, we have :

$$\left\langle \hat{R}_{ij}(t) \right\rangle_N = \left\langle \frac{\partial X_{t,i}}{\partial X_{0,j}} \right\rangle \tag{13}$$

Moreover, if we suppose the process to be already stationary at time $t = 0$ and in the case where we have a sufficient amount of measures, the empirical average can be seen as an approximation of an ensemble average. In this limiting case, the empirical response converges to the response defined in the linear response theory, as exposed in Eq (1) :

$$\lim_{N \to \infty} \left\langle \hat{R}_{ij}(t) \right\rangle_N \to R_{ij}(t) \tag{14}$$

2.4 Determination of the optimal empirical response function

When there are fewer number of measures than the number of parameters in a ML model, in our case $N < D$, we fall in the *undetermined* regime where the optimized values of our set of parameters is not unique. On the contrary, if you have a number of measures that is close but superior to the size of the matrix $\hat{R}(t)$, $N \gtrsim D$ one falls in the overfitting regime where your function inferred fits almost perfectly your dataset. However, this model would have high *variance* due to its high dependency to the dataset we are given.

More quantitatively, the usual protocol that one does to best determine $\hat{R}(t)$ is to minimize the overall euclidian distance between $\boldsymbol{X_t}$ and $\hat{R}(t)\boldsymbol{X_0}$, taking into account the N different measures. This can be done by defining a *loss function* built on the mean squared error :

$$\mathfrak{L}[\hat{R}(t)] = \frac{1}{2N}\sum_{s=1}^{N}\left(\boldsymbol{X_t^s} - \hat{R}(t)\boldsymbol{X_0^s}\right)^{\top}\left(\boldsymbol{X_t^s} - \hat{R}(t)\boldsymbol{X_0^s}\right) \tag{15}$$

and minimizing it with respect to every components of the inferred matrix $\hat{R}_{ij}(t)$. To counteract the possible overfitting, one can add another constraint on the norm of the elements of the matrix. This additional term is called a *regularizer* and is added to the loss function :

$$\mathfrak{L}_{reg}[\hat{R}(t)] = \alpha\mathfrak{L}[\hat{R}(t)] + L(\hat{R}) \tag{16}$$

Usually, what is used in the literature are function of the form $L(\hat{R}) = \sum_{i,j}|\hat{R}_{ij}|$ (L1 regularization) or $L(\hat{R}) = \frac{1}{2}Tr[\hat{R}(t)^{\top}\hat{R}(t)]$ (L2 regularization). L1 regularization tends to make the inferred matrix sparser than L2. Note that we introduced a parameter α in order to give more or less power to this regularization. This parameter has to be fine tuned empirically in order to see how good is the generalization of your model.
In the following, we will concentrate ourselves with the L2 regularizer, as the L1 one do not give a closed form expression for the coefficient of the response matrix. This is because the function $y = |x|$ is not differentiable at $x = 0$ [7].
To solve the inference problem, first we define the sample correlation matrix $\hat{C}_0$ and the sample autocorrelation matrix $\hat{C}_t$ as :

$$\hat{C}_0 \equiv \frac{1}{N}\sum_{s=1}^{N}\boldsymbol{X_0^s}\boldsymbol{X_0^{s\top}} \tag{17}$$

$$\hat{C}_t \equiv \frac{1}{N}\sum_{s=1}^{N}\boldsymbol{X_t^s}\boldsymbol{X_0^{s\top}} \tag{18}$$

By using the L2 regularizer and minimizing the loss function expressed in Eq (16), one finds (for a demonstration, see appendix B) :

$$\hat{R}(t) = \alpha\hat{C}_t\hat{G}^{(D)} \tag{19}$$

Where we defined $\hat{G}^{(D)} \equiv (\mathbb{1} + \alpha \hat{C}_0)^{-1}$. The superscript D is here specified to show the finite size of our system. This will be analyzed in more detail in the following.
We see in $\hat{G}^{(D)}$ the importance of the regularizer. Adding the identity to the correlation matrix make its diagonal elements bigger, which are themselves related to the variance of the vector state $\boldsymbol{X}$. We can interpret this addition as bringing less correlation between different directions, which means a higher possibility to have a fully-ranked matrix, a necessary criteria for its inversion.
Let us look at both the small and the big α regime. When α tends to $+\infty$, one has $\lim_{\alpha\to\infty} \hat{R}(t) = \hat{C}_t {\hat{C}_0}^+$, where ${\hat{C}_0}^+$ is the *pseudoinverse* of the matrix $\hat{C}_0$. The pseudoinverse is defined as $A^+ \equiv \left(A^\top A\right)^{-1} A^\top$. If the matrix is invertible, then the pseudoinverse reduces to the usual inverse operation. This formula is the one that we recover without adding the regularizer in the loss function.
On the contrary, $\lim_{\alpha\to 0} \hat{R}(t) = 0$. This is the regime where only the L2 regularization occurs and all the values of the matrix $\hat{R}(t)$ tends to be as small as possible, eventually converging to 0.

3 Study of the response function in the classical case, extension to the large dimensional limit

3.1 Causal links in the case of small sizes systems

In the frame of this thesis, we are interested in whether or not there is a causal link from the variable X_j, located at node j, to X_i. Therefore, we will focus ourselves in quantities of the form $\hat{R}_{ij}(t) = \alpha \boldsymbol{e_i}^\top \hat{C}_t \hat{G}^{(D)} \boldsymbol{e_j}$. If a causal link is present, the response is expected to show a non-null value at a time t^*, corresponding to the presence of a directed path from X_j to X_i of length t^* on the underlying graph. To give a clear example which was explored in a previous thesis [5], consider the 3×3 adjacency matrix A as being :

$$A = \begin{pmatrix} 0.5 & \epsilon & 0 \\ 0.5 & 0.5 & 0 \\ 0.5 & 0 & 0.5 \end{pmatrix}$$

