Tomaso Aste and T. Di Matteo

Sparse causality network retrieval from short time series

Article (Published version)
(Refereed)

WILEY | Hindawi

Research Article

Sparse Causality Network Retrieval from Short Time Series

Tomaso Aste[1,2,3] and T. Di Matteo[1,2,3,4]

[1]*Department of Computer Science, UCL, London, UK*
[2]*UCL Centre for Blockchain Technologies, UCL, London, UK*
[3]*Systemic Risk Centre, London School of Economics and Political Sciences, London, UK*
[4]*Department of Mathematics, King's College London, London, UK*

Correspondence should be addressed to Tomaso Aste; t.aste@ucl.ac.uk

We investigate how efficiently a known underlying sparse causality structure of a simulated multivariate linear process can be retrieved from the analysis of time series of short lengths. Causality is quantified from conditional transfer entropy and the network is constructed by retaining only the statistically validated contributions. We compare results from three methodologies: two commonly used regularization methods, Glasso and ridge, and a newly introduced technique, LoGo, based on the combination of information filtering network and graphical modelling. For these three methodologies we explore the regions of time series lengths and model-parameters where a significant fraction of true causality links is retrieved. We conclude that when time series are short, with their lengths shorter than the number of variables, sparse models are better suited to uncover true causality links with LoGo retrieving the true causality network more accurately than Glasso and ridge.

1. Introduction

Establishing causal relations between variables from observation of their behaviour in time is central to scientific investigation and it is at the core of data-science where these causal relations are the basis for the construction of useful models and tools capable of prediction. The capability to predict (future) outcomes from the analytics of (past) input data is crucial in modelling and it should be the main property to take into consideration in model selection, when the validity and meaningfulness of a model is assessed. From a high-level perspective, we can say that the whole scientific method is constructed around a circular procedure consisting in observation, modelling, prediction, and testing. In such a procedure, the accuracy of prediction is used as a selection tool between models. In addition, the principle of parsimony favours the simplest model when two models have similar predictive power.

The scientific method is the rational process that, for the last 400 years, has mostly contributed to scientific discoveries, technological progress, and the advancement of human knowledge. Machine learning and data-science are nowadays pursuing the ambition to mechanize this discovery process by feeding machines with data and using different methodologies to build systems able to make models and predictions by themselves. However, the automatisation of this process requires to identify, without the help of human intuition, the relevant variables and the relations between these variables out of a large quantity of data. Predictive models are methodologies, systems, or equations which identify and make use of such relations between sets of variables in a way that the knowledge about a set of variables provides information about the values of the other set of variables. This problem is intrinsically high-dimensional with many input and output data. Any model that aims to explain the underlying system will involve a number of elements which must be of the order of magnitude of the number of relevant relations between the system's variables. In complex systems, such as financial markets or the brain, prediction is probabilistic in nature and modelling concerns inferring the probability of the values of a set of variables given the values of another set. This requires the estimation of the joint probability of all

variables in the system and, in complex systems, the number of variables with potential macroscopic effects on the whole system is very large. This poses a great challenge for the model construction/selection and its parameter estimation because the number of relations between variables scales with, at least, the square of the number of variables but, for a given fix observation window, the amount of information gathered from such variables scales, at most, linearly with the number of variables [1, 2].

For instance, a linear model for a system with p variables requires the estimation from observation of $p(p+1)/2$ parameters (the distinct elements of the covariance matrix). In order to estimate $\mathcal{O}(p^2)$ parameters one needs a comparable number of observations requiring time series of length $q \sim p$ or larger to gather a sufficient information content from a number of observations which scales as $p \times q \sim \mathcal{O}(p^2)$. However, the number of parameters in the model can be reduced by considering only $\mathcal{O}(p)$ out of the $\mathcal{O}(p^2)$ relations between the variables reducing in this way the required time series length to $\mathcal{O}(p)$. Such models with reduced numbers of parameters are referred to in the literature as sparse models. In this paper we consider two instances of linear sparse modelling: Glasso [3] which penalizes nonzero parameters by introducing a ℓ_1 norm penalization and LoGo [4] which reduces the inference network to an $\mathcal{O}(p)$ number of links selected by using information filtering networks [5–7]. The results from these two sparse models are compared with the ℓ_2 norm penalization (nonsparse) ridge model [8, 9].

This paper is an exploratory attempt to map the parameter regions of time series length, number of variables, penalization parameters, and kinds of models to define the boundaries where probabilistic models can be reasonably constructed from the analytics of observation data. In particular, we investigate empirically, by means of a linear autoregressive model with sparse inference structure, the true causality link retrieval performances in the region of short time series and large number of variables which is the most critical region—and the most interesting—in many practical cases. Causality is defined in information theoretic sense as a significant reduction on uncertainty over the present values of a given variable provided by the knowledge of the past values of another variable obtained in excess to the knowledge provided by the past of the variable itself and—in the conditional case—the past of all other variables [10]. We measure such information by using transfer entropy and, within the present linear modelling, this coincides with the concept of Granger causality and conditional Granger causality [11]. The use of transfer entropy has the advantage of being a concept directly extensible to nonlinear modelling. However, nonlinearity is not tackled within the present paper. Linear models with multivariate normal distributions have the unique advantage that causality and partial correlations are directly linked, largely simplifying the computation of transfer entropy, and directly mapping the problem into the sparse inverse covariance problem [3, 4].

Results are reported for artificially generated time series from an autoregressive model of $p = 100$ variables and time series lengths q between 10 and 20,000 data points. Robustness of the results has been verified over a wider range

of p from 20 to 200 variables. Our results demonstrate that sparse models are superior in retrieving the true causality structure for short time series. Interestingly, this is despite considerable inaccuracies in the inference network of these sparse models. We indeed observe that statistical validation of causality is crucial in identifying the true causal links, and this identification is highly enhanced in sparse models.

The paper is structured as follows. In Section 2 we briefly review the basic concepts of mutual information and conditional transfer entropy and their estimation from data that will then be used in the rest of the paper. We also introduce the concepts of sparse inverse covariance, inference network and causality networks. Section 3 concerns the retrieval of causality network from the computation and statistical validation of conditional transfer entropy. Results are reported in Section 4 where the retrieval of the true causality network from the analytics of time series from an autoregressive process of $p = 100$ variables is discussed. Conclusions and perspectives are given in Section 5.

2. Estimation of Conditional Transfer Entropy from Data

In this paper causality is quantified by means of statistically validated transfer entropy. Transfer entropy $T(\mathbf{Z_i} \rightarrow \mathbf{Z_j})$ quantifies the amount of uncertainty on a random variable, $\mathbf{Z_j}$, explained by the past of another variable, $\mathbf{Z_i}$, conditioned to the knowledge about the past of $\mathbf{Z_j}$ itself. Conditional transfer entropy, $T(\mathbf{Z_i} \rightarrow \mathbf{Z_j} \mid \mathbf{W})$, includes an extra condition also to a set variables $\mathbf{W}$. These quantities are introduced in detail in Appendix A (see also [11–13]). Let us here just report the main expression for the conditional transfer entropy that we shall use in this paper:

$$
\begin{aligned}
T\left(\mathbf{Z_i} \longrightarrow \mathbf{Z_j} \mid \mathbf{W}\right) = {} & H\left(\mathbf{Z_{j,t}} \mid \left\{\mathbf{Z_{j,t}^{lag}}, \mathbf{W}_t\right\}\right) \\
& - H\left(\mathbf{Z_{j,t}} \mid \left\{\mathbf{Z_{i,t}^{lag}}, \mathbf{Z_{j,t}^{lag}}, \mathbf{W}_t\right\}\right),
\end{aligned}
\tag{1}
$$

where $H(\cdot \mid \cdot)$ is the conditional entropy and $\mathbf{Z_{j,t}}$ is a random variable at time t, whereas $\mathbf{Z_{i,t}^{lag}} = \{\mathbf{Z_{i,t-1}}, \ldots, \mathbf{Z_{i,t-\tau}}\}$ is the lagged set of random variable "$\mathbf{i}$" considering previous times $t-1 \cdots t-\tau$ and $\mathbf{W}_t$ are all other variables and their lags (see Appendix A, (A.5)).

In this paper we use Shannon entropy and restrict to linear modelling with multivariate normal setting (see Appendix B). In this context the conditional transfer entropy can be expressed in terms of the determinants of conditional covariances $\det(\mathbf{\Sigma}(\cdot \mid \cdot))$ (see (B.5) in Appendix B):

$$
\begin{aligned}
& T\left(\mathbf{Z_i} \longrightarrow \mathbf{Z_j} \mid \mathbf{W}\right) \\
& = \frac{1}{2} \log \det\left(\mathbf{\Sigma}\left(\mathbf{Z_{j,t}} \mid \left\{\mathbf{Z_{j,t}^{lag}}, \mathbf{W}_t\right\}\right)\right) \\
& \quad - \frac{1}{2} \log \det\left(\mathbf{\Sigma}\left(\mathbf{Z_{j,t}} \mid \left\{\mathbf{Z_{j,t}^{lag}}, \mathbf{Z_{i,t}^{lag}}, \mathbf{W}_t\right\}\right)\right).
\end{aligned}
\tag{2}
$$

Conditional covariances can be conveniently computed in terms of the inverse covariance of the whole set of variables

$\mathbf{Z}_t = \{\mathbf{Z}_{k,t}, \mathbf{Z}_{k,t-1}, \ldots, \mathbf{Z}_{k,t-\tau}\}_{k=1}^{p} \in \mathbb{R}^{p \times (\tau+1)}$ (see Appendix C). Such inverse covariance matrix, $\mathbf{J}$, represents the structure of conditional dependencies among all couples of variables in the system and their lags. Each subpart of $\mathbf{J}$ is associated with the conditional covariances of the variables in that part with respect to all others. In terms of $\mathbf{J}$, the expression for the conditional transfer entropy becomes

$$
\begin{aligned}
T\left(\mathbf{Z_i} \longrightarrow \mathbf{Z_j} \mid \mathbf{W}\right) = -\frac{1}{2} \\
\cdot \log \det\left(\mathbf{J_{1,1}} - \mathbf{J_{1,2}}\left(\mathbf{J_{2,2}}\right)^{-1}\mathbf{J_{2,1}}\right) + \frac{1}{2}\log \det\left(\mathbf{J_{1,1}}\right),
\end{aligned}
\tag{3}
$$

where the indices "$\mathbf{1}$" and "$\mathbf{2}$" refer to submatrices of $\mathbf{J}$, respectively, associated with the variables $\mathbf{Z}_{j,t}$ and $\mathbf{Z}_{i,t}^{\text{lag}}$.

