An Examination of Practical Granger Causality Inference

Mohammad Taha Bahadori* Yan Liu†

Abstract

Learning temporal causal structures among multiple time series is one of the major tasks in mining time series data. Granger causality is one of the most popular techniques in uncovering the temporal dependencies among time series; however it faces two main challenges: (i) the spurious effect of unobserved time series and (ii) the computational challenges in high dimensional settings. In this paper, we utilize the confounder path delays to find a subset of time series that via conditioning on them we are able to cancel out the spurious confounder effects. After study of consistency of different Granger causality techniques, we propose Copula-Granger and show that while it is consistent in high dimensions, it can efficiently capture non-linearity in the data. Extensive experiments on a synthetic and a social networking dataset confirm our theoretical results.

1 Introduction

In the era of data deluge, we are confronted with large-scale time series data, i.e., a sequence observations of concerned variables over a period of time. For example, terabytes of neural activity time series data are produced to record the collective response of neurons to different stimuli; petabytes of climate and meteorological data, such as temperature, solar radiation, and precipitation, are collected over the years; and exabytes of social media contents are generated over time on the Internet. A major data mining task for time series data is to uncover the temporal causal relationship among the time series. For example, in the climatology, we want to identify the factors that impact the climate patterns of certain regions. In social networks, we are interested in identification of the patterns of influence among users and how topics activate or suppress each other. Developing effective and scalable data mining algorithms to uncover temporal dependency structures between time series and reveal insights from data has become a key problem in machine learning and data mining.

There are two major challenges in discovering temporal causal relationship in large-scale data: (i) not all influential confounders are observed in the datasets and (ii) enormous number of high dimensional time series need to be analyzed. The first challenge stems from the fact that in most datasets not all confounders are measured. Some confounders cannot even be measured easily which makes the spurious effects of unobserved confounders inevitable. The question in these situations is how can we utilize the prior knowledge about the unmeasured confounders to take into account their impact. The second challenge requires us to design scalable discovery algorithms that are able to uncover the temporal dependency among millions of time series with short observations.

Granger Causality [10] is one of the earliest methods developed to quantify the temporal-causal effect among time series. It is based on the common conception that the cause usually occurs prior to its effect. Formally, X Granger causes Y if its past value can help to predict the future value of Y beyond what could have been done with the past value of Y only. It has gained tremendous success across many domains due to its simplicity, robustness, and extendability [3, 4, 12, 18, 21]. Granger causality, similar to other causality discovery algorithms is also posed to the two data challenges. Spirtes et al [27, ch. 12] in the open problems section of their book describe the challenges in Granger causality as following: “First, tests of regression parameters waste degrees of freedom at the cost in small samples of power against alternatives. Since in many cases the number of observations is of the order of the number of parameters, whatever can be done to increase reliability should be. Second, it appears that while the time series setting removes ambiguities about the direction of dependencies, or edges, it does not remedy problems about unmeasured common causes of the outcome and regressors, and thus even asymptotically regression may yield significant coefficients for variables that are neither direct nor indirect causes of the outcome.”

In this paper we address both issues. In attempt to cancel out the effects of unobserved confounders, authors in [6, 7, 8] have extended Pearl’s criteria [22] for determining a set of time series that via conditioning on them the spurious causation paths are blocked and the Granger causality identifies the true temporal dependency graph. As shown in Fig 1, by analysis of the effects of unobserved confounders in simple structures,
we derive a new set of criteria which utilizes the aggregate delay in the confounding paths. The new criteria requires smaller subset of time series to be observed; hence it is more likely to be able to guarantee that Granger causality results are the true temporal relationships among the time series.

Next we identify a key set of unobserved variables that there existence prevents any guarantee on accuracy of Granger causality results. We show that under causal sufficiency assumption which excludes these structures, the two main linear Granger causality inference techniques, Significance Test [17] and Lasso-Granger [2, 26, 28], are consistent. However, we observe that in higher dimensions only Lasso-Granger is consistent. Utilizing the high dimensional advantages of $L_1$ regularization, we design a semi-parametric Granger causality inference algorithm called Copula-Granger and show that while it is consistent in high dimensions, it can efficiently capture non-linearity in the data.