And the model we are looking at is given by Eq (2). By defining $\boldsymbol{X} = \begin{pmatrix} X_1 & X_2 & X_3 \end{pmatrix}^\top$. The system can be equally rewritten as :

$$\begin{cases} X_{1,t+1} = 0.5X_{1,t} + \epsilon X_{2,t} \\ X_{2,t+1} = 0.5X_{1,t} + 0.5X_{2,t} \\ X_{3,t+1} = 0.5X_{1,t} + 0.5X_{3,t} \end{cases} \tag{20}$$

We see that X_2 and X_3 are linked to the variable X_1. In the case where $\epsilon = 0$, we expect the response from X_2 to X_3 to be 0. However, if you set ϵ to be a small but non-zero value, a causal relationship should emerge by the intermediary of

X_1.
This can be explained as follow : at the first time step, $X_{1,t+1}$ gets modified depending on the value of $X_{2,t}$ then, at the next time step, the variable $X_{3,t+2}$ is changed depending on the value of $X_{1,t+1}$, and so by the value of$X_{2,t}$. In this case, the variable X_2 is *causal* to X_3. The theoretical response of the system without regularizer, $R(t) = C_t C_0^+$ is shown in Fig 1 for multiple values of ϵ. As expected, the bigger the value of ϵ, the bigger the causal link.

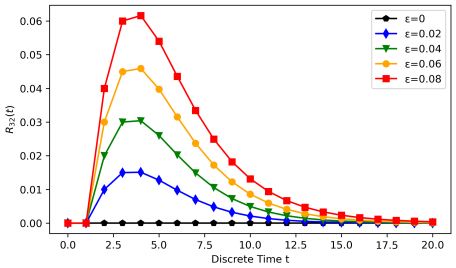


Figure 1: Theoretical response from 2 to 3 for different values of ϵ.

3.2 Derivation of the theoretical response function

By doing an inference, we are facing two sources of errors : the bias and the variance [8]. The bias usually appears because of a simplified model. In our case, we have D^2 parameters, which is exactly the good number of parameters to totally define our system. However, the bias here comes into account because of the regularization that we added in our loss function. This will modify the solution of the optimization as the global minimum of $\mathfrak{L}_{reg}[\hat{R}(t)]$ can only be bigger or equal than the one of $\mathfrak{L}[\hat{R}(t)]$.
Moreover, our model is also sensitive to the training data that we have at our disposal. This is called the variance of the regression. In the case where you have very few measures to fit your model, you expect it to be very data-dependant. On the opposite, if your number of measures in infinite, there should be no variance in your regression. We expect these two terms to show up in our calculation.
We saw in Section 2.3 that the response is averaged on the different measures available. However, our system is also subject to some noise, which has to be averaged. In order to keep a fingerprint of this noise in the computation, we will first look at the square of the response and then average over the noise.
Our objective is to get information on the causal link from X_j to X_i. Thus, the quantity that we will be focusing at is $\hat{R}^2_{ij}(t)$. This can be written as the trace of a scalar, so that the cyclicity property can be used :

$$\hat{R}^2_{j\to i}(t) = \alpha^2 Tr \hat{G}^{(D)} \boldsymbol{e_j} \boldsymbol{e_j}^\top \hat{G}^{(D)} \underbrace{\hat{C}_t{}^\top \boldsymbol{e_i} \boldsymbol{e_i}^\top \hat{C}_t}_{\text{(a)}} \tag{21}$$

Note that the sample autocorrelation matrix can be expressed as a function of the sample correlation matrix by plugging Eq (10) into Eq (18) (remember that $A(t) \equiv A^t$) :

$$\hat{C}_t = A(t)\hat{C}_0 + \frac{1}{N}\sum_{s=1}^{N}\sum_{\tau=0}^{t-1} A(t-\tau-1)\boldsymbol{\xi_\tau^s}\boldsymbol{X_0^s}^\top \tag{22}$$

For now, let us focus us on the last piece of the equation. Using Eq (22), one finds :

$$\begin{aligned}(a) = \hat{C}_0 A^\top(t)\boldsymbol{e_i}\boldsymbol{e_i}^\top A(t)\hat{C}_0 + &\frac{1}{N}\sum_{s,\tau}\boldsymbol{X_0^s}\boldsymbol{\xi_\tau^s}^\top A^\top(t-\tau-1)\boldsymbol{e_i}\boldsymbol{e_i}^\top A(t)\hat{C}_0 + \\ &\frac{1}{N}\sum_{s,\tau}\hat{C}_0 A(t)^\top \boldsymbol{e_i}\boldsymbol{e_i}^\top A(t-\tau-1)\boldsymbol{\xi_\tau^s}\boldsymbol{X_0^s}^\top + \\ &\frac{1}{N^2}\sum_{s,s',\tau,\tau'}\boldsymbol{X_0^s}\boldsymbol{\xi_\tau^s}^\top A^\top(t-\tau-1)\boldsymbol{e_i}\boldsymbol{e_i}^\top A(t-\tau'-1)\boldsymbol{\xi_{\tau'}^{s'}}\boldsymbol{X_0^{s'}}^\top\end{aligned} \tag{23}$$

By averaging the squared response with respect to the noise, $\left\langle R^2_{i\to j}(t)\right\rangle_\xi$ we have :

$$\begin{aligned}\left\langle \hat{R}^2_{j\to i}(t)\right\rangle_\xi = \alpha^2 Tr\hat{G}^{(D)}\boldsymbol{e_j}\boldsymbol{e_j}^\top\hat{G}^{(D)}\times &\\ \Bigg(\hat{C}_0 A^\top(t)\boldsymbol{e_i}\boldsymbol{e_i}^\top A(t)\hat{C}_0 + &\\ \frac{\sigma^2}{N}\sum_\tau \boldsymbol{e_i}^\top A(t-\tau-1)A^\top(t-\tau-1)\boldsymbol{e_i}\hat{C}_0\Bigg)&\end{aligned} \tag{24}$$

