Analysis of Large Market Data Using Neural Networks: A Causal Approach

M.-A. Divernois[a,b,c], J. Etesami[a], D. Filipovic[a,b], N. Kiyavash[a]

[a] *EPFL - Ecole Polytechnique Fédérale de Lausanne, Switzerland*
[b] *SFI - Swiss Finance Institute*
[c] *Corresponding author*

Abstract

We develop a data-driven framework to identify the interconnections between firms using an information-theoretic measure. This measure generalizes Granger causality and is capable of detecting nonlinear relationships within a network. Moreover, we develop an algorithm using recurrent neural networks and the aforementioned measure to identify the interconnections of high-dimensional nonlinear systems. The outcome of this algorithm is the causal graph encoding the interconnections among the firms. These causal graphs can be used as preliminary feature selection for another predictive model or for policy design. We evaluate the performance of our algorithm using both synthetic linear and nonlinear experiments and apply it to the daily stock returns of U.S. listed firms and infer their interconnections.

Keywords: Granger Causality, Directed Information, Recurrent Neural Networks, Network Inference

⋆

Email addresses: marc-aurele.divernois@epfl.ch (M.-A. Divernois), sjetesami@gmail.com (J. Etesami), damir.filipovic@epfl.ch (D. Filipovic), negar.kiyavash@epfl.ch (N. Kiyavash)

1. Introduction

The causal network of a dynamical system provides important information that may help to better understand its behavior and ultimately design better policies to predict and control it. Large number of banks, insurances, hedge funds, and other financial institutions around the globe are interacting daily and thus their causal network is of great importance in econometrics.

There have been many attempts during the past decades to capture and visualize the network of interconnections among a set of financial institutions. The most widely used concept of causality in time series econometrics is due to Granger [16]. This is based on statistical analysis of the financial series such as their stock prices over a finite time period. Granger's definition of causality states that a time series X is a cause of another time series Y, if the mean squared error of the 1-step ahead forecast for Y is smaller when the history of X is included in the forecasting information set. Otherwise, when the forecast does not improve by including the information of X, then it is declared that X does not cause Y. This idea is reflected in the information-theoretic measure that we use in this work to infer the causal interactions among a network of time series.

In the great majority of practical applications, Granger-causality has been studied in the context of Vector Autoregressive (VAR) models. For instance, [7] proposes several measures based on Granger-causality to capture the connections between the monthly returns of different financial institutions. It uses principle component analysis and "pairwise" Granger-causality tests to identify the causal networks. Other related works are [13] and [3] in which the authors propose connectedness measures based on generalized variance decomposition. However, the measures introduced in these works are again limited to linear systems and they are based on pairwise comparison which as we show in Section 2.2 fails to infer the true causal relationships.

Contributions of this paper are both in network identification literature as well in finance. Our contribution to network identification are as follows:

- We use an information-theoretic measure known as directed information (DI) to infer the Granger-causalities among a set of time series. This measure is non-parametric, i.e., it does not depend on the underlying model of the dynamics and it is capable of capturing causal relationships in both linear and nonlinear systems. The output of this approach is a directed graph known as Directed Information Graph (DIG) that visualizes the interconnections among a set of time series such as stock returns.

- Computing DI has both high computational and sample complexity which makes it not suitable for inferring the causal structure of large networks. To overcome this problem, we develop a novel approach based on Recurrent Neural Networks (RNNs) that reduces the complexity of evaluating DIs in high-dimensional settings.

Applications of DIG are various in finance. In particular, we recommend to use it as a preliminary feature selection of another predictive model. Feature selection is a process often used in machine learning and statistics which consists of keeping only a subset of relevant features, usually to avoid overfitting or to reduce dimensionality. For example, [33] and [20] show that extraneous features are prone to reduce model's performance measures. Finance applications of feature selection models are various and include credit scoring, stock market behavior analysis or even fraud detection ([2]). [40]

states that feature selection preprocessing is not addressed carefully enough in the bankruptcy prediction literature. They compare five feature selection methods used in bankruptcy prediction: t-test, correlation matrix, stepwise regression, Principle Component Analysis (PCA) and factor analysis and show that any of these methods improves performance. [42] uses a Sequential Forward Selection algorithm to select relevant features predicting the Turkish market index. The use of such feature selection model reduces model prediction error compared to the case where all features are used. This is due to information embedded in several economic factors already included in the market index. They show that only the lagged value of the market index is enough to predict the forthcoming value of the index.

1.1. Related Work

In recent years, several approaches have been developed to generalize the applicability of Granger-causality to non-linear and large dynamics. To mention a few, [34] that introduces different terminologies for causality based on Granger's ideas and provide a set of parametric non-causality constraints in the context of Markov switching VAR models. In a similar context, [5] investigates time-varying systemic risk based on a range of multi-factor asset pricing models and develops a Markov Chain Monte Carlo (MCMC) scheme to infer their model parameters and consequently obtain their corresponding networks. Another attempt is [9] in which the authors explore quantile-based methods of Granger causality. This method is consistent with [18] and [10] that focus on causality among tail events. These methods are suitable for capturing causal relationships that are not in the center of their distributions, or in the mean but they are in the tails of their distributions. It is important to emphasize that our proposed approach using DI is also capable of capturing such causal relations.

Most of the above aforementioned approaches are developed and tested for small size networks. Often, the problem of network identification in high dimensional settings requires more considerations and even its own techniques. For instance, [8] proposes a Bayesian non-parametric Lasso prior (BNP-Lasso) for high-dimensional VAR models that can improve efficiency and accuracy. In order to overcome overparametrization and overfitting issues in large VAR models, BNP-Lasso clusters the VAR coefficients into groups and shrinks the coefficients of each group toward a common location. However, this method is limited to linear models with Gaussian innovations. To overcome this limitation, [23] proposes a Bayesian non-parametric approach that allows for nonlinearity in the conditional mean, heteroskedasticity in the conditional variance, and non-Gaussian innovations. However, unlike the BNP-Lasso, it does not allow sparsity in the model. [32] proposes yet another non-parametric, quasi-Bayesian likelihood estimation methodology for high dimensional setting with time varying parameters. The work in [21] tackles the curse of dimensionality by a two-stage approach. In the first stage, a spike-and-slab prior distribution is used for each entry of the coefficient matrix which also identifies the interconnection network. In the second stage, it imposes prior dependence on the coefficients by specifying a Markov process for their random distribution. A closely related work is [4] that proposes a shrinkage and selection methodology designed for network inference in high-dimensional settings. It uses a regularized linear regression model with spike-and-slab prior on the parameters. However, both methods are limited to VAR models.

The rest of the paper is organized as follows. Section 2 reviews the notion of Granger causality and formally introduce directed information graphs which is suitable for linear and nonlinear systems. In Section 3, we introduce a novel approach for inferring the Granger-causal network of high dimensional nonlinear systems. Finally, in Section 4, we apply our method to learn the causal network of both synthetic and real-world dataset. For the real-world experiment, we used the daily stock prices of major US firms.

2. Causal Network

In this section, we present a statistical approach to learn the causal interconnections in a dynamical systems based on Granger causality [16]. We begin by introducing some notations. Plain capital letters denote random variables or processes, while lowercase letters denote their realizations. Bold letters are used for column vectors, matrices, and tensors and calligraphy letters are used for sets. We use $X_{j,t}$ to denote the value of a time series X_j at time t and X_j^t to denote the time series X_j up to time t. For a set $\mathcal{A} = \{a_1, ..., a_n\}$ and an index set $\mathcal{I} \subseteq \{1, ..., n\}$, we define $\mathcal{A}_{-\mathcal{I}} := \mathcal{A} \setminus \{a_i : i \in \mathcal{I}\}$.

2.1. Granger Causality

Researchers from different fields have developed various frameworks and graphical models to capture and represent interconnections among variables or processes. One of the most popular and widely used frameworks in economics is the notion of Granger causality. The basic idea in this framework was originally introduced by Wiener [41], and later formalized by Granger [16]. The idea is as follows: "we say that X is causing Y if we are better able to predict the future of Y using all available information than if the information apart from the past of X had been used."

Despite broad philosophical viewpoint of [17], his formulation for practical implementation was done using multivariate autoregressive (MVAR) models and linear regression which has been widely adopted in econometric and other disciplines. More precisely, in order to identify the influence of X_t on Y_t in a MVAR comprises of three time series $\{X, Y, Z\}$, Granger's idea is to compare the performance of two linear regressions: the first one predicts Y_t given $\{X^{t-1}, Y^{t-1}, Z^{t-1}\}$ and the second one predicts Y_t given $\{Y^{t-1}, Z^{t-1}\}$. Clearly, the performance of the second predictor is bounded by the first predictor. If they have the same performance, then we say X does not Granger cause Y. It is important to emphasize that this formulation is only applicable in linear systems.

To go beyond linear systems, works such as [35] and [28] use information-theoretical measures and generalize Granger causality. In this work, we introduce and apply directed information (DI) [35], an information-theoretical tools to measure interconnections among firms. DI has been used in many applications to infer causal relationships. For example, [36] and [24] used it for analyzing neuroscience data and [14] and [15] applied to market data.

In order to visualize the inferred interconnections among time series using DI, directed information graphs (DIGs) have been developed [35]. DIGs are a type of graphical models in which nodes represent time series and arrows indicate the direction of causation. We use DIG to represent the causal network among the covered firms.

2.2. Directed Information Graphs (DIGs)

In the rest of this section, we describe how the DI can capture the interconnections in causal[1] dynamical systems (linear or non-linear) and formally define DIGs.

Consider a dynamical system comprised of three time series $\{X, Y, Z\}$ that we assume they have a joint probability density function $p(X, Y, Z)$. To answer whether X has influence on Y or not over time horizon $[1, T]$, following the idea of Granger, we compare the average performance of two particular predictors over this time horizon. The first predictor uses the history of all three time series while the second one uses the history of all processes excluding process X. On average, the performance of the predictor with less information (the second one) is upper bounded by the performance of the predictor with more information (the first one). However, if the prediction of both predictors are close over time horizon $[1, T]$, it is an indication that X does not cause Y in this time horizon. To rigorously formalize this idea, we need the predictors and a measure to compare their performances.

In the definition of DI, the predictors belong to the space of probability measures. More precisely, the prediction of the first predictor at time t is $p(Y_t|Y^{t-1}, Z^{t-1}, X^{t-1})$ that is the conditional density function of Y_t given the history of all time series. Similarly, the prediction of the second predictor is $p(Y_t|Y^{t-1}, Z^{t-1})$ that is the conditional density function of Y_t given the history of all time series except time series X.

[1] In causal systems, given the full past of the system, the present of the processes become statistically independent. In other words, there are no simulations relationships between the time series.

Given the predictions of the first and the second predictors at time t for an outcome $y_t \in \mathcal{Y}$, the goodness of these predictions are measured by the log-loss that are defined respectively by

$$
\begin{aligned}
&-\log p(Y_t = y_t | Y^{t-1}, Z^{t-1}, X^{t-1}),\\
&-\log p(Y_t = y_t | Y^{t-1}, Z^{t-1}).
\end{aligned}
$$

According to the above measures of goodness, the better the predictor is, the smaller its log-loss will be. This loss function has meaningful information-theoretical interpretations. Namely, the log-loss is the Shannon's code length[2], i.e., the number of bits required to efficiently represent y_t.

At time t for an outcome $y_t \in \mathcal{Y}$, the difference between the log-losses of the two predictors compares their performances. This difference is also called *regret*,

$$
r_t := -\log p(Y_t = y_t | Y^{t-1}, Z^{t-1}) - \big(-\log p(Y_t = y_t | Y^{t-1}, Z^{t-1}, X^{t-1})\big) \tag{1}
$$

$$
= \log \frac{p(Y_t = y_t | Y^{t-1}, Z^{t-1}, X^{t-1})}{p(Y_t = y_t | Y^{t-1}, Z^{t-1})}. \tag{2}
$$

Note that the regrets are non-negative for all t and all outcomes y_t. The average regret over the time horizon $[1, T]$ is given by

$$
\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[r_t], \tag{3}
$$

where the expectation is taken over the joint density function[3] of X, Y, and

[2] It is also called the description length of y_t. For more information see [11].

