POLITECNICO DI TORINO
SORBONNE UNIVERSITÉ

TOWARDS THE ESTIMATION OF CAUSAL NETWORKS BY THE INFORMATION IMBALANCE APPROACH

AUTHOR

ALLIONE MATTEO
Student No. S309890
MSc PHYSICS OF COMPLEX SYSTEMS
Mathematical modelling for engineering (LM-44)

SUPERVISORS

ALESSANDRO LAIO, SISSA
ANDREA PAGNANI, POLITO

TRIESTE, JULY 2024
ACADEMIC YEAR 2023/2024

This page intentionally left blank.

Acknowledgements

I deeply thank Vittorio Del Tatto for his assistance in the development of this thesis. My heartfelt thanks also go to my friends in Trieste and in the PCS master's program for their unwavering support and for making every step of this journey amusing, pleasant, and enriching.

Finally, I thank my family, who have undoubtedly been my first and foremost inspiration.

This page intentionally left blank.

Abstract

Inferring the presence of causal links from observational data is a challenging task which typically goes under the name of causal discovery. After reviewing the basic concepts of this field and a standard algorithm used to infer the topology of causal graphs, we develop an approach specifically designed for times series data based on the Information Imbalance measure.

This estimator, first introduced in Glielmo et al., 2022 for ranking the information content of different distance spaces, has been applied in Del Tatto et al., 2024 to detect causal relationships between time-dependent variables. Recently, it has been reformulated in a differentiable version (Wild et al., 2024) which allows the automatic learning of the most informative distance function in a gradient-descent fashion. In the first part of this thesis, we further extend this measure, introducing a procedure to estimate its statistical error. We then use this measure in a framework for causal network reconstruction from time series data.

Specifically, we define a protocol to progressively find subsets of independent variables in complex non-linear dynamical systems. This allows generating a coarse grained graph showing the hierarchy of the interactions between different groups of features of the dataset. In contrast with standard causal discovery methods, the algorithm proposed here does not require any combinatorial search of conditioning sets, can be applied to high-dimensional systems and intrinsically retrieves multi-body interactions.

Keywords: causal discovery, dynamical systems, model free analysis

This page intentionally left blank.

Contents

This page intentionally left blank.

1

Introduction

Identifying the interactions between different parts of a complex system can improve deeply our understanding of it, suggesting, for example, how different external interventions would propagate their effect in the whole system. Furthermore, pointing out the presence of direct links between different subsystems allows building parsimonious models, able to describe physical interconnections with a reduced number of parameters. This is deemed useful in countless applications (from responsibility attribution to policy making, to mention a few).

Nowadays, the most standard strategies to reach this objective relate to the studies done, among many, by Pearl, 1995. Starting from purely observational data, such methods try to build a graphical model representing the statistical dependencies between the variables. The graph model can then be used per se or to impose physical constraints for more advanced techniques (see for example Brunton et al., 2016), effectively reducing the space of possible models to explore. Most of the available techniques for deriving these graphs models lose statistical power, or become too computationally expensive, if applied to high-dimensional systems, where the graph should include $O(100)$ nodes or more.

In this work we investigate the possibility of using a method recently introduced by Glielmo et al., 2022 to build in a causal graph even when dealing with high dimensional systems. Our primary focus is on systems composed of smaller subsystems. We aim to identify these subsystems and the hierarchy of interactions among them, resulting in what could be properly described as a directed hyper-graph.

We organise the thesis as follows: in Chapter 1 we introduce the state of the art of causal graph reconstruction with information theoretical tools, then we summarise the ideas underlying the Information Imbalance. In Chapter 2 we introduce a new estimator of the Information Imbalance which allows for an explicit calculation of standard errors. Finally, in Chapter 3 we describe a protocol, developed by us, which allows finding subsets of progressively independent variables and we apply this approach to some benchmark examples.

1.1 Information theory for causal discovery

A *causal model* with variables $\{X^i\}$ $(i = 1, ..., D)$ can be defined through the *structural causal equations*

$$X^i := f^i(\mathcal{P}(X^i), \boldsymbol{\theta}^i)\,, \tag{1.1}$$

where f^i is a function of a subset $\mathcal{P}(X^i)$ of all the variables, called *set of parents of i*, which includes all its direct causes. The parameters $\boldsymbol{\theta}^i$ may tune the strength of the interactions between $\mathcal{P}(X^i)$ and X^i, as well as the contribution of external (or *exogenous*) variables, typically modelled as noise terms. The form of Equation 1.1 naturally implies a factorization of the joint distribution of the $\{X^i\}$ into the conditional probabilities of each variable given its parents,

$$p(X^1, ..., X^D) = \prod_{i=1}^{D} p\left(X^i \mid \mathcal{P}(X^i)\right)\,, \tag{1.2}$$

which is known as *Markov property*. An intuitive definition of the causal graph representing Equation 1.1 would assign a node to each variable and a directed link connecting each element in $\mathcal{P}(X^i)$ to X^i for all i[1].

Different assumptions allows to retrieve the causal graph from observations (see for example Assaad et al., 2022 or Spirtes et al., 1993 for an introduction to different approaches). In this part we will focus mostly on the framework of conditional independence testing, following what discussed in Runge, 2018.

The rationale behind this approach is to perform independence tests to find a skeleton of the causal graph and eventually use some logical rules to orient the arrows. This implies that, for each pair of nodes X and Y the null hypothesis $X \perp\!\!\!\perp Y|\mathcal{S}$ (X and Y are *conditionally independent* given $\mathcal{S}$) is tested against the alternative hypothesis $X \not\perp\!\!\!\perp Y|\mathcal{S}$ (X and Y are *conditionally dependent* given $\mathcal{S}$), with a combinatorial search over all the possible conditioning sets $\mathcal{S}$. In the following part we will use as a measure of conditional dependence the mutual information, but other measures such as the partial correlation may be employed (Runge, 2018). If a set of variables X is found to be independent from a set of variables Y when conditioned on a set of variables $\mathcal{S}$, i.e. if

$$\begin{cases} I(X, Y|\mathcal{S}) = \iiint d\mathcal{S}dxdy\; p(x, y, \mathcal{S}) \log \frac{p(x,y|\mathcal{S})}{p(x|\mathcal{S})p(y|\mathcal{S})} = 0 \\ I(X, Y|\emptyset) = \iint dxdy\; p(x, y) \log \frac{p(x,y)}{p(x)p(y)} \neq 0 \end{cases}\,, \tag{1.3}$$

then three possible causal motifs can be identified:

- X causes $\mathcal{S}$ which causes Y: $X \to \mathcal{S} \to Y$;
- Y causes $\mathcal{S}$ which causes X: $Y \to \mathcal{S} \to X$;
- $\mathcal{S}$ causes both X and Y: $X \leftarrow \mathcal{S} \to Y$.

[1] Notice that the conditional probabilities can be factorised as in Equation 1.2 provided that the underlying causal graph is acyclic.

These three causal motifs form a so called *Markov equivalence class*, i.e. a set of causal structures that correspond to the same conditional independencies, and are therefore observationally equivalent.

Example 1.1.1. The idea can be illustrated in a simple system with only three nodes X, Y and Z. Considering the presence of two causal links, the possible patterns that might appear are $X \to Z \to Y, Y \to Z \to X, X \leftarrow Z \to Y, X \to Z \leftarrow Y$, which are shown in Fig. 1.1.

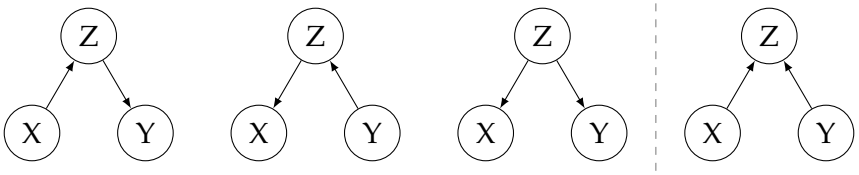

Figure 1.1: *Example of simple causal motifs in a system with three variables*

The first two diagrams show the case in which two variables are *indirect causes* one of the other. Take the first diagram, where the variable X causes Z which then causes Y. In this case you'll find that the two variables X and Y are not *marginally* independent (i.e. $I(X, Y|\emptyset) \neq 0$). When conditioning on Z, instead, the information flow from one variable to the other gets blocked and this results in the conditional independence $I(X, Y|Z) = 0$.

In the third case Z is a common cause of X and Y. Z is said to be a *confounder*. We can easily see that when conditioning on Z the probabilities factorise ($p(x, y|z) = p(x|z)p(y|z)$) leading to $I(X, Y|Z) = 0$. Instead, $p(x, y) = \sum_z p(x, y, z) = \sum_z p(x|z)p(y|z)p(z)$ in general does not factorise, resulting again in $I(X, Y|\emptyset) \neq 0$. Using the independence tests described here, is then impossible to recognise these three causal motifs among themselves.

The last pattern, instead, clearly falls in a different scenario. Indeed here one gets that $p(x, y) = p(x)p(y)$, but $p(x, y|z) = \frac{p(z|x,y)}{p(z)}p(x)p(y)$ using the Bayes Theorem. So $I(X, Y|\emptyset) = 0, I(X, Y|Z) \neq 0$. In this case Z is said to be a *collider* and conditioning on it actually opens a path for the information flow from X to Y and viceversa.

The framework described above allows to retrieve the correct equivalence class for the causal graphs if we work under the hypothesis that there are no unobserved common causes (*causal sufficiency*). Furthermore, we need the joint distribution of the observed variables to truly reflects the conditional independencies embedded in the graph structure (*causal Markov condition*) and viceversa (*faithfulness*).

This suggests an algorithmic implementation in which we test, for all couples (X, Y), whether a subset $\mathcal{S}$ which makes them independent exists. If it does exist, then X and Y are not directly linked. This is one of the main ingredients of the *PC algorithm* (Spirtes et al., 1991)(and of many other methods, see for example Verma et al., 2022, Assaad et al., 2022). Notice that, after this procedure, only the arrows of the colliders can be directed exactly. Without further considerations the outcome is then a *partially* directed graph representing a Markov equivalence class.

Example 1.1.2. We now present a slightly more complex system with six variables, which we represent in Fig. 1.2. We can start drawing a fully connected graph representing all possible interactions and then progressively prune it. We focus here on the edge between X and Y, and we will assume to perform exact independence tests checking for Equation 1.3. Then we can see for different subsets $\mathcal{S}$ that[2]

$$
\begin{array}{ccccc}
 & & X \not\perp\!\!\!\perp Y \mid \emptyset & & \\
 & X \not\perp\!\!\!\perp Y \mid A & & X \not\perp\!\!\!\perp Y \mid B & \\
 & X \not\perp\!\!\!\perp Y \mid C & & X \not\perp\!\!\!\perp Y \mid D & \\
X \not\perp\!\!\!\perp Y \mid \{A,B\} & & X \not\perp\!\!\!\perp Y \mid \{A,C\} & & X \not\perp\!\!\!\perp Y \mid \{A,D\} \\
X \not\perp\!\!\!\perp Y \mid \{B,C\} & & X \perp\!\!\!\perp Y \mid \{B,D\} & &
\end{array}
$$

and we stopped here because we can finally safely remove the link $X - Y$. In this example we we stopped after checking $\mathcal{S}$ up to size two, clearly this might not always happen to be the case.

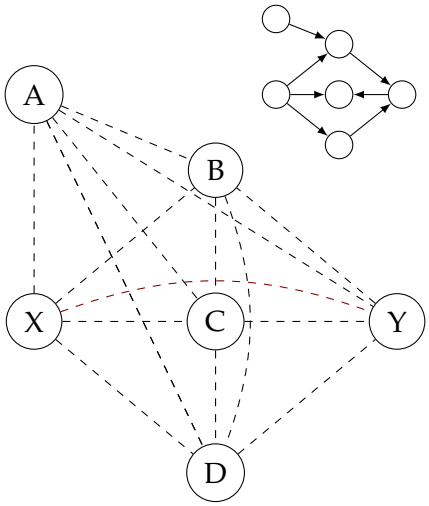

Figure 1.2: *The fully connected graph used as initialisation for the PC algorithm and the true underlying causal graph on the top right. In dark red we highlight the edge $X - Y$ of which we discuss in the example.*

This approach can be easily used also for time dependent data. In this case it is enough to consider each time step as an independent node and verify the relationship between variables at different time lags. Moreover, in this situation, causal ordering allows to direct the links following the arrow of time, so that if also instantaneous links are excluded the true causal graph is fully identifiable.

[2] This can be inferred looking at the path connecting X and Y in the true graph in Fig. 1.2. In particular, only conditioning on both D and B allows to *d-separate* X and Y. A description of d-separation, which allows to connect graphical models to conditional independencies, is presented in Appendix A. Notice that the vertex C is a *collider*, hence conditioning on it opens a path connecting the two variables of our interest. This was inserted on purpose to show that, in general, it is not enough to perform a test conditioning on all the variables (i.e. checking $I(X, Y|A, B, C, D)$), which at a first look might have seemed the fastest option to remove the link.

Remark (Conditional independence tests)**.** *In the examples we assumed we had the chance of performing exact estimations of conditional independence; unfortunately, the tests currently available strongly suffer from the curse of dimensionality. This makes them extremely hard to use when a large number of variables are present.*

Remark (On the necessity of causal sufficiency)**.** *Even though the independence tests allow in principle to build soundly the causal graph from data, when dealing with real dataset some problems might arise. In particular, if the assumption of causal sufficiency does not hold (namely, if a common cause of two variables is not observed), then the links drawn by the PC algorithm could include spurious links.*

The causal sufficiency assumption can also be violated in presence of observational noise coming from the data collection process, i.e. when the observed variables are a noisy version of the truly interacting ones. This case is depicted in Fig. 1.3, where the addition of noise is equivalent to an additional link from the true interacting variables (round black nodes), which are unobserved, and their noisy versions (square red nodes), which are observed. In this scenario no measurable mutual information would ever be found equal to zero, independently of the conditioning set. However, since the conditional independence tests are usually implemented with a threshold tolerance, the effects of this violation could be practically irrelevant in presence of small observational noise.