2.1. Causality and Inference Networks.

The inverse covariance $\mathbf{J}$, also known as precision matrix, represents the structure of conditional dependencies. If we interpret the structure of $\mathbf{J}$ as a network, where nodes are the variables and nonzero entries correspond to edges of the network, then we shall see that any two subsets of nodes that are not directly connected by one or more edges are conditionally independent. Condition is with respect to all other variables.

Links between variables at different lags are associated with causality with direction going from larger to smaller lags. The network becomes therefore a directed graph. In such a network entropies can be associated with nodes, conditional mutual information can be associated with edges between variables with the same lag, and conditional transfer entropy can be associated with edges between variables with different lags. A nice property of this mapping of information measures with directed networks is that there is a simple way to aggregate information which is directly associated with topological properties of the network. Entropy, mutual information, and transfer entropies can be defined for any aggregated subset of nodes with their values directly associated with the presence, direction, and weight of network edges between these subparts.

Nonzero transfer entropies indicating, for instance, variable $\mathbf{i}$ causing variable $\mathbf{j}$ are associated with some nonzero entries in the inverse covariance matrix $\mathbf{J}$ between lagged variables $\mathbf{i}$ (i.e., $\mathbf{Z}_{i,t-\tau}$, with $\tau > 0$) and variable $\mathbf{j}$ (i.e., $\mathbf{Z}_{j,t}$). In linear models these nonzero entries define the estimated *inference network*. However, not all edges in this inference network correspond to transfer entropies that are significantly different from zero. To extract the structure of the *causality network* we shall retain only the edges in the inference network which correspond to statistically validated transfer entropies.

Conditioning eliminates the effect of the other variables retaining only the exclusive contribution from the two variables in consideration. This should provide estimations of transfer entropy that are less affected by spurious effects from other variables. On the other hand, conditioning in itself can introduce spurious effects; indeed two independent variables can become dependent due to conditioning [13]. In this paper we explore two extreme conditioning cases: (i) conditioned to all other variables and their lags; (ii) unconditioned.

In principle, one would like to identify the maximal value of $T(\mathbf{Z_i} \rightarrow \mathbf{Z_j} \mid \mathbf{W})$ over all lags and all possible conditionings $\mathbf{W}$. However, the use of multiple lags and conditionings increases the dimensionality of the problem making estimation of transfer entropy very hard especially when only a limited amount of measurements is available (i.e., short time series). This is because the calculation of the conditional covariance requires the estimation of the inverse covariance of the whole set of variables and such an estimation is strongly affected by noise and uncertainty. Therefore, a standard approach is to reduce the number of variables and lags to keep dimensionality low and estimate conditional covariances with appropriate penalizers [3, 8, 9, 14]. An alternative approach is to invert the covariance matrix only locally on low dimensional subsets of variables selected by using information filtering networks [5–7] and then reconstruct the global inversion by means of the LoGo approach [4]. Let us here briefly account for these two approaches.

2.2. Penalized Inversions.

The estimate of the inverse covariance is a challenging task to which a large body of literature has been dedicated [14]. From an intuitive perspective, one can say that the problem lies in the fact that uncertainty is associated with nearly zero eigenvalues of the covariance matrix. Variations in these small eigenvalues have relatively small effects on the entries of the covariance matrix itself but have major effects on the estimation of its inverse. Indeed small fluctuations of small values can yield to unbounded contributions to the inverse. A way to cure such near-singular matrices is by adding finite positive terms to the diagonal which move the eigenvalues away from zero: $\widehat{\mathbf{J}} = ((1 - \gamma)\mathbf{S} + \gamma\mathbf{I}_N)^{-1}$, where $\mathbf{S} = \text{cov}(\mathbf{Z})$ is the covariance matrix of the set of variables $\mathbf{Z} \in \mathbb{R}^N$ estimated from data and $\mathbf{I}_N \in \mathbb{R}^{N \times N}$ is the identity matrix (where $N = p \times (\tau + 1)$; see later). This is what is performed in the so-called ridge regression [9], also known as shrinkage mean-square-error estimator [15] or Tikhonov regularization [8]. The effect of the additional positive diagonal elements is equivalent to compute the inverse covariance which maximizes the log-likelihood: $\log \det(\widehat{\mathbf{J}}) - \text{tr}(\mathbf{S}\widehat{\mathbf{J}}) - \gamma \| \widehat{\mathbf{J}} \|_2$, where the last term penalizes large off-diagonal coefficients in the inverse covariance with a ℓ_2 norm penalization [16]. The regularizer parameter γ tunes the strength of this penalization. This regularization is very simple and effective. However, with this method insignificant elements in the precision matrix are penalized toward small values but they are never set to zero. By using instead ℓ_1 norm penalization $\log \det(\widehat{\mathbf{J}}) - \text{tr}(\mathbf{S}\widehat{\mathbf{J}}) - \gamma \| \widehat{\mathbf{J}} \|_1$, insignificant elements are forced to zero leading to a sparse inverse covariance. This is the so-called lasso regularization [3, 14, 17]. The advantage of a sparse inverse covariance consists in the provision of a network representing a conditional dependency structure. Indeed, let us recall that in linear models zero entries in the inverse covariance are associated with couples of nonconditionally dependent variables.

2.3. Information Filtering Network Approach: LoGo.

An alternative approach to obtain sparse inverse covariance is by using information filtering networks generated by keeping

the elements that contribute most to the covariance by means of a greedy process. This approach, named LoGo, proceeds by first constructing a chordal information filtering graph such as a Maximum Spanning Tree (MST) [18, 19] or a Triangulated Maximally Filtered Graph (TMFG) [7]. These graphs are built by retaining edges that maximally contribute to a given gain function which, in this case, is the log-likelihood or—more simply—the sum of the squared correlation coefficients [5–7]. Then, this chordal structure is interpreted as the inference structure of the joint probability distribution function with nonzero conditional dependency only between variables that are directly connected by an edge. On this structure the sparse inverse covariance is computed in such a way to preserve the values of the correlation coefficients between couples of variables that are directly connected with an information filtering graph edge. The main advantage of this approach is that inversion is performed at local level on small subsets of variables and then the global inverse is reconstructed by joining the local parts through the information filtering network. Because of this Local-Global construction this method is named LoGo. It has been shown that LoGo method yields to statistically significant sparse precision matrices that outperform the ones with the same sparsity computed with lasso method [4].

3. Causality Network Retrieval

3.1. Simulated Multivariate Autoregressive Linear Process.

In order to be able to test if causality measures can retrieve the true causality network in the underlying process, we generated artificial multivariate normal time series with known sparse causality structure by using the following autoregressive multivariate linear process [20]:

$$\mathbf{Z}_t = \sum_{\lambda=1}^{\tau} \mathbf{A}_\lambda \mathbf{Z}_{t-\lambda} + \mathbf{U}_t, \tag{4}$$

where $\mathbf{A}_\lambda \in \mathbb{R}^{p \times p}$ are matrices with random entries drawn from a normal distribution. The matrices are made upper diagonal (diagonal included) by putting to zero all lower diagonal coefficients and made sparse by keeping only a $\mathcal{O}(p)$ total number of entries different from zero in the upper and diagonal part. $\mathbf{U}_t \in \mathbb{R}^p$ are random normally distributed uncorrelated variables. This process produces autocorrelated, cross-correlated, and causally dependent time series. We chose it because it is among the simplest processes that can generate this kind of structured datasets. The dependency and causality structure is determined by the nonzero entries of the matrices $\mathbf{A}_\lambda$. The upper-triangular structure of these matrices simplifies the causality structure eliminating causality cycles. Their sparsity reduces dependency and causality interactions among variables. The process is made autoregressive and stationary by keeping the eigenvalues of $\mathbf{A}_\lambda$ all smaller than one in absolute value. For the tests we used $\tau = 5$, $p = 100$ and sparsity is enforced to have a number of links approximately equal to p. We reconstructed the network from time series of different lengths q between 5 and 20,000 points. To test statistical reliability the process was repeated 100 times with every time a different set of randomly generated matrices

$\mathbf{A}_\lambda$. We verify that the results are robust and consistent by varying sample sizes from $p = 20$ to 200, by changing sparsity with number of links from $0.5p$ to $5p$ and for τ from 1 to 10. We verified that the presence of isolated nodes or highly connected hub nodes does not affect results significantly.

3.2. Causality and Inference Network Retrieval.

We tested the agreement between the causality structure of the underlying process and the one inferred from the analysis of p time series of different lengths q, $\mathbf{Z}_t \in \mathbb{R}^p$ with $t = 1 \cdots q$, generated by using (4) We have p different variables and τ lags. The dimensionality of the problem is therefore $N = p \times (\tau + 1)$ variables at all lags including zero.