Finally, using the identity $\alpha\hat{C}_0\hat{G}^{(D)} = \mathbb{1} - \hat{G}^{(D)}$, one can rewrite the response as :

$$\begin{aligned}\left\langle \hat{R}^2_{j\to i}(t)\right\rangle_\xi = Tr\left\{\boldsymbol{e_i}^\top A(t)\left(\mathbb{1}-\hat{G}^{(D)}\right)\boldsymbol{e_j}\boldsymbol{e_j}^\top\left(\mathbb{1}-\hat{G}^{(D)}\right)A(t)^\top\boldsymbol{e_i}\right\} + &\\ + \alpha\frac{\sigma^2}{N}Tr\left[\left(\mathbb{1}-\hat{G}^{(D)}\right)\boldsymbol{e_j}\boldsymbol{e_j}^\top\hat{G}^{(D)}\right]\boldsymbol{e_i}^\top\left(\sum_\tau A(t-\tau-1)A^\top(t-\tau-1)\right)\boldsymbol{e_i}&\end{aligned} \tag{25}$$

As was said earlier intuitively, the response is now decomposed as a sum of two terms : the first one is associated with the bias due to the regularizer. This can be seen by computing $\lim_{\alpha\to\infty}\hat{G}^{(D)} = 0$. This means that in the case where no regularizer is present, the bias term is exactly given by $A^2_{ij}(t)$. The second term is linked to the variance : we see explicitly that it decays as a inverse function of the number of measure. Thus, this term should be negligible in the scenario where we have at our disposition an infinite number of measures.

3.3 Limiting case : from classical multivariate statistics to the curse of dimensionality.

Now that we got some insights on the different terms of this equation, let us see some other limiting case. For the following, we may define $\rho = \frac{N}{D}$ as being our control parameter.
In the case of a finite size system, and a big number of measures, that is $\rho \to \infty$, one is said to be in the classical multivariate statistics regime, where our intuition usually do not fail. In this regime, we expect a phenomenon of concentration of measure, meaning that $\hat{C}_0$ should converge (almost surely) to the population correlation matrix C_0 [9].
Moreover, by looking back at Eq (6) and remembering that a symmetric semidefinite matrix only has positive eigenvalues, one can see that the spectrum of C_0 is positive and strictly bigger than 1. Thus, its inverse is defined and we can give up on the regularizer introduced earlier. This is the same as taking the limit $N, \alpha \to \infty$. In this regime, one end up with the unbiased response function :

$$\begin{aligned}\left\langle \hat{R}^2_{i\to j}(t) \right\rangle_\xi = Tr\left\{ \boldsymbol{e_i}^\top A(t) \boldsymbol{e_j} \boldsymbol{e_j}^\top A(t)^\top \boldsymbol{e_i} \right\} + \\ + \frac{\sigma^2}{N} Tr\left\{ \boldsymbol{e_i} \boldsymbol{e_j} \boldsymbol{e_j}^\top C_0^{-1} \boldsymbol{e_i}^\top \sum_\tau A(t-\tau-1) A^\top (t-\tau-1) \right\}\end{aligned} \tag{26}$$

However, the interesting regime that we will put ourself in is the so-called *large dimensional limit* (LDL), that is the regime where $N, D \to \infty$ while $\rho = \mathcal{O}(1)$. In this high dimensional frame, our system exhibits some interesting properties and should be analyzed with caution. Let us look at some non-intuitive properties emerging from the LDL.
Consider the easiest exemple for the covariance matrix $\hat{C}_0 = \frac{1}{N} \sum_{s=1}^N \boldsymbol{X^s} \boldsymbol{X^s}^\top$ based on N samples $\boldsymbol{X^s} \sim \mathcal{N}(0, \mathbb{1}), s \in [\![1; N]\!]$. For simultaneously large N, D, the sample covariance matrix $\hat{C}_0$ is only an *entry-wise* consistent estimator of the population covariance matrix $\mathbb{1}$, that is :

$$||\hat{C}_0 - \mathbb{1}||_{ij} \overset{a.s.}{\to} 0, \forall i, j \in [\![1; d]\!] \tag{27}$$

But a poor estimator in the Frobenius norm sense :

$$||\hat{C}_0 - \mathbb{1}||_F \not\to 0 \tag{28}$$

This is because the Frobenius norm, defined as $||M||_F \equiv \left(\sum_{i,j} |M_{ij}|^2 \right)^{1/2}$ consists of a sum of D^2 term. As a consequence, if the convergence of every component of $\hat{C}_0$ to $\mathbb{1}$ is smaller than $O(1/D^2)$, $\hat{C}_0$ cannot be considered as a good estimator to look at.

Now let us put ourselves in the case where $D > N$. Then, $\hat{C}_0$ is the sum of N rank-1 matrices, the rank of $\hat{C}_0$ is at most D and it so has at least $D - N$ null eigenvalues. In this frame, convergence to the identity matrix cannot be attained, even in the entry-wise sense.

These kind of problems that one is dealing in the high-dimensional frame is called the *curse of dimensionality*, which arises nowadays in the context of the Big data era that we are in.

3.4 Determination of the deterministic equivalent

We just saw that in the LDL, $\hat{C}_0$ is a poor estimator. However, it can have a controlled asymptotic behavior. This is where Random Matrix Theory (RMT) arises. It has been shown for the first time in one paper from Marchenko and Pastur [10] that, even if the convergence of a correlation matrix to a deterministic matrix is not given, the spectrum of eigenvalue is universal and converges to a specific distribution, now called the Marchenko-Pastur distribution. In order to show this, they used the Stieltjes transform method, which elegantly links operator theory and complex analysis.
However, this technique does not allow us to have any information on the eigenvectors. Some refined techniques emerged, such as the field of free probability, which is a larger frame than probability theory, as it studies non-commutative random variables, a framework which is particularly suited for random matrices. Another framework of RMT that we will look at in the following is the one where we look at *deterministic equivalents*.
By being able to find deterministic scalar properties of a matrix, such as its distribution of eigenvalues or the expected values of any quadratic form, defined as $\boldsymbol{u}^\top Q \boldsymbol{v}$ for some vectors $\boldsymbol{u}$, $\boldsymbol{v}$, one could be tempted to look for a deterministic matrix which has all the same scalar properties of the random matrix we are interested in.
More formally and to get a suitable example, if one look at a matrix $\hat{G}^{(D)} = \left(\mathbb{1} + \hat{C}_0\right)^{-1}$, where $\hat{C}_0 = \frac{1}{N}\sum_{s=1}^{N} \boldsymbol{X^s}\boldsymbol{X^s}^\top$, then, in the LDL, that is $N, D \to \infty, \rho$ finite, it is possible to find a deterministic matrix G so that :

$$\lim_{\substack{N,D\to\infty \\ N/D=\rho}} Tr\left(\hat{G}^{(D)} - G\right) \overset{a.s.}{\to} 0 \tag{29}$$

or, by looking at the expected value :

$$\lim_{\substack{N,D\to\infty \\ N/D=\rho}} Tr\left(\mathbb{E}[\hat{G}^{(D)}] - G\right) \to 0 \tag{30}$$