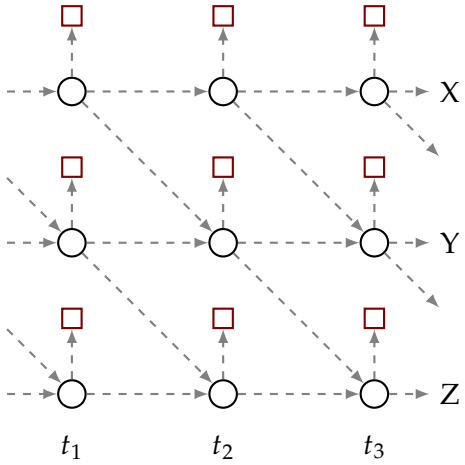

Figure 1.3: *When dealing with observational noise, you only get access to the square nodes.*

Remark (On multi-body interactions)**.** *An algorithm such as the one presented above might not be able to retrieve all the causal relationships. Take for example $Y = X_1 \oplus X_2 + \eta^Y$, where X_i are binary random processes with equal probabilities for each of their values and $\oplus$ is the XOR-gate (Runge, 2018). Then, even though $I((X_1, X_2), Y) > 0$, one can easily verify that $I(X_1, Y) = 0 = I(X_2, Y)$, so no link would be inserted. Indeed, the PC algorithm would conclude that no causal arrow between X_1 (X_2) and Y exists, as such variables are not even marginally dependent. The main problem here is that the concept of pairwise-dependencies is not sufficient to describe the causal structure of the system. Hyper-graphs becomes necessary, but this comes with a combinatorial explosion of the number of tests to be performed. In Section B.1 we show the results obtained with our method in a case similar to the one presented here.*

1.1.1 Transfer entropy and Granger causality

A simple implementation of the principles presented above needs an exponential number of independence tests. Indeed, given a graph of N nodes, for each true link 2^{N-2} subsets $\mathcal{S}$ are checked as conditioning sets. In general a maximum size of $\mathcal{S}$ can be fixed in advance and smart strategies are devised to reduce the number of suitable sets to test for (e.g. building $\mathcal{S}$ out of the nodes adjacent to those for which we are checking the independence) (Runge et al., 2019; Spirtes et al., 1993).

Despite these stratagems, the computational demands could be excessive for moderate to large networks. When dealing with time series, one way to address this problem is to use directly an asymmetric measure of causality, relying on the ideas by Granger, 1980. Following Assaad et al., 2022, we can provide the following definition:

Definition (Granger causality)**.** A time series X^p Granger-causes X^q if past values of X^p provide unique, statistically significant information about future values of X^q.

In its simplest implementation this criterion results in the comparison of prediction errors of different linear models. In particular one considers a maximal time lag τ_{max} and the following autoregressive model for X^q:

$$X_t^q = a_{q,0} + \sum_{i=1}^{\tau_{max}} a_{q,i} X_{t-i}^q + \eta_t^q \,.$$

Its augmented version, which also includes the past of X^p, can be written analogously:

$$X_t^q = a_{q,0} + \sum_{i=1}^{\tau_{max}} a_{q,i} X_{t-i}^q + \sum_{i=1}^{\tau_{max}} a_{p,i} X_{t-i}^p + \eta_t^q \,.$$

Using a statistical test such as the F-test (Cannelli, 2004) it is then possible to determine if the variance of the residuals is significantly decreased in the augmented model or not. If it is, then we can reject the hypothesis that X^p is not Granger-causing X^q (Assaad et al., 2022).

A stronger formulation of the criterion would not assume any underlying model. For example, it is possible to check on the so called *transfer entropy* (Schreiber, 2000),

$$T(X_t^p \to X_{t+1}^q) = h(X_{t+1}^q | X_t^q) - h(X_{t+1}^q | X_t^q, X_t^p) \,,$$

which directly quantifies the difference between the entropy $h(\cdot)$ of the two conditional distributions $p(X_{t+1}^q | X_t^q)$ and $p(X_{t+1}^q | X_t^q, X_t^p)$. Still both these methods do not fully capture the idea underlying Granger causality because they specifically focus only on pairwise (bivariate) relations. Multivariate versions can be devised (Runge, 2018; Assaad et al., 2022) but usually they are extremely expensive computationally and lead to a smaller effect size with consequent lower detection power (Runge et al., 2019).

The method that we will introduce in this work will broadly take inspiration from this idea of estimating causality through the presence of an information flow between variables.

1.2 Information Imbalance

This section reviews a statistical measure named Information Imbalance, as introduced in Glielmo et al., 2022.

The Information Imbalance was initially developed to evaluate informativeness of different metrics for a given dataset. Suppose to measure different properties of a system leading to two different distance measures d_A and d_B among the data points. If some information on d_A could be easily retrieved from d_B, but not viceversa, then one would be prone to claim that d_B is more informative than d_A. This intuition can be quantified by studying how distance ranks among data points change when using the two different metrics. An illustrative example is presented in Fig. 1.4. In practice, one checks how the nearest neighbours according to the metric remain close if the second metric is used to measure their distance. Ranks are considered instead of pure distances, in order to allow comparing metrics with different units of measure. The Information Imbalance is defined as

$$\Delta(d_A \to d_B) = \frac{2}{N^2} \sum_{i,j}^{N} \delta_{r_{ij}^A,1} r_{ij}^B \tag{1.4}$$

where N is the number of points in the dataset, δ is the Kronecker delta and r_{ij} is the rank obtained after sorting in ascending order the pairwise distances between i and rest of the points. The superscript in the ranks refers to the distance used. For example, $r_{ij}^A = 2$ if j is the second nearest neighbor of i in distance space d_A. The normalization factor is chosen in order to get $\Delta(d_A \to d_B) = 1$ in the least informative case (which corresponds to the case in which the ranks computed with d_A are completely independent from those computed with d_B). The case in which nearest neighbours are kept exactly the same using the different metrics gives $\Delta(d_A \to d_B) = \frac{2}{N}$. It must be noticed that by definition the Information Imbalance is asymmetric ($\Delta(d_A \to d_B) \neq \Delta(d_B \to d_A)$). In general, the lower the Information Imbalances are in both directions, the more the distances can be considered equivalent one to the other.

Even though only a similarity measure between data points is needed to compute $\Delta(d_A \to d_B)$ we will focus in the rest of the work on distances d_A and d_B built on the metric spaces A and B, for which the variables (or features) entering the distances are explicitly known.

If the number of data is large, one can estimate the Information Imbalance using the first k nearest neighbours, instead of the first only:

$$\Delta(d_A \to d_B) = \frac{2}{k_{max} N^2} \sum_{k=1}^{k_{max}} \sum_{i,j=1}^{N} \delta_{r_{ij}^A,k} r_{ij}^B \,. \tag{1.5}$$

This estimator is affected by a smaller statistical error than the one defined in Equation 1.4.

The method has already been used in a framework of causal discovery (Del Tatto et al., 2024) to test whether a dynamic variable (or a group of dynamic variables) X

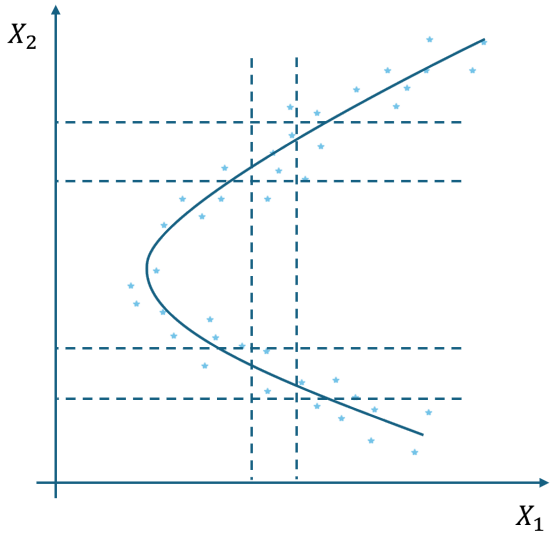

Figure 1.4: *An example in which the feature X_1 is less informative on X_2 than viceversa. Indeed, points which are close considering the Euclidean distance built using X_1 (d_{X_1}), are not necessarily close according to d_{X_2}, while the viceversa is true.*

causes another variable (or another group of variables) Y. The underlying intuition is that if X indeed causes Y, then predicting the future state of Y using a distance measure that incorporates the present state of X and Y together will be more accurate than using the state of Y alone. This implies that for some $\tau > 0$, representing the time lag of information transfer, we have:

$$\Delta\left(d_{(w\cdot X(0),Y(0))} \rightarrow d_{Y(\tau)}\right) > \Delta\left(d_{Y(0)} \rightarrow d_{Y(\tau)}\right) \; \forall w \neq 0$$

Here, w scales the units of X, accounting for the strength of the coupling.

Causality can then be assessed using a variational scheme, which involves testing many values of w to find $\min_w \Delta\left(d_{(w\cdot X(0),Y(0))} \rightarrow d_{Y(\tau)}\right)$ and verifying that the minimum is not obtained for $w = 0$.

The connection between this measure and causal discovery will be further elaborated upon in the following chapters. In particular, we will extend the ideas in Del Tatto et al., 2024 to address two main limitations: the use of only one weight to build a distance measure and the use of a grid search procedure to approximate the minimum. Introducing multiple weights will create a more expressive metric, enhancing the quality of predictions and enabling the selection of subsets of causes. This will also necessitate a new tool to explore the space of the possible weights, which is presented in the next section.

1.2.1 A Differentiable Information Imbalance

As stated above, a complete exploration of the parameter space would be computationally impossible. A different approach, which at least allows to find local minima, is to devise a gradient descent protocol.

Following the ideas in Wild et al., 2024, we define a differentiable version of the Information Imbalance as

$$\Delta^{\lambda}(d_{\boldsymbol{w}\odot\boldsymbol{A}} \to d_B) = \lim_{\lambda\to 0} \frac{2}{N^2} \sum_{i,j} c_{ij}(\lambda, \boldsymbol{w}) r^B_{ij} \,, \tag{1.6}$$

where the coefficients c_{ij} represent a smooth and differentiable version of the constraints $\delta_{r^A_{ij},1}$:

$$c_{ij}(\lambda, \boldsymbol{w}) = \frac{e^{-d_{\boldsymbol{w}\odot\boldsymbol{A}}(X_i,X_j)/\lambda}}{\sum_{m\neq i} e^{-d_{\boldsymbol{w}\odot\boldsymbol{A}}(X_i,X_m)/\lambda}} = \frac{e^{-d_{\boldsymbol{w}\odot\boldsymbol{A}}(X_i,X_j)/\lambda}}{Z_i} \,. \tag{1.7}$$

In particular, $c_{ij} \xrightarrow{\lambda\to 0} \delta_{r^A_{ij},1}$, so that the usual expression is recovered in the limit of small λ. In this way, in a short notation:

$$\frac{\partial}{\partial w_\alpha}\Delta^{\lambda} = \frac{2}{N^2} \sum_{i,j} r^B_{ij} \left[\frac{-\frac{Z_i}{\lambda} e^{-d_{\boldsymbol{w}\odot\boldsymbol{A}}(X_i,X_j)/\lambda} \partial_{w_\alpha} d_{\boldsymbol{w}\odot\boldsymbol{A}}(X_i, X_j) - e^{-d_{\boldsymbol{w}\odot\boldsymbol{A}}(X_i,X_j)} \partial_{w_\alpha} Z_i}{Z_i^2} \right] ,$$

where

$$\partial_{w_\alpha} Z_i = \sum_{m\neq i} -\frac{1}{\lambda} e^{-d_{\boldsymbol{w}\odot\boldsymbol{A}}(X_i,X_j)/\lambda} \partial_{w_\alpha} d_{\boldsymbol{w}\odot\boldsymbol{A}}(X_i, X_j)$$

and for an Euclidean distance squared:

$$\partial_{w_\alpha} d_{\boldsymbol{w}\odot\boldsymbol{A}}(X_i, X_j) = \partial_{w_\alpha} \sum_{\beta} w_\beta^2 (X_i^\beta - X_j^\beta)^2 = 2 w_\alpha (X_i^\alpha - X_j^\alpha)^2 \,.$$

It is interesting to observe the following properties, valid when using Euclidean squared distances:

$$\begin{gathered} \Pi_\alpha : \boldsymbol{w} \to (\cdots, -w_\alpha, \cdots) \,, \\ \Delta^{\lambda}(d_{(\Pi_\alpha \boldsymbol{w})\odot\boldsymbol{A}} \to d_B) = \Delta^{\lambda}(d_{\boldsymbol{w}\odot\boldsymbol{A}} \to d_B) \,, \\ \lim_{w_\alpha\to 0} \partial_{w_\alpha}\Delta^{\lambda} = 0 \,, \end{gathered} \tag{1.8}$$

meaning, in particular, that the information imbalance approaches the axes flatly.

In practice, the coefficients c_{ij} replace the selection of close points in the first space with a Gaussian weighting, with fixed variance. Since it might seem more reasonable to have different local weighting according to the local density, in order to effectively select the same number of neighbors around each point, an adaptive variance λ_i can be considered for each data point. Since in a point-adaptive scheme λ_i depends on the distances from point i in space A, which are updated during the gradient descent optimization, a dependence on the weights of the $\lambda_i(\boldsymbol{w})$ is introduced. This leads to a correction for the gradient computed above involving $\partial_{w_\alpha}\lambda_i(\boldsymbol{w})$.

A possible choice for the adaptive λ_i is suggested by the similarity of Equation 1.7 to that of the probabilities in a t-SNE procedure (see Mehta et al., 2019). Then, in a similar fashion to what is commonly done in that case, we can define a local entropy

$$H_i = -\sum_j c_{ji} \log c_{ji} \,,$$

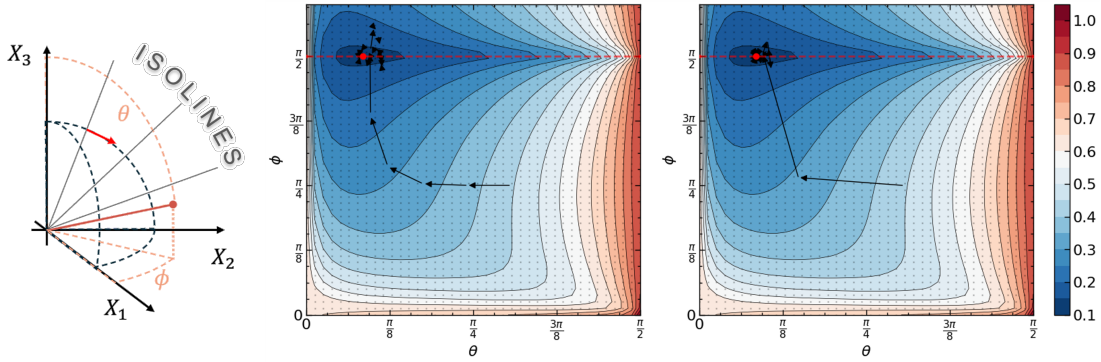

Figure 1.5: *The results of the gradient descent procedure for the coupled logistic maps as in Equation 2.4, using polar coordinates for the parametrization of the 2-sphere the dynamics effectively evolves on. The graph in the center is obtained with standard GD, the one on the right using Adam optimizer keeping fixed S' while sampling 5 batches of S at each iteration. The axis $\phi = \pi/2$ is highlighted: thanks to Equation 1.8, the Information Imbalance is symmetric above and below it.*

and set it equal across all data points to a perplexity $\Sigma = 2^{H_i}$. Derivatives can then be computed exactly making use of the implicit function theorem.