To estimate the inference and causality networks we started by computing the inverse covariance, $\mathbf{J} \in \mathbb{R}^{N \times N}$, for all variables at all lags $\mathbf{Z} \in \mathbb{R}^{N \times q}$ by using the following three different estimation methods:

(1) ℓ_1 norm penalization (Glasso [14]);

(2) ℓ_2 norm penalization (ridge [8]);

(3) information filtering network (LoGo [4]).

We retrieved the inference network by looking at all couples of variables, with indices $\mathbf{i} \in [1, \ldots, p]$ and $\mathbf{j} \in [1, \ldots, p]$, which have nonzero entries in the inverse covariance matrix $\mathbf{J}$ between the lagged set of $\mathbf{j}$ and the nonlagged $\mathbf{i}$. Clearly, for the ridge method the result is a complete graph but for the Glasso and LoGo the results are sparse networks with edges corresponding to nonzero conditional transfer entropies between variables $\mathbf{i}$ and $\mathbf{j}$. For the LoGo calculation we make use of the regularizer parameter as a local shrinkage factor to improve the local inversion of the covariance of the 4-cliques and triangular separators (see [4]).

We then estimated transfer entropy between couples of variables, $\mathbf{i} \rightarrow \mathbf{j}$ conditioned to all other variables in the system. This is obtained by estimating of the inverse covariance matrix (indicated with an "hat" symbol) by using (C.7) (see Appendix C.2) with

$$\begin{aligned} \mathbf{Z_1} &= \mathbf{Z}_{\mathbf{j},t} \\ \mathbf{Z_2} &= \left\{ \mathbf{Z}_{\mathbf{i},t-1} \cdots \mathbf{Z}_{\mathbf{i},t-\tau} \right\} \\ \mathbf{Z_3} &= \left\{ \mathbf{Z}_{\mathbf{j},t-1} \cdots \mathbf{Z}_{\mathbf{j},t-\tau}, \mathbf{W} \right\}, \end{aligned} \tag{5}$$

with $\mathbf{W}$ a conditioning to all variables $\mathbf{Z}$ except $\mathbf{Z_1}, \mathbf{Z_2}$, and $\{ \mathbf{Z}_{\mathbf{j},t-1} \cdots \mathbf{Z}_{\mathbf{j},t-\tau} \}$. The result is a $p \times p$ matrix of conditional transfer entropies $T(\mathbf{Z}_{\mathbf{i},t} \rightarrow \mathbf{Z}_{\mathbf{j},t})$. Finally, to retrieve the causality network we retained the network of statistically validated conditional transfer entropies only. Statistical validation was performed as follows.

3.3. Statistical Validation of Causality.

Statistical validation has been performed from likelihood ratio statistical test. Indeed, entropy and likelihood are intimately related: entropy measures uncertainty and likelihood measures the reduction in uncertainty provided by the model. Specifically, the Shannon entropy associated with a set of random variables, $\mathbf{Z_i}$, with probability distribution $p(\mathbf{Z_i})$ is $H(\mathbf{Z_i}) = -\mathbb{E}[\log p(\mathbf{Z_i})]$

(see (B.1)) whereas the log-likelihood for the model $\widehat{p}(\mathbf{Z_i})$ associated with a set of independent observations $\widehat{\mathbf{Z}}_{i,t}$ with $t = 1 \cdots q$ is $\log \mathscr{L}(\widehat{\mathbf{Z}}_\mathbf{i}) = \sum_{t=1}^{q} \log \widehat{p}(\widehat{\mathbf{Z}}_{i,t})$ which can be written as $\log \mathscr{L}(\widehat{\mathbf{Z}}_\mathbf{i}) = q\mathbb{E}_{\widehat{p}}[\log \widehat{p}(\mathbf{Z_i})]$. Note that q is the total available number of observations which, in practice, is the length of the time series minus the maximum number of lags. It is evident from these expressions that entropy and the log-likelihood are strictly related though this link might be nontrivial. In the case of linear modelling this connection is quite evident because the entropy estimate is $H = (1/2)(-\log|\widehat{\mathbf{J}}| + p\log(2\pi) + p)$ and the log-likelihood is $\log \mathscr{L} = (q/2)(\log|\widehat{\mathbf{J}}| - \mathrm{Tr}(\widehat{\mathbf{\Sigma}}\widehat{\mathbf{J}}) - p\log(2\pi))$. For the three models we study in this paper we have $\mathrm{Tr}(\widehat{\mathbf{\Sigma}}\widehat{\mathbf{J}}) = p$ and therefore the log-likelihood is equal to q times the opposite of the entropy estimate. Transfer entropy and conditional transfer entropy are differences between two entropies: the one of a set of variables conditioned to their own past minus the one conditioned also to the past of another variable. This, in turns, is the difference of the unitary log-likelihood of two models and therefore it is the logarithm of a likelihood ratio. As Wilks pointed out [21, 22] the null distribution of such model is asymptotically quite universal. Following the likelihood ratio formalism, we have $\lambda = qT$ and the probability of observing a transfer entropy larger than T, estimated under null hypothesis, is given by $p_v \sim 1 - \chi_c^2(rqT, d)$ with $r \simeq 2$ and χ_c^2 the chi-square the cumulative distribution function with d degrees of freedom which are the difference between the number of parameters in the two models. In our case the two models have, respectively, $\tau(p_j^2 + 1)$ and $\tau(p_j^2 + 1) + \tau(p_j p_i)$ parameters.

3.4. Statistical Validation of the Network.
The procedures described in the previous two subsections produce the inference network and causality network. Such networks are then compared with the known network of true causalities in the underlying process which is defined by the nonzero elements in the matrices A_λ (see (4)). The overlapping between the retrieved links in the inference or causality networks with the ones in the true network underlying the process is an indication of a discovery of a true causality relation. However some discoveries can be obtained just by chance or some methodologies might discover more links only because they produce denser networks. We therefore tested the hypothesis that the matching links in the retrieved networks are not obtained just by chances by computing the null-hypothesis probability to obtain the same or a larger number of matches randomly. Such probability is given by the conjugate cumulative hypergeometric distribution for a number equal or larger than TP of "true positive" matching causality links between an inferred network of n links and a process network of K true causality links, from a population of $p^2 - p$ possible links:

$$P(X \geq \mathrm{TP} \mid n, K, p) = 1 - \sum_{k=0}^{\mathrm{TP}-1} \frac{\binom{K}{k}\binom{p^2-p-K}{n-k}}{\binom{p^2-p}{n}}. \quad (6)$$

Small values of P indicate that the retrieved TP links out of K are unlikely to be found by randomly picking n edges from

$p^2 - p$ possibilities. Note that in the confusion matrix notation [23] we have $n = \mathrm{TP} + \mathrm{FP}$ and $K = \mathrm{TP} + \mathrm{FN}$ with TP number of true positives, FP number of false positives, FN number of false negatives, and TN number of true negatives. The total number of "negatives" (unlinked couples of vertices) in the true model is instead $m = \mathrm{FP} + \mathrm{TN}$.

4. Results

4.1. Computation and Validation of Conditional Transfer Entropies.
By using (4) we generated 100 multivariate autoregressive processes with known causality structures. We here report results for $p = 100$ but analogous outcomes were observed for dimensionalities between $p = 20$ and 200 variables. Conditional transfer entropies between all couples of variables, conditioned to all other variables in the system, were computed by estimating the inverse covariances by using tree methodologies, ridge, Glasso, and LoGo and applying (3). Conditional transfer entropies were statistically validated with respect to null hypothesis (no causality) at $p_v = 1\%$ p value. Results for Bonferroni adjusted p value at 1% (i.e., $p_v = 0.01/(p^2 - p) \sim 10^{-6}$ for $p = 100$) are reported in Appendix E. We also tested other values of p_v from 10^{-8} to 0.1 obtaining consistent results. We observe that small p_v reduce the number of validated causality links but increase the chance that these links match with the true network in the process. Conversely large values of p_v increase the numbers of mismatched links but also of the true links discoveries. Let us note that here we use p_v as a thresholding criteria and we are not claiming any evidence of statistical significance of the causality. We assess the goodness of this choice a posteriori by comparing the resulting causality network with the known causality network of the process.

4.2. Statistical Significance of the Recovered Causality Network.
Results for the contour frontiers of significant causality links for the three models are reported in Figure 1 for a range of time series with lengths q between 10 and 20,000 and regularizer parameters γ between 10^{-8} and 0.5. Statistical significance is computed by using (6) and results are reported for both $P < 0.05$ and $P < 10^{-8}$ (continuous and dotted lines respectively). As one can see, the overall behaviours for the three methodologies are little affected by the threshold on P. We observe that LoGo significance region extends well beyond the Glasso and ridge regions.

The value of the regularizer parameter γ affects the results for the three models in a different way. Glasso has a region in the plane γ-p/q where it has best performances (in this case it appears to be around $\gamma \simeq 0.1$ and $p/q \simeq 2.5$). Ridge appears instead to be little affected with mostly constant performances across the range of γ. LoGo has best performances for small, even infinitesimal, values of γ. Indeed, different from Glasso in this case γ does not control sparsity but instead acts as local shrinkage parameter. Very small values can be useful in some particular cases to reduce the effect of noise but large values have only the effect to reduce information.