2 Preliminaries and Related Work

Granger Causality is one of the most popular approaches to quantify causal relationships for time series observations. It is based on two major principles: (i) The cause happens prior to the effect and (ii) The cause makes unique changes in the effect [10, 11]. There have been extensive debates on the validity and generality of these principles. In this paper, we omit the lengthy discussion and simply assume their correctness for the rest of the discussion.

Given two stationary time series $X = \{X(t)\}_{t \in \mathbb{Z}}$ and $Y = \{Y(t)\}_{t \in \mathbb{Z}}$, we can consider the following information sets: (i) $\mathcal{I}^*(t)$, the set of all information in the universe up to time $t$, and (ii) $\mathcal{I}^{*\perp}_X(t)$, the set of all information in the universe excluding $X$ up to time $t$. Under the two principles of Granger causality, the conditional distribution of future values of $Y$ given $\mathcal{I}^{*\perp}_X(t)$ and $\mathcal{I}^*(t)$ should differ. Therefore $X$ is defined to Granger cause $Y$ [10, 11] if

$$
\mathbb{P}[Y(t + 1) \in A|\mathcal{I}^*(t)] \neq \mathbb{P}[Y(t + 1) \in A|\mathcal{I}^{*\perp}_X(t)],
$$

for some measurable set $A \subseteq \mathbb{R}$ and all $t \in \mathbb{Z}$. As we can see, the original definition of Granger causality is very general and does not have any assumptions on the data generation process. However, modeling the distributions for multivariate time series could be extremely difficult while linear models are a simple yet robust approach, with strong empirical performance in practical applications. As a results, Vector Autoregression (VAR) models have evolved to be one of the dominate approaches for Granger causality.
Up to now, two major approaches based on VAR model have been developed to uncover Granger causality for multivariate time series. One approach is the significance test [17, ch. 3.6.1]: given multiple time series \( X_1, \ldots, X_V \), we run a VAR model for each time series \( X_j \), i.e.,

\[
X_j(t) = \sum_{i=1}^{V} \beta_{j,i}^t X_i^{t,\text{Lagged}} + \epsilon_j(t),
\]

where \( X_i^{t,\text{Lagged}} = [X_i(t-L), \ldots, X_i(t-1)] \) is the history of \( X_i \) up to time \( t \), \( L \) is the maximal time lag, and \( \beta_{j,i} = [\beta_{j,i}(1), \ldots, \beta_{j,i}(L)] \) is the vector of coefficients modeling the effect of time series \( X_i \) on the target time series. We can determine that time series \( X_i \) Granger causes \( X_j \) if at least one value in the coefficient vector \( \beta_i \) is nonzero by statistical significant tests. The second approach is the Lasso-Granger approach [2, 26, 28], which applies lasso-type VAR model to obtain a sparse and robust estimate of the coefficient vectors for Granger causality tests. Specifically, the regression task in eq (2.2) can be achieved by solving the following optimization problem:

\[
\min_{\beta} \sum_{t=L+1}^{T} \left\| X_j(t) - \sum_{i=1}^{P} \beta_{j,i}^t X_i^{t,\text{Lagged}} \right\|_2^2 + \lambda \| \beta \|_1,
\]

where \( \lambda \) is the penalty parameter, which determines the sparsity of the coefficients \( \beta \).

The Lasso-Granger technique addresses the first challenge regarding the high dimensional learning to a great extent. However, it is applicable only to linear systems and the challenge still remains for the nonlinear systems. Several approaches have been proposed for identification of Granger Causality in non-linear systems; among the notable ones, kernelized regression [18], non-parametric techniques such as [12, 21, 25], Non-Gaussian Structural VAR [14] and generalized linear autoregressive models [15]. However these methods either perform poorly in high dimensions or do not scale to large datasets. In this paper we propose a semi-parametric approach based on the copula approach [16] to retain the scalability of linear VAR and high dimensional accuracy of Lasso methods and at the same time efficiently cancel out the effect of non-linearity of the data with no prior assumption on the marginal distribution of the data.