In the following parts, though, for the sake of simplicity we will use $\lambda_i = \min_j d_{\boldsymbol{w}\odot\boldsymbol{A}}(X_i, X_j)$ and we will employ the automatic differentiation performed by the Python package JAX (Bradbury et al., 2018) (using, step by step, the current nearest neighbour of each point).

2

Error estimation of information imbalance

In this chapter we introduce an extension of the Information Imbalance which allows to compute analytically a statistical error. The cumulative mean of this estimator converges to the same value obtained with the standard version in Equation 1.4 (see Fig. 2.1), namely it gives the same result on average.

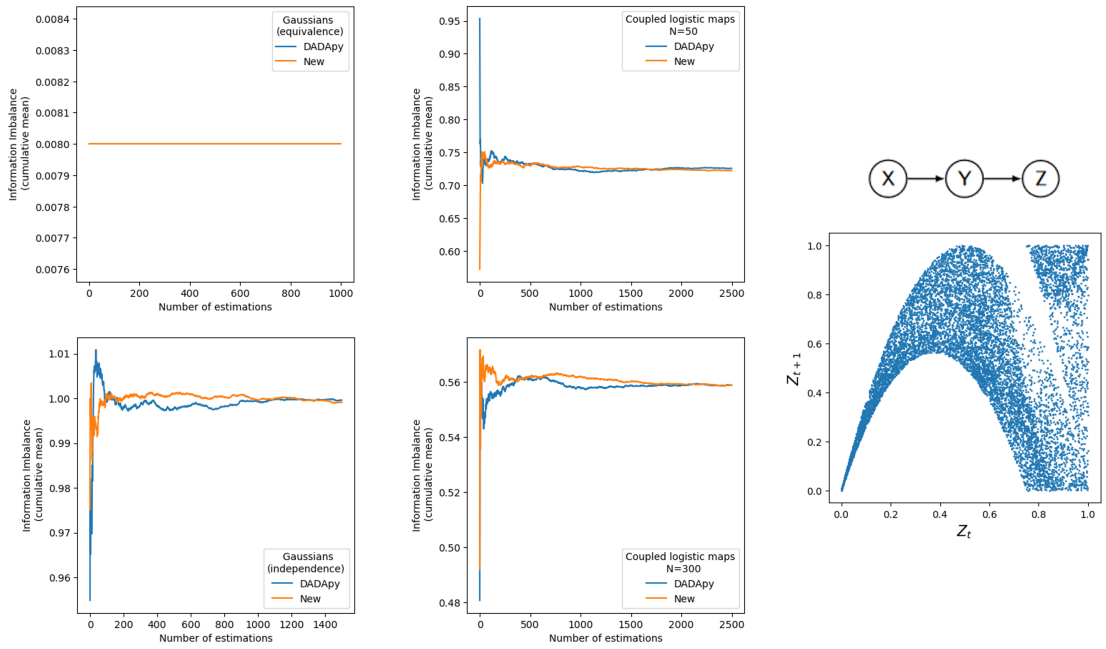

Figure 2.1: *The convergence of both the current and the previous estimate to the same value is shown for different systems and settings. On the left we compute the information imbalance between two equivalent (top) or independent (bottom) components of data sampled from a multivariate Gaussian distribution. On the right, we compute* $\Delta(d_{(Y_t,Z_t)} \to d_{(Y_{t+1},Z_{t+1})})$*, using coupled logistic maps (as in Section 2.1.1).*

The main idea is making independent the computed ranks by separating the dataset in two groups, as summarised in Fig. 2.2. Specifically, the ranks in space A and B are computed considering the distances of the elements of the first group from those in the second (hence avoiding to consider distances between elements of the same group).

This allows to remove the dependence between ranks that arises when considering the whole dataset. Indeed, if distances are computed between all pairs of points, the ranks r_{ij} and r_{ji} are not independent, due to the symmetry property of any distance function, and r_{ik} is not independent of r_{ij} and r_{jk}, as a consequence of the triangular inequality.

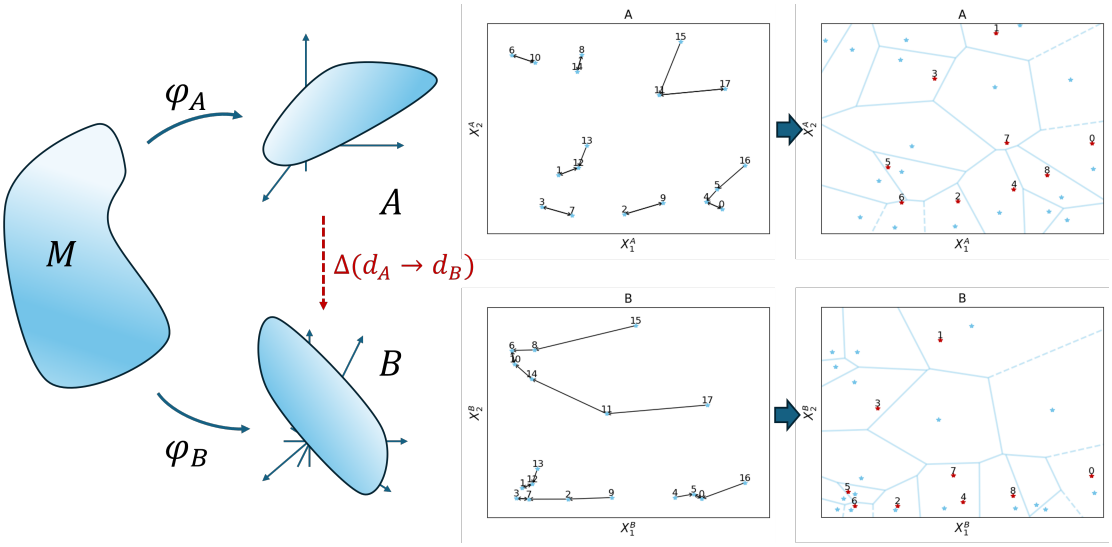

Figure 2.2: *While the normal information imbalance checks how the neighbouring relationships change when going from space A to space B (central graphs), the version proposed here first uses part of the dataset to divide the two spaces into cells, then checks how the remaining data change its position in them (graphs on the right).*

2.1 Mathematical details

Consider a dynamical system living in a space M and two different observations given by $\varphi_A : M \to A$ an $\varphi_B : M \to B$. In proper conditions, we can define a probability density function $\rho : M \to \mathbb{R}$ and our samples will look like:

$$\begin{aligned} X &\sim \rho \,, \\ X^A = \varphi_A(X) &\sim \rho_A \,, \\ X^B = \varphi_B(X) &\sim \rho_B \,, \end{aligned}$$

where ρ_A and ρ_B are the corresponding densities in spaces A and B. Given a set $S' = \{X_\alpha\}_{\alpha=1,\dots,|S'|}$ where X_α are i.i.d. from ρ, we can automatically access the observations $S'^A = \{X^A_\alpha\}$ and $S'^B = \{X^B_\alpha\}$ by applying the maps φ_A and φ_B.

We define $p(r^A_{i\alpha} = 1|S')$ as the probability that a new point X_i sampled from ρ has X^A_α as the first neighbour in S'^A, when mapped in space A. This probability can be written as

$$\begin{aligned} p(r^A_{i,\alpha} = 1|S') &= \int_I \rho_A(x)\, dx \,, \\ I &= \left\{ x \in A | \operatorname*{argmin}_{\tilde{\alpha}} d(x, X^A_{\tilde{\alpha}}) = \alpha \right\} . \end{aligned}$$

Here we call $r^A(X_i, X_\alpha) = r^A_{i,\alpha}$ the rank of $d(X^A_i, X^A_\alpha)$ in the list of ordered distances $\{d(X_i, X_{\tilde{\alpha}})\}_{\tilde{\alpha}=1,\dots,|S'|}$ [1] In a case in which ρ_A is uniform, this is exactly like computing the volume of a specific cell of the Voronoi tassellation generated by the set S'^A.

Suppose now to sample N new points ($S = \{X_i\}_{i=1,\dots,N}$) independently from S' and among themselves from ρ. Then

$$p(\{r^A_{i,\alpha}\}_{i=1,\dots,N}|S') = \prod_{i=1}^{N} p(r^A_{i,\alpha}|S')^{[2]} \tag{2.1}$$

Now consider the following expectation value:

$$\begin{aligned}\mathbb{E}[r^B|r^A = 1, S'] &= \sum_{k=1}^{N} k\, p(r^B = k|r^A = 1, S') \\ &= \sum_{k=1}^{N} k\, \frac{p(r^B = k, r^A = 1|S')}{p(r^A = 1|S')}\,,\end{aligned}$$

where

$$\begin{aligned}p(r^B = k, r^A = 1|S') &= \int_{I''} \rho(x)dx\,, \\ I'' &= \bigcup_{X_\alpha} \{x \in M | r^A(x, X_\alpha) = 1, r^B(x, X_\alpha) = k\}\,.\end{aligned}$$

This expectation value can be estimated by the right-hand side of the following equation:

$$\frac{|S'|+1}{2}\Delta_{S'}(d_A \to d_B) := \frac{1}{N}\sum_{i=1}^{N}\sum_{\alpha=1}^{|S'|} r^B_{i,\alpha}\delta_{r^A_{i,\alpha},1}\,,$$

that we use to define $\Delta_{S'}(d_A \to d_B)$. The multiplicative factor is inserted as a normalization element adjusting for the size of S'. It is chosen in such a way that if $p(r^B|r^A = 1)$ is uniform over the rank values $\{1, 2, \dots, |S'|\}$ then $\mathbb{E}[r^B|r^A = 1, S'] = 1$, as it was done in the standard Information Imbalance. From now on we will consider $\Delta_{S'}(d_A \to d_B)$ as a central element in our discussion. To simplify the notation, we will write $\Delta_{S'}(d_A \to d_B)$ simply as $\Delta_{S'}$, and we will write $\mathbb{E}[r^B|r^A = 1, S']$ for $\frac{2}{|S'|+1}\mathbb{E}[r^B|r^A = 1, S']$, such that $\mathbb{E}[r^B|r^A = 1, S'] \approx \Delta_{S'}$. Similarly, we will drop the prefactor dependent on $|S'|$ also when writing variances $\mathbb{V}[\cdot]$.

Since the sample mean is an unbiased estimator of the expectation value $\Delta_{S'}$ is an unbiased estimator of $\mathbb{E}[r^B|r^A = 1, S']$ and, since each of the elements $e_i = \frac{2}{|S'|+1}\sum_\alpha r^B_{i,\alpha}\delta_{r^A_{i,\alpha},1}$ is independent from the others because of 2.1, we have that

$$\mathbb{V}[\Delta_{S'}|S'] = \frac{1}{N}\mathbb{V}[r^B|r^A = 1, S']$$

[1] Notice that the probabilities of higher ranks can also be written in a similar way to what we have here. Take for example $p(r^A_{i,\alpha} = 2)$, it can be computed as an integral over a new set $I' = \left[\bigcup_{\beta\in(S\setminus\{\alpha\})}\{x \in A|\,\mathrm{argmin}_{\tilde{\alpha}\setminus\beta}\, d(x, X^A_{\tilde{\alpha}}) = \alpha\}\right] \setminus \{x \in A|\,\mathrm{argmin}_{\tilde{\alpha}}\, d(x, X^A_{\tilde{\alpha}}) = \alpha\}$.