Marchenko and Pastur explored the case where all of the entries of each vector $\boldsymbol{X^s}$ are independent and identically distributed gaussian variables . Nowadays, the Marchenko Pastur distribution has been recovered for much more relaxed conditions.
No formal derivation for the deterministic equivalent of $\hat{G}^{(D)}$ will be given in this manuscript, as it involves a big mathematical machinery in order to prove the convergence and some non-intuitive conditions on the moments of $\boldsymbol{X}$. However for a heuristic demonstration, one can check the review of Romain Couillet [11],

Section 2.2.2.
Another method, based on the diagrammatic expansion of the free probability theory has been done in the appendix of [12]. It is a non-rigorous method, allowing however for much more interpretation. This method tells us that, due to the fluctuations induced in our model by the noise, the parameter α, telling us how negligible is the regularizer in our model, gets renormalized to a value α^* and ends up being smaller, depending only on the parameter ρ. The computation of the value of α^* involves solving a self-consistent equation, given by :

$$(\alpha^*)^{-1} = (\alpha)^{-1} + \frac{1}{\rho}\int \nu(x)x\,(1+\alpha^* x)^{-1}\,dx \tag{31}$$

Where $\nu(x)$ is the spectral density of the population matrix C_0. If ρ is big enough, $\alpha^* \approx \alpha$. This means that in the case where your number of measures is much bigger than the parameter that one has to infer, the fluctuations have no real impact on the inferred response matrix $\hat{R}$. On the countrary, if ρ gets close to one, the importance of the renormalization gets important and one must be careful about the effects of fluctuations on our model. Our deterministic equivalent will be decomposed on the modes of the population matrix and will be of the form :

$$G = (\mathbb{1} + \alpha^* C_0)^{-1} \tag{32}$$

The interesting point here is that we found a formula of the inferred response function, based only on the population matrix, which is a quantity that can be computed for a large range of models. This means that we have a quantitative method to judge the precision of our inference, in the case where the underlying model is a linear one. Note that the regime that we are putting ourselves in is a theoretical one. Meaning, that we should have corrections in our formula of order $\mathcal{O}(1/N)$. This has to be studied more carefully.

4 Numerical simulations

4.1 Implementation of the model

In order to validate our theory, one needs a procedure in order to implement a numerical simulation of the computation of the response matrices and compare quantitatively the response function using the method of empirical correlation matrices and the one based only on the population matrix.
The first step is to sample an adjacency matrix A based on a random directed Erdős-Rényi graph with parameter(D,p), D being the number of nodes of our system while p is the average number of edges coming out of a vertex. Our graph has to be weighted and directed. Thus, for every edges created, we sample a weight uniformly between 0 and 1. Finally, we compute the eigenvalues of the non-symmetric matrix A and rescale the whole matrix so that its spectral radius is less than 1, ensuring statistical convergence. In our case, we decided to set it to 0.9.
We set a time T for which we want to study the response. Our objective will

be to compute T different matrices $\hat{R}(\tau)$, $\tau \in [\![1;T]\!]$ through the computation of the correlation matrices from one side, and from the theoretical formula from the other side.
To compute the empirical correlation matrices, we apply iteratively the autoregressive model, that is, for each t, we sample a noise term $\boldsymbol{\xi_t} \in \mathbb{R}^D$, $\boldsymbol{\xi_t} \sim \mathcal{N}(0, \mathbb{1})$ and compute

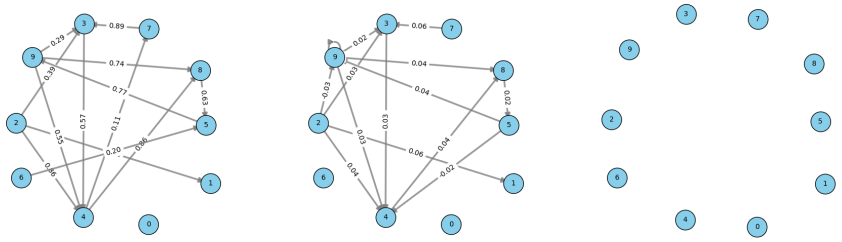
$$\boldsymbol{X_{t+1}} = A\boldsymbol{X_t} + \boldsymbol{\xi_t} \tag{33}$$

We iterate this process at each timestep until $t+T$ and add values of correlations to $\hat{C}_0, \hat{C}_\tau$. At the step $t+\tau+1$, one is tempted to compute new values of the correlation matrices directly. However, there are some correlation between the measures done at time t and at time $t+\tau+1$. Thus, we must be careful and have to wait during a characteristic time, that we fixed by looking at the biggest eigenvalues of the population correlation matrix, computed numerically using Eq (9). We finally divide the two computed matrices by the number of values we did the average on. Note that for the process, the initial condition is given by $\boldsymbol{X_0} \sim \mathcal{N}(0, \mathbb{1})$, that is pure noise. Also, the process first runs for a sufficiently long amount of time so that it forgets about initial condition and converges to an attractor of the system. We are then able to compute $\hat{R}_{ij}(t) = \alpha \hat{C}_t \hat{G}^{(D)}$.
The theoretical response presented in Eq () must be computed by finding a good candidate for $\hat{G}^{(D)}$, which is a random matrix. We are able to compute it theoretically using Eq (32). Note that the matrices we will use in our simulation will not be of infinite size, meaning that we should have corrections in our formula of order $\mathcal{O}(1/N)$. This will be studied more carefully in the following month. We now have a procedure in order to compare our two different responses.