[3] For the sake of notational simplicity, we use $p(y^t, z^{t-1}, x^{t-1})$ to denote $p(Y^t = y^t, Z^{t-1} = z^{t-1}, X^{t-1} = x^{t-1})$.

Z, i.e.,

$$\mathbb{E}[r_t] = \int p(y^t, z^{t-1}, x^{t-1}) \log \frac{p(y_t|y^{t-1}, z^{t-1}, x^{t-1})}{p(y_t|y^{t-1}, z^{t-1})} dy^t dx^{t-1} dz^{t-1}. \tag{4}$$

The average regret in (3) is called *directed information* (DI) and will be our measure of causation in this work. This measure is always positive and if it is zero, it indicates that the history of time series X contains no significant information that would help in predicting the future of time series Y given the history of Y and Z. This definition can be generalized to more than three time series as follows,

Definition 1. *Consider a network of m time series $\mathcal{R} = \{R_1, ..., R_m\}$ with the joint probability density function p. The directed information from R_i to R_j over time horizon $[1, T]$ is given by*

$$I(R_i \to R_j || \mathcal{R}_{-\{i,j\}}) := \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left[\log \frac{p(R_{j,t}|\mathcal{R}_{-\{i,j\}}^{t-1}, R_j^{t-1}, R_i^{t-1})}{p(R_{j,t}|\mathcal{R}_{-\{i,j\}}^{t-1}, R_j^{t-1})}\right], \tag{5}$$

where $\mathcal{R}_{-\{i,j\}}^{t-1} := \{R_1^{t-1}, ..., R_m^{t-1}\} \setminus \{R_i^{t-1}, R_j^{t-1}\}$. We declare R_i influences R_j over time horizon $[1, T]$, if and only if

$$I(R_i \to R_j || \mathcal{R}_{-\{i,j\}}) > 0. \tag{6}$$

An interpretation of R_i influencing R_j is that varying R_i will change the value of R_j even if all the other variables within the network remains unchanged. In another words, if R_i does not influence R_j, then varying R_i would not change R_j when the values of the remaining times series are fixed. This can be seen from the fact that DI compares two conditional distributions of R_j over a time horizon of length T; one is given the history of all the time series while the other one is given all the history except the history of R_i. Thus, if DI in (5) is zero, then these two conditional distributions are equal over this time horizon. This implies that the history of R_i does not contain any useful information for R_j.

It is important to emphasize that the definition of DI does not rely on any model assumption, thus DI is capable of inferring the causal relationships in general (linear or non-linear) dynamical systems. Next, we define the graphical model that we use in this work to visualize the causal network among firms.

Definition 2. *Directed information graph (DIG) of a set of m time series $\mathcal{R} = \{R_1, ..., R_m\}$ is a directed graph $G = (\mathcal{V}, \mathcal{E})$, where nodes represent time series ($\mathcal{V} = \mathcal{R}$) and arrow $(R_i, R_j) \in \mathcal{E}$ denotes that R_i influences R_j.*

A simple way to represent the DIG G of a dynamical system is via the adjacency matrix $\mathbf{DIG} = [d_{i,j}]_{m\times m}$ that is defined by

$$d_{j,i} = \begin{cases} 1 & \text{if } I(R_i \to R_j || \mathcal{R}_{-\{i,j\}}) > 0, \\ 0 & \text{otherwise.} \end{cases} \tag{7}$$

Given a DIG $G = (\mathcal{V}, \mathcal{E})$, we define the parent set of node R_j denoted by $\mathcal{PA}_j \subset \mathcal{V}$ to be the set of all times series that have direct influences on R_j, i.e., $\mathcal{PA}_j := \{R_k : d_{j,k} = 1\}$. Similarly, the children set of node R_j is given by $\mathcal{CH}_j := \{R_k : d_{k,j} = 1\}$. Next example demonstrates the DIG of a simple linear system.

Example 1. *Consider a network of three times series $\{X, Y, Z\}$ with the following linear dynamic,*

$$\begin{pmatrix} X_t \\ Y_t \\ Z_t \end{pmatrix} = \begin{pmatrix} 0.5 & 0 & 0 \\ 0.4 & 0.5 & 0 \\ 0 & -0.2 & 0 \end{pmatrix} \begin{pmatrix} X_{t-1} \\ Y_{t-1} \\ Z_{t-1} \end{pmatrix} + \begin{pmatrix} N_{X_t} \\ N_{Y_t} \\ N_{Z_t} \end{pmatrix}, \tag{8}$$

where N_X, N_Y, and N_Z are three independent stationary Gaussian processes with zero mean and a diagonal covariance matrix (1, 0.9, 1). Since the

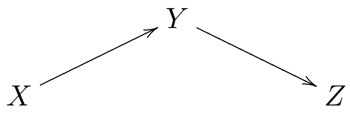

Figure 1: DIG of the system in (8).

dynamic is linear and the exogenous noises are Gaussian, we can compute the DIs using the following expression[4] *[14].*

$$I(Z \to Y||X) = \frac{1}{2T}\sum_{t=1}^{T} \log \frac{|\Sigma_{Y_{t-1},Y_t,X_{t-1}}||\Sigma_{Z_{t-1},Y_{t-1},X_{t-1}}|}{|\Sigma_{Y_{t-1},X_{t-1}}||\Sigma_{Z_{t-1},Y_{t-1},Y_t,X_{t-1}}|}, \tag{9}$$

where $|\Sigma_{Y_{t-1},Y_t,X_{t-1}}|$ *denotes the determinant of the covariance matrix of* $\{Y_{t-1}, Y_t, X_{t-1}\}$*. Using (9), we computed the DIs of this system,*

$$I(Y \to X||Z) = 0, \quad I(Z \to X||Y) = 0, \quad I(Z \to Y||X) = 0, \quad I(Z \to Z||X, Y) = 0,$$
$$I(X \to Z||Y) = 0, \qquad I(X \to Y||Z) \approx 0.1, \qquad I(Y \to Z||X) \approx 0.03.$$

Figure 1 illustrates the DIG of this system. In this example, $\mathcal{PA}_Z = \{Y\}$ *and* $\mathcal{CH}_Z = \{\}$*.*

Inference methods based on pairwise comparison has been developed and applied in the literature to identify the causal structure of time series. The methods in [7], [6], and [1] are three such examples. However, pairwise comparison is not a correct approach in general and may fail to capture the true underlying network. For instance, considering the pairwise comparison in Example 1 between X and Z leads to a conclusion that X directly influences Z, which would be inaccurate. More precisely, without conditioning on Y, we obtain

$$I(X \to Z) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\log \frac{p(Z_t|Z^{t-1}, X^{t-1})}{p(Z_t|Z^{t-1})}\right] \approx 0.002 > 0.$$

[4] Equation (9) does not hold in general setting.

Notice that the DI in (5) is not a measure based on pairwise comparison. On the contrary, it measures the influence by conditioning on the remaining time series within the network.

Remark 1. *A causal model allows a factorization of the joint density function in some specific ways. It was shown in [35] that under a mild assumption, the joint density function of a causal discrete-time dynamical system with DIG $G(\mathcal{R}, \mathcal{E})$ can be factorized as follows,*

$$p(R_1, ..., R_m) = \prod_{i=1}^{m} \prod_{t=1}^{T} p(R_{i,t}|\mathcal{R}^{t-1}_{\mathcal{PA}_i \cup \{i\}}). \tag{10}$$

Such factorization is called a generative model.

2.3. Inferring DIGs

Inferring the DIG of a dynamical system requires estimating the DIs between all ordered pairs of time series within that system. More precisely, inferring the DIG of a network of m time series requires computing $m(m-1)$ number of DIs. On the other hand, estimating DI requires estimating all the expectation terms in (5). In information theory this expectation is known as *conditional mutual information*[5], i.e.,

$$I(R_{j,t}; R_i^{t-1}|\mathcal{R}^{t-1}_{-\{i,j\}}, R_j^{t-1}) := \mathbb{E}\left[\log \frac{p(R_{j,t}|\mathcal{R}^{t-1}_{-\{i,j\}}, R_j^{t-1}, R_i^{t-1})}{p(R_{j,t}|\mathcal{R}^{t-1}_{-\{i,j\}}, R_j^{t-1})}\right]. \tag{11}$$

Using this notation, (5) can be written as follows

$$I(R_i \to R_j||\mathcal{R}_{-\{i,j\}}) = \frac{1}{T}\sum_{t=1}^{T} I(R_{j,t}; R_i^{t-1}|\mathcal{R}^{t-1}_{-\{i,j\}}, R_j^{t-1}), \tag{12}$$

Therefore, parametric and non-parametric estimators for the conditional mutual information can be used to estimate the DIs. There are different

[5] For more details see [11].

methods that can be used to estimate the terms in (12) given i.i.d. samples of the time series such as plug-in empirical estimator and k-nearest neighbor estimator. For an overview of such estimators see the Appendix and the articles in [31], [30], and [22].

In general, estimating the DI in (12) is a complicated task and has high sample complexity. This is due to the fact that it requires estimating high dimensional conditional distributions. However, knowing some side information about the underlying dynamic can simplify the learning task of the DIG. For instance, in Example 1, since the underlying dynamic is linear with Gaussian exogenous noises, the DIs can be computed via the covariance matrices (9). Clearly, the covariance matrix can be estimated with lower complexity compared to conditional mutual information. For our experimental results, we used (9) for the linear Gaussian experiment and the k-nearest method in [39] for the non-linear experiment. The main reason for selecting k-nearest method is because it has shown relatively better performance compared to the other non-parametric estimators. For the sake of completeness, we describe the steps of this method in Appendix.

Side information can also help to infer the DIG of a dynamical system without directly estimating the DIs but instead it provides an alternative approach to identify the DIG. For example, if it is given that the underlying dynamic is linear, i.e., $\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \mathbf{N}_t$, then it has been shown in [14] that the support[6] of the coefficient matrix $\mathbf{A}$ is equal to the adjacency matrix of its corresponding DIG. This result implies that in linear systems, one can obtain the DIG by estimating the coefficient matrix. The latter problem

[6] The support of a matrix $\mathbf{B} = [b_{i,j}]$ is a binary matrix of the same dimension as $\mathbf{B}$ such that its entry (i, j) is one if and only if $b_{i,j} \neq 0$.

has lower complexity and it can be done using e.g., linear regression. For similar examples in econometric models see [15].

2.4. DIG in High-dimensional Settings

For large networks with thousands nodes and millions of edges such as social or financial networks, DIGs become too complex to infer and analyze. The main reason is that without any side information, estimating the DI has high computational and sample complexity. Furthermore, the estimating complexity of DI increase with the dimension of the network. This is due to the fact that the DI in (5) measures the influence from R_i to R_j by conditioning on the information from the remaining network $\mathcal{R}_{-\{i,j\}}$. Therefore, the size of the conditioning set grows with the size of the network. This motivates the prior works to reduce the complexity of estimating DIs and thus make it more suitable for inferring the DIG of large networks by reducing the size of the conditioning set.

One such approaches is proposed by [37], in which they developed an efficient algorithm to identify the best directed tree approximation of a given network. This means reducing the size of the conditioning set to zero, i.e., no conditioning. However, this approach comes with the price of an approximation error and furthermore it fails to identify many interconnections between the processes.

The authors in [38] presented a more generalized version of the above approximation in which they identify the optimal connected bounded in-degree[7] approximations. This method reduces the size of the conditioning

[7] Connected bounded in-degree graphs with bound k are connected directed graphs in which each node has at most k number of parents.

set in (5) to some constant value (bound of the in-degrees) which is independent of the network size. Although, this approach improves upon the approximation error but there is still a trade-off between the sample complexity and the approximation error. In another words, as the in-degree bound increases, the sample complexity increases but the approximation error decreases.