The remarkable success of Granger causality via the VAR approach in different applications [3, 4, 12, 18, 21] has led to definition of Granger Graphical models [5, 7] and Directed Information Graphs [23]. Both graphical models are obtained via graphical representation of each time series with a node and the dependency of the future of a time series \( X_i(t) \) to past values of another time series \( X_j(t) \) via a directed edge \( X_j \rightarrow X_i \) in the graph.

Granger graphical models are similar to the Causal Graphs [22, 27], however they have several significant differences: (i) they are not necessarily acyclic; a Granger graph can even have bidirectional edges; i.e. both of \( X_i \rightarrow X_j \) and \( X_j \rightarrow X_i \) edges and (ii) they are not irreflexive; i.e. a node can have an edge into itself \( X_i \rightarrow X_i \), the situation that is usually can be interpreted as memory in the system. The differences between Granger graphical models and causal graphs pose the question of how one can handle the effect of spurious causation due to unobserved confounders in Granger graphical models.

Several key steps have been taken by Eichler in analysis of effects of unobserved confounders in Granger graphical models, see [7] and the references therein. He introduced the m-separation criteria, as the counterpart of Pearl’s d-separation in causal graphs [22], for detection of connectivity of spurious paths in Granger graphs using causal priors on the unobserved time series. In this work, we show that often times, when the delay values of the edges are available in the causal prior information, many directionally connected paths, identified by the m-connectivity criteria, are disconnected considering the delay values. As a result, coping with effects of unobserved confounders is simpler in the Granger graphical models.

3 Coping with effects of unobserved confounders in Granger networks

In response to the second challenge, in this section, we show that coping with hidden confounders’ effect is easier in Granger networks. In particular, in Granger networks, many directionally connected paths are disconnected considering the delay associated with the edges. Thus, often times we require conditioning on fewer variables to block the spurious causation paths. We start with a canonical example to introduce the main concept of path delays. Via demonstration of the effect of path delays in the three basic graphical structures, we extend the “m-separation” criteria to include path delays in identification of connectivity of the paths. We show that the generalized criteria are able to detect more blocked paths which yields to higher possibility of successful causal identification. Note that the results in this section are general and not limited to the linear VAR models.

Consider the following set of linear autoregressive equations:

\[
X_1(t) = \alpha X_4(t-2) + \epsilon_1(t), \quad X_3(t) = \epsilon_3(t), \\
X_2(t) = \beta X_4(t-1) + \gamma X_3(t-1) + \epsilon_2(t), \quad X_4(t) = \epsilon_4(t),
\]
process, the following spurious edges are detected:

\[
\begin{align*}
\tau_1 &< \tau_2 \quad \Rightarrow \quad X \rightarrow Y \\
\tau_1 &> \tau_2 \quad \Rightarrow \quad Y \rightarrow X \\
\tau_1 &= \tau_2 \quad \Rightarrow \quad \text{No Causality}
\end{align*}
\]

In other words, the path from \(X\) to \(Y\), when \(Z\) is unobserved, is blocked if \(\tau_1 \geq \tau_2\) while the path \(X \leftarrow Z \rightarrow Y\) is connected in \(m\)-connectivity criteria.

**Proof.** A proof is given in the supplementary materials.

**The Collider Structure** Before delving into the theories, we first formally define the collider structure in Granger causality. Suppose the time series \(Z\) are generated from two independent time series \(X\) and \(Y\) as follows:

\[
Z_t = f(X(t), \ldots, X(t-L), Y(t-1), \ldots, Y(t-L)) + \varepsilon_Z,
\]

where the noise term \(\varepsilon_Z\) is \(N(0, \sigma)\). The causal relationships between \(X, Y\) and \(Z\) include \(X \rightarrow Z\) and \(Y \rightarrow Z\), where \(Z\) is called the Collider Node. Fig. 3b shows an example of the collider structure where the effects of \(X\) and \(Y\) reach \(Z\) with \(\tau_1\) and \(\tau_2\) delays, respectively. Next, we discuss our results on the collider structure in Lemma 3.2.