4.2 Classical limit : small matrices and high number of measures

In the case where the number of measures is much bigger than the number of parameters to infer, that is $\rho \gg 1$, the response calculated using correlation matrices should be really close to the theoretical one. As a consequence, we would be able to effectively determine if a causal link is present or not, depending on the shape of the curve $\hat{R}(t)$.
Let us look at a system of size $D = 10$, with $N = 10^6$ measures and an average out-degree of 2. We sampled an adjacency matrix, according to the rules exposed in Subsection 4.1. We inferred the empirical adjacency matrix using the correlation matrices in the case of a big regularization ($\alpha = 10$) and a small one ($\alpha = 10^3$). The graphs are depicted in Fig **??**. One can see the importance of the parameter α on the inference that we are doing. Even with a big number of measures, when the regularization is strong enough, there is a residual error in the inferred graph. To confirm that this is not a finite size effect due to the noise inherent to our model, we now have to compare the response functions that we computed. This will be done by centering our analysis around vertex 5, abbreviated as 5 in the following. One can distinguish three different cases :

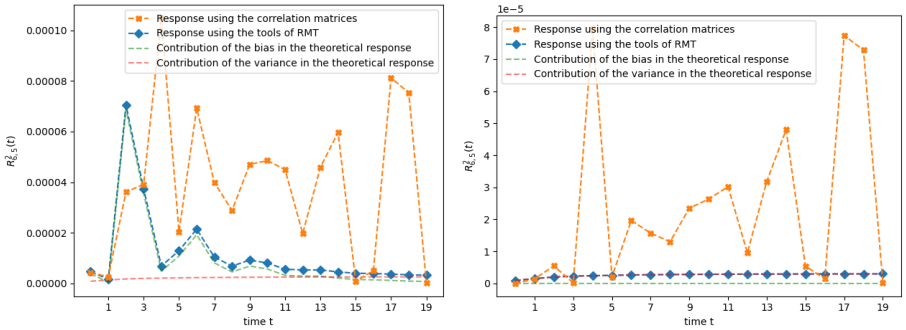
i) pure noise : Consider the response from 5 to 6. No causal link is present.



(a) True graph (b) $\alpha = 10$ (c) $\alpha = 1000$

Figure 2: True graph and the errors in our inference for multiple values of α. A threshold has been set to 0.02 so that the small residual links do not show up.

However, due to the finite α, the bias imposes a non-null response function. This is depicted in Fig 3.

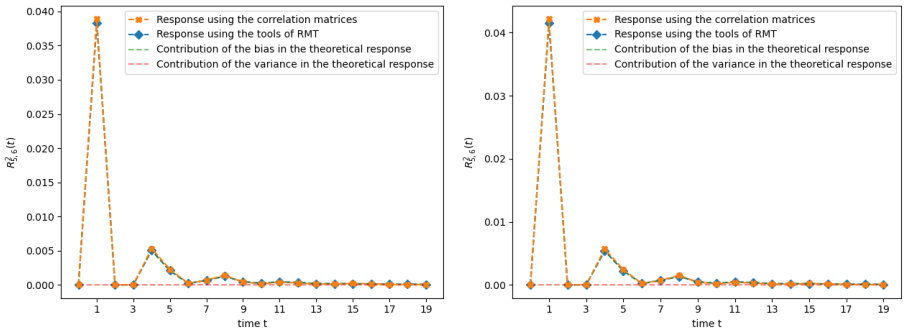


(a) $\alpha = 10$ (b) $\alpha = 1000$

Figure 3: Response from 5 to 6 for 2 different values of alpha.

ii) first order causation : As seen on the true graph of our system **??**, there is a unidirectional direct response from 6 to 5, meaning that the edge from 6 to 5 exists and that 6 is not caused by any other vertices. This makes it impossible for directed loops to go back to 6. This situation is depicted in Fig 4. In this specific case, the regularization doesn't seem to be that important. However, the system predicts a loop of order 4 which is not present in the true graph.

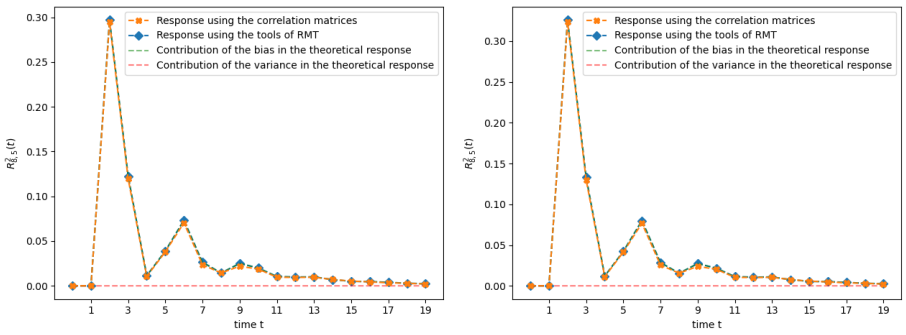
iii) higher order causation : Let us look at a more complex situation. 5 is directed only toward 9 and only 8 is directly causal to 5. However, we have different sized-loop that start and end on 5, namely the 3-loop 5-9-8-5, the 4-loop 5-9-4-8-5 and the 5-loop 5-9-3-4-8-5. The response functions from 5 to 8, illustrated in fig 5 show multiple spike. The spikes at times $t = 2, 3, 4$ are



(a) $\alpha = 10$ (b) $\alpha = 1000$

Figure 4: Response from 6 to 5 for 2 different values of alpha.

present due to the paths we characterized above. The spikes at later time are characteristic of the loops and are qualified as being harmonics of our system. For example, the spike at $t = 6$ corresponds to the sum of the 4-loop and the path 5-9-8 plus the 3-loop and the path 5-9-4-8.