In this work, we propose a new method that reduces the size of the conditioning set in (5) to only one for any given network while introducing less approximation error compared to the prior works. In this method, we estimate the directed information from R_i to R_j by conditioning on an auxiliary time series. This auxiliary time series is defined such that it comprises the information that the remaining of the network $\mathcal{R}_{-\{i,j\}}$ has about R_j. Next section explains this idea in more details.

3. Methodology

In order to present our method, we need the following preliminary result that characterizes an important property of DI in (5). All the proofs are presented in Appendix A.

Lemma 1. *Consider a network of m time series $\mathcal{R} = \{R_1, ..., R_m\}$ with corresponding DIG $G = (\mathcal{V}, \mathcal{E})$. Let $\mathcal{C}$ be a subset of $\mathcal{R}_{-\{i,j\}}$ such that $\mathcal{PA}_j \subseteq \mathcal{C}$. If $R_i \notin \mathcal{PA}_j$, then we have*

$$I(R_i \to R_j || \mathcal{C}) = 0. \tag{13}$$

Note that if $\mathcal{C} = \mathcal{R}_{-\{i,j\}}$ and R_i is not a parent of R_j, then by the definition of DIG, Equation (13) holds. On the other hand, this result states that to detect whether there is an influence from R_i to R_j in a network of time series, it suffices to find a subset of time series that either contains the parents of R_j or their information. In the remaining of this section, we first clarify the above statement via a simple linear system and later generalize it to non-linear models using neural networks.

Remark 2. *It is important to emphasize that the reverse of Lemma 1 does not hold. In another words, if there exists a subset $\mathcal{C} \subset \mathcal{R}_{-\{i,j\}}$ such that (13) holds, we cannot conclude that R_i has no direct influence on R_j.*

3.1. Linear Systems

Consider a first order vector autoregression model (VAR) with m time series,

$$\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \mathbf{N}_t, \tag{14}$$

where $\mathbf{X}_t, \mathbf{N}_t \in \mathbb{R}^m$, $\mathbf{A} \in \mathbb{R}^{m\times m}$, and $\mathbf{N}_t$ is a vector of m independent exogenous noises. As we discussed earlier, the result in [14] implies that the DIG of this VAR model is encoded in the support of its coefficient matrix $\mathbf{A} = [a_{i,j}]$, i.e.,

$$I(X_i \to X_j || \mathcal{X}_{-\{i,j\}}) = 0 \Longleftrightarrow a_{j,i} = 0. \tag{15}$$

In another words, the parents of time series X_j are the ones whose corresponding coefficients are non-zero in the j-th row of matrix $\mathbf{A}$. This also can be seen from the j-th row of the matrix equation in (14),

$$X_{j,t} = \sum_{k=1}^{m} a_{j,k} X_{k,t-1} + N_{j,t}. \tag{16}$$

Another way to interpret the above equation is to say that the information of the network about time series X_j is in the form of a “portfolio”, i.e., a linear combination of the other time series. Therefore, it is possible to summarize the network's information about X_j into only one time series, namely a well-designed portfolio. Next result shows the form of such portfolio.

Lemma 2. *In the linear system of* (14)*, X_i has no direct influence on X_j if and only if*

$$I(X_i \to X_j || Q) = 0, \tag{17}$$

where Q is a time series which we call the ideal portfolio and it is defined by $Q_{t-1} := \boldsymbol{u}_t^T \boldsymbol{X}_{-\{i\},t-1}$, where

$$\boldsymbol{u}_t := \arg \min_{\boldsymbol{w} \in \mathbb{R}^{m-1}} \mathbb{E}\left[||X_{j,t} - \boldsymbol{w}^T \boldsymbol{X}_{-\{i\},t-1}||_2^2 \right],$$

$$\boldsymbol{X}_{-\{i\},t-1} := [X_{1,t-1}, \cdots, X_{i-1,t-1}, X_{i+1,t-1}, \cdots, X_{m,t-1}]^T.$$

According to the above Lemma, projecting X_j on $\mathcal{X}_{-\{i\}}$ results in an ideal portfolio Q that contains all the information for deciding whether there is an influence from X_i to X_j. Hence, instead of estimating $I(X_i \to X_j || \mathcal{X}_{-\{i,j\}})$ whose complexity depends on the network size, one can estimate $I(X_i \to X_j || Q)$. Note that the sample complexity of the latter DI does not grow with the size of the network and thus it is suitable for estimating the DIG of large networks.

3.2. Non-linear Systems with Additive Noise

Inferring the causal network of non-linear systems is a challenging problem that its complexity increases exponentially with the dimension of the network. In this section, we study the causal inference problem in non-linear systems whose dynamic can be captured by

$$X_{j,t} = F_j(\mathcal{X}^{t-1}) + \varepsilon_{j,t}, \quad j = 1, ..., m, \tag{18}$$

where $\mathcal{X}^{t-1} = \{X_1^{t-1}, ..., X_m^{t-1}\}$, $\{F_j(\cdot)\}$ is a set of non-linear continuous functions, and $\{\varepsilon_{j,t}\}$ is a set of independent exogenous noises. We call this model non-linear with *additive noise* due to the noise term that is added to the non-linear term[8]. This is a general non-linear dynamic that can be used to model the behavior of wide range of physical dynamical systems. The dynamic is called *Markovian* if $\mathcal{X}^{t-1}$ is replaced by $\mathcal{X}_{t-1} = \{X_{1,t-1}, ..., X_{m,t-1}\}$.

Below, we generalize the result of Lemma 2 to the non-linear system in (18) by showing that in such systems, it is possible to reduce the conditioning set in the DI to one time series.

Lemma 3. *In (18), X_i has no direct influence on X_j if and only if*

$$I(X_i \rightarrow X_j || Q) = 0, \tag{19}$$

where Q is a time series defined by $Q_{t-1} := F_j(\mathcal{X}^{t-1}_{-\{i\}})$.

In the remaining of this section, we propose two methods to obtain the time series Q introduced in the above Lemma.

Koopman-based lifting technique. Consider a particular sub-class of the non-linear system in (18) whose dynamic is defined by

$$F_j(\mathcal{X}^{t-1}) = \sum_{k=1}^{K} w_{j,k} h_k(\mathcal{X}_{t-1}), \quad j = 1, ..., m, \tag{20}$$

where $\{w_{j,k} \in \mathbb{R}\}$ are the weights and $\{h_k(\cdot)\}$ denotes a set of library functions that are assumed to be known. This model is Markovian and the library functions can be seen as a set of basis that are used to approximated

[8] In contrary to additive noise, there are systems in which the exogenous terms are multiplicative, e.g., $X_{j,t} = X_{i,t-1}\varepsilon_{j,t} + X_{j,t-1}$.

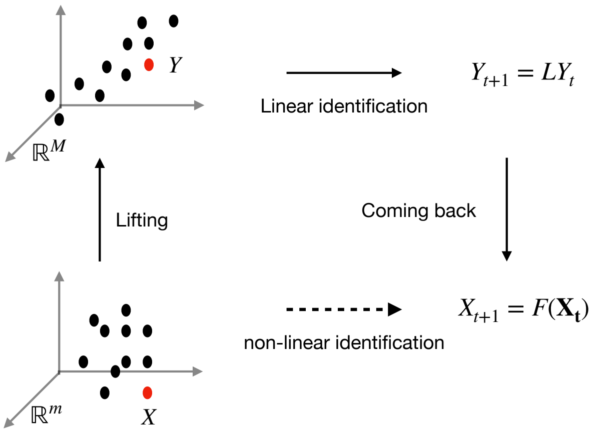

Figure 2: Koopman lifting technique compared to classical non-linear identification.

$F_j(\cdot)$. Examples of such library functions are monomials and Gaussian radial basis functions.

In this setting, the results of Lemma 3 implies that the following time series can be substituted in the conditioning of the DI.

$$Q_t = \sum_{k=1}^{K} w_{j,k} h_k(\mathcal{X}_{-\{i\},t-1}). \tag{21}$$

However, in this formulation, the weights $\{w_{,k}\}$ are unknown. An approach to obtain the weights is a non-linear filtering technique known as Koopman-based lifting [25]. This technique takes observational data and a set of library functions as inputs and obtains the unknown coefficients $\{w_{,k}\}$. The main steps of this technique are; transforming the data (lifting the data), applying a linear identification on the lifted data, and finally applying another transformation to bring down the results into the original vector field. Figure 2 illustrates the main steps. For more details see Appendix and [29].

Although, the Koopman-based lifting technique is theoretically sound but it has some shortcomings facing real-world applications. First, the

Koopman's performance depends on the choice of the library functions and second, it often fails to estimate the real time series Q. More precisely, this technique involves the computation of matrix $\mathbf{L} := log(\mathbf{P}_x^{\dagger}\mathbf{P}_y)/T_s$, where $\mathbf{P}_x$ and $\mathbf{P}_y$ are estimated from the observational data[9]. Matrix $\mathbf{P}_x^{\dagger}$ denotes the pseudo-inverse, and the function $\log(\cdot)$ denotes the (principal) matrix logarithm. On the other hand, Koopman lifting technique is applicable for estimating the time series Q only when the resulting matrix $\mathbf{L}$ is real[10]. However, this is not always the case in real-world applications due to observational noises and lack of sufficient data. To overcome such shortcomings, we propose an alternative approach to estimate Q using recurrent neural networks (RNNs).

RNNs method. Recurrent neural networks are a specific class of Neural Networks well suited to learn time series. They are distinguished by their memory as they are able to remember information from prior inputs to influence their current outputs. The universal approximation theorem states that a neural network with enough hidden layers can approximate any non-linear continuous function such as $F_j(\cdot)$ in (18) (see [19]).

Given the aforementioned result, we train an RNN using the observational data to estimate the time series Q defined in Lemma 3. More precisely, our RNN maps $\mathcal{X}_{-\{i\}}^{t-1}$ as the inputs to $X_{j,t}$ as the output. Let $R_j(\mathcal{X}_{-\{i\}}^{t-1}; \Theta^*)$ denotes the trained RNN with parameters Θ^*. In this case, the time series Q can be written as $Q_{t-1} = R_j(\mathcal{X}_{-\{i\}}^{t-1}; \Theta^*)$. Finally, we use (19) to detect whether X_i has influence on X_j or not. Algorithm Infer-DIG in 1 summarizes the steps of our RNN method.

[9] See Appendix for more details.

[10] See [12] for conditions under which a real matrix has a real logarithm.

```
Algorithm 1: Infer-DIG
  Input: Observational data of $m$ time series up to time $T$, $\mathcal{X}^T$,
   Threshold $\alpha > 0$;
  Output: Adjacency matrix of $\mathbf{DIG} = [d_{i,j}]$;
  for  $i, j = 1, ..., m$ do
      Train an RNN $R_j(\cdot; \Theta^*)$ that maps $\mathcal{X}^{t-1}_{-\{i\}}$ to $X_{j,t}$;
      Define $Q_{t-1} = R_j(\mathcal{X}^{t-1}_{-\{i\}}; \Theta^*)$;
      if $I(X_i \to X_j || Q) > \alpha$ then
          $d_{j,i} = 1$
      else
          $d_{j,i} = 0$
```

4. Experimental Results

Since the true empirical DIG of firms is unknown, to evaluate the performance of our approach, we use different simulated environment. In this section, we first describe the simulation methodology in a linear Gaussian framework. We then show that our results generalize well to nonlinear setting by conducting an experiment on a nonlinear system. Finally, we apply our approach to a set of empirical data describing the daily stock prices of US firms and obtain their corresponding causal network.

4.1. Linear Gaussian Framework

In this experiment, we consider a linear system, a VAR(1) model whose dynamic is given by

$$\mathbf{X}_t = \mathbf{A}\mathbf{X}_{t-1} + \mathbf{N}_t \tag{22}$$

with m being the number of asset returns, $\mathbf{X}_t = (X_{1,t}, X_{2,t}, ..., X_{m,t})^\top$ being the vector of returns at time t, $\mathbf{A} = [a_{i,j}]$ being a $m \times m$ matrix and $\mathbf{N}_t$ being a $\mathcal{N}(0, \mathbf{I})$ vector of noises. As we discussed earlier, in such linear systems, $a_{i,j}$ captures the influence of asset j on asset i, i.e., there is an influence from j to i if and only if $a_{i,j} \neq 0$.