**Lemma 3.2.** In the inference of Granger causality, observing the collider node does not create spurious edge between the parents of the collider node.

**Proof.** The formal proof is given in the supplementary materials.

**The Chain Structure** The third structure the chain structure as shown in Fig. 3c. It is already known that given the variable \(Z\), no edges from \(X\) to \(Y\) will be detected; while when \(Z\) is not given, the path \(X \rightarrow Y\) is connected.

To summarize the results of observations in the fundamental structures, consider the following definition of path delay:

**Definition 3.1.** Consider a path \(P\) of length \(p-1\) from \(X_j\) to \(X_i\) defined by a set of ordered nodes \(\{X_{(k)}\}_{k=1}^p\) where \(X_{(1)} = X_j\) and \(X_{(p)} = X_i\) is given. Define the path delay as \(T_{ji}(P) = \sum_{k=1}^{p-1} \alpha_{(k),(k+1)} \tau_{(k),(k+1)}\) where \(\alpha_{(k),(k+1)} = +1\) if the edge between \(X_{(k)}\) and \(X_{(k+1)}\) is oriented as \(X_{(k)} \rightarrow X_{(k+1)}\) and \(\alpha_{(k),(k+1)} = -1\) otherwise.

In other words, start from \(X_j\) and add the delay of edges if they are towards \(X_i\) and subtract otherwise. For example, in the example given in Fig. 3.4 the path delay from \(X_3\) to \(X_1\) is computed as \(\tau_3 - \tau_2 + \tau_1\). Using the definition of path delay, we can state the following general theorem.

- **Lemma 3.1.** In the co-parent structure in Fig. 3a, when \(Z\) is unobserved and generated from a white noise process, the following spurious edges are detected:

- **Figure 2:** In this toy Granger graphical model, according to \(m\)-separation criteria, when \(X_4\) is unobserved, a spurious edge \(X_1 \leftarrow X_3\) is detected. However, the spurious edge is not detected when \(\tau_3 - \tau_2 + \tau_1 \leq 1\), where \(L\) is the maximum lag in Granger causality test.

where \(\varepsilon_i(t), i = 1, \ldots, 4\) are independent noise processes. The corresponding Granger graphical model is shown in Fig. 2, with \(\tau_1 = 2, \tau_2 = 1\) and \(\tau_3 = 1\). The direction of the edges and the values of delays on them are causal priors, obtained from field knowledge, which are necessary for analysis of effects of hidden variables. For example, consider three events defined as following: \(X\) : rain in Los Angeles, CA, \(Y\) : rain in Riverside, CA and \(Z\) : the approach of the coastal air masses. One can observe that the effect of coastal air masses cannot reach Riverside earlier than Los Angeles; consequently, the edge \(Z \rightarrow X\) must have smaller delay than \(Z \rightarrow Y\).

In analysis of the structure in Fig. 2, Eichler [5] showed that in absence of \(X_4\) an spurious edge is detected from \(X_3\) to \(X_1\). This spurious causation is the result of the spurious path \(X_1 \leftarrow X_4 \rightarrow X_2 \leftarrow X_3\) which is connected under the “m-separation” criteria. However, a quick inspection shows that when \(\tau_1 \leq 1\) the spurious edge \(X_3 \rightarrow X_1\) is never inferred. This implies that in Granger networks, we might inspect not only for graphical connectivity, but also for the delays in the connected paths. This idea is scrutinized via the three basic structures of directed graphs possible with three time series (see Fig. 3).