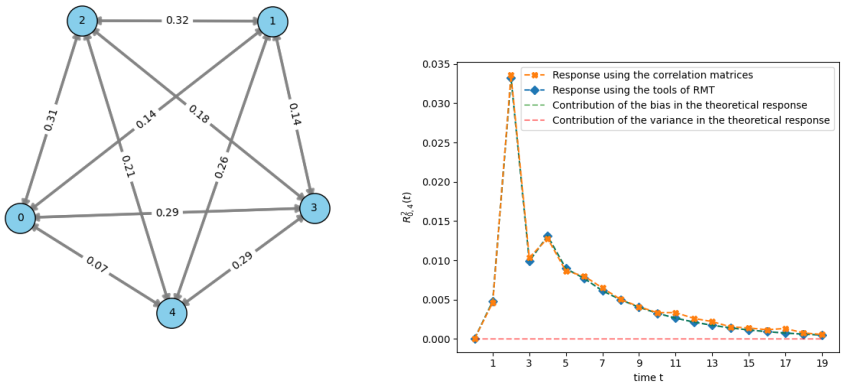


(a) $\alpha = 10$ (b) $\alpha = 1000$

Figure 5: Response from 5 to 8 for 2 different values of alpha.

The apparition of harmonics becomes more and more common as we increase the average out-degree of the Erdos-Renyi directed graph. This is because we also increase the probability of creating loops in our graph. For example, consider a directed graph with $D = 100$ nodes and an average out-degree of 5. To analyze the likelihood of forming a 3-cycle, we consider the number of nodes reachable within 3 steps, which is approximately $5^3 = 125$. Since this exceeds the total number of nodes, it implies a high probability of revisiting nodes and forming cycles during the exploration process. We expose in Fig 6b a limiting case. If one look at a fully connected graph with $D = 5$ and $N = 10^6$, the response should be maximum at $t = 1$ and decay slowly after. However, since

the graph is weighted, one can look at the link between 4 and 0, which is small (in this case 0.07), while the response is bigger at $t = 2$. This is because you have the contribution from 3 different paths, namely 4-1-0, 4-3-0 and 4-2-0.



(a) Sampled fully connected graph. (b) Response from 4 to 0, $\alpha = 100$.

Figure 6: Limiting case of a fully connected graph.

4.3 Large dimensional limit

In the LDL, we expect the identification of the causal links to be much more difficult than in the classical limit. However, we may think that, since we were able to separate the bias and the variance term in the theoretical formula, looking only at the bias term would allow us to get some information on the graph we are working on.
We sampled a graph with $D = 10^3$, $N = 10^3$ and an average out-degree of 2. In the following, we set $\alpha = 10^3$. To begin with, let us look at the response between two points which are not causal one to the other, namely the response from 0 to 417. This is exposed in Fig 7. One can see the importance of the variance, that doesn't have the same order of magnitude with respect to the empirical response. This is still being explored nowadays and we expect a mistake in the numerical simulation.

Because of this, we will now concentrate ourselves only on the theoretical bias and the empirical response in order to confirm if our method may be suitable to recover the causality between our different variables. We were able to find 1, 2 and 3-order causation by looking at the adjacency matrix. This will help us selecting the good edges to look at in the response function. In Fig **??** is exposed 4 different responses function, accounting only for the bias term and the empirical response. We can see that the theoretical bias computed by our formula indeed recover first and second order causal links in the case where

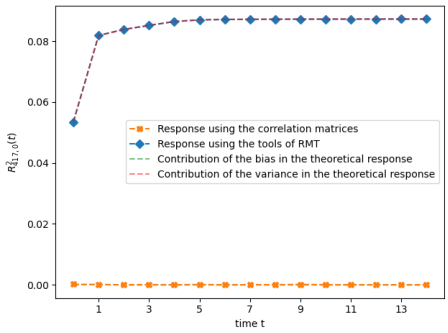
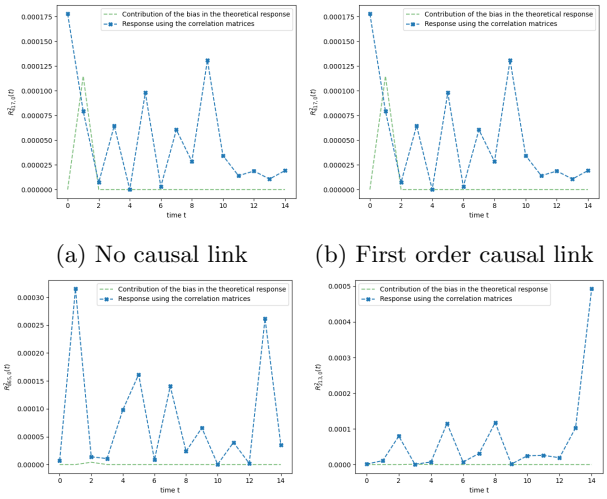


Figure 7: Plot of the usual response one get in the LDL.



(a) No causal link (b) First order causal link

(c) Second order causal link (d) Third order causal link

Figure 8: Empirical response function and bias term of the theoretical formula. One can see that the bias is non zero in subfigures (b) and (c), while for (d) the third order causality cannot be observed.

$\rho = 1$. This means that one would expect the possibility to theoretically recover the underlying dependence between the variables, even in the case where the number of measures is of the same order of magnitude than the number of variables in the LDL.

5 Conclusion

In this master's thesis, we explored the integration of causality into machine learning models with a particular focus on high-dimensional data. Beginning

with an in-depth analysis of linear regression, we linked it to linear response theory, deriving a closed-form solution for the empirical response function while accounting for bias and variance. Our investigation extended to the complexities of identifying causal relationships in high-dimensional settings, addressing the challenges posed by the curse of dimensionality.

Through numerical simulations, we saw that *observational causality* can be inferred in high-dimensional contexts. This means that, by focusing only on the bias term, valuable insights can be found for a general high-dimensional linear system.

Our research bridged the gap between statistical mechanics and machine learning, offering new perspectives on understanding and inferring causal relationships in complex systems. This work contributes to the field of machine learning by providing methods to uncover causality in high-dimensional scenarios where traditional approaches often fail, ultimately paving the way for more robust and interpretable AI systems.
The following works will focus on :

i) manage to numerically correct for the variance term.

ii) see if the causality can also be recovered in the case where we do not observe all of the variables but only a fixed fraction of them. We may look for an eventual theoretical formula connecting the number of measures, the dimension of our system and the minimal number of variables that one would have to observe in order to recover the causality.

iii) explore other types of dependency between variables, namely nonlinear systems. This may help us to recover the causal link in any situation, as soon as the system is of sufficiently large dimension, which is usually the framework that we are in in the era of *Big Data*.