To reflect an important property of the market that some firms are more connected than others in our experiment, we divided the m time series into two parts. First part $(1 \leq i \leq s)$ indicates assets with high degrees of connectedness and the second part $(1 + s \leq i \leq m)$ are the ones with low degrees of connectedness. Parameter $1 < s < m$ denotes the numbers of assets with high degrees of connectedness. Afterward, for every entry (i, j) of $\mathbf{A}$, we independently generated a random number $x \sim U(-0.9, 0.9)$ and decided on value $a_{i,j}$ as follows,

$$a_{i,j} = \begin{cases} x1_{|x|>\underline{\epsilon}}, & \text{if } \ 1 \leq i \leq s, 1 \leq j \leq s, \\ x1_{|x|>\bar{\epsilon}}, & \text{if } \ 1 + s \leq i \leq m, 1 \leq j \leq m, \\ 0, & \text{if } \ 1 \leq i \leq s, 1 + s \leq j \leq m, \end{cases} \tag{23}$$

where $1_{a>b}$ denotes the indicator function which is equal to 1 when $a > b$ and 0 otherwise and $\underline{\epsilon}$ and $\bar{\epsilon}$ are thresholds to define non-zero entries in the upper-left and the lower part of $\mathbf{A}$, respectively. Figure 3 illustrates the structure of the resulting $\mathbf{A}$. We select these thresholds such that $\underline{\epsilon} < \bar{\epsilon}$. This ensures that the upper-left of $\mathbf{A}$ is denser than its lower part or equivalently, assets with indices $\{1, ..., s\}$ are more connected than the ones with indices $\{1 + s, ..., m\}$. In our experiment, we select $(s, m) = (85, 100)$ and $(\underline{\epsilon}, \bar{\epsilon}) = (0.4, 0.7)$. Finally, to guarantee the stability of the time series, we rescale[11] $\mathbf{A}$ such that its spectral radius is strictly less than one, i.e., $\rho(\mathbf{A}) < 1$. Once the matrix $\mathbf{A}$ is defined, we simulate the time series using (22) for a period of $T = 30000$ and use the resulting data for our estimations.

[11] Formally, we use $\mathbf{A}/(\rho(\mathbf{A}) + \epsilon)$, where $0 < \epsilon$.

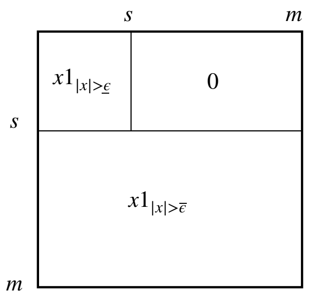
s
m
$x1_{|x|>\underline{\epsilon}}$
0
s
$x1_{|x|>\overline{\epsilon}}$
m

Figure 3: Structure of matrix $\mathbf{A}$ in (22) that is build using (23).

To study the effect of the conditioning set on detecting the influences, in our experiments, we consider four different conditioning sets. More precisely, to measure whether asset i influences asset j, we estimate $I(X_i \to X_j||\mathcal{C}_j)$ for the following choices of the conditioning set;

1. *True parents:* In this approach, we select $\mathcal{C}_j$ to be the true parents of X_j excluding X_i, i.e., $\mathcal{C}_j = \mathcal{PA}_j \setminus \{X_i\}$. Note that this approach is not practical[12] and we use it only as the benchmark to better understand the performances of the other approaches.
2. *Most correlated:* In this case, we define $\mathcal{C}_j$ to be the set of k most correlated assets with X_j (except X_i).
3. *Ideal portfolio:* In this scenario, $\mathcal{C}_j$ contains the portfolio Q, where Q is defined in Lemma 2. For further discussion see Appendix.
4. *RNN:* This method applies Algorithm 1 to estimate the time series Q and defines $\mathcal{C}_j = \{Q\}$.

Note that we also applied the Koopman-based lifting techniques but due

[12] This is because in structural learning problems, we do not know the true parents of each asset. In another words, if we had access to the true parents of each asset, we would have the DIG of the system and there is no need to compute the DIs.

to its mentioned shortcomings, it was unable to robustly identify the interconnections. Hence, we could not compare its performance with the other methods. In this experiment, since the dynamic is linear and the noises are Gaussian, we use Equation (9) to estimate the DIs. Finally, we obtain the adjacency matrix of the corresponding DIGs by comparing the estimated DIs with a threshold $\alpha > 0$, i.e.,

$$[\mathbf{DIG}]_{j,i} = \begin{cases} 1 & \text{if } \widehat{I}(X_i \to X_j || \mathcal{C}_j) > \alpha, \\ 0, & \text{otherwise,} \end{cases} \tag{24}$$

where $\widehat{I}(\cdot \to \cdot || \cdot)$ denotes the estimated DI from the data. In order to compare the performances of the aforementioned four approaches, we use the precision and recall measure between the true DIG (obtained from $\mathbf{A}$) and their estimated DIGs. Formally, the precision and the recall are defined by

$$Precision := \frac{TP}{TP+FP}, \quad Recall := \frac{TP}{TP+FN},$$

where

$$TP := \sum_{i,j=1}^{m} 1_{a_{j,i}\neq 0} 1_{[\mathbf{DIG}]_{j,i}\neq 0}, \quad FP := \sum_{i,j=1}^{m} 1_{a_{j,i}=0} 1_{[\mathbf{DIG}]_{j,i}\neq 0},$$
$$FN := \sum_{i,j=1}^{m} 1_{a_{j,i}\neq 0} 1_{[\mathbf{DIG}]_{j,i}=0}.$$

Figure 4 shows the performances of the four aforementioned approaches in the linear framework. It is not surprising that the *true parents* approach achieves 100% accuracy, as it is anticipated by Lemma 1. The *ideal portfolio*'s performance is guaranteed by Lemma 2 and it is verified by our experiment. However, it is important to emphasize that the *ideal portfolio* shows ideal performance because the underlying model is linear. As we will see in

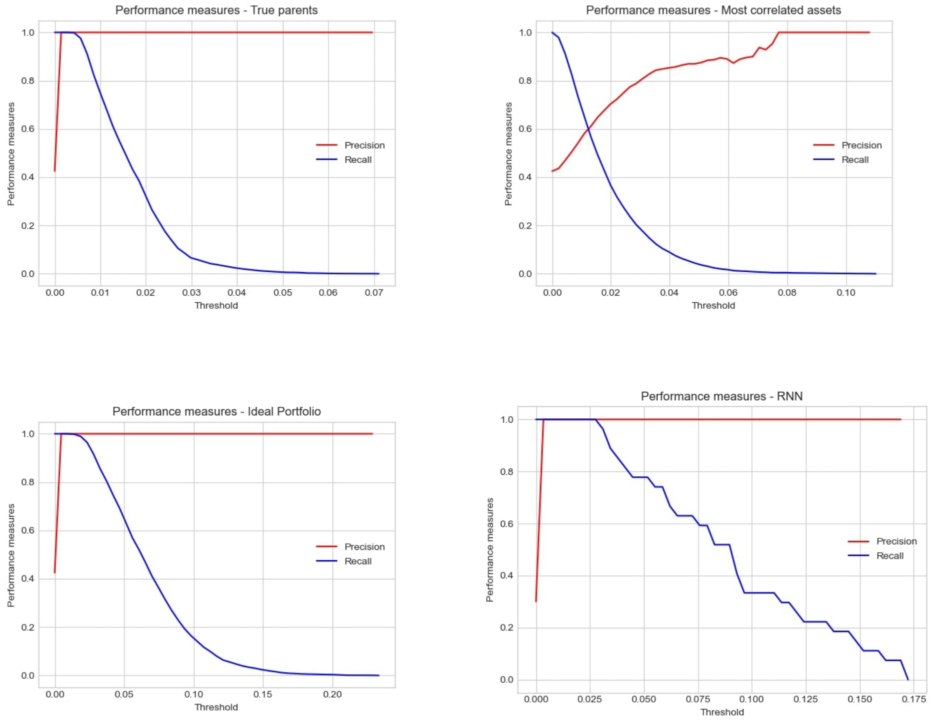

Figure 4: Precision and recall curves in the linear framework for the *True parents*, *Most correlated*, *Ideal portfolio*, and the *RNN*, respectively.

the next section, its performance declines when the underlying model deviates from being linear. For the *most correlated* approach, we used $k = 10$ but as it is shown in Figure 4, it has the worst performance among the four conditioning methods. This is due to the fact that the set of the ten most correlated assets with a given asset j does not necessarily contains the true parents of asset j. On the other hand, we observe high accuracy from the *RNN* approach which is a striking result. This result is an evidence that an RNN is capable of estimating the ideal portfolio, i.e., the time series Q in Lemma 2 without any side information about the underlying model.

4.2. Non-Linear framework

To compare the performances of the different approaches from the previous section in non-linear environment, we simulate a set of quadratic processes whose dynamic is given below,

$$X_{i,t} = b_i \mathbf{X}_{t-1}^T \mathbf{A}_i \mathbf{X}_{t-1} + N_{i,t},\ i = 1, ..., m, \tag{25}$$

where $\mathbf{A}_i \in \mathbb{R}^{m\times m}$, $\mathbf{X}_t = (X_{1,t}, X_{2,t}, ..., X_{m,t})^T$, $N_{i,t} \sim \mathcal{N}(0, \sigma^2)$, and $b_i \sim U(-0.9, 0.9)$. Note that the term $|[\mathbf{A}_i]_{j,k} + [\mathbf{A}_i]_{k,j}|$ captures the effect of $X_{j,t-1}X_{k,t-1}$ on $X_{i,t}$. Thus, it is possible to obtain the true parents of asset i as follows,

$$\mathcal{PA}_i = \{X_j : [\mathbf{1}^T \cdot \big(|\mathbf{A}_i^T + \mathbf{A}_i|\big)]_j > 0\}, \tag{26}$$

where $\mathbf{1}$ denotes all-one vector of length m. Each matrix $\mathbf{A}_i$ is simulated independently by following the similar procedure as in Section 4.1. In this experiment, since the model is non-linear, we could not apply (9) to estimate the DIs but instead we used the k-nearest method to estimate the mutual information and applied Equation (12).

Herein, we again compare the performances of the four different conditioning approaches. Figure 5 shows the precision-recall curves for these approaches in the quadratic model with $m = 15$. Precision-recall curves are a standard tools to illustrate and compare the performances of different learning methods. In this curve the precision is demonstrated in the y-axis vs. the recall on the x-axis for all potential values of the threshold α.

Similar to the linear setting, we use the *true parents* as a benchmark since it has the ideal performance. It is however important to emphasize that this conditioning approach has higher complexity compared to the others. This is because in the *true parent* approach, the size of the conditioning set is relatively larger than the other approaches.

For the *most correlated* approach, we use $k = 5$, i.e., the size of the conditioning set is five. With this method, we could slightly reduce the estimation complexity of the DIs compared to the *true parent* approach but this comes with the price of losing the performance. clearly, the performance of the *most correlated* approach can be improved by increasing k but this will increase the complexity.

The performance of the *ideal portfolio* approach (using the time series in Lemma 2 as the conditioning) is worse than all others which is not surprising as the model is no longer linear. This means that the information embedded in the linear portfolio is not sufficient to decide the non-linear influences among the time series.