4.3. Causality Links Retrieval.
Once identified the parameter regions where the retrieved causality links are statistically
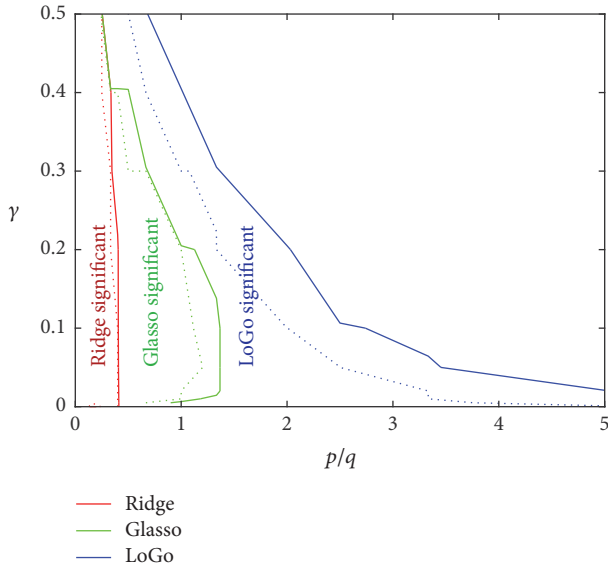
FIGURE 1: *Regions in the p/q-γ space where causality networks for the three models are statistically significant*. The significance regions are all at the left of the corresponding lines. Tick line reports the boundary $P < 0.05$ (see (6)) and dotted lines indicate $P < 10^{-8}$ significance levels (P is averaged over 100 processes). The plots refer to $p = 100$ and report the region where the causality networks are all significant for 100 processes.

significant, we also measured the fraction of true links retrieved. Indeed, given that the true underlying causality network is sparse, one could do significantly better than random by discovering only a few true positives. Instead, from any practical perspective we aim to discover a significant fraction of the edges. Figure 2 shows that the fraction of causality links correctly discovered (true positive, TP) with respect to the total number of causality links in the process (n) is indeed large reaching values above 50%. This is the so-called true positive rate or sensitivity, which takes values between 0 (no links discovered) and 1 (all links discovered). Reported values are averages over 100 processes. We observe that the region with discovering of 10% or more true causality links greatly overlaps with the statistical validity region of Figure 1.

We note that when the observation time becomes long, $p/q \lessapprox 0.25$, ridge discovery rate becomes larger than LoGo. However, statistical significance is still inferior to LoGo, indeed the ridge network becomes dense when q increases and the larger discovery rate of true causality links is also accompanied by a larger rate of false links incorrectly identified (false positive FP).

The fraction of false positives with respect to the total number of causality links in the process (n) are reported in Table 1 together with the true positive rate for comparison. This number can reach values larger than one because the process is sparse and there are much more possibilities to randomly chose false links than true links. Note that this is not the false positive rate, which instead is FP/m, and cannot be larger than one. Consistent with Figure 1 we observe that, for short time series, up to $p/q \sim 0.5$, the sparse
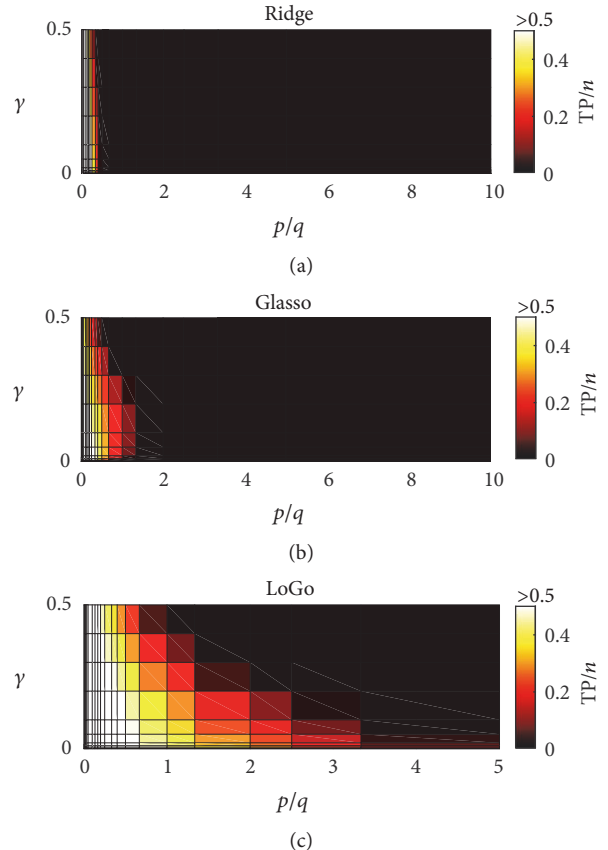

FIGURE 2: *True positive rate: fraction of retrieved true causality links (TP) with respect to the total number of links in the process (n)*. The three panels refer to ridges, Glasso, and LoGo ((a), (b), and (c)). Data are average fractions over 100 processes.

models have better capability to identify true causality links and to discard the false ones with LoGo being superior to Glasso. Remarkably, LoGo can identify a significant fraction of causality links already from time series with lengths of 30 data points only. p value significance, reported in the table with one or two stars, indicates when all values of $P(X \geq$ TP | $n, K, p)$ from (6) for all 100 processes have, respectively, $P < 0.05$ or $P < 10^{-8}$. Again we observe that LoGo discovery rate region extends well beyond the Glasso and ridge regions.

4.4. Inference Network. We have so far empirically demonstrated that a significant part of the true causality network can be retrieved from the statistically validated network of conditional transfer entropies. Results depend on the choice of the threshold value of p_v at which null hypothesis is rejected. We observed that lower p_v are associated with network with fewer true positives but also fewer false positives and conversely larger p_v yield to causality networks with larger true positives but also larger false positives. Let us here report on the extreme case of the inference network which contains all causality channels with no validation. For the ridge model this network is the complete graph with all variables connected to each other. Instead, for Glasso and LoGo the inference network is sparse.

TABLE 1: *Causality network validation*. Comparison between fraction of true positive (TP/n) and fraction of false positive (FP/n), statistically validated, causality links for the three models, and different time-series lengths. The table reports only the case for the parameter $\gamma = 0.1$. Statistical validation of conditional transfer entropy is at $p_v = 1\%$ p value. Note that LoGo can perform better than reported in this table for smaller values of γ (see Figures 1 and 2).

q	10	20	30	50	200	300	1000	20000
Ridge TP/n	0.00	0.00	0.00	0.00	0.23**	0.49**	0.76**	0.93**
Ridge FP/n	0.00	0.00	0.00	0.00	0.00	0.10	0.65	1.06
Glasso TP/n	0.00	0.00	0.00	0.13**	0.48**	0.53**	0.62**	0.74**
Glasso FP/n	0.00	0.00	0.00	0.00	0.06	0.10	0.23	0.54
LoGo TP/n	0.00	0.08*	0.21**	0.37**	0.61**	0.65**	0.75**	0.90**
LoGo FP/n	0.00	0.00	0.00	0.01	0.06	0.08	0.15	0.34

* $P < 0.05$; ** $P < 10^{-8}$.

TABLE 2: *Inference network validation: comparison between fraction of true positive (TP/n) and fraction of false positive (FP/n)*. Data for ridge are only for comparison because it is a complete graph with all links present. The table reports only the case for the parameter $\gamma = 0.1$.

q	10	20	30	50	200	300	1000	20000
Ridge TP/n	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Ridge FP/n	97.84	97.84	97.84	97.84	97.84	97.84	97.84	97.84
Glasso TP/n	0.61*	0.74*	0.79*	0.85*	0.87**	0.84**	0.80**	0.80**
Glasso FP/n	28.39	38.11	45.79	53.58	40.61	26.60	1.54	0.92
LoGo TP/n	0.31*	0.50**	0.58**	0.63**	0.75**	0.78**	0.85**	0.93**
LoGo FP/n	4.53	4.27	4.18	4.03	3.72	3.63	3.44	3.21

* $P < 0.05$; ** $P < 10^{-8}$.

Results are summarized in Table 2. In terms of true positive rate we first notice that they are all larger than the ones in Table 1. Indeed, the network of statistically validated conditional transfer entropies is a subnetwork of the inference network. On the other hand we notice that the false positive fraction is much larger than the ones in Table 2. Ridge network has a fraction of 1 because, in this case, the inference network is the complete graph.

Glasso also contains a very large number of false positives reaching even 55 times the number of links in the true network and getting to lower fractions only from long time series with $q > 1000$. These numbers also indicate that Glasso networks are not sparse. LoGo has a sparser and more significant inference network with smaller fractions of false positives which stay below $5n$, which is anyway a large number of misclassifications. Nonetheless, we observe that, despite such large fractions of FP, the discovered true positives are statistically significant.

4.5. Unconditioned Transfer Entropy Network. We last tested whether conditioning to the past of all other variables gives better causality network retrievals than the unconditioned case. Here, transfer entropy, $T(\mathbf{Z_i} \rightarrow \mathbf{Z_j})$, is computed by using (3) with $\mathbf{W} = \emptyset$, the empty set. For the ridge case this unconditioned transfer entropy depends only from the time series, $\mathbf{Z}_{i,t}, \{\mathbf{Z}_{i,t-1}, \ldots, \mathbf{Z}_{i,t-\tau}\}$ and $\{\mathbf{Z}_{j,t-1}, \ldots, \mathbf{Z}_{j,t-\tau}\}$ (with $\tau = 5$ in this case). Glasso and LoGo cases are instead hybrid because a conditional dependency has been already introduced in the sparse structure of the inverse covariance $\mathbf{J}$ (the inference network). Results are reported in Table 3 where we observe that these networks retrieve a larger quantity of

true positives than the ones constructed from conditional entropy. However, the fraction of false positive is also larger than the ones in Table 1 although it is smaller than what observed in the inference network in Table 2. Overall, these results indicate that conditioning is effective in discarding false positives.