6 What has been done after

This report had to be sent in the mid-june 2024. However, my internship ended on the 31st of july. Some more work has been done on this specific topic and a summary is needed in the case where this work has to be continued.

6.1 addition in the simulation

The variance of the random variable $\boldsymbol{\xi}$ can now be chosen. Note however that one has to be careful in the convergence of the process to a fixed point, since no criteria has been set to check whether the time chosen is long enough or not. Interestingly, one could think that setting $\alpha \gg 1$ is enough for the regularizer to be negligible. However, since the correlation matrix scales as σ^2 (see 7), the condition reads $\alpha\sigma^2 \gg 1$.

6.2 corrections in simulation

The variance term in the simulation was not the good one. This has been corrected by rechecking the implementation. We now have consistent results in the LDL.

[figures]

Note however that the implementation of the formulation is still not perfect. If one looks again to the formula of the response function 4.1, some terms are quadratic in the resolvent. In the implementation, I naively put each deterministic equivalent in the equation and took the trace.
This however cannot be done because one is facing the trace of a product of random matrices. One must be careful when doing the computation. To simplify a little bit the problem, let us look at the *total response function*, that is the response from every node to all the others. This part will be exhaustive. For all the derivations of the formulae used, see [12] More formally, the quantity that we will look at is :

$$\begin{aligned}\left\langle \hat{R}^2(t) \right\rangle &= \sum_{i,j} \left\langle \hat{R}^2_{j\to i}(t) \right\rangle \\ &= Tr\left[A(t)\left(\mathbb{1} - G^{(D)}\right)\left(\mathbb{1} - G^{(D)}\right)A^\top(t)\right] + \\ &+ \alpha\frac{\sigma^2}{N} Tr\left[\left(\mathbb{1} - G^{(D)}\right)G^{(D)}\right] \times Tr\left[\sum_{\tau=0}^{t-1} A(t-\tau-1)A^\top(t-\tau-1)\right]\end{aligned} \tag{34}$$

Moreover, by using the identity $\frac{\partial}{\partial\alpha}G^{(D)} = -\frac{1}{\alpha}\left(\mathbb{1} - G^{(D)}\right)G^{(D)}$, the equation reads :

$$\begin{aligned}\left\langle \hat{R}^2(t) \right\rangle = Tr\left[A(t)A^\top(t)\right] - Tr\left[A(t)G^{(D)}A^\top(t)\right] + \alpha Tr\left[A(t)\frac{\partial}{\partial\alpha}G^{(D)}A^\top(t)\right] + \\ - \alpha^2\frac{\sigma^2}{N} Tr\left[\frac{\partial}{\partial\alpha}G^{(D)}\right] \times Tr\left[\sum_{\tau=0}^{t-1} A(t-\tau-1)A^\top(t-\tau-1)\right]\end{aligned} \tag{35}$$

$G^{(D)}$ is a random matrix. Let us now take the thermodynamic limit. In this case, $G^{(D)} \to G[\Lambda]$. By applying the chain rule and using back the identity, we have :

$$\begin{aligned}\left\langle \hat{R}^2(t) \right\rangle \to Tr\left[A(t)A^\top(t)\right] - Tr\left[A(t)G[\Lambda]A^\top(t)\right] - \frac{\alpha}{\Lambda}\frac{\partial\Lambda}{\partial\alpha} Tr\left[A(t)G[\Lambda]\left(\mathbb{1} - G[\Lambda]\right)A^\top(t)\right] + \\ + \frac{\alpha^2}{\Lambda}\frac{\sigma^2}{N}\frac{\partial\Lambda}{\partial\alpha} Tr\left[G[\Lambda]\left(\mathbb{1} - G[\Lambda]\right)\right] \times Tr\left[\sum_{\tau=0}^{t-1} A(t-\tau-1)A^\top(t-\tau-1)\right]\end{aligned} \tag{36}$$

We see that the only difference that one faces when dealing with power of G is to introduce a prefactor $\frac{\partial\Lambda}{\partial\alpha}$.

This can now be computed. Indeed, it can be shown that :

$$\frac{\partial \Lambda}{\partial \alpha} = \frac{\Lambda^2}{\alpha^2} \frac{\rho}{\rho - Q[\Lambda]} \tag{37}$$

where :

$$g[\Lambda] = \int dx \frac{\nu_\infty(x)}{1 + \Lambda x} \tag{38}$$

$$Q[\Lambda] = 1 - 2g[\Lambda] + \int dx \frac{\nu_\infty(x)}{[1 + \Lambda x]^2} \tag{39}$$

$$\rho = N_{eff}/D \tag{40}$$

Appendix A Solution of the Fokker-Planck equation

In the following, we will follow a derivation from [13]. The intuition behind this computation is that the time independent probability distribution related to a Fokker-Planck equation is known to be a gaussian. One may so expect that the time-dependent distribution also has a gaussian form. One thus need only the first two moments in order to fully determine the law.
We first write the Fokker-Planck equation associated to the process :

$$\frac{\partial}{\partial t}p(\boldsymbol{X},t) = -\nabla_{\boldsymbol{X}}\left[(A-\mathbb{1})\,\boldsymbol{X}p(\boldsymbol{X},t)\right] + \frac{1}{2}\nabla_{\boldsymbol{X}}\nabla_{\boldsymbol{X}}p(\boldsymbol{X},t) \tag{41}$$

In the following and for clarity, we will set $A-\mathbb{1} \equiv A$. By multyplying by $\boldsymbol{X}$ and integrating over $\boldsymbol{X}$ Eq (41), one gets, after an integration by part and assuming that the probability distribution decays sufficiently fast on the extremes of the domain of definition of $\boldsymbol{X}$:

$$\frac{\partial}{\partial t}\langle \boldsymbol{X}\rangle = A\langle \boldsymbol{X}\rangle \tag{42}$$