Finally, as it is shown in Figure 5, the *RNN* approach outperforms the *most correlated* and the *ideal portfolio* approaches and it shows close performance to the *true parents* but with the size of the conditioning set equal to one. This result once more fortifies our claim that with an RNN we can summarize the information of the network into one time series and use it for

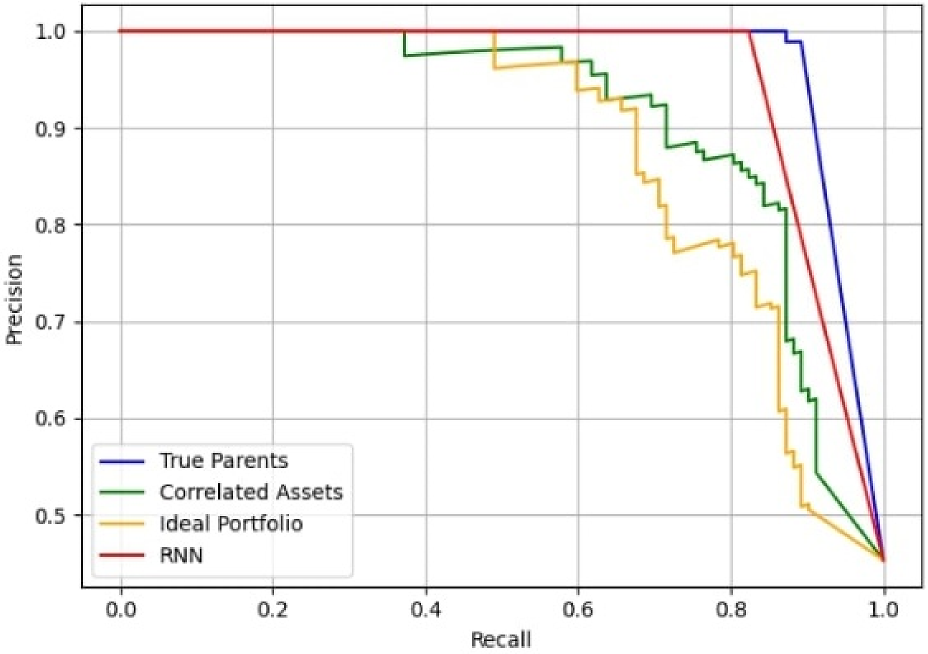

Figure 5: Precision-Recall curves for the quadratic model.

detecting the causal relationships. This claim is due to Lemma 3 and the universal approximation theorem which states that a neural network with enough hidden layers can approximate any non-linear function [19]. The slight difference between the performance of the *RNN* and the *true Parents* is because of the estimation error in the recurrent neural network.

4.3. Empirical DIG

This section describes how to apply our approach to empirical data and obtain the DIG of some US firms. We extracted the daily stock prices and the daily US Treasury rate as risk free returns from the CRSP database from 1990 to 2020. As the market is likely to evolve through these years, we chose to divide the dataset into six subsets, each of which has a length of five years and estimate the corresponding DIG of each subset separately.

Herein, we assume that the causal structure of the market evolved but its rate was slow enough such that during a period of five years, the DIG of the market remained unchanged.

For every subset, we keep the data of the 1000 firms with the highest maximum market capitalization and compute their excess return time series $X_{i,t}$, using the following relationship,

$$X_{i,t} = ln(P_{i,t}) - ln(P_{i,t-1}) - r_t, \tag{27}$$

where $P_{i,t}$ denotes the stock price of the firm i at time t and r_t is the risk free rate at time t. Afterwards, we apply Algorithm 1 with the excess returns as the input to estimate the corresponding DIG of each subset. We use the k-nearest neighbor method to estimate the DIs. We define the threshold α to be the unconditional mean across the estimated DIs. Note that in this experiment, the true DIGs of the market are not known, hence, we could not compute the precision-recall curves.

For the sake of presentation, instead of the complete DIGs with 1000 nodes, we draw the sub-graphs consisting of the 30 largest firms in Figures[13] 6, 7, and 8 . Each graph consists of 60 nodes illustrating the cause firm on the top hemisphere and the effect firm on the bottom hemisphere. For instance, if there is an edge between "from: AAPL" on the top and "to: GOOGL" on the bottom, it means that Apple influences Google. The dynamic evolution of the DIGs through time can often be explained by real events that happened in the market. For instance, in the DIG 2010-2014, Apple was not influencing General Electric (GE). However, on the 17th Oc-

[13] For a better presentation, interactive plots are available at https://marcaureledivernois.github.io/firm-network/

tober 2017, Apple announced a partnership with GE to bring Predix, GE's data and analytics platform, to their iPhones and iPads. We are able to capture this partnership in the DIG 2015-2019 as an edge is now present from Apple to GE. Another example is the announced collaboration between AT&T and Cisco to manage IoT devices and launch 5G service at the end of the 2010s: there was neither an edge from AT&T to Cisco nor from Cisco to AT&T during the first half of the 2010s, but the DIG for the second half of the 2010s shows a mutual influence, reflecting an increased relationship between the two companies.

Table 1 shows the Degree of Granger Causality (DGC) defined as the fraction of relationships in the network among all potential relationships. Formally,

$$DGC = \frac{1}{N^2} \sum_i \sum_j [\mathbf{DIG}]_{j,i}, \tag{28}$$

These results show that the DGC increased both in the DotCom bubble and in the Subprime Crisis, suggesting an increase of the connectedness in turmoil periods. This finding is consistent with [27] stating that correlation increases in bear markets.

Tables 2 and 3 show the outdegree and indegree of every firm in the six subsets. Outdegree is defined as the number of edges going out of a specific node. Indegree is the number of edges going to a specific node. These tables also reveal interesting facts. For instance, the SPY ticker, an ETF launched in 1993 and aiming at tracking the S&P500 return, enters in the 30 biggest market capitalizations in 2010 and has the highest number of outdegrees in the periods 2010-2014 and 2015-2019 but relatively low number of indegrees. This result suggests that the market return is influencing a high number of firms, but the converse is not necessarily true. One way to

1990-1994	1995-1999	2000-2004	2005-2009	2010-2014	2015-2019
0.23	0.27	0.18	0.28	0.32	0.25

Table 1: Degree of Granger Causality (DGC) for each sub-graph. DGC is defined as the fraction of relationships in the network among all potential relationships.

intuitively interpret this finding is that bullish and/or bearish markets are likely to influence next period stock returns (as a persistence effect known in business cycles), but individual stock returns struggle at predicting next period market returns.

5. Conclusion

In this work, we introduce an information-theoretic measure known as directed information that is capable of capturing nonlinear Granger-causality in an interactive system. We develop a novel algorithm based on recurrent neural network utilized with directed information. This algorithm can infer the interconnections within a large network with less complexity than previous works. As a proof of concept, we show that our approach performs well both in a linear and in a non-linear simulated environments. Finally, we apply this algorithm to infer the causal relationships among the major US firms during 1990 to 2020.

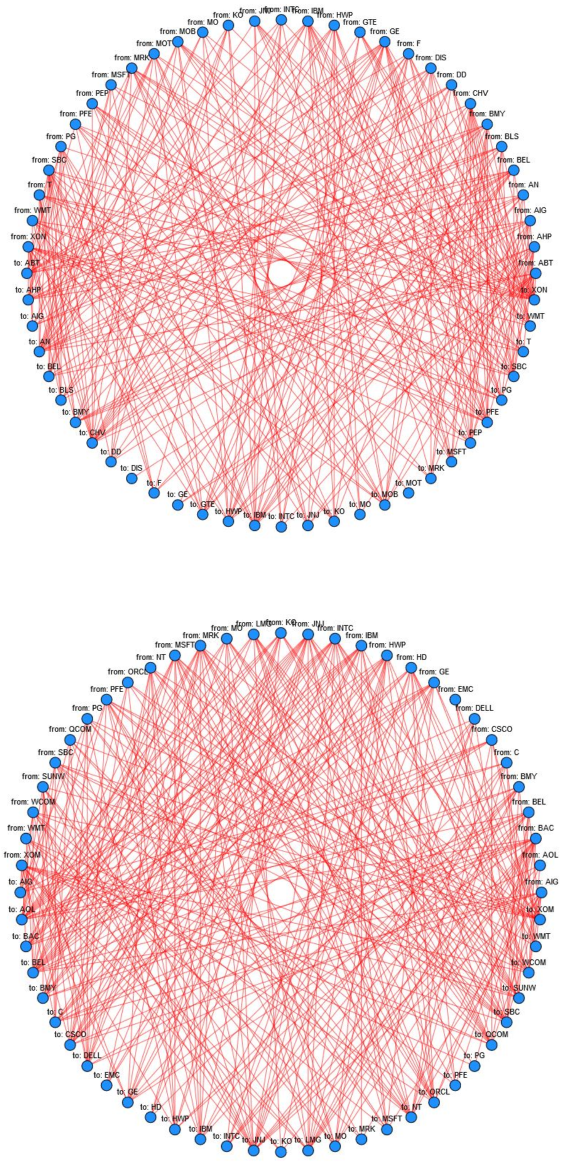

Figure 6: Empirical DIG for the periods 1990-1994 (top) and 1995-1999 (bottom). Interactive graphs can be found at https://marcaureledivernois.github.io/firm-network/

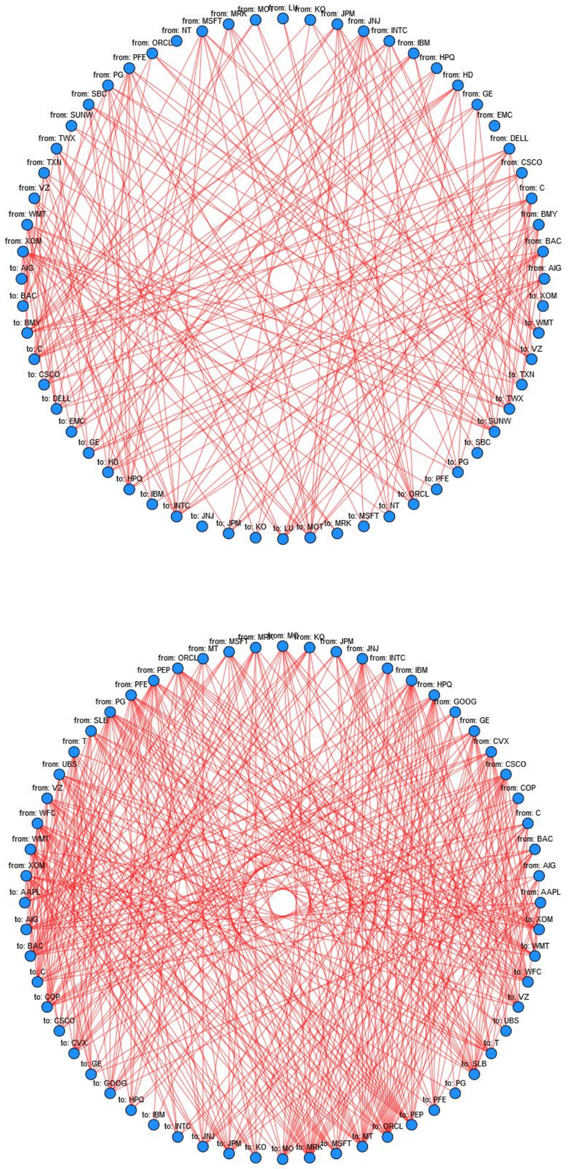

Figure 7: Empirical DIG for the periods 2000-2004 (top) and 2005-2009 (bottom). Interactive graphs can be found at https://marcaureledivernois.github.io/firm-network/

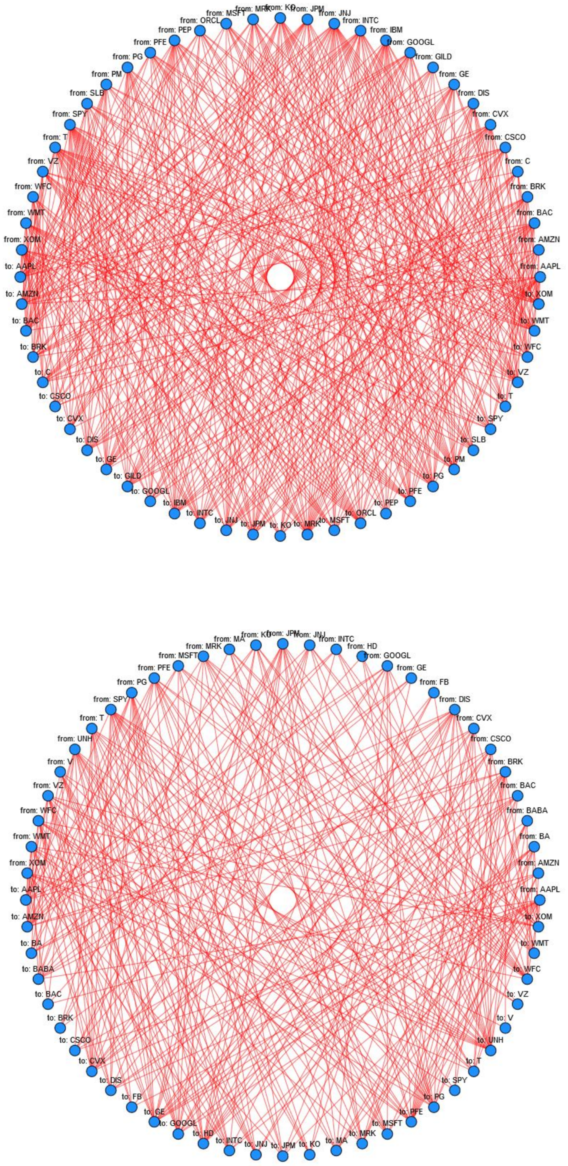

Figure 8: Empirical DIG for the periods 2010-2014 (top) and 2015-2019 (bottom). Interactive graphs can be found at https://marcaureledivernois.github.io/firm-network/