4.6. Summary of All Results in a Single ROC Plot. In summary, we have investigated the networks associated with conditional transfer entropy, unconditional transfer entropy, and inference for three models under a range of different parameters. In the previous subsections we have provided some comparisons between the performances of the three models in different ranges of parameters. Let us here provide a summary of all results within a single ROC plot [23]. Figure 3 reports the ROC values, for each model and each parameter combination, x-axis is false positive rates (FP/m), and y-axis is true positive rates (TP/n). Each point is an average over 100 processes. Points above the diagonal line are associated with relatively well performing models with the upper left corner representing the point where models correctly discover all true causality links without any false positive. The plot reports with large symbols the cases for $\gamma = 0.1$ and validation at p value $p_v = 0.01$, which can be compared with the data reported in the tables. We note that, by construction, LoGo models are sparse (with a number of edges $\sim 3p$ [4]). This restrains the ROC results to the left-hand side of the plot. For this reason an expanded view of the figure is also proposed with the x-axis scaled. Note that this ROC curve is provided as a visual tool for intuitive comparison between models.

TABLE 3: *Unconditioned transfer entropy network: comparison between fraction of true positive (TP/n) and fraction of false positive (FP/n). Statistical validation of transfer entropy is at $p_v = 1\%$ p value. The table reports only the case for the parameter $\gamma = 0.1$.*

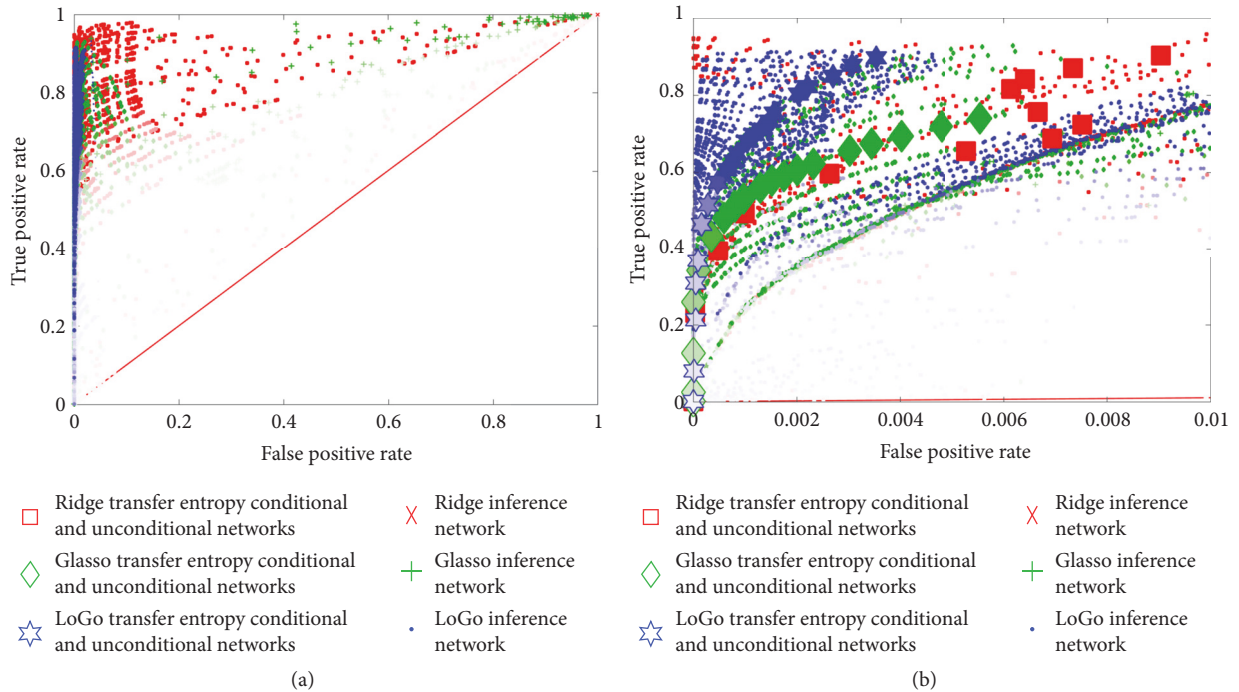| q	10	20	30	50	200	300	1000	20000
Ridge TP/n	0.02	0.39**	0.45**	0.51**	0.65**	0.69**	0.78**	0.92**
Ridge FP/n	0.07	1.06	0.95	0.85	0.93	0.99	1.20	1.73
Glasso TP/n	0.00	0.24**	0.35**	0.43**	0.57**	0.60**	0.67**	0.77**
Glasso FP/n	0.00	0.10	0.20	0.29	0.51	0.56	0.73	1.66
LoGo TP/n	0.11	0.34**	0.41**	0.47**	0.63**	0.66**	0.76**	0.89**
LoGo FP/n	0.02	0.16	0.25	0.34	0.59	0.66	0.87	1.49

**$P < 10^{-8}$.

FIGURE 3: *ROC values, for each model and each parameter combination. x-axis false positive rates (FP/m); y-axis true positive rates (TP/n). (a) and (b) are the same with x-axis expanded on the low values only for (b) to better visualise the differences between the various models. Large symbols refer to $\gamma = 0.1$ and validation at p value $p_v = 0.01$. Color intensity is proportional to time series length. Inference network results are all outside the range of the plot. Reported values are averages over 100 processes.*

Overall from Tables 1, 2, and 3 and Figure 3 we conclude that all models obtain better results for longer time series and that conditional transfer entropy overperforms the unconditional counterparts (see, Tables 1 and 3 and the two separated ROC figures for conditional and unconditional transfer entropies reported in Figure 5 in Appendix D). In the range of short time series, when $q \leq p$, which is of interest for this paper, LoGo is the best performing model with better performances achieved for small $\gamma \lesssim 10^{-4}$ and validation with small p values $p_v \lesssim 10^{-4}$. LoGo is consistently the best performing model also for longer time series up to lengths of $q \sim 1000$. Instead, above $q = 2000$ ridge begins to provide better results. For long time series, at $q = 20,000$, the best performing model is ridge with parameters $\gamma = 10^{-5}$, p value $p_v = 5 \ 10^{-6}$. LoGo is also performing well when time series are long with best performance obtained at $q = 20,000$ for

parameters $\gamma = 10^{-10}$, p value $p_v = 5 \ 10^{-6}$. We note that LoGo instead performs poorly in the region of parameters with $\gamma \leq 0.1$ and $p_v \leq 0.01$ for short time series $q \leq p/2$.

5. Conclusions and Perspectives

In this paper we have undertaken the challenging task to explore models and parameter regions where the analytics of time series can retrieve significant fractions of true causality links from linear multivariate autoregressive process with known causality structure. Results demonstrate that sparse models with conditional transfer entropy are the ones who achieve best results with significant causality link retrievals already for very short time series even with $q \leq p/5 = 20$. This region is very critical and general considerations would suggest that no solutions can be discovered. Indeed, this result

is in apparent contradiction with a general analytical results in [24, 25] who find that no significant solutions should be retrieved for $q \leq N/2 = 150$. However, we notice that the problem we are addressing here is different from the one in [24, 25]. In this paper we have been considering an underlying sparse true causality structure and such a sparsity changes considerably the condition of the problem yielding to significant solutions even well below the theoretical limit from [24, 25] which is instead associated with nonsparse models.

Unexpectedly, we observed that the structure of the inference networks in the two sparse models, Glasso and LoGo, has excessive numbers of false positives yielding to rather poor performances. However, in these models false positive can be efficiently filtered out by imposing statistical significance of the transfer entropies.

Results are affected by the choice of the parameters and the fact that the models depend on various parameters (q, p, γ, p_v, P) make the navigation in this space quite complex. We observed that the choice of p values, p_v, for valid transfer entropies affects results. Within our setting we obtained best results with the smaller p values especially in the regions of short time series. We note that the regularizer parameter γ also plays an important role and best performances are obtained by combination of the two parameters γ and p_v. Not surprisingly, longer time series yield to better results. We observe that conditioning to all other variables or unconditioning is affecting the transfer entropy estimation with better performing causality network retrieval obtained for conditioned transfer entropies. However, qualitatively, results are comparable. Other intermediate cases, such as conditioning to past of all other variables only, have been explored again with qualitatively comparable results. It must be said that in the present system results are expected to be robust to different conditionings because the underlying network of the investigated processes is sparse. For denser inference structures, conditioning could affect more the results.

Consistently with the findings in [4] we find that LoGo outperforms the other methods. This is encouraging because the present settings of LoGo is using a simple class of information filtering networks, namely, the TMFG [7], obtained by retaining largest correlations. There are a number of alternative information filtering networks which should be explored. In particular, given the importance of statistical validation emerged from the present work, it would be interesting to explore statistical validation within the process of construction of the information filtering networks themselves.

In this paper we investigate a simple case with a linear autoregressive multivariate normal process analysed by means of linear models. Both LoGo and Glasso can be extended to the nonlinear case with LoGo being particularly suitable for nonparametric approaches as well [4].

There are alternative methods to extract causality networks from short time series, in particular Multispatial CCM [26, 27] appears to perform well for short time series. A comparison between different approaches and the application of these methods to real data will be extremely interesting. However this should be the object of future works.

Appendix

A. Conditional Transfer Entropy

Let us here briefly review two of the most commonly used information theoretic quantities that we use in this paper, namely, mutual information (quantifying dependency) and transfer entropy (quantifying causality) for the multivariate case [11–13].