Which is solved by:

$$\langle \boldsymbol{X}(t)\rangle = e^{At}\langle \boldsymbol{X}(0)\rangle \tag{43}$$

In the same way, By multyplying by $\boldsymbol{X}\boldsymbol{X}^\top$ and integrating over $\boldsymbol{X}$ Eq 41, we have :

$$\frac{\partial}{\partial t}\left\langle \boldsymbol{X}\boldsymbol{X}^\top\right\rangle = A\left\langle \boldsymbol{X}\boldsymbol{X}^\top\right\rangle + \left\langle \boldsymbol{X}\boldsymbol{X}^\top\right\rangle A^\top + \mathbb{1} \tag{44}$$

Next, using the vocabulary of quantum mechanics, one goes to the interaction picture. This is usually done to find the time dependence of the creation and annihilation operators. We so define :

$$\left\langle \boldsymbol{X}\boldsymbol{X}^\top\right\rangle \equiv e^{Rt}\left\langle \boldsymbol{X}\boldsymbol{X}^\top\right\rangle^* e^{R^\top t} \tag{45}$$

By taking the time derivative of Eq (45), one finds the constraint :

$$\frac{\partial}{\partial t}\left\langle \boldsymbol{X}\boldsymbol{X}^\top\right\rangle^* = e^{-t(R+R^\top)} \tag{46}$$

By solving and plugging back into 45, one finds :

$$\left\langle \boldsymbol{X}\boldsymbol{X}^\top(t)\right\rangle = e^{Rt}\left\langle \boldsymbol{X}\boldsymbol{X}^\top(0)\right\rangle e^{R^\top t} + e^{Rt}(R+R^\top)^{-1}\left(1-e^{-t(R+R^\top)}\right)e^{R^\top t} \tag{47}$$

We can now define the time-dependant covariance matrix $\Sigma(t) \equiv \left\langle \boldsymbol{X}\boldsymbol{X}^\top(t)\right\rangle - \langle \boldsymbol{X}(t)\rangle\langle \boldsymbol{X}(t)\rangle^\top$. One can check that the probability distribution :

$$P(\boldsymbol{X},t) = (2\pi)^{-d/2}(det\Sigma)^{-1/2}\exp\left[-\frac{1}{2}\left(\boldsymbol{X}-\langle \boldsymbol{X}(t)\rangle\right)^\top\Sigma^{-1}\left(\boldsymbol{X}-\langle \boldsymbol{X}(t)\rangle\right)\right] \tag{48}$$

indeed solves the Fokker-Planck equation of our system.

Appendix B Loss function

We set a loss function for $\hat{R}$ that is composed of the L2 regularizer and of the mean squared error :

$$\mathfrak{L}[\hat{R}(t)] = Tr\frac{[\hat{R}(t)^\top \hat{R}(t)]}{2} + \frac{\alpha}{2N}\sum_s (\boldsymbol{X_t^s} - \hat{R}(t)\boldsymbol{X_0^s})^\top (\boldsymbol{X_t^s} - \hat{R}(t)\boldsymbol{X_0^s}) \tag{49}$$

$$= \sum_i \frac{\hat{R}_{ii}^2(t)}{2} + \frac{\alpha}{2N}\sum_s \sum_i \left(X_{t,i}^s - \sum_j \hat{R}_{ij}(t)X_{0,j}^s\right)\left(X_{t,i}^s - \sum_j \hat{R}_{ij}(t)X_{0,j}^s\right) \tag{50}$$

We want to infer on the matrix $\hat{R}$ that is the closest to explain the model. Thus , we look at the derivative :

$$\frac{\partial \mathfrak{L}}{\partial \hat{R}_{ij}(t)} = \hat{R}_{ij}(t)\delta_{ij} + \frac{\alpha}{N}\sum_s -X_{0,j}^s\left(X_{t,i}^s - \sum_j \hat{R}_{ij}(t)X_{0,j}^s\right) \tag{51}$$

$$= \hat{R}_{ij}(t)\delta_{ij} - \frac{\alpha}{N}\sum_s X_{0,j}^s X_{t,i}^s + \frac{\alpha}{N}\sum_s X_{0,j}^s\left(\hat{R}(t)\boldsymbol{X_0^s}\right)_i \tag{52}$$

This equation can be put in matrix form. Moreover, we set the derivative to be equal to 0 :

$$0 \stackrel{!}{=} \hat{R}(t)\mathbb{1} - \frac{\alpha}{N}\sum_s \boldsymbol{X_t^s}\boldsymbol{X_0^s}^\top + \frac{\alpha}{N}\hat{R}(t)\sum_s \boldsymbol{X_0^s}\boldsymbol{X_0^s}^\top \tag{53}$$

The final formula for the inferred response function is recovered :

$$\hat{R}(t) = \alpha \hat{C}_t \hat{G}^{(D)} \tag{54}$$

References

[1] S. G. H. Philander. El Niño Southern Oscillation phenomena. , 302(5906):295–301, March 1983.

[2] Alain Aspect, Philippe Grangier, and Gérard Roger. Experimental realization of einstein-podolsky-rosen-bohm gedankenexperiment: A new violation of bell's inequalities. *Phys. Rev. Lett.*, 49:91–94, Jul 1982.

[3] Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

[4] Lars Onsager. Reciprocal relations in irreversible processes. i. *Phys. Rev.*, 37:405–426, Feb 1931.

[5] Davide Rossetti. Machine learning on causality. Master's thesis, Politecnico di Torino, 2023. Available at `http://webthesis.biblio.polito.it/id/eprint/29026`.

[6] P. W. Anderson. More is different. *Science*, 177(4047):393–396, 1972.

[7] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[8] Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

[9] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998.

[10] V A Marčenko and L A Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, apr 1967.

[11] Romain Couillet and Zhenyu Liao. *Random Matrix Methods for Machine Learning*. Cambridge University Press, 2022.

[12] Cyril Furtlehner. Free dynamics of feature learning processes. *Journal of Statistical Physics*, 190(3), January 2023.

[13] N.G. van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier Science Publishers, Amsterdam, 1992.