1990-1994		1995-1999		2000-2004		2005-2009		2010-2014		2015-2019	
Ticker	Out	Ticker	Out	Ticker	Out	Ticker	Out	Ticker	Out	Ticker	Out
SBC	13	JNJ	12	XOM	12	PEP	15	SPY	17	SPY	14
CHV	11	HWP	12	JNJ	10	WFC	13	T	15	DIS	12
XON	10	INTC	12	MSFT	9	CSCO	12	JNJ	15	BRK	12
PG	10	BAC	12	DELL	9	WMT	11	CVX	13	PFE	11
GE	10	PFE	11	WMT	9	IBM	11	AAPL	12	UNH	11
AN	9	IBM	11	C	9	PG	11	INTC	12	JPM	11
BEL	9	MSFT	11	HD	8	CVX	11	CSCO	12	AAPL	10
MRK	8	SBC	11	JPM	8	HPQ	10	GE	12	GOOGL	10
IBM	8	GE	10	PFE	8	VZ	10	IBM	11	WMT	9
BMY	8	MRK	10	PG	7	GE	9	GOOGL	11	CSCO	9
ABT	7	LMG	9	INTC	7	UBS	9	XOM	11	VZ	9
AHP	7	MO	9	BAC	7	PFE	9	JPM	10	BA	9
T	7	NT	9	BMY	6	C	9	KO	10	CVX	8
BLS	7	XOM	9	SBC	6	SLB	9	MRK	10	WFC	8
AIG	7	KO	9	IBM	5	MSFT	9	BRK	10	V	8
F	7	HD	8	TWX	5	ORCL	9	SLB	10	PG	8
HWP	7	BEL	8	AIG	5	MRK	8	WMT	10	T	7
MOT	7	BMY	8	CSCO	4	KO	8	VZ	9	JNJ	7
DD	6	SUNW	8	GE	4	GOOG	7	DIS	9	MRK	7
DIS	6	AIG	7	VZ	4	XOM	7	PFE	9	KO	6
MOB	6	WCOM	7	MRK	4	JNJ	7	BAC	8	XOM	5
PEP	6	CSCO	6	MOT	4	MO	7	PM	8	MSFT	5
MSFT	6	QCOM	6	TXN	4	JPM	6	ORCL	7	MA	5
INTC	6	C	6	HPQ	3	COP	6	PG	7	BABA	5
WMT	6	PG	6	ORCL	3	T	6	PEP	6	FB	4
GTE	5	DELL	5	KO	2	INTC	6	MSFT	6	BAC	4
MO	5	ORCL	4	SUNW	1	AAPL	6	C	5	INTC	4
JNJ	4	WMT	4	LU	1	BAC	5	WFC	5	GE	3
PFE	4	EMC	2	EMC	0	MT	4	AMZN	5	HD	3
KO	2	AOL	2	NT	0	AIG	4	GILD	5	AMZN	2

Table 2: Outdegrees (Out) ranked for each sub-graph. Outdegree is defined as the number of edges going out of a specific node.

1990-1994		1995-1999		2000-2004		2005-2009		2010-2014		2015-2019	
Tic	In	Tic	In	Tic	In	Tic	In	Tic	In	Tic	In
XON	22	BEL	15	MOT	12	MRK	19	MRK	18	UNH	17
CHV	17	JNJ	15	INTC	11	ORCL	19	PM	15	PFE	16
AN	13	AOL	14	C	11	BAC	16	JPM	15	HD	14
SBC	12	XOM	14	SUNW	11	SLB	14	ORCL	14	AMZN	14
HWP	11	NT	12	ORCL	10	WFC	13	IBM	14	GOOGL	12
PEP	10	QCOM	12	BMY	9	JPM	12	XOM	13	GE	12
AHP	10	SUNW	11	DELL	9	COP	12	PG	13	PG	12
MOB	10	ORCL	11	TWX	7	MT	11	JNJ	12	CVX	11
GTE	9	CSCO	11	HPQ	7	PEP	10	DIS	12	BABA	9
IBM	9	C	11	GE	7	C	10	INTC	11	MRK	9
BMY	8	DELL	10	CSCO	7	MO	10	WFC	10	DIS	9
ABT	8	SBC	10	LU	6	CVX	10	C	10	WFC	8
KO	7	BMY	9	HD	6	AIG	9	VZ	10	BA	7
BEL	7	WCOM	9	TXN	6	T	9	WMT	10	MA	7
MSFT	7	GE	9	EMC	6	JNJ	8	GE	10	AAPL	7
BLS	6	IBM	9	BAC	5	GOOG	7	KO	9	T	7
PFE	6	MO	9	PG	5	KO	7	BAC	9	WMT	6
DD	6	LMG	8	WMT	5	VZ	6	PFE	9	MSFT	6
JNJ	5	HWP	6	VZ	5	MSFT	6	AAPL	9	INTC	6
WMT	4	MSFT	6	JPM	4	INTC	6	SPY	8	FB	5
F	4	BAC	5	KO	3	AAPL	6	CVX	7	KO	5
MRK	4	INTC	5	NT	3	GE	6	AMZN	7	JPM	4
PG	4	WMT	5	XOM	2	WMT	6	GILD	7	XOM	4
AIG	4	PFE	4	MSFT	2	XOM	5	CSCO	7	JNJ	4
T	3	AIG	4	AIG	2	PFE	5	GOOGL	7	SPY	4
GE	2	KO	3	MRK	1	CSCO	4	MSFT	6	BRK	4
MO	2	MRK	2	PFE	1	HPQ	4	PEP	5	CSCO	3
INTC	2	PG	2	SBC	1	UBS	2	SLB	5	BAC	2
DIS	1	HD	2	IBM	0	IBM	1	T	4	V	1
MOT	1	EMC	1	JNJ	0	PG	1	BRK	4	VZ	1

Table 3: Indegrees (In) ranked for each sub-graph. Indegree is defined as the number of edges going to a specific node.

Appendix A: Technical proofs

This section presents the proofs.

Proof of Lemma 1:

From the definition of DIG, we know that for any $R_k \notin \mathcal{PA}_j$, $I(R_k \rightarrow R_j || \mathcal{R}_{-\{k,j\}}) = 0$. This implies that for all t,

$$p(R_{j,t}|\mathcal{R}^{t-1}) = p(R_{j,t}|\mathcal{R}^{t-1}_{-\{k\}}). \tag{.1}$$

On the other hand, by the assumption of the Lemma, $R_i \notin \mathcal{PA}_j$, we have

$$I(R_i \rightarrow R_j || \mathcal{R}_{-\{i,j\}}) = 0, \tag{.2}$$

or equivalently, for all t,

$$p(R_{j,t}|\mathcal{R}^{t-1}) = p(R_{j,t}|\mathcal{R}^{t-1}_{-\{i\}}). \tag{.3}$$

Combining (.1) and (.3) imply that for any pair $\{R_i, R_k\}$ that are not in the parent set of R_j, we have

$$p(R_{j,t}|\mathcal{R}^{t-1}_{-\{k\}}) = p(R_{j,t}|\mathcal{R}^{t-1}_{-\{i\}}). \tag{.4}$$

To prove the claim of this lemma, we use (.4) to show that all the time series in $\mathcal{R}_{-\{i,j\}} \setminus \mathcal{C}$ can be removed from the conditioning in (.2). Let $R_k \in \mathcal{R}_{-\{i,j\}} \setminus \mathcal{C}$, by multiplying the above equality with $p(R_i^{t-1}|\mathcal{R}^{t-1}_{-\{i,k\}})$ and marginalizing over R_i^{t-1}, we obtain

$$\int p(R_{j,t}|\mathcal{R}^{t-1}_{-\{k\}}) p(R_i^{t-1}|\mathcal{R}^{t-1}_{-\{i,k\}}) dR_i^{t-1} = p(R_{j,t}|\mathcal{R}^{t-1}_{-\{i,k\}})$$
$$= \int p(R_{j,t}|\mathcal{R}^{t-1}_{-\{i\}}) p(R_i^{t-1}|\mathcal{R}^{t-1}_{-\{i,k\}}) dR_i^{t-1} = p(R_{j,t}|\mathcal{R}^{t-1}_{-\{i\}}).$$

The above equalities and (.3) imply that for all t,

$$p(R_{j,t}|\mathcal{R}^{t-1}) = p(R_{j,t}|\mathcal{R}^{t-1}_{-\{i,k\}}),$$

or equivalently,

$$I(R_i \to R_j||\mathcal{R}_{-\{i,j,k\}}) = 0.$$

By repeating the above procedure, we obtain

$$I(R_i \to R_j||\mathcal{C}) = 0.$$

Proof of Lemma 2:

Consider the VAR model in (14). First, we assume that X_i has no influence on X_j, i.e., $I(X_i \to X_j||\mathcal{X}_{-\{i,j\}}) = 0$ or equivalently $a_{j,i} = 0$ and show that (17) holds. Given this assumption, we have that for all t,

$$p(X_{j,t}|\mathcal{X}^{t-1}_{-\{i\}}) = p(X_{j,t}|\mathcal{X}^{t-1}).$$

Using the equations in (14) and the assumption that $a_{j,i} = 0$, we obtain

$$p(X_{j,t}|\mathcal{X}^{t-1}) = p(N_{j,t} + \sum_{k} a_{j,k}X_{k,t-1}|\mathcal{X}^{t-1}) = p(N_{j,t} + \sum_{k\neq i} a_{j,k}X_{k,t-1}|\mathcal{X}^{t-1})$$

$$= p(N_{j,t} + \sum_{k\neq i} a_{j,k}X_{k,t-1}|\sum_{k\neq i} a_{j,k}X_{k,t-1}, X_i^{t-1}, X_j^{t-1})$$

$$= p(N_{j,t} + \sum_{k\neq i} a_{j,k}X_{k,t-1}|\sum_{k\neq i} a_{j,k}X_{k,t-1}, X_j^{t-1}).$$

Note that we could replace $\mathcal{X}^{t-1}$ by $\{\sum_{k\neq i} a_{j,k}X_{k,t-1}, X_i^{t-1}, X_j^{t-1}\}$ or $\{\sum_{k\neq i} a_{j,k}X_{k,t-1}, X_j^{t-1}\}$ in the above equations, because given either of them $\sum_{k\neq i} a_{j,k}X_{k,t-1}$ becomes a constant and independent of $N_{j,t}$. By defining $Q_{t-1} := \sum_{k\neq i} a_{j,k}X_{k,t-1}$, the above equations can be rewritten as follows

$$p(X_{j,t}|\mathcal{X}^{t-1}) = p(X_{j,t}|Q_{t-1}, X_i^{t-1}, X_j^{t-1}) = p(X_{j,t}|Q_{t-1}, X_j^{t-1}), \forall t,$$

or equivalently,

$$\mathbb{E}\left[\log\frac{p(X_{j,t}|Q_{t-1},X_i^{t-1},X_j^{t-1})}{p(X_{j,t}|Q_{t-1},X_j^{t-1})}\right]=0,\forall t.$$

Using the definition of DI, the above equalities can be written in terms of DI as follows

$$I(X_i\to X_j||Q)=0.$$

On the other hand, we have

$$[a_{j,1},...,a_{j,i-1},a_{j,i+1},...,a_{j,m}]=\arg\min_{\mathbf{w}\in\mathbb{R}^{m-1}}\mathbb{E}\left[||X_{j,t}-\mathbf{w}^T\mathbf{X}_{-\{i\},t-1}||_2^2\right]:=\mathbf{u}_t.$$

where $\mathbf{X}_{-\{i\},t-1}:=[X_{1,t-1},...,X_{i-1,t-1},X_{i+1,t-1},...,X_{m,t-1}]^T$. This means that $Q_{t-1}=\mathbf{u}_t^T\mathbf{X}_{-\{i\},t-1}$.