[2] This is a substantial difference from what we had with the previous imbalance, where for example $p(r^A_{ij} = 1, r^A_{ji} = 1|S \setminus \{i, j\}) \neq p(r^A_{ij} = 1|S \setminus \{i\})\, p(r^A_{ji} = 1|S \setminus \{j\})$.

which can be estimated in an unbiased way as[3]

$$\mathbb{V}[r^B|r^A = 1, S'] \approx \frac{1}{N-1}\sum_{i=1}^{N}(e_i - \Delta_{S'})^2 \tag{2.2}$$

It is easy to generalize the same statement also for

$$\tilde{e}_i = \frac{2}{(|S'|+1)k_{max}}\sum_{k=1}^{k_{max}}\sum_{\alpha=1}^{|S'|} r^B_{i,\alpha}\delta_{r^A_{i,\alpha},k}$$

In real applications one would be interested in a slightly different quantity, i.e. $\mathbb{E}_{S'}\mathbb{E}[r^B|r^A = 1, S'] \approx \Delta$, where the first expected value is computed over the distribution of possible S'. Supposing that S' can be independently sampled R times, we can estimate this double expected value as:

$$\mathbb{E}[r^B|r^A = 1, S'] \approx \frac{1}{R}\sum_{r=1}^{R}\Delta^r_{S'} = \overline{\Delta_{S'}}\,,$$

where the superscript r runs over multiple estimation of the $\Delta_{S'}$ with different S'.
We can then see that for this quantity we have:

$$\mathbb{E}_{S'}\left[\frac{1}{R-1}\sum_{r=1}^{R}(\Delta^r_{S'} - \overline{\Delta_{S'}})^2\right] = \mathbb{V}_{S'}[\mathbb{E}[r^B|r^A = 1, S']] + \frac{1}{N}\mathbb{E}_{S'}[\mathbb{V}[r^B|r^A = 1, S']]\,. \tag{2.3}$$

Proof. We can see that in general for the empirical averages

$$\overline{X}^r = \frac{1}{N_r}\sum_{i=1}^{N_r} X^r_i$$

$$\overline{\overline{X}} = \frac{1}{R}\sum_{r=1}^{R}\overline{X}^r$$

we can write

$$\begin{aligned}
&\frac{1}{R-1}\sum_{r=1}^{R}\left(\overline{X}^r - \overline{\overline{X}}\right)^2 = \\
=\ &\frac{1}{R-1}\sum_{r=1}^{R}\left[\frac{\left(\sum_{i=1}^{N_r} X^r_i\right)^2}{N_r^2} + \frac{\left(\sum_{r'=1}^{R}\overline{X}^{r'}\right)^2}{R^2} - \frac{2\left(\sum_{i=1}^{N_r} X^r_i\right)\left(\sum_{r'=1}^{R}\overline{X}^{r'}\right)}{RN_r}\right] \\
=\ &\frac{1}{R-1}\left[\sum_{r=1}^{R}\left(\frac{\sum_{i=1}^{N_r} X^r_i}{N_r}\right)^2 - \frac{\left(\sum_{r'=1}^{R}\overline{X}^{r'}\right)^2}{R}\right] \\
=\ &\frac{1}{R-1}\left[\sum_{r=1}^{R}\left(\frac{\sum_i X^r_i}{N_r}\right)^2 - \frac{1}{R}\left(\sum_{r'=1}^{R}\left(\frac{\sum_i X^{r'}_i}{N^2_{r'}}\right)^2 + \sum_{r''\neq r'}\frac{\left(\sum_i X^{r'}_i\right)}{N_{r'}}\frac{\left(\sum_i X^{r''}_i\right)}{N_{r''}}\right)\right] \\
=\ &\frac{1}{R}\sum_{r=1}^{R}\frac{1}{N^2_r}\left(\sum_i (X^r_i)^2 + \sum_{i\neq j} X^r_i X^r_j\right) - \frac{1}{R(R-1)}\sum_{r''\neq r'}\frac{\left(\sum_i X^{r'}_i\right)}{N_{r'}}\frac{\left(\sum_i X^{r''}_i\right)}{N_{r''}}
\end{aligned}$$

[3] Another way to reach the same conclusion is to see that, given S', e_i are sampled independently from a multinomial distribution with values equally spaced from $\frac{2}{|S'|+1}$ to $\frac{2|S'|}{|S'|+1}$

Now, since in our case $\overline{X}^r = \Delta_{S'}$ and all X_i^r are drawn independently, assuming that all N_r are equal, we can then write:

$$\begin{aligned}
&\mathbb{E}_{S'}\mathbb{E}[\cdot] = \\
= \;& \frac{1}{R}\mathbb{E}_{S'}\sum_{r=1}^{R}\left[\frac{1}{N_r}\mathbb{E}\left[X^2|S'\right] + \mathbb{E}\left[X|S'\right]^2\frac{N_r(N_r-1)}{N_r^2}\right] - \frac{1}{R(R-1)}\mathbb{E}_{S'}\left[\sum_{r''\neq r'}\mathbb{E}\left[X|S'\right]\mathbb{E}\left[X|S''\right]\right] \\
= \;& \frac{1}{N_r}\mathbb{E}_{S'}\left[\mathbb{E}\left[X^2|S'\right]\right] + \frac{N_r(N_r-1)}{N_r^2}\mathbb{E}_{S'}\left[\mathbb{E}\left[X|S'\right]^2\right] - \mathbb{E}_{S'}\left[\mathbb{E}\left[X|S'\right]\right]^2 \\
= \;& \mathbb{V}_{S'}\mathbb{E}\left[X|S'\right] + \frac{1}{N_r}\mathbb{E}_{S'}\mathbb{V}\left[X|S'\right]
\end{aligned}$$

□

The behaviour of these two terms is briefly discussed in the next subsection.

In Appendix B some extensions of the previous calculations are presented for the case of categorical variables.

2.1.1 Numerical estimation of the error

It is easy to construct unbiased estimators for at least two of the quantities appearing in Equation 2.3. In particular the quantity on the left-hand side can be explicitly estimated as it is, while the second term in the right-hand side can be estimated in an unbiased way making use of Equation 2.2.

We now empirically investigate the relevance of the different terms in Equation 2.3 varying only $|S'|$. In particular, in Fig. 2.3 we depict the results obtained for the following coupled logistic maps:

$$\begin{cases} X_{t+1} = X_t(r - rX_t + \sigma\eta_t^X) \; mod \; 1, \\ Y_{t+1} = Y_t(r - rY_t - aX_t + \sigma\eta_t^Y) \; mod \; 1, \\ Z_{t+1} = Z_t(r - rZ_t - bY_t + \sigma\eta_t^Z) \; mod \; 1, \end{cases} \tag{2.4}$$

with $r = 4, a = 1, b = 1, \sigma = 0$ (a two dimensional Poincaré plot of the map Z is shown on the right panel of Fig. 2.1). Increasing the size of S' allows to make the second term on the right dominant. This is particularly important because sampling a lot of sets S' to estimate the variance would be a very data-hungry procedure. Instead, thanks to the emergence of some concentration properties for large $|S'|$, we can reliably estimate the error of Δ as that of $\Delta_{S'}$ in this regime (see Fig. 2.3).

2.2 Gradient descent algorithm

Given the new version of the Information Imbalance, we rewrite Equation 1.6 as follows:

$$\Delta^{\lambda}(d_{\boldsymbol{w}\odot\boldsymbol{A}} \rightarrow d_B) = \lim_{\lambda\to 0}\frac{2}{N(|S'|+1)}\sum_{i,j} c_{i,j}(\lambda, \boldsymbol{w}) r_{ij}^B \,, \tag{2.5}$$

We also use a version of stochastic gradient descent in which S' is fixed, while subsets of S are sampled at each step.

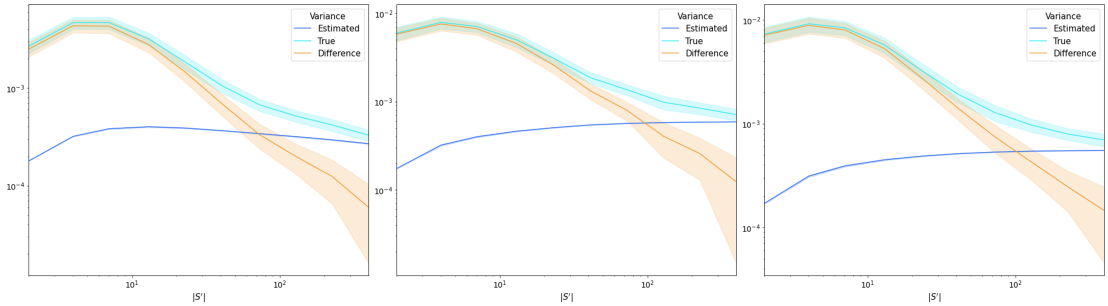

Figure 2.3: *From left to right the variances of* $\Delta(d_{Y_t} \to d_{Y_{t+1}})$, $\Delta(d_{Z_t} \to d_{Z_{t+1}})$, $\Delta(d_{(X_t,Y_t,Z_t)} \to d_{(X_{t+1},Y_{t+1},Z_{t+1})})$ *computed with* $|S| = 600$ *fixed. Standard deviations are computed with 100 repetitions of the calculations. The light blue line represent an estimate of the left-hand side of Equation 2.3, while the dark blue one represents the second term on the right-hand side of the same equation, which becomes the dominant one for large* $|S'|$.

3

Causal discovery of autonomous sets

In this chapter we finally deploy the Information Imbalance for causal discovery in time series data.

When dealing with high-dimensional dynamical systems, the problem of detecting causal links is considered computationally demanding (Runge, 2018). We introduce a method which allows partially mitigating the complexity of further analyses, in a divide-and-conquer spirit. The idea is that, if an *autonomous set* of variables (namely a set of variables whose time evolution depends only on the variables inside the set itself) can be identified, then, considering this group as a single "node" will not mar the whole analysis, but will make it computationally simpler, reducing the number of statistical tests that one should perform. By progressively finding the minimal autonomous subsets it is in principle possible to rebuild step by step the dynamical relations of the whole system.

3.1 Preliminary observations

Consider a graph representing a time series graph (*microscopic graph*) with all times collapsed: each node represents a variable and a link is present if at any time delay a link was present in the microscopic graph. This is what is commonly called a *macroscopic graph* (Chicharro et al., 2014). For example, the microscopic and macroscopic graphs of the map in Equation 2.4 can be represented as in Fig. 3.1.

Given a node X^i in the macroscopic graph, it is easy to see that the minimal autonomous set $\mathcal{A}^i$ to which it belongs contains all and only its ancestors, aside from X^i itself. Equivalently, $\mathcal{A}^i$ contains X^i and all and only its causes (direct and indirect). Always in the example of Equation 2.4 we have $\mathcal{A}^Z = \{X, Y, Z\}$, $\mathcal{A}^Y = \{X, Y\}$, $\mathcal{A}^X = \{X\}$. Following the ideas in Del Tatto et al., 2024 already presented at the end of Section 1.2, we observe that the distance space $d_{\boldsymbol{w}\odot\boldsymbol{X}(t)}$ which optimally predicts future distances in space $d_{X^i(t+\tau)}$, using all the dynamical variables at time t, should assign zero weight to all the variables that do not belong to $\mathcal{A}^i$, which are neither direct nor indirect causes of X^i. It is then possible to reconstruct $\mathcal{A}^i$ for each variable by computing, for different

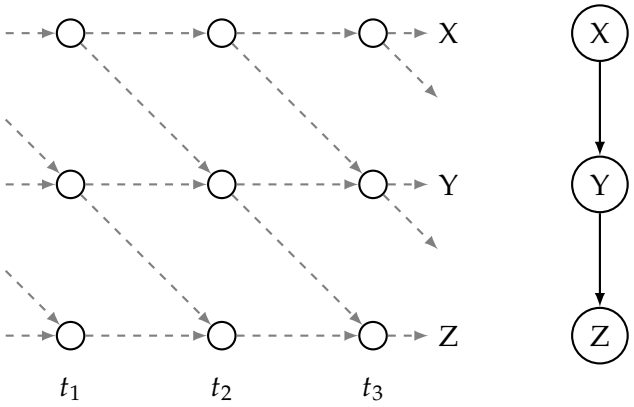

Figure 3.1: *On the left the microscopic causal graph of the map in Equation 2.4, in which we have a new node for each time step. On the right the corresponding macroscopic graph.*

values of τ,

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmin}} \; \Delta\left(d_{\boldsymbol{w}\odot\boldsymbol{X}(t)} \to d_{X^i(t+\tau)}\right) \,, \tag{3.1}$$

where $\boldsymbol{w}$ is a vector of parameters weighting each dynamical variable, and $\odot$ denotes the element-wise product. If the weight w_j is found to be significantly different from zero (see Section 3.2), we conclude that $X^j(t)$ is a (direct or indirect) cause of $X^i(t+\tau)$ in the microscopic graph. Equivalently, $w_j \neq 0$ implies that X^j is a direct or indirect cause of X^i in the macroscopic graph, namely that X^j belongs to the minimal autonomous set of X^i. In principle, this procedure can identify multi-body interactions just as it identifies individual links, an example is shown in Section B.1.

As an example, we show in Fig. 3.2 the landscape of the Information Imbalance as a function of the weights $\boldsymbol{w}$ for the three coupled logistic maps of Equation 2.4.

Using $\tau = 1$ we recover, for example, the direct links $Z(t) \to Z(t+1)$ and $Y(t) \to Z(t+1)$, as their weights are non-zero in the minimum of the Information Imbalance, while we do not infer the presence of any link $X(t) \to Z(t+1)$, as the weight of X is found to be zero in the same minimum (right panel of Fig. 3.2). However, the minimal autonomous set of Z also contains X, for which an (indirect) link can be observed only by setting $\tau > 1$. For example, setting $\tau = 2$ in the same test results in a global minimum where all the three weights are non-zero. This link, then, can be detected by considering other time lags in the trial distance which is the left argument of Equation 3.1. This can be achieved using the differentiable version of the Information Imbalance presented in the previous chapter.

3.2 Algorithmic implementation

Consider now N observations of the couple $(\mathbf{X}(t), \mathbf{X}(t+\tau))$ with $\mathbf{X} \in \mathbb{R}^d$. As we outlined, it is possible to understand whether a variable at time t is directly or indirectly causing another variable X^i at time $t+\tau$ by minimizing the quantity $\Delta\left(d_{\boldsymbol{w}\odot\boldsymbol{X}(t)} \to d_{X^i(t+\tau)}\right)$. After dividing the dataset in two subsets S and S', we can use the differentiable version of the Information Imbalance for a fast search of local minima also in high-dimensional