**The Co-parent Structure** In the co-parent structure (Fig. 3a), an unobserved time series \(Z\) causes two observed time series \(X\) and \(Y\). The effect of the cause \(Z\) reaches \(X\) and \(Y\) with possibly different delays \(\tau_1\) and \(\tau_2\), respectively. A simple inspection shows that the identified direction of causality between \(X\) and \(Y\) depends on the relative value of \(\tau_1\) and \(\tau_2\). In particular,

\[
\begin{align*}
\tau_1 &= \tau_2 \quad \Rightarrow \quad \text{No Causality}
\end{align*}
\]
Theorem 3.1. Consider a Granger network $G(V,E)$ with set of nodes $V = \{X_i\}$ for $i = 1, \ldots, n$, set of directed edges $E$ and the edge delays $\tau_{i,j} \in \mathbb{Z}^+$ for every edge $X_i \rightarrow X_j \in E$. Suppose the unobserved time series are generated from white processes. Then, every path $P$ from an arbitrary node $X_j$ pointing to $X_i$ is connected if it is both $m$-connected and the path delay $T_{j,i}(P) > 0$.

Proof. A proof based on step by step reduction of the path using the three fundamental structures is provided in the supplementary materials. The intuition behind the theorem is rather simple if we accept the directional information transfer interpretation of Granger graphical models: a spurious edge is detected whenever the information from the effect reaches the cause with a positive delay.

Note that the profound implication of Theorem 3.1 is that the time order information that is usually assumed available in confounder analysis can be used more efficiently in the Granger causality analysis. If the time order between hidden variables are given, we can make stricter rules for the connectivity of paths in the Granger causality framework by ruling out many paths that would be identified as connected by m-separation. This makes the unidentifiability problem less likely in Granger networks with hidden variables. The next example demonstrates the advantages implied by Theorem 3.1.

Example 3.1. Consider the Granger graph in Fig. 4. Time series $X_1, X_2$ and $X_3$ are observed while $X_4$ and $X_5$ are unobserved. The goal is to find the Granger causal effect of $X_1$ on $X_3$.

Solution. The true causal path is $X_1 \rightarrow X_5 \rightarrow X_3$, however the $X_1 \leftarrow X_4 \rightarrow X_2 \leftarrow X_3$ path is a potential confounding path. The $m$-connectivity criteria states that unless $X_4$ is observed, the causality from $X_1$ to $X_3$ will not be identifiable. However, utilizing the delay of the path $X_1 \leftarrow X_4 \rightarrow X_2 \leftarrow X_3$, unidentifiability only occurs when the path delay $T_{3,1} > 0$ and we have higher possibility of successful causal inference.

While whiteness of the unobserved variables is satisfied in many applications, even in the cases that the hidden time series are not white, the analysis in the supplementary materials shows that the unidentified spurious causations need to propagate through long paths and undergo significant attenuation which makes Theorem 3.1 approximately hold.

4 Consistency of Granger causality methods
After the analysis of spurious causality in Granger causality given its fundamental assumptions, we are ready to analyze the consistency results of different approaches to uncover Granger causality. In this section, we first review the consistency of two main Granger causality analysis techniques, significance tests and Lasso-Granger, in low dimensional regime where sufficient number of observations are given. First we show that in presence of hidden coparent variables, no Granger causality test can be consistent. To solve the problem, we show that in the causally sufficient system, the consistency results are established for both approaches. Next we show that in high dimensional regime, unlike Lasso-Granger, the significance test is inconsistent; leading to the main incentive for the use of $L_1$ regularized methods in high dimensional regimes. Thus, we introduce the semi-parametric approach Copula-Granger and show that while it is consistent in high dimensions, it can capture non-linearity in the data.

First we have the following inconsistency result in
Proof. Similar to [24], consider the common cause scenario as shown in Fig. 3a with \( \tau_2 > \tau_1 \). It can be easily seen that in absence of \( Z \), \( X \) will be identified as the Granger cause for \( Y \).

In order to avoid the situations described in Proposition 4.1, a common practice is to make the following Causal Sufficiency assumption, [27, ch. 5].

**Assumption 4.1.** A causal system is Causally Sufficient if no common cause of any two observed variables in the system is left out.