Next, we show the reverse direction, i.e., we assume (17) holds, then we show $I(X_i\to X_j||\mathcal{X}_{-\{i,j\}})=0$. To do so, it suffices to show $a_{j,i}=0$. Since (17) holds, we have

$$p(X_{j,t}|\mathbf{u}_t^T\mathbf{X}_{-\{i\},t-1},X_i^{t-1},X_j^{t-1})=p(X_{j,t}|\mathbf{u}_t^T\mathbf{X}_{-\{i\},t-1},X_j^{t-1}),\forall t.$$

Using the j-th equation of (14) and the above equalities, for any instances $(\mathbf{u}_t^T\mathbf{x}_{-\{i\},t-1},x_i^{t-1},x_j^{t-1})$ of $(\mathbf{u}_t^T\mathbf{X}_{-\{i\},t-1},X_i^{t-1},X_j^{t-1})$, we obtain $\forall t$,

$$\mathbb{E}\left[X_{j,t}|\mathbf{u}_t^T\mathbf{x}_{-\{i\},t-1},x_i^{t-1},x_j^{t-1}\right]=\mathbb{E}\left[X_{j,t}|\mathbf{u}_t^T\mathbf{x}_{-\{i\},t-1},x_j^{t-1}\right],$$

which implies

$$\begin{aligned}&\mathbb{E}\left[N_{j,t}\right]+\sum_{k\neq i}a_{j,k}x_{k,t-1}+a_{j,i}x_{i,t-1}=\\&\mathbb{E}\left[N_{j,t}\right]+\sum_{k\neq i}a_{j,k}x_{k,t-1}+a_{j,i}\mathbb{E}\left[X_{i,t-1}|\mathbf{u}_t^T\mathbf{x}_{-\{i\},t-1},x_j^{t-1}\right].\end{aligned}$$

This simplifies to

$$a_{j,i}x_{i,t-1}=a_{j,i}\mathbb{E}\left[X_{i,t-1}|\mathbf{u}_t^T\mathbf{x}_{-\{i\},t-1},x_j^{t-1}\right],\forall t.$$

This equation should hold for any $x_{i,t-1}$. This is only possible if $a_{j,i}=0$.

Proof of Lemma 3:

The proof is similar to the linear version and uses the fact that exogenous noises $\{\varepsilon_{j,t}\}$ are independent. More precisely, we have

$$p(X_{j,t}|\mathcal{X}^{t-1}) = p(F_j(\mathcal{X}^{t-1}) + \varepsilon_{j,t}|F_j(\mathcal{X}^{t-1})).$$

Since there is no influence from X_i to X_j, we can eliminate it from the conditioning and the argument of function F_j and obtain

$$p\big(F_j(\mathcal{X}^{t-1}) + \varepsilon_{j,t}|\mathcal{X}^{t-1}\big) = p\big(F_j(\mathcal{X}^{t-1}_{-\{i\}}) + \varepsilon_{j,t}|\mathcal{X}^{t-1}_{-\{i\}}\big).$$

On the other hand, because given either $\{F_j(\mathcal{X}^{t-1}_{-\{i\}}), X_i^{t-1}, X_j^{t-1}\}$ or $\{F_j(\mathcal{X}^{t-1}_{-\{i\}}), X_j^{t-1}\}$, the value of $F_j(\mathcal{X}^{t-1}_{-\{i\}})$ is no longer a random variable. Using this relationship and the fact that $\varepsilon_{j,t}$ is independent of $\mathcal{X}^{t-1}$, we obtain

$$p\big(F_j(\mathcal{X}^{t-1}_{-\{i\}}) + \varepsilon_{j,t}|F_j(\mathcal{X}^{t-1}_{-\{i\}}), X_i^{t-1}, X_j^{t-1}\big) = p\big(F_j(\mathcal{X}^{t-1}_{-\{i\}}) + \varepsilon_{j,t}|F_j(\mathcal{X}^{t-1}_{-\{i\}}), X_j^{t-1}\big).$$

By defining $Q_{t-1} := F_j(\mathcal{X}^{t-1}_{-\{i\}})$, the above equations can be rewritten in terms of DI as follows,

$$I(X_i \to X_j||Q) = 0.$$

To show the reverse, we need to prove that $I(X_i \to X_j||\mathcal{X}_{-\{i,j\}}) = 0$ if Equation (19) holds. Because $I(X_i \to X_j||Q) = 0$ and using Equation (18), for all t, we have

$$p\big(F_j(\mathcal{X}^{t-1}) + \varepsilon_{j,t}|Q_{t-1}, X_i^{t-1}, X_j^{t-1}\big) = p\big(F_j(\mathcal{X}^{t-1}) + \varepsilon_{j,t}|Q_{t-1}, X_j^{t-1}\big),$$

where $Q_{t-1} = F_j(\mathcal{X}^{t-1}_{-\{i\}})$. Note that the conditioning on the right-hand-side distribution is independent of X_i^{t-1}. This implies that function F_j does not depend on X_i. Therefore, we can remove X_i^{t-1} from the argument of F_j, i.e.,

$$X_{j,t} = F_j(\mathcal{X}^{t-1}) + \varepsilon_{j,t} = F_j(\mathcal{X}^{t-1}_{-\{i\}}) + \varepsilon_{j,t},$$

which further implies

$$p(F_j(\mathcal{X}^{t-1}_{-\{i\}}) + \varepsilon_{j,t}|\mathcal{X}^{t-1}_{-\{i,j\}}, X_i^{t-1}, X_j^{t-1}) = p(F_j(\mathcal{X}^{t-1}_{-\{i\}}) + \varepsilon_{j,t}|\mathcal{X}^{t-1}_{-\{i,j\}}, X_j^{t-1}).$$

This is equivalent to

$$I(X_i \to X_j||\mathcal{X}_{-\{i,j\}}) = 0.$$

Appendix B : k-nearest Estimator

Suppose that $N+M$ i.i.d. realizations $\{\mathbf{x}_1, ..., \mathbf{x}_{N+M}\}$ are available from $\mathbb{P}_{X,Y,Z}$, where $\mathbf{x}_i$ denotes the i-th realization of (X, Y, Z). The data is randomly divided into two subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ of N and M points, respectively. The estimator has two main stages: In the first stage, a k-nearest density estimator $\widehat{\mathbb{P}}_{X,Y,Z}$ at the N points of $\mathcal{S}_1$ is estimated using the M realizations of $\mathcal{S}_2$ as follows:

Let $d(\mathbf{x}, \mathbf{y}) \in \mathbb{R}$ denote the Euclidean distance between points $\mathbf{x}$ and $\mathbf{y}$ and $d_k(\mathbf{x}) \in \mathbb{R}$ denotes the Euclidean distance between a point $\mathbf{x}$ and its k-th nearest neighbor among $\mathcal{S}_2$. The k-nearest region is $\mathcal{S}_k(\mathbf{x}) := \{\mathbf{y} : d(\mathbf{x}, \mathbf{y}) \leq d_k(\mathbf{x})\}$ and the volume of this region is $V_k(\mathbf{x}) := \int_{\mathcal{S}_k(\mathbf{x})} 1 dv$. The standard k-nearest density estimator of [39] is defined as

$$\widehat{\mathbb{P}}_{X,Y,Z}(\mathbf{x}) := \frac{k-1}{M \cdot V_k(\mathbf{x})}.$$

Similarly, we obtain k-nearest density estimators $\widehat{\mathbb{P}}_{X,Z}$, $\widehat{\mathbb{P}}_{Y,Z}$, and $\widehat{\mathbb{P}}_Z$. Subsequently, the N samples of $\mathcal{S}_1$ is used to approximate the conditional mutual information:

$$\widehat{I}(X;Y|Z) := \frac{1}{N}\sum_{i\in\mathcal{S}_1} \log \widehat{\mathbb{P}}_{X,Y,Z}(\mathbf{x}_i) + \log \widehat{\mathbb{P}}_{Z}(\mathbf{x}_i) - \log \widehat{\mathbb{P}}_{X,Z}(\mathbf{x}_i) - \log \widehat{\mathbb{P}}_{Y,Z}(\mathbf{x}_i).$$

For more details corresponding this estimator including its bias, variance, and confidence, please see the works by [39] and [26].

Appendix C : Koopman-based Lifting

Let $\mathbf{X}_t := \{X_{1,t}, ..., X_{m,t}\}$ denote a network of m time series such that

$$\dot{\mathbf{X}}_t = F(\mathbf{X}_t), \tag{.5}$$

where the vector field $F(\mathbf{X}) = (F_1(\mathbf{X}), ..., F_m(\mathbf{X}))$ is of the form

$$F_j(\mathbf{X}) = \sum_{k=1}^{K} w_{j,k} h_k(\mathbf{X}). \tag{.6}$$

In the above equation, $w_{j,k} \in \mathbb{R}$ are unknown weights and $\{h_k(\mathbf{X})\}$ denote a set of known library functions, e.g., monomials. Furthermore, let $\varphi^t(\mathbf{X}_0)$ denote the solution to (.5) associated with the initial condition $\mathbf{X}_0$.

Now, suppose that we have N noisy observations $\{(\mathbf{x}_1, \mathbf{y}_1), ..., (\mathbf{x}_N, \mathbf{y}_N)\}$ of the system trajectory, where $\mathbf{x}_i$ is the initial point and $\mathbf{y}_i$ is the final point after T_s steps, i.e.,

$$\mathbf{y}_i - \epsilon_i = \varphi^{T_s}(\mathbf{x}_i - \varepsilon_i), \quad i = 1, ..., N,$$

where ϵ_i and ε_i are the measurement noises. The goal is to estimate the weights $\{w_{j,k}\}$ using these observations and consequently infer the causal network among the time series. To do so, we use the Koopman approach [29] that lifts the observation space to another space in which the relationships are linear. More precisely, the steps are as follows:

- Select a set of M basis lifting functions $\{p_1(\mathbf{x}), ..., p_M(\mathbf{x})\}$, and lift the observations,

$$\mathbf{P}_x := \begin{pmatrix} p_1(\mathbf{x}_1) & \cdots & p_M(\mathbf{x}_1) \\ p_1(\mathbf{x}_2) & \cdots & p_M(\mathbf{x}_2) \\ \vdots & \ddots & \vdots \\ p_1(\mathbf{x}_N) & \cdots & p_M(\mathbf{x}_N) \end{pmatrix}, \mathbf{P}_y := \begin{pmatrix} p_1(\mathbf{y}_1) & \cdots & p_M(\mathbf{y}_1) \\ p_1(\mathbf{y}_2) & \cdots & p_M(\mathbf{y}_2) \\ \vdots & \ddots & \vdots \\ p_1(\mathbf{y}_N) & \cdots & p_M(\mathbf{y}_N) \end{pmatrix}. \tag{.7}$$

- Identify the Koopman operator $\mathbf{L} := \frac{1}{T_s} \log(\mathbf{P}_x^\dagger \mathbf{P}_y)$, where $\mathbf{P}_x^\dagger$ denotes the pseudo-inverse of $\mathbf{P}_x$ and the function log denotes the (principal) matrix logarithm.

- Identify the weights using the following equations: $\widehat{w}_{k,j} := [\mathbf{L}]_{k,l}$, with l such that $p_l(\mathbf{x}) = x_j$, where $\mathbf{x} = (x_1, ..., x_m)$.