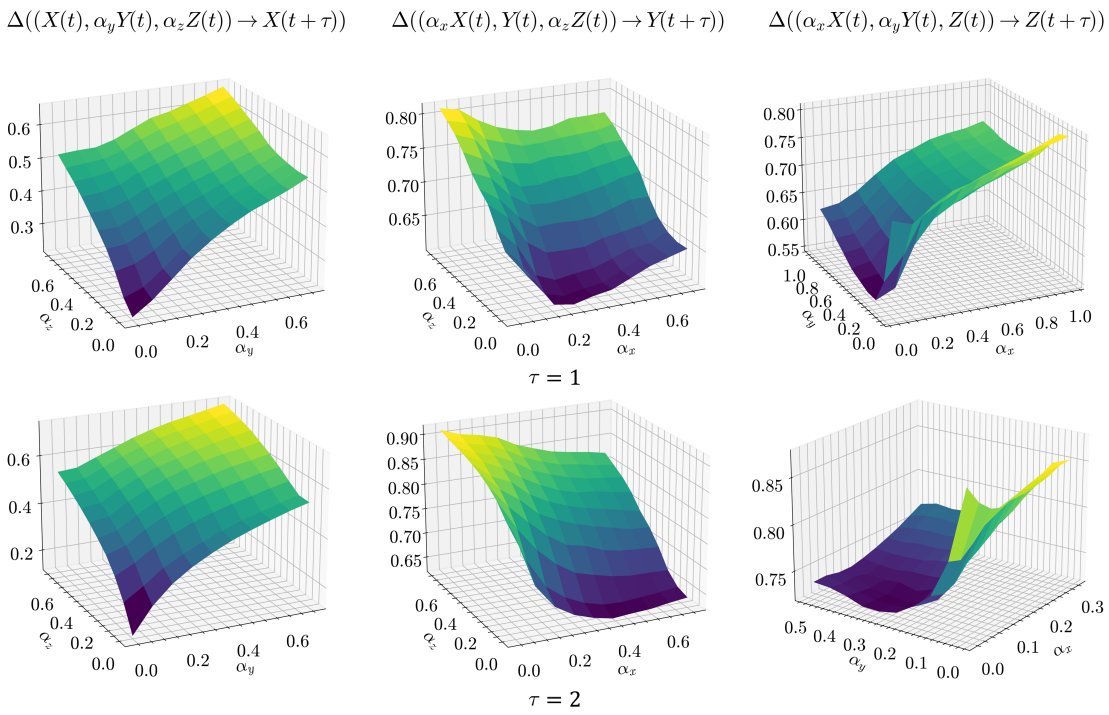

Figure 3.2: *Since the distance $d_{\boldsymbol{w}\odot\boldsymbol{X}(t)}$ in Equation 3.1 is invariant under a global scaling of all weights $\boldsymbol{w}$, for the three logistic maps it is possible to reduce the number of parameters employed in the test from 3 to 2, allowing for a 3d representation. The results in the first line are obtained for $\tau = 1$. As expected, only $X(t)$ is relevant to predict $X(t+1)$ (left panel, the global minimum is found in $\alpha_Z = \alpha_Y = 0$, $\alpha_X = 1$), both $Y(t)$ and $X(t)$ for $Y(t+1)$ (central panel, the global minimum is in $\alpha_X \neq 0$, $\alpha_Z = 0$,$\alpha_Y = 1$) and $Z(t), Y(t)$ for $Z(t+1)$ (right panel, the global minimum is in $\alpha_X = 0$, $\alpha_Y \neq 0$ $\alpha_Z = 1$). In the second line we plot also the results for $\tau = 2$. The main difference is that the variable $X(t)$ becomes relevant to predict $Z(t+\tau)$*

spaces. A local minimum of the Information Imbalance, we recall, corresponds to a combination of variables which provides maximal information on another combination of variables (in our case, a single variable observed with a time lag τ in the future).We can then scan multiple time lags to take into account faster and slower transmission of information.

The causal links in our approach are stored in a $n \times n$ matrix that we denote by G. The entries of G store the weights found with the minimisation procedure. In particular the element G_{ij} will be the absolute value of the i-th component of the weight vector $\hat{\boldsymbol{w}}$ found by minimizing $\Delta\left(d_{\boldsymbol{w}\odot\boldsymbol{X}(t)} \to d_{X^j(t+\tau)}\right)$ (namely $G_{ij} = |\hat{w}_i|$). For each new tested value of τ, we minimise d times the Information Imbalance, using a weighted combination of all the variables at time t in space A and one variable at time $t+\tau$ in space B. We will update the elements G_{ij} in such a way that they correspond to the highest weight found for any value of τ. The general procedure is illustrated in Algorithm 1. The outcome is a fully connected weighted graph.

Notice that the number of optimizations that one should perform to estimate the matrix G scales linearly with the number of variables d. An algorithm that computes the mutual informations or the transfer entropies among all the possible couples of

Algorithm 1 `Graph updating`

if G `hasn't been initialised yet` **then**
  `set` G `as a` $\mathbb{0}$ `valued` $d \times d$ `matrix`
end if
for $j \in \{1, ..., d\}$ **do**
  $\hat{w} \leftarrow \text{argmin}_{\{w\}} \; \Delta\left(d_{w \odot X(t)} \rightarrow d_{X^j(t+\tau)}\right)$
  for $i \in \{1, ..., d\}$ **do**
    $G_{ij} \leftarrow \max\left\{G_{ij}, |\hat{w}_i|\right\}$
  end for
end for

variables at time t and $t + \tau$ scales as d^2. If the computation of multi-body interactions is attempted, the scaling is instead exponential. The matrix G does not directly represent the coarse grained causal graph, but contains all the necessary information to construct it. In order to identify the autonomous sets, we first remove all the links whose weights are lower than a specified threshold. This allows to remove the weights which are numerically different from zero but can be safely neglected. In the next parts and in Appendix C we show that the results are robust with respect to the choice of this threshold. Applying this procedure, the G matrix becomes sparse (as we expect in the macroscopic representation of a system composed by smaller subsystems). We will call A the unweighted adjacency matrix of the resulting graph: $A_{ij} = 1$ if $G_{ij} >$ threshold and $A_{ij} = 0$ otherwise. Given the adjacency matrix A the autonomous sets can be found applying a recursive procedure.

In the first step for each variable we look for the set of all its ancestors. This can be done efficiently with an algorithm similar in spirit to that of breadth first search (Cormen et al., 2017): in particular, one starts with a set containing only the node itself, then considers all the links pointing into a node in the set, and adds to the set itself all the nodes from which these links are exiting. This procedure is repeated recursively until the set is not changed anymore. Then, the smallest of these groups of ancestors will certainly be a minimal autonomous set. We can remove from A the rows and lines referred to the autonomous set identified at the previous step and repeat recursively the procedure until no elements in A are left.

This allows to find groups with a well defined hierarchy of "autonomy": if a group $\mathcal{G}_k$ is identified at step k of the algorithm, then it can be caused (directly or indirectly) only by groups $\mathcal{G}_j$ identified at a step $j < k$, and, in turn, it can cause only groups $\mathcal{G}_{j'}$ identified in the next steps $j' > k$. We will call the group of autonomous sets for the whole system the first *shell* of autonomy. Then, all the groups which depend only on variables in the first shell will be part of the second shell and so on.

A "coarse grained" macroscopic graph can be finally drawn by representing the variables in the same groups as single nodes, and drawing the links between such nodes using the microscopic adjacency matrix A. For example, a link between node $\mathcal{G}_j$ and $\mathcal{G}_k$

will be drawn if at least a link from any variable of $\mathcal{G}_j$ to any variable of $\mathcal{G}_k$ is present in the all-variable representation. Particular care, though, should be put on the fact that these links may represent both direct and indirect connections. Indeed, with the minimisation of the Information Imbalance we can only state if there is an information flow between variables at different time lags or not. Identifying whether this flow occurs through other variables requires additional tests, involving appropriate conditioning on different subsets (Chicharro et al., 2014).

As we will show in specific examples the procedure presented above still allows to identify some direct links with high statistical confidence. Indeed, if a link among two subsets is identified and if it is also the only path connecting them, then that link is surely a direct one. For example, each link between groups of consecutive shells is necessarily a direct link, in this case it is easy to see that the information flow cannot be blocked by any conditioning on any variable of the system.

3.3 Benchmark tests

We test this approach on trajectories generated by deterministic chaotic systems of different complexity. First of all we consider a system of three Rössler oscillators X,Y and Z, described each by the following three non-linear ordinary differential equations

$$\begin{cases} \frac{dx_i}{dt} = -\omega_i y_i - z_i \\ \frac{dy_i}{dt} = \omega_i x_i + 0.15 y_i \\ \frac{dz_i}{dt} = 0.2 + z_i(x_i - 10)\,, \end{cases}$$

with $\omega_X = 1.015$, $\omega_Y = 0.985$, $\omega_Z = 1.005$. We considered these coupling scenarios:

- unidirectional couplings $X \to Z$ and $Y \to Z$;
- unidirectional couplings $X \to Z$ and $Z \to Y$;
- unidirectional coupling $X \to Y$, Z uncoupled.

realised similarly to those in Del Tatto et al., 2024.

We then consider two unidirectionally coupled Lorenz 96 systems of 40 variables each, defined for $i = 1, \cdots, 40$ by

$$\frac{dx_i}{dt} = (x_{i+1} - x_{i-2})x_{i-1} - x_i + F_{X/Y}\,,$$

where $x_{-1} = x_{40}$, $x_{41} = x_1$, $F_X = 5$, $F_Y = 6$. As above, the coupling between the systems of the kind presented in Del Tatto et al., 2024. These systems represent a challenging test of a high-dimensional setting.

After integrating the previous equations we used our procedure to retrieve the coarse grained graph. In Fig. 3.3 we plot our results. All the nodes belonging to the same autonomous set are painted with the same color and then depicted in the coarse grained version of the graph in the small box in each figure.

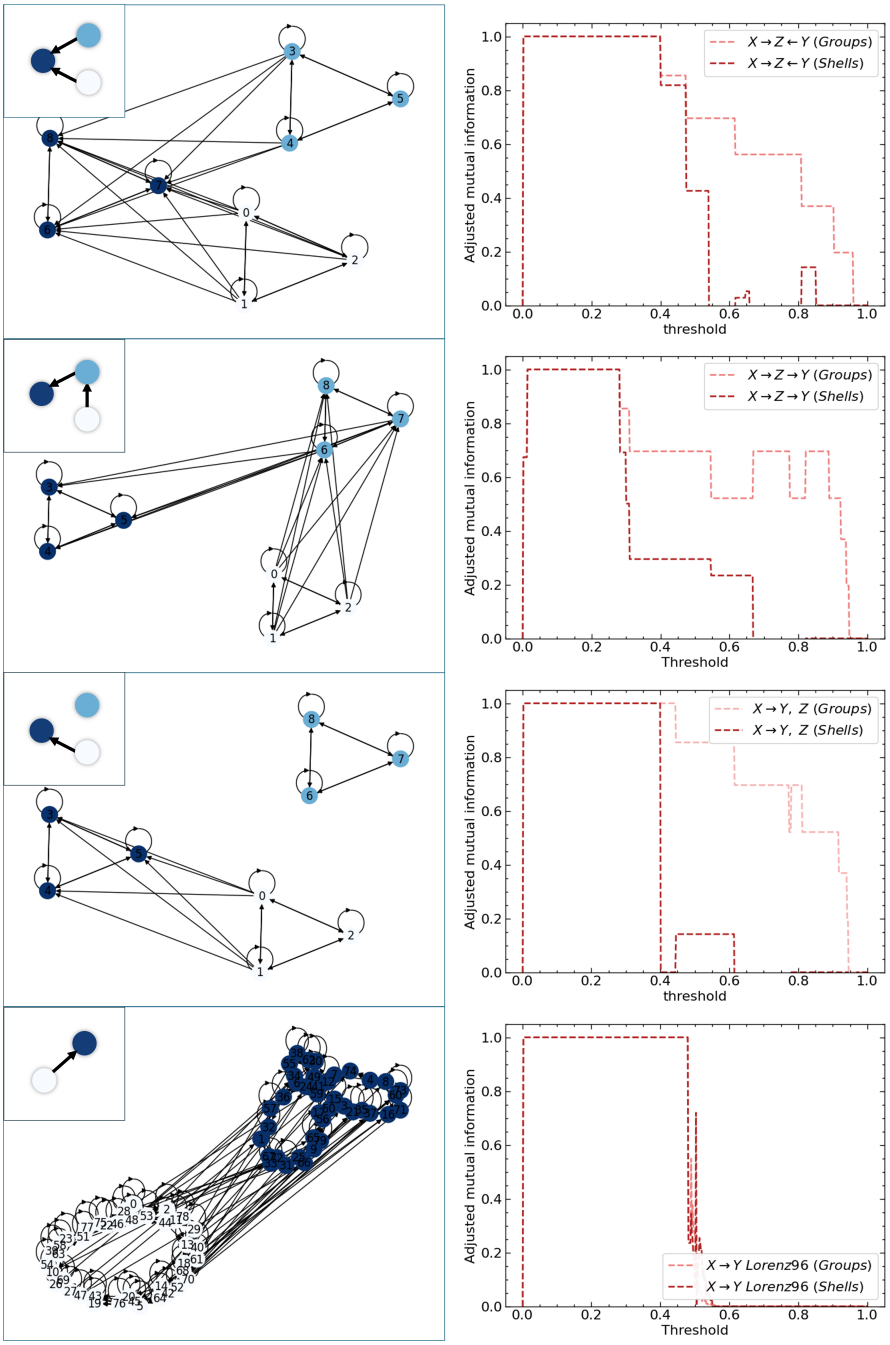

Figure 3.3: *We plot on the left column the graphs retrieved using our procedure for different systems and using a threshold of 0.1. Nodes of the same color are found to belong to the same autonomous set in different steps of the procedures. Colors range from the lightest to the darkest, showing the order of retrieval and the hierarchy of the autonomous sets. On the right column we show how the (adjusted) mutual information between the true groups (and shells) and the one found following our procedure change when modifying the threshold. In general, as can be seen from the initial plateau with mutual information 1, the solutions are consistent for quite a large range of the threshold. From the top to the bottom: 3 Rössler oscillators coupled as $X \longrightarrow Z \longleftarrow Y$; 3 Rössler oscillators coupled as $X \longrightarrow Z \longrightarrow Y$; 3 Rössler oscillators coupled as $X \longrightarrow Y, Z$ autonomous; two 40-dimensional Lorenz 96 coupled unidirectionally $X \longrightarrow Y$.*

Since the outcome of the classification in autonomous sets (but also that of the division in shells) depends on the threshold used to build the matrix A from matrix G, we study the effect of changing this parameter. In particular on the right column of Fig. 3.3 we plot the adjusted mutual information (AMI) (Vinh et al., 2010) between the true groups and the recollected ones. The AMI modifies the estimation of the mutual information to correct for the fact that generally its values tends to be higher for two classifications with a larger number of clusters. A value of 1 is obtained when the two partitions are perfectly consistent. Remarkably, our method appears to robustly deduce both the correct groups and shells for a huge range of the threshold parameter, even for the high-dimensional systems.

Further analyses on the change of other parameters of the algorithm are presented in Appendix C and in the next section.

3.4 Stability of the results

After having shown the accuracy of the algorithm in high dimensional setups and its robustness to variations of the threshold for the construction of the matrix A, we now focus on the effect of the time lag , of statistical noise and of observational noise.

In order to perform our analysis, we consider five logistic maps coupled as in Equation 2.4, with links $X_t^1 \to X_{t+1}^2$, $X_t^2 \to X_{t+1}^3$, $X_t^3 \to X_{t+1}^4$ and $X_t^4 \to X_{t+1}^5$. The advantage of using maps of this kind is that the true links at any time lag can be deduced unambiguously from the equations. This is the perfect playground to check exactly our results with an unequivocal ground truth.