A.1. Mutual Information. Let us first start from the simplest case of two random variables, $X \in \mathbb{R}^1$ and $Y \in \mathbb{R}^1$, where dependence can be quantified by the amount of shared information between the two variables, which is called mutual information: $I(X; Y) = H(X) + H(Y) - H(X, Y)$, where $H(X)$ is the entropy of variable X, $H(Y)$ is the entropy of variable Y, and $H(X, Y)$ is the joint entropy of variables X and Y [13]. Extending to the multivariate case, the shared information between a set of n random variables $\mathbf{X} = (X_1, \ldots, X_n)^T \in \mathbb{R}^n$ and another set of m random variables $\mathbf{Y} = (Y_1, \ldots, Y_m)^T \in \mathbb{R}^m$ is

$$I(\mathbf{X}; \mathbf{Y}) = H(\mathbf{X}) + H(\mathbf{Y}) - H(\mathbf{X}, \mathbf{Y}) \qquad (A.1)$$

with $H(\mathbf{X})$, $H(\mathbf{Y})$ being the entropies, respectively, for the set of variables $\mathbf{X}$ and $\mathbf{Y}$ and $H(\mathbf{X}, \mathbf{Y})$ being their joint entropy. It must be stressed that this quantity is the mutual information between two sets of multivariate variables and it is not the multivariate mutual information between all variables $\{\mathbf{X}, \mathbf{Y}\}$ which instead measures the intersection of information between all variables. Mutual information in (A.1) can also be written as

$$I(\mathbf{X}; \mathbf{Y}) = H(\mathbf{Y}) - H(\mathbf{Y} \mid \mathbf{X}) = H(\mathbf{X}) - H(\mathbf{X} \mid \mathbf{Y}) \quad (A.2)$$

which makes use of the conditional entropy of $\mathbf{Y}$ given $\mathbf{X}$: $H(\mathbf{Y} \mid \mathbf{X}) = H(\mathbf{Y}, \mathbf{X}) - H(\mathbf{X}) = \mathbb{E}(H(\mathbf{Y}) \mid \mathbf{X})$.

Conditioning to a third set of variables $\mathbf{W}$ can also be applied to mutual information itself and its expression is a direct extension of (A.1) and it is called conditional mutual information:

$$I(\mathbf{X}; \mathbf{Y} \mid \mathbf{W}) = H(\mathbf{X} \mid \mathbf{W}) + H(\mathbf{Y} \mid \mathbf{W}) \\ - H(\mathbf{X}, \mathbf{Y} \mid \mathbf{W}). \qquad (A.3)$$

Equations (A.1) and (A.3) coincide in the case of an empty set $\mathbf{W} = \emptyset$. Mutual information and conditional mutual information are symmetric measures with $I(\mathbf{X}; \mathbf{Y} \mid \mathbf{W}) = I(\mathbf{Y}; \mathbf{X} \mid \mathbf{W})$ always. Let us note that symmetry is unavoidable for information measures that quantify the simultaneous effect of a set of variables onto another. Indeed, in a simultaneous interaction cause and effect cannot be distinguished from the exchange of information and direction cannot be established. To quantify causality one must investigate the transmission of information not only between two sets of variables but also through time.

A.2. Conditional Transfer Entropy. Causality between two random variables, $X \in \mathbb{R}^1$ and $Y \in \mathbb{R}^1$, can be quantified by means of the so-called transfer entropy which quantifies

the amount of uncertainty on Y explained by the past of X given the past of Y. Let us consider a series of observations and denote with X_t being the random variable X at time t and with $X_{t-\tau}$ being the random variable at a previous time, τ lags before t. Using this notation, we can define transfer entropy from variable X to variable Y in terms of the following conditional mutual information: $T(X \rightarrow Y) = I(Y_t; X_{t-\tau} \mid Y_{t-\tau})$ [11, 13].

For the multivariate case, given two sets of random variables $\mathbf{X} \in \mathbb{R}^n$ and $\mathbf{Y} \in \mathbb{R}^m$, the transfer entropy is the conditional mutual information between the set of variables $\mathbf{Y}_t$ at time t and the past of the other set of variables, $\mathbf{X}_{t-\tau}$ conditioned to the past of the first variable $\mathbf{Y}_{t-\tau}$. That is, $T(\mathbf{X} \rightarrow \mathbf{Y}) = I(\mathbf{Y}_t; \mathbf{X}_{t-\tau} \mid \mathbf{Y}_{t-\tau})$ [13]. In general, the influence from the past can come from more than one lag and we can therefore extend the definition including different sets of lags for the two variables: $\tau_1, \ldots, \tau_k, \lambda_1, \ldots, \lambda_h$:

$$
\begin{aligned}
&T(\mathbf{X} \longrightarrow \mathbf{Y}) \\
&= I\left(\mathbf{Y}_t \left\{\mathbf{Y}_{t-\lambda_1} \cdots \mathbf{Y}_{t-\lambda_h}\right\}; \left\{\mathbf{X}_{t-\tau_1} \cdots \mathbf{X}_{t-\tau_k}\right\}\right) \\
&= H\left(\mathbf{Y}_t \left\{\mathbf{Y}_{t-\lambda_1} \cdots \mathbf{Y}_{t-\lambda_h}\right\}\right) \\
&\quad - H\left(\mathbf{Y}_t \left\{\mathbf{X}_{t-\tau_1} \cdots \mathbf{X}_{t-\tau_k}, \mathbf{Y}_{t-\lambda_1} \cdots \mathbf{Y}_{t-\lambda_h}\right\}\right);
\end{aligned}
\tag{A.4}
$$

a further generalization, which we use in this paper, includes conditioning to any other set of variables $\{\mathbf{W}_{t-\theta_1} \cdots \mathbf{W}_{t-\theta_g}\}$ lagged at $\theta_1, \ldots, \theta_g$:

$$
\begin{aligned}
T(\mathbf{X} \longrightarrow \mathbf{Y} \mid \mathbf{W}) = I\Big(\mathbf{Y}_t; &\left\{\mathbf{X}_{t-\tau_1} \cdots \mathbf{X}_{t-\tau_k}\right\} \mid \\
&\left\{\mathbf{Y}_{t-\lambda_1} \cdots \mathbf{Y}_{t-\lambda_h}, \mathbf{W}_{t-\theta_1} \cdots \mathbf{W}_{t-\theta_g}\right\}\Big).
\end{aligned}
\tag{A.5}
$$

In this paper we simplify notation using $\mathbf{X}_t^{\text{lag}} = \{\mathbf{X}_{t-\tau_1} \cdots \mathbf{X}_{t-\tau_k}\}$, $\mathbf{Y}_t^{\text{lag}} = \{\mathbf{Y}_{t-\lambda_1} \cdots \mathbf{Y}_{t-\lambda_h}\}$, and $\mathbf{W}_t = \{\mathbf{W}_{t-\theta_1} \cdots \mathbf{W}_{t-\theta_g}\}$.

In the literature, there are several examples that use adaptations of (1) to compute causality and dependency measures [28]. A notable example is the directed information, introduced by Massey in [29], where τ spans all lags in a range between 0 and $s - 1$ and λ spans the lags from 1 to $s - 1$. The directed information is then defined as the sum over transfer entropies from $s = 1$ to present:

$$
\begin{aligned}
&I\left(\{\mathbf{X}\}_1^t \longrightarrow \{\mathbf{Y}\}_1^t \mid \mathbf{W}\right) \\
&= \sum_{s=1}^{t} I\left(\mathbf{Y}_s; \{\mathbf{X}\}_1^s \mid \{\mathbf{Y}\}_1^{s-1}, \mathbf{W}\right),
\end{aligned}
\tag{A.6}
$$

where we adopted the notations $\{\mathbf{X}\}_1^t = \{\mathbf{X}_1 \cdots \mathbf{X}_t\}$ and $\{\mathbf{Y}\}_1^t = \{\mathbf{Y}_1 \cdots \mathbf{Y}_t\}$. Interestingly, this definition includes the conditional synchronous mutual information contributions

between $\mathbf{X}_s$ and $\mathbf{Y}_s$. Following Kramer et al. [30, 31] we observe that for stationary processes

$$
\begin{aligned}
\lim_{t \to \infty} \frac{1}{t} I\left(\{\mathbf{X}\}_1^t \longrightarrow \{\mathbf{Y}\}_1^t\right) &= \lim_{t \to \infty} I\left(\{\mathbf{X}\}_1^t; \mathbf{Y}_t \mid \{\mathbf{Y}\}_1^{t-1}\right) \\
&= T\left(\{\mathbf{X}\}_1^{t-1} \longrightarrow \mathbf{Y}_t\right) + I\left(\{\mathbf{X}\}_1^t; \{\mathbf{Y}\}_1^t \mid \{\mathbf{X}\}_1^{t-1}\right),
\end{aligned}
\tag{A.7}
$$

with $T(\{\mathbf{X}\}_1^{t-1} \rightarrow \mathbf{Y}_t) = I(\mathbf{Y}_t; \{\mathbf{X}_1 \cdots \mathbf{X}_{t-1}\} \mid \{\mathbf{Y}_1 \cdots \mathbf{Y}_{t-1}\})$. This identity supports the intuition that the directed information accounts for the transfer entropy plus an instantaneous term.

B. Shannon-Gibbs Entropy

The general expression for the transfer entropy reported di in Section A, (1), is independent of the kind of entropy definition. In this paper we use Shannon entropy, which is defined as

$$
H(\mathbf{X}) = -\mathbb{E}\left[\log p(\mathbf{X})\right],
\tag{B.1}
$$

$$
H(\mathbf{Y}) = -\mathbb{E}\left[\log p(\mathbf{Y})\right],
\tag{B.2}
$$

where $p(\mathbf{X})$ and $p(\mathbf{Y})$ are the probability distribution function for the set of random variables $\mathbf{X}$ and $\mathbf{Y}$. Similarly, the joint Shannon entropy for the variables $\mathbf{X}$ and $\mathbf{Y}$ is defined as

$$
H(\mathbf{X}, \mathbf{Y}) = -\mathbb{E}\left[\log p(\mathbf{X}, \mathbf{Y})\right]
\tag{B.3}
$$

with $p(\mathbf{X}, \mathbf{Y})$ being the joint probability distribution function of $\mathbf{X}$ and $\mathbf{Y}$. This is the most common definition of entropy. It is a particularly meaningful and suitable entropy for linear modelling, as we focus in the paper.