Dual Lifting Method

An alternative approach to obtain the weights is the dual lifting method which executes the following steps instead of the above last step. At first, it finds matrix $\widehat{\mathbf{F}}$ using the following equation,

$$\widehat{\mathbf{F}}_{N\times m} := \mathbf{L}_{N\times N} \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{pmatrix}_{N\times m}.$$

Next, it constructs

$$\mathbf{H}_x := \begin{pmatrix} p_1(\mathbf{x}_1) & \cdots & p_M(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ p_1(\mathbf{x}_N) & \cdots & p_M(\mathbf{x}_N) \end{pmatrix}_{N\times M},$$

and for each j, solve the following regression problem to get the weights

$$\widehat{\mathbf{w}}_j := \arg\min_{\mathbf{w}\in\mathbb{R}^M} \left|\left|\mathbf{H}_x\mathbf{w} - \widehat{\mathbf{F}}_{:,j}\right|\right|_2^2 + \rho||\mathbf{w}||_1,$$

where $\widehat{\mathbf{F}}_{:,j}$ denotes the j-th column of matrix $\widehat{\mathbf{F}}$ and $\widehat{\mathbf{w}}_j = [\hat{w}_{j,1}, ..., \hat{w}_{j,M}]^T$.

Appendix D : Ideal portfolio

In this section, we show how the ideal portfolio is related to the coefficients of the linear system in (14). Recall the optimization problem in Lemma 2.

$$\begin{aligned}&\mathbf{u}_t := \arg\min_{\mathbf{w}\in\mathbb{R}^{m-1}} \mathbb{E}\left[||X_{j,t} - \mathbf{w}^T\mathbf{X}_{-\{i\},t-1}||_2^2\right],\\ &\mathbf{X}_{-\{i\},t-1} := [X_{1,t-1}, \cdots, X_{i-1,t-1}, X_{i+1,t-1}, \cdots, X_{m,t-1}]^T.\end{aligned}$$

Consider the j-th Equation in (14), i.e.,

$$X_{j,t} = \sum_{k=1}^{m} a_{j,k}X_{k,t-1} + N_{j,t}.$$

If $a_{j,i} = 0$, by substituting the above equation into the optimization, we obtain

$$\begin{aligned}&\min_{\mathbf{w}\in\mathbb{R}^{m-1}} \mathbb{E}\left[||\sum_{k\neq i}(a_{j,k} - w_k)X_{k,t-1} + N_{j,t}||_2^2\right] = \min_{\mathbf{w}\in\mathbb{R}^{m-1}} \mathbb{E}\left[||N_{j,t}||_2^2\right]\\ &+ \mathbb{E}\left[||\sum_{k\neq i}(a_{j,k} - w_k)X_{k,t-1}||_2^2\right] + 2\mathbb{E}\left[||\left(\sum_{k\neq i}(a_{j,k} - w_k)X_{k,t-1}\right)N_{j,t}||_2^2\right]\\ &= \min_{\mathbf{w}\in\mathbb{R}^{m-1}} \mathbb{E}\left[||\sum_{k\neq i}(a_{j,k} - w_k)X_{k,t-1}||_2^2\right] + \mathbb{E}\left[||N_{j,t}||_2^2\right]\end{aligned}$$

The last equality is due to the fact $N_{j,t}$ is independent of $\{X_{k,t-1}\}$ and have zero mean. This implies that the solution is $w_k = a_{j,k}$ for $k \in \{1, ..., i-1, i+1, ..., m\}$.

References

[1] F. Allen, A. Babus, and E. Carletti. Financial connections and systemic risk. Technical report, National Bureau of Economic Research, 2010. 12

[2] H. Altinbas and O. T. Biskin. Selecting macroeconomic influencers on stock markets by using feature selection algorithms. *Procedia Economics and Finance*, 2015. doi: http://dx.doi.org/10.1016/S2212-5671(15)01251-4. 3

[3] M. Barigozzi and M. Hallin. A network analysis of the volatility of high dimensional financial series. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2016. doi: https://doi.org/10.1111/rssc.12177. 2

[4] M. Bernardi and M. Costola. High-dimensional sparse financial networks through a regularised regression model. *SAFE Working Paper*, 2019. doi: https://dx.doi.org/10.2139/ssrn.3342240. 5

[5] D. Bianchi, M. Billio, R. Casarin, and M. Guidolin. Modeling systemic risk with markov switching graphical sur models. *Journal of Econometrics*, 2019. doi: https://doi.org/10.1016/j.jeconom.2018.11.005. 4

[6] M. Billio, M. Getmansky, A. W. Lo, and L. Pelizzon. Measuring systemic risk in the finance and insurance sectors. *MIT Sloan School of Management Working Paper*, 2010. doi: http://hdl.handle.net/1721.1/66679. 12

[7] M. Billio, M. Getmansky, A. W. Lo, and L. Pelizzon. Econometric measures of connectedness and systemic risk in the finance and in-

surance sectors. *Journal of Financial Economics*, 2012. doi: https://doi.org/10.1016/j.jfineco.2011.12.010. 2, 12

[8] M. Billio, R. Casarin, and L. Rossini. Bayesian nonparametric sparse var models. *Journal of Econometrics*, 2019. doi: https://doi.org/10.1016/j.jeconom.2019.04.022. 5

[9] G. Bonaccolto, M. Caporin, and R. Panzica. Estimation and model-based combination of causality networks among large us banks and insurance companies. *Journal of Empirical Finance*, 2019. doi: https://doi.org/10.1016/j.jempfin.2019.08.008. 4

[10] F. Corsi, F. Lillo, D. Pirino, and L. Trapin. Measuring the propagation of financial distress with granger-causality tail risk networks. *Journal of Financial Stability*, 2018. doi: https://doi.org/10.1016/j.jfs.2018.06.003. 4

[11] T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley & Sons, 2012. 9, 13

[12] W. J. Culver. On the existence and uniqueness of the real logarithm of a matrix. *Proceedings of the American Mathematical Society*, 1966. doi: https://doi.org/10.2307/2036109. 21

[13] F. X. Diebold and K. Yılmaz. On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of Econometrics*, 2014. doi: https://doi.org/10.1016/j.jeconom.2014.04.012. 2

[14] J. Etesami and N. Kiyavash. Directed information graphs: A generalization of linear dynamical graphs. In *American Control Conference*.

IEEE, 2014. doi: https://doi.org/10.1109/ACC.2014.6859362. 7, 12, 14, 17

[15] J. Etesami, A. Habibnia, and N. Kiyavash. Econometric modeling of systemic risk: A time series approach. Unpublished results. doi: https://doi.org/10.475/123_4. 7, 15

[16] C. W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 1969. doi: https://doi.org/10.2307/1912791. 2, 6

[17] C. W. J. Granger. Economic processes involving feedback. *Information and control*, 1963. doi: https://doi.org/10.1016/S0019-9958(63)90092-5. 7

[18] Y. Hong, Y. Liu, and S. Wang. Granger causality in risk and detection of extreme risk spillover between financial markets. *Journal of Econometrics*, 2009. doi: https://doi.org/10.1016/j.jeconom.2008.12.013. 4

[19] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 1989. doi: https://doi.org/10.1016/0893-6080(89)90020-8. 21, 29

[20] C.-L. Huang and C.-J. Wang. A ga-based feature selection and parameters optimization for support vector machines. *Expert Systems with Applications*, 2006. doi: https://doi.org/10.1016/j.eswa.2005.09.024. 3

[21] M. Iacopini and L. Rossini. Bayesian nonparametric graphical models for time-varying parameters var. Unpublished results. doi: http://dx.doi.org/10.2139/ssrn.3400078. 5

[22] J. Jiao, H. H. Permuter, L. Zhao, Y.-H. Kim, and T. Weissman. Universal estimation of directed information. *Information Theory*, 2013. doi: https://doi.org/10.1109/TIT.2013.2267934. 14

[23] M. Kalli and J. E. Griffin. Bayesian nonparametric vector autoregressive models. *Journal of econometrics*, 2018. doi: https://doi.org/10.1016/j.jeconom.2017.11.009. 5

[24] S. Kim, D. Putrino, S. Ghosh, and E. N. Brown. A granger causality measure for point process models of ensemble neural spiking activity. *PLoS computational biology*, 2011. doi: https://doi.org/10.1371/journal.pcbi.1001110. 7

[25] B. O. Koopman. Hamiltonian systems and transformation in hilbert space. *Proceedings of the national academy of sciences of the united states of america*, 1931. doi: https://dx.doi.org/10.1073/pnas.17.5.315. 20

[26] D. O. Loftsgaarden, C. P. Quesenberry, et al. A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics*, 1965. doi: https://doi.org/10.1214/aoms/1177700079. 43

[27] F. Longin and B. Solnik. Extreme correlation of international equity markets. *The Journal of Finance*, 2001. doi: https://doi.org/10.1111/0022-1082.00340. 31

[28] J. Massey. Causality, feedback and directed information. In *Proc. Int. Symp. Inf. Theory Applic.(ISITA-90)*, 1990. 7

[29] A. Mauroy and J. Goncalves. Koopman-based lifting techniques for nonlinear systems identification. *IEEE Transactions on Automatic*

Control, 2019. doi: https://doi.org/10.1109/TAC.2019.2941433. 20, 43

[30] M. Noshad, Y. Zeng, and A. O. Hero. Scalable mutual information estimation using dependence graphs. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019. doi: https://doi.org/10.1109/ICASSP.2019.8683351. 14

[31] L. Paninski. Estimation of entropy and mutual information. *Neural computation*, 2003. doi: https://doi.org/10.1162/089976603321780272. 14

[32] K. Petrova. A quasi-bayesian local likelihood approach to time varying parameter var models. *Journal of Econometrics*, 2019. doi: https://doi.org/10.1016/j.jeconom.2019.04.031. 5

[33] S. Piramuthu. Evaluating feature selection methods for learning in data mining applications. *European Journal of Operational Research*, 2004. doi: https://doi.org/10.1016/S0377-2217(02)00911-6. 3

[34] Z. Psaradakis, M. O. Ravn, and M. Sola. Markov switching causality and the money–output relationship. *Journal of Applied Econometrics*, 2005. doi: https://doi.org/10.1002/jae.819. 4

[35] C. Quinn, N. Kiyavash, and T. P. Coleman. Directed information graphs. *Transactions on Information Theory*, 2015. doi: https://doi.org/10.1109/TIT.2015.2478440. 7, 13

[36] C. J. Quinn, T. P. Coleman, N. Kiyavash, and N. G. Hatsopoulos. Estimating the directed information to infer causal relationships in en-

semble neural spike train recordings. *Journal of computational neuroscience*, 2011. doi: https://doi.org/10.1007/s10827-010-0247-2. 7

[37] C. J. Quinn, N. Kiyavash, and T. P. Coleman. Efficient methods to compute optimal tree approximations of directed information graphs. *IEEE Transactions on Signal Processing*, 2013. doi: https://doi.org/10.1109/TSP.2013.2259161. 15

[38] C. J. Quinn, A. Pinar, and N. Kiyavash. Bounded degree approximations of stochastic networks. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 2017. doi: https://doi.org/10.1109/TMBMC.2017.2686387. 15

[39] K. Sricharan, R. Raich, and A. O. Hero. k-nearest neighbor estimation of entropies with confidence. In *Information Theory Proceedings*, 2011. doi: https://doi.org/10.1109/ISIT.2011.6033726. 14, 42, 43

[40] C.-F. Tsai. Feature selection in bankruptcy prediction. *Knowledge-Based Systems*, 2009. doi: https://doi.org/10.1016/j.knosys.2008.08.002. 3

[41] N. Wiener. The theory of prediction. *Modern mathematics for engineers*, 1956. 6

[42] H. Yuqinq, F. Kamaladdin, and W. Lipo. Feature selection for stock market analysis. *Lecture Notes in Computer Science*, 2013. doi: http://dx.doi.org/10.1007/978-3-642-42042-9_91. 4