In Fig. 3.4 we show the weights retrieved when minimizing $\Delta\left(d_{\boldsymbol{w}\odot\boldsymbol{X}(t)} \to d_{X^5(t+\tau)}\right)$ varying the time lag. First of all, we verify that considering longer time lags allows to check for slower information transfer. In particular, we were able to see (Fig. 3.4, upper panel) that before the information can physically pass from one part of the system to the other, no link is retrieved by minimizing $\Delta\left(d_{\boldsymbol{w}\odot\boldsymbol{X}(t)} \to d_{X^i(t+\tau)}\right)$. For example, the link $X_t^3 \to X_{t+\tau}^5$ only appears for $\tau \geq 2$ (the ground-truth lag of this indirect link is 2), while the link $X_t^1 \to X_{t+\tau}^5$ is detected for $\tau \geq 5$ (the ground-truth lag of this indirect link is 4). It also turns out that the variance of the position of the retrieved minimum of the Information Imbalance increases when increasing the time lag. In this case also $\min \Delta$ itself grows approaching the value of one (Fig. 3.4, lower panel). In this regime it is actually quite difficult to infer causality, as the system has lost almost all the memory of its initial state and the dependencies among the variables are completely washed away.

Next, we benchmarked the effect of modifying the size of the dataset. In particular, in Fig. 3.5 we check the effect of changing both $|S|$ and $|S'|$ for a fixed $\tau = 3$. As expected, very small datasets (always with less than ~ 100 points) lead to spurious estimations.

Finally in Fig. 3.6 we illustrate the effect of the insertion of a dynamical and an observational noise. The observational noise is a Gaussian noise of standard deviation σ_o added a posteriori to the trajectory. The dynamical one is added at each step of integration of the map, as from Equation 2.4. In the specific case we were considering,

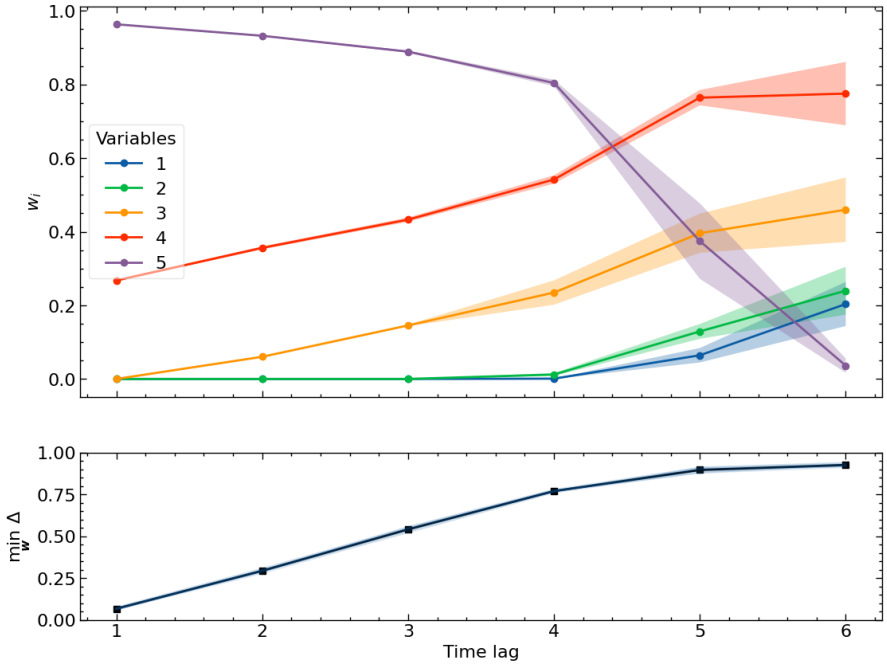

Figure 3.4: *We minimize* $\Delta\left(d_{\boldsymbol{w}\odot\boldsymbol{X}(t)} \longrightarrow d_{X^5(t+\tau)}\right)$ *for different time lags* τ*. We perform 20 estimations for every time-lag and plot the averages. The colored area shows the standard deviation of the estimates. In the graph in the bottom we also show how the minimal Information Imbalance increases with the considered lag.*

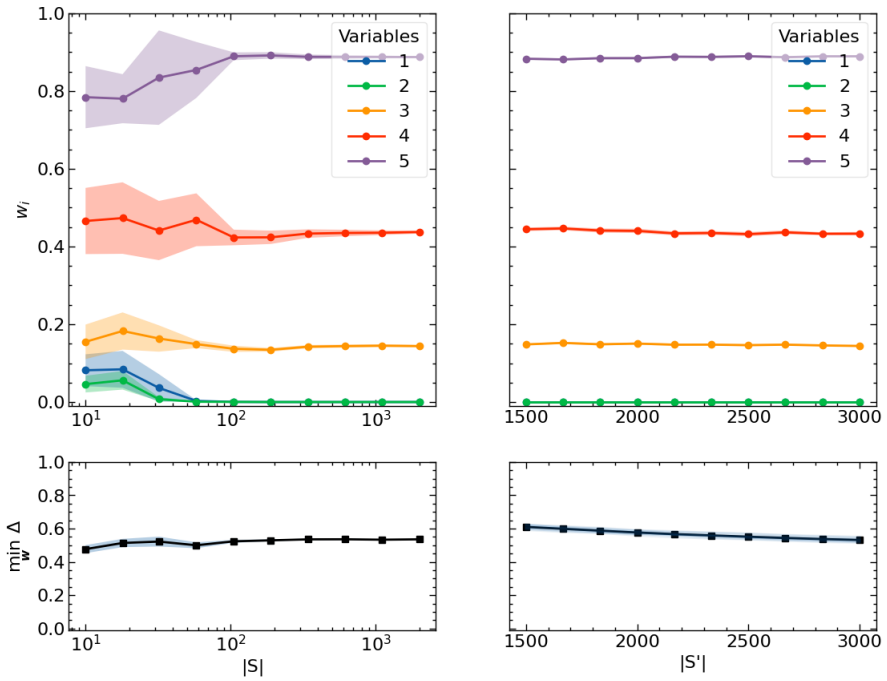

Figure 3.5: *We minimize* $\Delta\left(d_{\boldsymbol{w}\odot\boldsymbol{X}(t)} \longrightarrow d_{X^5(t+3)}\right)$*. On the left we depict the effect of changing* $|S|$ *with* $|S'| = 3000$*, on the right the effect of a change in* $|S'|$ *with* $|S| = 1500$*. As above, we perform 20 estimations for every value of the parameters and plot the averages. The colored area shows the standard deviation of the estimates.*

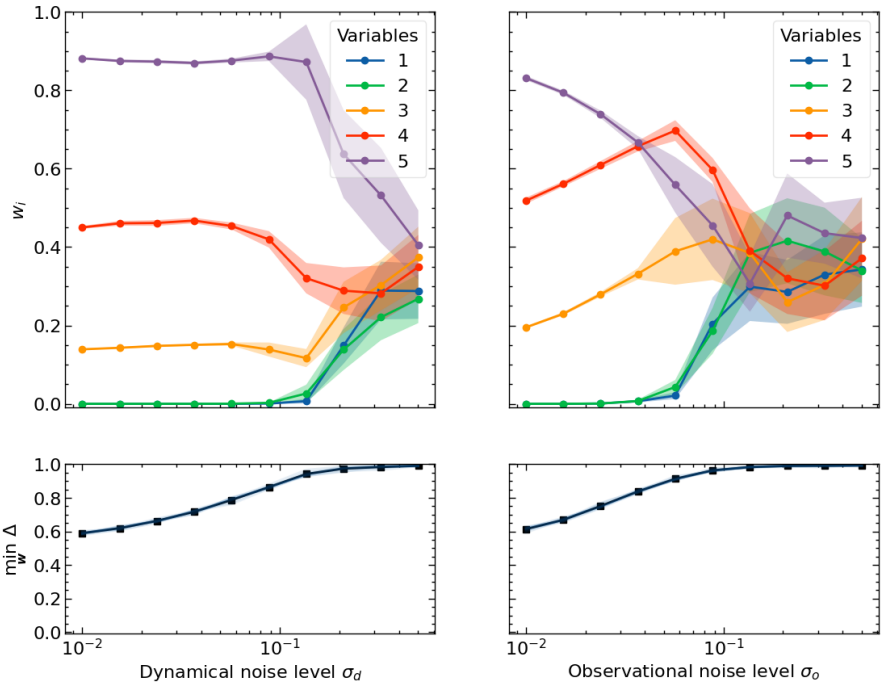

Figure 3.6: *We minimize* $\Delta\left(d_{\boldsymbol{w}\odot\boldsymbol{X}(t)} \to d_{X^5(t+3)}\right)$. *On the left we depict the effect of inserting a dynamical noise, on the right the effect of an observational one. We perform 20 estimations for every value of the parameters and plot the averages. The colored area shows the standard deviation of the estimations.*

in both cases, no spurious link is detected up to a noise of about $\sim 10\%$ of the maximum value of the features. In regimes strongly dominated by the noise a true form of causality is hardly found.

3.5 Alternative approaches for obtaining the causal graph

As a final remark we notice that the approach used for analyzing the systems described in Fig. 3.3 can be considered a first attempt for building a coarse grained graph from the matrix G. In the following, we outline an eigenvalue-based approach which we are currently testing, and which, we hope, might bring to even more robust results.

Given the adjancency matrix A, it is possible to define a stochastic process with transition rates given by the matrix A^T, normalized by rows. This matrix defines jumps between nodes in the opposite direction of the links, namely from effect to causes. Its non-ergodicity can be exploited to recover the minimal autonomous sets, since the dynamics that it generates, starting from a given node, cannot leave the minimal autonomous set to which the starting node belongs. Then, from a natural extension of Perron-Frobenius theorem, the autonomous sets can be recovered from the non-zero components of the left eigenvectors of A^T with eigenvalue 1. Indeed, such eigenvectors allows to define the submatrices in which the matrix A^T can be reduced into a block upper triangular form. This suggests an alternative way to compute the autonomous sets as the connected components of the subgraph of A which contains only the variables which appear in the eigenvectors with eigenvalue 1. Furthermore, this procedure allows

to understand directly the hierarchy of autonomy of the subsets in the graph. Indeed each iteration of this procedure identify a new *shell* of autonomous variables.

Following the previous observations we deduce that the spectral properties of A can be linked to the autonomous set. If A is obtained from G by applying a sufficiently small threshold, then G itself can be seen as a small perturbation of the matrix A (in particular G^T normalized by rows will be "close" to A^T normalised by rows). In the hypothesis that small perturbations in the entrances do not change dramatically the spectrum of the matrix, one can expect that eigenvectors with eigenvalues "close" to one obtained from G will give similar information to those obtained directly from A. In principle, this approach could be used to devise an alternative to the threshold on the parameters. Obviously, though, in the case in which A itself does not show a relevant spectral gap between 1 and the following eigenvalue, it could be hard to isolate the correct eigenvectors. Other possibilities will be commented in Section 4.1.

4

Conclusion

In summary, we first introduced a new version of the Information Imbalance which allows for an easier estimation of its error, and then we employed it to develop a powerful technique for discovering a hierarchy of causal relationships between different groups of features of the dataset. Even though the relationship between the Information Imbalance and the mutual information has been shown (Del Tatto et al., 2024), we verified that the method is able to retrieve the correct results for all the systems we tested, obtaining coarse grained graphs which are consistent with the ground truth, and therefore compliant with those which would be obtained using standard (and more computationally expensive) causal inference methods based on information theory. Of particular interest are the results obtained with the coupled Lorenz 96 systems, which show promising performances of the algorithm in a challenging high-dimensional setting.

4.1 Further perspectives

Even though at the end of the previous chapter we showed the robustness of the results modifying the threshold for the selection of the presence of links, we still believe that some improvements could be made in this direction. Making use of the results obtained in Equation 2.3, it is possible to compute the expected standard error of the Information Imbalance. This can help defining a region in the parameter space around the minimum in which the Information Imbalance does not change relevantly (working in a similar way to what was done by James et al., 1975). By this approach it might be possible to check directly whether any of the parameters is significantly close to zero and to remove links after selecting a specified significance level. Covariances between parameters can also be considered in order to avoid eliminating all of them. We plan to explore this reserach direction in the near future.

The method developed here potentially opens the way for many applications with real-world data. In particular, the fact that it is suitable for high-dimensional systems makes it a remarkable tool to study causality in this specific regime. Among the countless applications, one that strikes into mind is to extend the analysis of the low-frequency variability of the atmosphere presented in Springer et al., 2024. Deepening

the understanding of the topic could help dealing with weather and climate-related risks.

As already suggested in Chapter 1 the method could be efficiently coupled to other tools of system discovery to extend their usage for more complex systems. Some applications could arise in neuroscience where the potential of pairing parsimonious models and machine learning has already been shown (Luo et al., 2023).

This page intentionally left blank.

Bibliography

Assaad, Charles K., Emilie Devijver, and Eric Gaussier (May 2022). "Survey and Evaluation of Causal Discovery Methods for Time Series". In: *J. Artif. Int. Res.* 73. ISSN: 1076-9757. DOI: 10.1613/jair.1.13428. URL: https://doi.org/10.1613/jair.1.13428.

Bradbury, James et al. (2018). *JAX: composable transformations of Python+NumPy programs*. Version 0.3.13. URL: http://github.com/google/jax.

Brunton, Steven L., Joshua L. Proctor, and J. Nathan Kutz (2016). "Discovering governing equations from data by sparse identification of nonlinear dynamical systems". In: *Proceedings of the National Academy of Sciences* 113.15, pp. 3932–3937. DOI: 10.1073/pnas.1517384113. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.1517384113. URL: https://www.pnas.org/doi/abs/10.1073/pnas.1517384113.

Cannelli, G. (2004). *Metodologie sperimentali in fisica*. Edises. ISBN: 9788879592789. URL: https://books.google.it/books?id=Nd13AAAACAAJ.

Chicharro, Daniel and Stefano Panzeri (2014). "Algorithms of causal inference for the analysis of effective connectivity among brain regions". In: *Frontiers in Neuroinformatics* 8. ISSN: 1662-5196. DOI: 10.3389/fninf.2014.00064. URL: https://www.frontiersin.org/articles/10.3389/fninf.2014.00064.

Cormen, H. Thomas et al. (2017). *Introduction to Algorithms*. PHI Learning Private Limited.