B.1. Multivariate Normal Modelling. For multivariate normal variables the Shannon-Gibbs entropy is

$$
H(\mathbf{X}) = \frac{1}{2} \log\left(\det \boldsymbol{\Sigma}(\mathbf{X})\right) + \frac{n}{2} \log(2\pi e)
\tag{B.4}
$$

and its conditional counterpart is

$$
H(\mathbf{X} \mid \mathbf{W}) = \frac{1}{2} \log\left(\det \boldsymbol{\Sigma}(\mathbf{X} \mid \mathbf{W})\right) + \frac{n}{2} \log(2\pi e)
\tag{B.5}
$$

with $\boldsymbol{\Sigma}$ being the covariance matrix and $\det(\cdot)$ being the matrix determinant. In the paper we use these expressions to compute mutual information and conditional transfer entropy.

C. Computing Conditional Covariances for Subsets of Variables from the Inverse Covariance

Let us consider three sets of variables $\mathbf{Z}_1 \in \mathbb{R}^{p_1}$, $\mathbf{Z}_2 \in \mathbb{R}^{p_2}$, and $\mathbf{Z}_3 \in \mathbb{R}^{p_3}$ and the associated inverse covariance $\mathbf{J} \in \mathbb{R}^{(p_1+p_2+p_3) \times (p_1+p_2+p_3)}$ for $\{\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3\} \in \mathbb{R}^{(p_1+p_2+p_3)}$. The conditional covariance of $\mathbf{Z}_1$ given $\mathbf{Z}_2$ and $\mathbf{Z}_3$ is the inverse of the $p_1 \times p_1$ upper left part of $\mathbf{J}$ with indices in $V_1 = (1, \ldots, p_1)$ (see Figure 4):

$$
\boldsymbol{\Sigma}(\mathbf{Z}_1 \mid \mathbf{Z}_2, \mathbf{Z}_3) = (\mathbf{J}_{1,1})^{-1}.
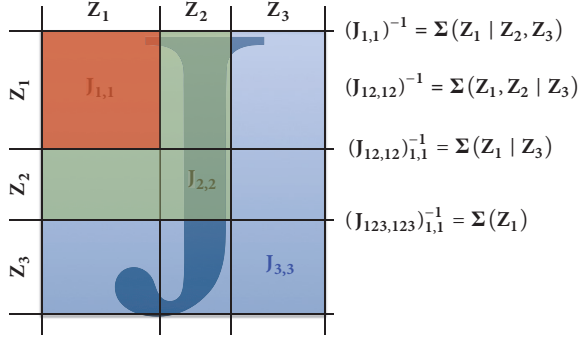\tag{C.1}
$$

FIGURE 4: The inverse of parts the inverse covariance $\mathbf{J}$ gives the covariance of the variables corresponding to that part conditioned to the other variables.

Instead, the conditional covariance of $\mathbf{Z_3}$ given $\mathbf{Z_3}$ is obtained by inverting the larger upper left part $\mathbf{J_{12,12}}$ with both indices in $\{V_1, V_2\}$ with $V_2 = (p_1 + 1, \ldots, p_1 + p_2)$, and then taking the inverse of the part with indices in V_1 which, using the Schur complement [13], is

$$\Sigma\left(\mathbf{Z_1} \mid \mathbf{Z_3}\right) = \left(\mathbf{J_{1,1}} - \mathbf{J_{1,2}}\left(\mathbf{J_{2,2}}\right)^{-1}\mathbf{J_{2,1}}\right)^{-1}. \tag{C.2}$$

Figure 4 schematically illustrates these inversions and their relations with conditional covariances. Let us note that these conditional covariances can also be expressed directly in terms of subcovariances by using again the Schur complement:

$$\Sigma\left(\mathbf{Z_1 Z_2}, \mathbf{Z_3}\right) = \Sigma_{1,1} - \Sigma_{1,23}\left(\Sigma_{23,23}\right)^{-1}\Sigma_{23,1},$$
$$\Sigma\left(\mathbf{Z_1} \mid \mathbf{Z_3}\right) = \Sigma_{1,1} - \Sigma_{1,3}\left(\Sigma_{3,3}\right)^{-1}\Sigma_{3,1}. \tag{C.3}$$

However, when p_3 (cardinality of V_3) is much larger than p_1 and p_2 (cardinalities of V_1 and V_2) then the equivalent expressions, (C.1) and (C.2), that use the inverse covariance involve matrices with much smaller dimensions. This can become computationally crucial when very large dimensionalities are involved. Furthermore, if the inverse covariance $\mathbf{J}$ is estimated by using a sparse modelling tool such as Glasso or LoGo [4, 14] (as we do in this paper), then computations in expressions (C.1) and (C.2) have to handle only a few nonzero elements providing great computational advantages over (C.3).

In the paper we make use of (C.1)-(C.2) to compute mutual information and conditional transfer entropy for the system of all variables and their lagged versions.

C.1. Mutual Information. Let us consider the mutual information between any two subsets $\mathbf{X} \in \mathbb{R}^n$ and $\mathbf{Y} \in \mathbb{R}^m$ of variables conditioned to all other variables, which we shall call $\mathbf{W} \in \mathbb{R}^{p-n-m}$. For these three sets of variables $\{\mathbf{X}, \mathbf{Y}, \mathbf{W}\} \in \mathbb{R}^p$ the conditional mutual information, $I(\mathbf{X}, \mathbf{Y}, \mathbf{W}) = H(\mathbf{X}, \mathbf{Y} \mid \mathbf{W}) - H(\mathbf{X} \mid \mathbf{Y}, \mathbf{W})$ (see (A.3)), can be expressed in terms of the conditional covariances by using (B.5):

$$I\left(\mathbf{X}; \mathbf{Y} \mid \mathbf{W}\right) = \frac{1}{2}\log\det\Sigma\left(\mathbf{X} \mid \mathbf{W}\right)$$
$$- \frac{1}{2}\log\det\Sigma\left(\mathbf{X} \mid \mathbf{Y}, \mathbf{W}\right). \tag{C.4}$$

Given the inverse covariance $\mathbf{J} \in \mathbb{R}^{p \times p}$, by using (C.1) and (C.2) and substituting

$$\mathbf{Z_1} = \mathbf{X},$$
$$\mathbf{Z_2} = \mathbf{Y}, \tag{C.5}$$
$$\mathbf{Z_3} = \mathbf{W},$$

we can express the conditional mutual information, (C.4), directly in terms of the parts of $\mathbf{J}$:

$$I\left(\mathbf{X}; \mathbf{Y} \mid \mathbf{W}\right) = -\frac{1}{2}\log\det\left(\mathbf{J_{1,1}} - \mathbf{J_{1,2}}\left(\mathbf{J_{2,2}}\right)^{-1}\mathbf{J_{2,1}}\right)$$
$$+ \frac{1}{2}\log\left(\det\mathbf{J_{1,1}}\right). \tag{C.6}$$

Note that although this is not directly evident, (C.6) is symmetric by exchanging **1** and **2** (i.e., **X** and **Y**).

C.2. Conditional Transfer Entropy. Conditional transfer entropy (see (1)) is conditional mutual information between lagged sets of variables and therefore it can be computed directly from (C.6). In this case we shall name

$$\mathbf{Z_1} = \mathbf{Y}_t,$$
$$\mathbf{Z_2} = \left\{\mathbf{X}_{t-\tau_1}\cdots\mathbf{X}_{t-\tau_k}\right\},$$
$$\mathbf{Z_3} = \left\{\mathbf{Y}_{t-\lambda_1}\cdots\mathbf{Y}_{t-\lambda_h}, \mathbf{W}_{t-\theta_1}\cdots\mathbf{W}_{t-\theta_g}\right\},$$
$$T\left(\mathbf{X}\longrightarrow\mathbf{Y} \mid \mathbf{W}\right) \tag{C.7}$$
$$= -\frac{1}{2}\log\det\left(\mathbf{J_{1,1}} - \mathbf{J_{1,2}}\left(\mathbf{J_{2,2}}\right)^{-1}\mathbf{J_{2,1}}\right)$$
$$+ \frac{1}{2}\log\det\left(\mathbf{J_{1,1}}\right)$$

obtaining an expression which is formally identical to (C.6) but with indices **1** and **2** referring to the above sets of variables instead.

Note that index **3** does not appear in this expression. Information from variables **3** (**W**) has been used to compute **J** but then only the subparts **1** and **2** are required to compute the conditional transfer entropy. The fact that these expressions for conditional mutual information and conditional transfer entropy involve only local parts (**1** and **2**) of the inverse covariance can become extremely useful when high-dimensional datasets are involved.