Del Tatto, Vittorio et al. (2024). "Robust inference of causality in high-dimensional dynamical processes from the Information Imbalance of distance ranks". In: *Proceedings of the National Academy of Sciences* 121.19, e2317256121. DOI: 10.1073/pnas.2317256121. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.2317256121. URL: https://www.pnas.org/doi/abs/10.1073/pnas.2317256121.

Glielmo, Aldo et al. (Apr. 2022). "Ranking the information content of distance measures". In: *PNAS Nexus* 1.2, pgac039. ISSN: 2752-6542. DOI: 10.1093/pnasnexus/pgac039. eprint: https://academic.oup.com/pnasnexus/article-pdf/1/2/pgac039/53406126/pgac039.pdf. URL: https://doi.org/10.1093/pnasnexus/pgac039.

Granger, C.W.J. (1980). "Testing for causality: A personal viewpoint". In: *Journal of Economic Dynamics and Control* 2, pp. 329–352. ISSN: 0165-1889. DOI: https://doi.org/10.1016/0165-1889(80)90069-X. URL: https://www.sciencedirect.com/science/article/pii/016518898090069X.

James, F. and M. Roos (1975). "Minuit - a system for function minimization and analysis of the parameter errors and correlations". In: *Computer Physics Communications* 10.6, pp. 343–

367. ISSN: 0010-4655. DOI: https://doi.org/10.1016/0010-4655(75)90039-9. URL: https://www.sciencedirect.com/science/article/pii/0010465575900399.

Luo, Thomas Zhihao et al. (2023). "Transitions in dynamical regime and neural mode underlie perceptual decision-making". In: *bioRxiv*. DOI: 10.1101/2023.10.15.562427. eprint: https://www.biorxiv.org/content/early/2023/11/20/2023.10.15.562427.full.pdf. URL: https://www.biorxiv.org/content/early/2023/11/20/2023.10.15.562427.

Mehta, Pankaj et al. (May 2019). "A high-bias, low-variance introduction to Machine Learning for physicists". In: *Physics Reports* 810, pp. 66–68. ISSN: 0370-1573. DOI: 10.1016/j.physrep.2019.03.001. URL: http://dx.doi.org/10.1016/j.physrep.2019.03.001.

Neuberg, Leland Gerson (Aug. 2003). "CAUSALITY: MODELS, REASONING, AND INFERENCE, by Judea pearl, Cambridge university press, 2000". In: *Econ. Theory* 19.04.

Pearl, Judea (Dec. 1995). "Causal diagrams for empirical research". In: *Biometrika* 82.4, p. 669.

— (Sept. 2009). *Causality*. Cambridge: Cambridge University Press.

Runge, Jakob (July 2018). "Causal network reconstruction from time series: From theoretical assumptions to practical estimation". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 28, p. 075310. DOI: 10.1063/1.5025050.

Runge, Jakob et al. (Nov. 2019). "Detecting and quantifying causal associations in large nonlinear time series datasets". en. In: *Sci. Adv.* 5.11, eaau4996.

Schreiber, Thomas (July 2000). "Measuring Information Transfer". In: *Phys. Rev. Lett.* 85 (2), pp. 461–464. DOI: 10.1103/PhysRevLett.85.461. URL: https://link.aps.org/doi/10.1103/PhysRevLett.85.461.

Spirtes, Peter and Clark Glymour (1991). "An Algorithm for Fast Recovery of Sparse Causal Graphs". In: *Social Science Computer Review* 9.1, pp. 62–72. DOI: 10.1177/089443939100900106. eprint: https://doi.org/10.1177/089443939100900106. URL: https://doi.org/10.1177/089443939100900106.

Spirtes, Peter, Clark Glymour, and Richard Scheines (Jan. 1993). *Causation, Prediction, and Search*. Vol. 81. ISBN: 978-1-4612-7650-0. DOI: 10.1007/978-1-4612-2748-9.

Springer, Sebastian et al. (May 2024). "Unsupervised detection of large-scale weather patterns in the northern hemisphere via Markov State Modelling: from blockings to teleconnections". en. In: *Npj Clim. Atmos. Sci.* 7.1.

Verma, TS and Judea Pearl (2022). "Equivalence and Synthesis of Causal Models". In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. 1st ed. New York, NY, USA: Association for Computing Machinery, pp. 221–236. ISBN: 9781450395861. URL: https://doi.org/10.1145/3501714.3501732.

Vinh, Nguyen, Julien Epps, and James Bailey (Oct. 2010). "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance". In: *Journal of Machine Learning Research* 11, pp. 2837–2854.

Wild, Romina et al. (2024). "Automatic feature selection and weighting using Differentiable Information Imbalance". In: *Still to be published*.

This page intentionally left blank.

Appendices

This page intentionally left blank.

A

Graphical causal models

Since the dawn of scientific reasoning, people tried to draw connections among the different parts of the complex world they were observing. For the people of our century it is no more obscure at all that not all connections are truly relevant in practice. For example, pure and simple correlations might often arise even when no deeper connection is present. Then, to Pearl, 1995 we attribute the introduction of a distinction between $p(y|do(x))$ and $p(y|x)$. The operator "do" implies an explicit action on the system, setting a variable X to a given value x, which is in contrast with simply observing $X = x$ (the simple conditioning). Since in many cases performing direct experiments is just unethical or impossible it might look very hard the true estimation of these causal effects. The great realisation, though, was that two graphical criteria can be used to deduce estimate $p(y|do(x))$ only from observations (over assumptions on the shape of the graphical model). These criteria are known as *front-door criterion* and *back-door criterion*. In order to understand their definition, let's first introduce some useful concepts for causal graph theory (Spirtes et al., 1993).

Given a directed graph G, we call an *undirected path* a sequence of adjacent vertices in the graph. Similarly, a *directed path* is an ordered sequence of vertices in which any two consecutive elements are connected in the graph with a direct link from the first to the second.

A *descendant* of a vertex is any other vertex in G such that a directed path from the first to the second exist in G.

We also call a vertex V a *collider* on a path U if and only if there are two distinct edges on U containing V as an endpoint.

Finally, we say that the vertices X and Y are *d-separated* given a distinct subset Z of vertices in G if and only if no undirected path U between X and Y exists, such that

- all the colliders on U have a descendant in Z
- no other vertices in U are also in Z

Relating to the last definition we can finally state (Neuberg, 2003):

Criterion (Back-door criterion)**.** *Relative to the ordered pair of nodes (X,Y), Z satisfies the back-door criterion in G if*

- *none of the nodes in Z are descendants of X*
- *Z d-separates all the paths from X and Y which contains an arrow into X*

In this case we can compute

$$p(y|do(x)) = \sum_z p(y|x,z)p(z)$$

Criterion (Front-door criterion)**.** *Relative to the ordered pair of nodes (X,Y), Z satisfies the front-door criterion if*

- *Z intercepts all paths from X to Y*
- *no back-door path from X to Z exists*
- *all back-door paths from Z to Y are blocked by X*

In this case, if additionally $p(x,z) \neq 0$*, we can compute*

$$p(y|do(x)) = \sum_z p(z|x) \sum_{x'} p(y|x',z)p(x')$$

Further insights can be found, for example, in Pearl, 2009. These criteria give another reason for the interest in learning graphical models, as they can employed after recovering the causal structure of the system to *quantify* causal effects.

B

Categorical variables in the new framework

Given the change in the information imbalance described above, we present here also how this affect the results for categorical variables. In this part we also consider a correction to the Information Imbalance which makes $\Delta = 0$ in the most informative case.

Consider to compute the rank in a discrete space as

$$\tilde{r}_{i,j} = \begin{cases} 0 \text{ if i and j in the same class} \\ 1 \text{ oth.} \end{cases}$$

Then if the discrete space is the second one, you can just write the information imbalance as (notice $N = |S|$):

$$\Delta(d_A \to d_B) = \mathcal{N} \frac{1}{|S|} \sum_{i,j} \delta_{r^A_{i,j},1} \tilde{r}^B_{i,j}$$

and the normalization $\mathcal{N}$ should be fixed in such a way that the least informative case gives as a result 1 (in agreement with the usual case). In particular

$$\begin{aligned} p(\tilde{r}^B_{i,j} = 1 | r^A_{i,j} = 1) &= p(\tilde{r}^B_{i,j}) = \\ &= \sum_\alpha p(i \notin C_\alpha, j \in C_\alpha) \\ &= \sum_\alpha p(i \notin C_\alpha) p(j \in C_\alpha) \\ &= \sum_\alpha (1 - p^S_\alpha) p^{S'}_\alpha = 1 - \sum_\alpha p^S_\alpha p^{S'}_\alpha \end{aligned}$$

where the index α cycles over the different classes. Then the normalization is just $\mathcal{N} = (\mathbb{E}[\tilde{r}^B])^{-1} = (\sum_\alpha (1 - p_\alpha) p_\alpha)^{-1}$. Now, it is easy to see that $\sum_\alpha (1 - p_\alpha) p_\alpha$ can be easily estimated in an unbiased way since

$$\mathbb{E}\left[\sum_\alpha \frac{N^S_\alpha}{|S|} \frac{N^{S'}_\alpha}{|S'|} \right] = \sum_\alpha \mathbb{E}\left[\frac{N^S_\alpha}{|S|} \right] \mathbb{E}\left[\frac{N^{S'}_\alpha}{|S'|} \right] = \sum_\alpha p^S_\alpha p^{S'}_\alpha$$

It must be noticed, though, that this is not enough to get an unbiased estimate of the normalization. Indeed, consider the case in which at least two different classes are seen among S and S', then you can expand in series

$$\frac{1}{1-\sum_\alpha \frac{N_\alpha^S}{|S|}\frac{N_\alpha^{S'}}{|S'|}} = 1 + \sum_\alpha \frac{N_\alpha^S}{|S|}\frac{N_\alpha^{S'}}{|S'|} + \sum_\alpha \frac{N_\alpha^S}{|S|}\frac{N_\alpha^{S'}}{|S'|}\sum_\beta \frac{N_\beta^S}{|S|}\frac{N_\beta^{S'}}{|S'|} + \cdots$$

and after taking the expected value of this you get:

$$\mathbb{E}[\cdot] = 1 + \sum_\alpha p_\alpha^S p_\alpha^{S'} + \sum_{\alpha,\beta} \frac{\mathbb{E}[N_\alpha^S N_\beta^S]\mathbb{E}[N_\alpha^{S'} N_\beta^{S'}]}{|S|^2|S'|^2} + \cdots \tag{B.1}$$

which should be compared with

$$\frac{1}{1-\sum_\alpha p_\alpha^S p_\alpha^{S'}} = 1 + \sum_\alpha p_\alpha^S p_\alpha^{S'} + \sum_{\alpha,\beta} p_\alpha^S p_\beta^S p_\alpha^{S'} p_\beta^{S'} + \cdots \tag{B.2}$$

Considering that N_αs turns out to be sampled from a multinomial distribution, after recalling that

$$\begin{aligned} Cov[X_i, X_j] = -Np_ip_j \quad &\implies \quad \mathbb{E}[X_iX_j] = p_ip_j(N^2-N) \\ \mathbb{V}[X_i] = Np_i(1-p_i) \quad &\implies \quad \mathbb{E}[X_i^2] = Np_i + p_i^2(N^2-N) \end{aligned}$$

and exploiting the symmetry to write $\sum_{\alpha,\beta}\cdot = 2\sum_{\beta>\alpha}\cdot + \sum_{\alpha=\beta}\cdot$, we can see that the last term in equation B.1 becomes:

$$\begin{aligned} & 2\sum_{\beta>\alpha} p_\alpha^S p_\beta^S p_\alpha^{S'} p_\beta^{S'}\left[1 + \frac{1}{|S||S'|} - \frac{1}{|S|} - \frac{1}{|S'|}\right] + \\ + & \sum_\alpha (p_\alpha^S)^2 (p_\alpha^{S'})^2\left[1 + \frac{1}{|S||S'|} - \frac{1}{|S|} - \frac{1}{|S'|}\right] + \\ + & \sum_\alpha (p_\alpha^S)^2 p_\alpha^{S'}\left[\frac{1}{|S'|} - \frac{1}{|S'||S|}\right] + \sum_\alpha p_\alpha^S (p_\alpha^{S'})^2\left[\frac{1}{|S|} - \frac{1}{|S'||S|}\right] + \\ + & \sum_\alpha p_\alpha^S p_\alpha^{S'}\frac{1}{|S||S'|} \end{aligned}$$

where in red we highlighted the terms forming the last term in equation B.2. Meaning that

$$\frac{\mathcal{N}}{|S|} \approx \frac{|S'|}{\sum_\alpha \left(|S| - N_\alpha^S\right) N_\alpha^{S'}}$$

is at least a consistent estimator (at least up to the second order expansion). It should be underlined that, in principle, these considerations can be used only if $\frac{\mathcal{N}}{|S|}$ is estimated independently from the rest of the information imbalance. Otherwise, the average would not factorise, not allowing to compute on their own the terms in B.1.

If the normalisation is indeed estimated in an independent way, then it is also possible to estimate an error which will look a little be more complex than the one used for continuous variables. Specifically, we can use that for independent variables

$$\mathbb{V}[XY] = \mathbb{E}[X^2]\mathbb{V}[Y] + \mathbb{E}[Y]^2\mathbb{V}[X] = \mathbb{V}[X]\mathbb{V}[Y] + \mathbb{E}[Y]^2\mathbb{V}[X] + \mathbb{E}[X]^2\mathbb{V}[Y]$$

and estimate the different terms for $X = \mathcal{N}$ and Y the remaining terms in the information imbalance.