D. Comparison between Conditional and Unconditional Transfer Entropies

The two ROC plots for conditional and unconditional transfer entropies are displayed in Figure 5. Form the comparison it is evident that, for the process studied in this paper, conditional transfer entropy provides best results. This is in line with what observed in Tables 1, 3, 4, and 5.
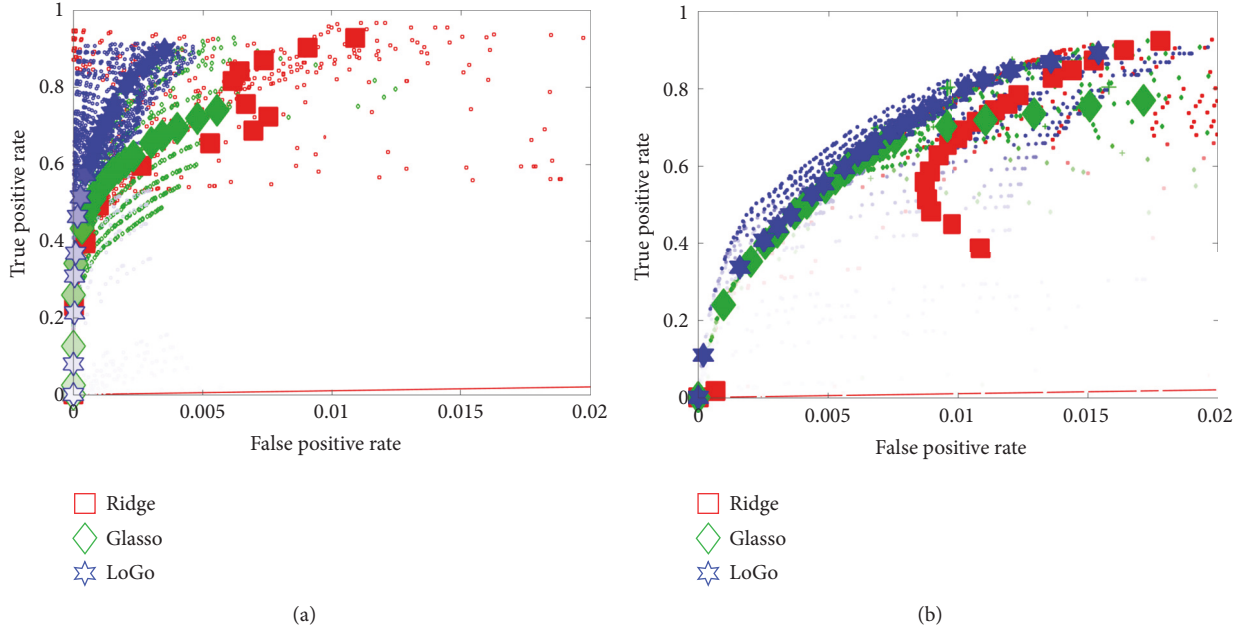
(a)

(b)

FIGURE 5: *ROC values, for conditional (a) and unconditional (b) transfer entropies. x-axis false positive rates (FP/m); y-axis true positive rates (TP/n). Large symbols refer to $\gamma = 0.1$ and validation at p value $p_v = 0.01$. Color intensity is proportional to time series length. Inference network results are all outside the range of the plot. Reported values are averages over 100 processes.*

TABLE 4: *Causality network validation with conditional transfer entropy validation at 1% Bonberroni adjusted p value. Fraction of true positive (TP/n) and fraction of false positive (FP/n), statistically validated, causality links for the three models, and different time series lengths. The table reports only the case for the parameter $\gamma = 0.1$.*

q	10	20	30	50	200	300	1000	20000
Ridge TP/n	0.00	0.00	0.00	0.00	0.00	0.30**	0.67**	0.89**
Ridge FP/n	0.00	0.00	0.00	0.00	0.00	0.01	0.18	0.75
Glasso TP/n	0.00	0.00	0.00	0.00	0.35**	0.43**	0.57**	0.71**
Glasso FP/n	0.00	0.00	0.00	0.00	0.01	0.03	0.13	0.45
LoGo TP/n	0.00	0.00	0.02	0.17**	0.50**	0.56**	0.69**	0.87**
LoGo FP/n	0.00	0.00	0.00	0.00	0.01	0.03	0.09	0.28

**$P < 10^{-8}$.

TABLE 5: *Causality network validation with unconditional transfer entropy validation at 1% Bonberroni adjusted p value. Fraction of true positive (TP/n) and fraction of false positive (FP/n), statistically validated, causality links for the three models, and different time-series lengths. The table reports only the case for the parameter $\gamma = 0.1$.*

q	10	20	30	50	200	300	1000	20000
Ridge TP/n	0.00	0.00	0.22**	0.36**	0.55**	0.59**	0.70**	0.88**
Ridge FP/n	0.00	0.00	0.09	0.21	0.47	0.55	0.77	1.32
Glasso TP/n	0.00	0.00	0.00	0.27**	0.48**	0.53**	0.62**	0.75**
Glasso FP/n	0.00	0.00	0.00	0.11	0.37	0.43	0.61	1.41
LoGo TP/n	0.00	0.00	0.22**	0.35**	0.53**	0.58**	0.69**	0.86**
LoGo FP/n	0.00	0.00	0.05	0.16	0.42	0.49	0.71	1.27

**$P < 10^{-8}$.

E. Causality Network Results for Transfer Entropy Validation with 1% Bonferroni Adjusted p Values

In Tables 4 and 5, true positive rates (TP/n) and fraction of false positives (FP/m) statistically validated and causality links with validation at 1% Bonberroni adjusted p value (i.e., $p_v \leq 10^{-6}$) are reported. These tables must be compared with Tables 1 and 3, in the main text where causality links are validated at $p_v = 1\%$ nonadjusted p value.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

References

[1] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM Review*, vol. 51, no. 1, pp. 34–81, 2009.

[2] S. Theodoridis, Y. Kopsinis, and K. Slavakis, "Sparsity-aware learning and compressed sensing: an overview," https://arxiv.org/abs/1211.5231.

[3] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[4] W. Barfuss, G. P. Massara, T. Di Matteo, and T. Aste, "Parsimonious modeling with information filtering networks," *Physical Review E*, vol. 94, no. 6, Article ID 062306, 2016.

[5] T. Aste, T. Di Matteo, and S. T. Hyde, "Complex networks on hyperbolic surfaces," *Physica A: Statistical Mechanics and its Applications*, vol. 346, no. 1-2, pp. 20–26, 2005.

[6] M. Tumminello, T. Aste, T. di Matteo, and R. N. Mantegna, "A tool for filtering information in complex systems," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 30, pp. 10421–10426, 2005.

[7] G. P. Massara, T. Di Matteo, and T. Aste, "Network filtering for big data: triangulated maximally filtered graph," *Journal of Complex Networks*, vol. 5, no. 2, article 161, 2017.

[8] A. Tikhonov, "Solution of incorrectly formulated problems and the regularization method," *Soviet Mathematics Doklady*, vol. 4, no. 4, pp. 1035–1038, 1984.

[9] A. E. Hoerl and R. W. Kennard, "Ridge regression: biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970.

[10] A. Zaremba and T. Aste, "Measures of causality in complex datasets with application to financial data," *Entropy*, vol. 16, no. 4, pp. 2309–2349, 2014.

[11] T. Schreiber, "Measuring information transfer," *Physical Review Letters*, vol. 85, no. 2, pp. 461–464, 2000.

[12] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[13] T. W. Anderson, *Multivariate Statistical Analysis*, Willey and Sons, New York, NY, USA, 1984.

[14] J. Friedman, T. Hastie, and R. Tibshirani, "Sparse inverse covariance estimation with the graphical lasso," *Biostatistics*, vol. 9, no. 3, pp. 432–441, 2008.

[15] M. H. Gruber, *Improving Efficiency by Shrinkage: The James–Stein and Ridge Regression Estimators*, vol. 156 of *CRC Press*, 1998.

[16] D. M. Witten and R. Tibshirani, "Covariance-regularized regression and classification for high dimensional problems," *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, vol. 71, no. 3, pp. 615–636, 2009.

[17] N. Meinshausen and P. Bühlmann, "High-dimensional graphs and variable selection with the lasso," *The Annals of Statistics*, vol. 34, no. 3, pp. 1436–1462, 2006.

[18] J. Kruskal Jr., "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proceedings of the American Mathematical Society*, vol. 7, no. 1, pp. 48–50, 1956.

[19] R. N. Mantegna, "Hierarchical structure in financial markets," *The European Physical Journal B—Condensed Matter and Complex Systems*, vol. 11, no. 1, pp. 193–197, 1999.

[20] J. D. Hamilton, *Time Series Analysis*, vol. 2, Princeton University Press, Princeton, NJ, USA, 1994.

[21] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, 1938.

[22] Q. H. Vuong, "Likelihood ratio tests for model selection and nonnested hypotheses," *Econometrica. Journal of the Econometric Society*, vol. 57, no. 2, pp. 307–333, 1989.

[23] J. A. Swets, *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers*, Psychology Press, 2014.

[24] I. Varga-Haszonits, F. Caccioli, and I. Kondor, "Replica approach to mean-variance portfolio optimization," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2016, no. 12, Article ID 123404, 2016.

[25] G. Papp, F. Caccioli, and I. Kondor, "Fluctuation-bias trade-off in portfolio optimization under expected shortfall with l_2 regularization," https://arxiv.org/abs/1602.08297.

[26] G. Sugihara, R. May, H. Ye et al., "Detecting causality in complex ecosystems," *Science*, vol. 338, no. 6106, pp. 496–500, 2012.

[27] A. T. Clark, H. Ye, F. Isbell et al., "Spatial convergent cross mapping to detect causal relationships from short time series," *Ecology*, vol. 96, no. 5, pp. 1174–1181, 2015.

[28] J. Pearl, *Causality*, Cambridge University Press, 2009.

[29] J. Massey, *Causality, Feedback And Directed Information*, Citeseer, 1999.

[30] G. Kramer, *Directed information for channels with feedback [PhD thesis]*, University of Manitoba, Winnipeg, Canada, 1998.

[31] P.-O. Amblard and O. J. Michel, "The relation between Granger causality and directed information theory: a review," *Entropy. An International and Interdisciplinary Journal of Entropy and Information Studies*, vol. 15, no. 1, pp. 113–143, 2013.