In the case in which you start from a discrete space and go to a continuous one, you can write:

$$\Delta(d_A \to d_B) = \mathcal{N}\frac{1}{|S|}\sum_{i,j}\delta_{r^A_{i,j},1}r^B_{i,j} = \frac{\mathcal{N}}{|S|}\sum_{\alpha^A}\sum_{i,j}\mathbb{I}[i \in C_{\alpha^A}, j \in C_{\alpha^A}]r^B_{i,j} \tag{B.3}$$

And this means that in the least informative case:

$$\mathbb{E}\left[\sum_{\alpha^A}\sum_{i,j}\mathbb{I}[i \in C_{\alpha^A}, j \in C_{\alpha^A}]r^B_{i,j}\right] =$$
$$\sum_{\alpha^A}\mathbb{E}_A\left[\sum_{i,j}\mathbb{I}[i \in C_{\alpha^A}]\mathbb{I}[j \in C_{\alpha^A}]\mathbb{E}_B\left[r^B_{i,j}\right]\right] =$$
$$\mathbb{E}_B\left[r^B_{i,j}\right]\sum_{\alpha^A}\mathbb{E}[N^S_{\alpha^A}]\mathbb{E}[N^{S'}_{\alpha^A}] =$$
$$\mathbb{E}_B\left[r^B_{i,j}\right]\sum_{\alpha^A}p^S_\alpha|S|p^{S'}_\alpha|S'| =$$

where $\mathbb{E}_B\left[r^B_{i,j}\right] = \frac{1}{|S'|}\sum_{i=0}^{|S'|-1} i = \frac{|S'|-1}{2}$ (ranks starting from 0).

It must be noticed, though, that when using B.3 for maximally informative variables you do not get 0 (even after making the ranks starting themselves from 0). In order to get effectively 0 a correction should be added:

$$\begin{aligned}\tilde{\Delta} &= \mathcal{N}\left[\frac{1}{|S|}\sum_\alpha\sum_{i\in C_\alpha, j\in C_\alpha} r^B_{i,j} - \tilde{C}\right]\\ \tilde{C} &= \frac{1}{|S|}\sum_\alpha\sum_{i\in C_\alpha}\left[\sum_{j\in C_\alpha} r^B_{i,j}\right]_{min} = \frac{1}{2|S|}\sum_\alpha N^S_\alpha(N^{S'}_\alpha - 1)N^{S'}_\alpha\end{aligned}$$

Which, combined with the term computed above allows to write

$$\begin{aligned}\mathcal{N}^{-1} &= \frac{1}{2|S|}\left[\sum_\alpha(|S'|-1)\sum_\alpha N^S_\alpha N^{S'}_\alpha - \sum_\alpha N^S_\alpha(N^{S'}_\alpha - 1)N^{S'}_\alpha\right]\\ &= \frac{1}{2|S|}\sum_\alpha N^S_\alpha N^{S'}_\alpha(|S'| - N^{S'}_\alpha)\end{aligned}$$

So finally

$$\tilde{\Delta} = \frac{2}{\sum_\alpha N^S_\alpha N^{S'}_\alpha(|S'| - N^{S'}_\alpha)}\left[\sum_\alpha\sum_{i,j\in C_\alpha} r^B_{i,j} - \frac{1}{2}\sum_\alpha N^S_\alpha(N^{S'}_\alpha - 1)N^{S'}_\alpha\right]$$

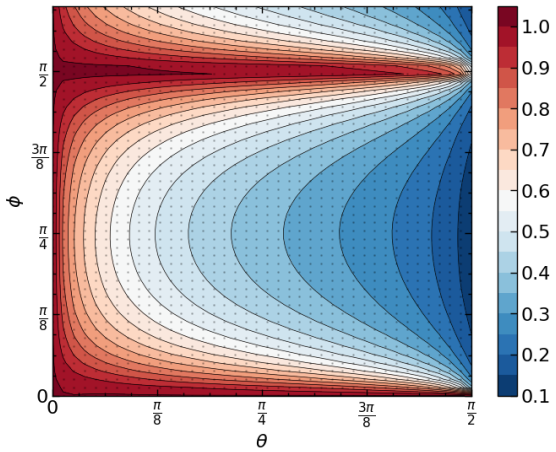

Figure B.1: *In this example we consider X, Y, Z sampled from a Gaussian distribution centered in the origin. As already done in the main body, we then consider polar coordinates for the plot. As expected, nor X nor Y alone ($\varphi = 0$ and $\varphi = \pi/2$ lines respectively) happens to be predictive on B. The minimum of the Information Imbalance is then found for $\theta = \pi/2$ (i.e. giving weight 0 to the variable Z, which is indeed not relevant to describe B).*

B.1 Synergical dependencies

Once introduced the concept of the Informtation Imbalance for discrete variables, it is possible to devise a system to analyse explicitly multi-body interactions (commonly described as *synergical dependencies* (Runge, 2018)). We consider three random variables $A = \{X, Y, Z\}$ independently sampled from a symmetric distribution and a binary variable $B = (X > 0) \oplus (Y > 0)$. We then compute the Information Imbalance $\Delta(d_{w \odot A} \to d_B)$. Because of the way B is constructed we have $X \perp\!\!\!\perp B$, $Y \perp\!\!\!\perp B$ but $\{X, Y\} \not\perp\!\!\!\perp B$. In Fig. B.1 we show using some Gaussian random variables that the Information Imbalance is able to retrieve these multi-body interactions. Indeed, the minimum is found for X and Y simulataneously different from zero.

C

Robustness test

In this section we show the effect of modifying some of the parameters of our algorithm with the coupled Rössler oscillators. In particular, in Fig. C.1 and Fig. C.2 we show the change in the weights when considering different time lags and strength for an observational noise. In Fig. C.3 and Fig. C.4 we first show the effect on the resulting graph of applying a threshold and then the statistical significance of the procedure in an alternative way to that presented in the main body. Finally, in Fig. C.5 we present the true positive rate against the false positive rate for the detection of direct links at each threshold setting.

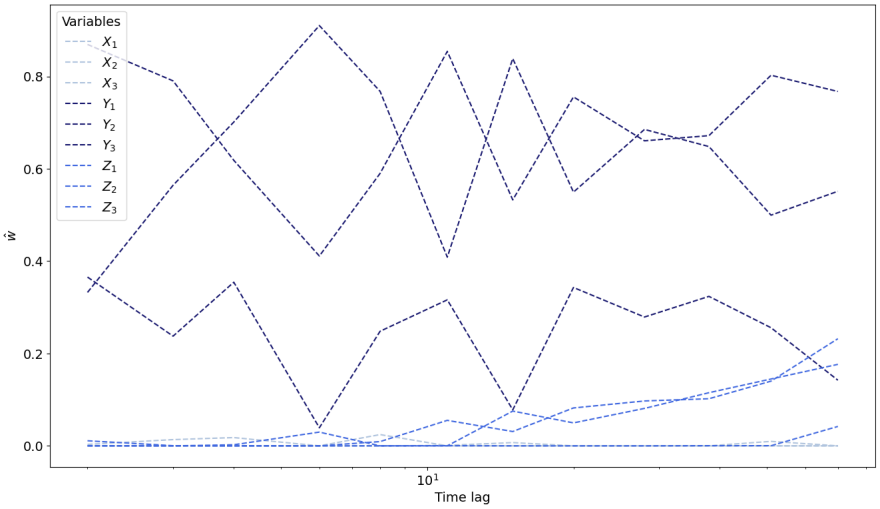

Figure C.1: *We consider the system of three coupled Rössler oscillators with $X \longrightarrow Z \longrightarrow Y$ and perform the minimization for the prediction of Y_3. Modifying the value of the time lag different variables show their relevance. With a small lag only the variables from the same Rössler are relevant, increasing the lag other variables starts being more relevant.*

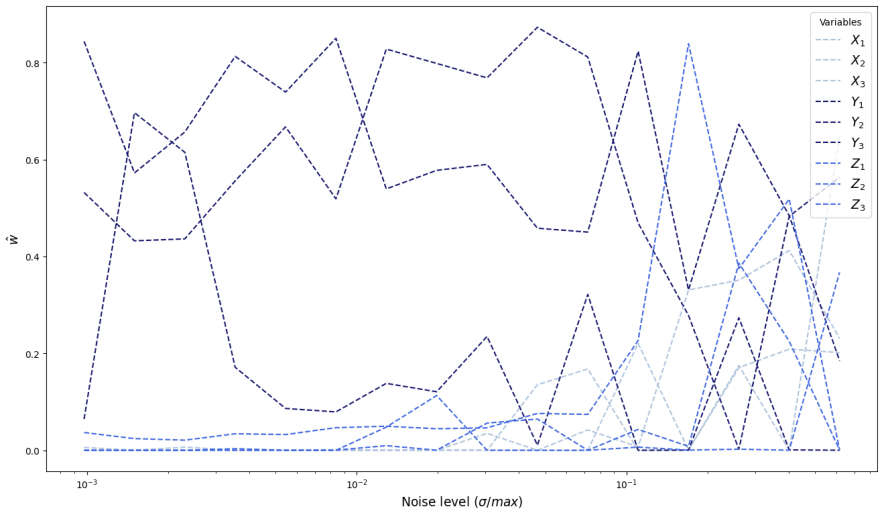

Figure C.2: *We consider again the same case seen in the previous plot. This time after selecting the datapoints, we add a gaussian noise whose's variance is progressively increased. We plot the results obtained the different weights found at the end of the minimization procedures. On the x axis we rescale the standard deviation of the noise by the maximum value found in the dataset.*

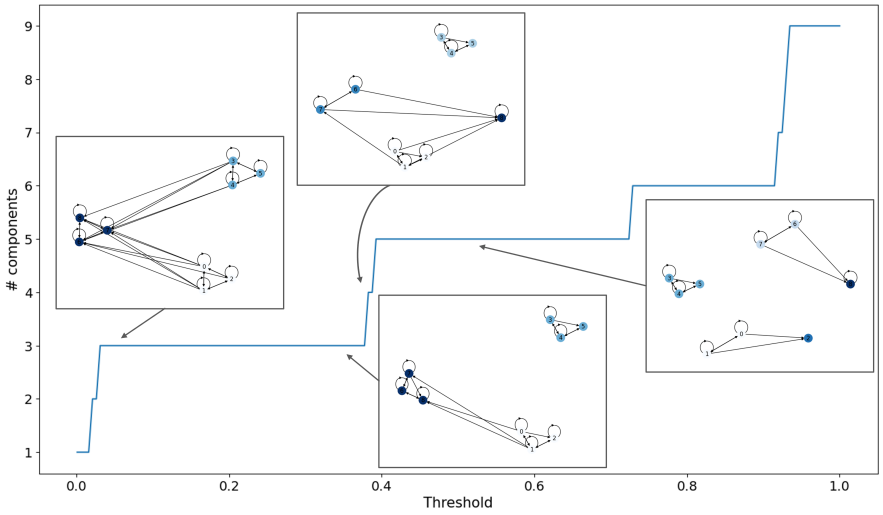

Figure C.3: *Varying a global threshold, different graph structures appears, starting from a fully connected graph to end with a graph in which each node is independent from the others. In the plot we show the results for 3 Rössler oscillators coupled as $X \longrightarrow Z \longleftarrow Y$.*

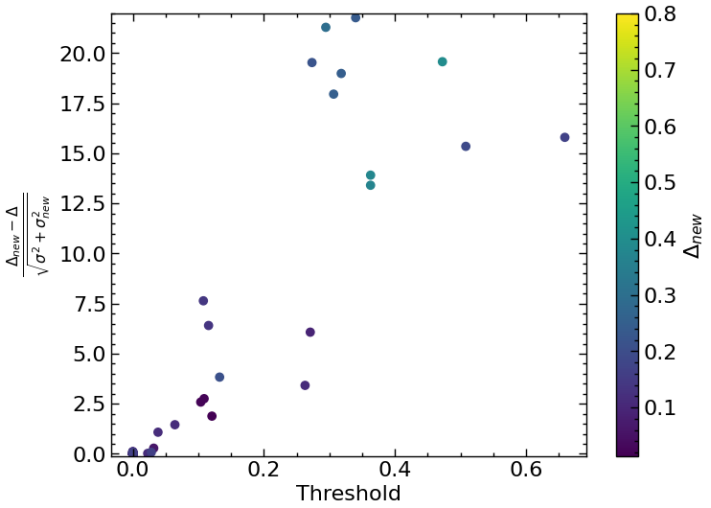

Figure C.4: *After minimizing* $\Delta\left(d_{\boldsymbol{w}\odot\boldsymbol{X}(t)} \to d_{X^i(t+\tau)}\right)$ *it is possible to apply a threshold on the weights and restart the minimization algorithm. It is then possible to compute* Δ *and* σ_Δ *with the initial weights and with the final ones* Δ_{new} *and* $\sigma_{\Delta_{new}}$*. One can then check if* Δ_{new} *is significantly different from* Δ*. We consider again 3 Rössler oscillators coupled as* $X \to Z \leftarrow Y$ *and* $\tau = 5$*, then for each variable* X^i *in the second space and any threshold which eliminates a different weight we follow the procedure above and plot the value of the Z-test. As expected, the highest the threshold the hardest it is to find an Information Imbalance comparable to the original one. Indeed, the variables with higher weights are necessary for the predictions.*

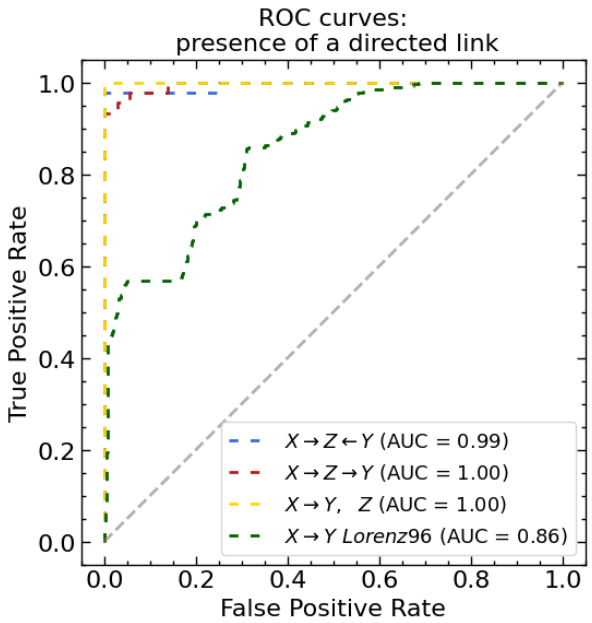

Figure C.5: *The receiver operating characteristic (ROC) curve is a common tool for graphically displaying the performance of a binary classifier as its threshold varies. A larger area under the curve (AUC) indicates better model performance in distinguishing between the two possible states. Although our method is not specifically designed to construct adjacency matrices (which can be seen as thresholded matrices with 0-1 entries based on the presence of links) for causal graphs, it still yields good results when we focus on small enough time lags (for which a ground truth can truly be established). As shown here, this is evident across all the cases presented in Fig. 3.3.*