The next proposition studies the consistency of significance tests and Lasso-Granger to uncover Granger causal relationships. Following [31], we define the Model Selection consistency for Granger causality tests using the following probability

\[
P[\text{Error}] = P \{ (i, j) : \hat{\beta}_{i,j} \neq 0 \neq (i, j) : \beta_{i,j} \neq 0 \}
\]

where \( \hat{\beta} \) is the coefficient vector inferred via a Granger causality inference algorithm. We say that a method is consistent if its probability of errors goes to zero as the number of observations increase.

**Proposition 4.2.** Given the causal sufficiency in a VAR system, both of significance test and Lasso-Granger do not include spurious causation. Furthermore, given sufficient number of observations \( T/L > n + 1 \), the causal estimates are consistent; i.e. for significance tests \( P[\text{Error}] \leq cL/T - L \exp \left( -\frac{c^2}{2} (T - L) \right) \) for some constant \( c \), where \( T \) is the length of time series, and \( L \) is the maximal lag; and for Lasso-Granger, subject to the Irrepresentable Condition in [31], the model selection error decays with rate \( o(c'L \exp(-T^\nu)) \) for some \( 0 \leq \nu < 1 \) and some constant \( c' \).

**Proof.** Proofs via different approaches can be found in the literature, see [17, ch. 2.3] and [31]. For completeness, we also provide a proof in the supplementary materials using asymptotic normality of maximum likelihood estimation.

**Remarks 1.** The result in Proposition 4.2 states that the error decreases exponentially as the length of the time series increase for both approaches. Also it states that when \( L \ll T \) large value of \( L \) linearly degrades the performance, whereas in the case of \( L \sim T \) the exponential term will be dominant and the error will increase exponentially with \( L \).

2. The consistency results also imply that learning linear Granger causal relationships is a simpler task than learning undirected graphical models [19]. This is intuitive since learning the edges for one node is a variable selection process isolated from that for other nodes and therefore no constraint on the neighborhood nodes is required.

Moving to high dimensional regime, we will show that the significance tests are inconsistent in high dimensions.

**Proposition 4.3.** In high dimensions, where \( T/L < n + 1 \), the significant test is inconsistent. The inferred coefficients using ridge regression decay according to following rate:

\[
\mathbb{E}_{X, \varepsilon}[\hat{\beta}_\lambda] = \begin{cases} 
\beta & \text{if } (n + 1)L \leq T \\
\left( \frac{T}{n} \right)^{-1} \beta & \text{if } (n + 1)L > T.
\end{cases}
\]

as the penalization parameter in the ridge regression \( \lambda \to 0 \). The expectation is over outcomes of the data \( X \) and noise \( \varepsilon \). The \( L_1 \) variable selection methods are consistent subject to incoherence conditions, [20].

**Proof.** A proof based on properties of random design matrix is provided in the supplementary materials. Several other authors also have pointed out the inconsistency of the ridge regression, and consequently significance tests, in high dimensions before, see [20, 29] and the references therein; however to the best knowledge of the authors, the above small sample result is novel.

Proposition 4.3 highlights the fact that the inadvertent choice of large lag length \( L \) can move the system to high dimensional regime and result in inconsistency of the significance test.

All the consistency results so far are for linear Granger causality inference techniques. Here we propose the Granger Non-paranormal (G-NPN) model and design the Copula-Granger inference technique to capture the non-linearity of the data while retaining the high dimensional consistency of Lasso-Granger.

**Definition 4.1.** Granger Non-paranormal (G-NPN) model We say a set of time series \( X = (X_1, \ldots, X_n) \) has Granger-Nonparanormal distribution \( G-NPN(X, B, F) \) if there exist functions \( \{F_j\}_{j=1}^n \) such that \( F_j(X_j) \) for \( j = 1, \ldots, n \) are jointly Gaussian and can be factorized according to the VAR model with coefficients \( B = \{ \beta_{i,j} \} \). More specifically, the joint distribution for the transformed random variables \( Z_j = F_j(X_j) \)
can be factorized as following
\[
p_Z(z) = N(\mathbf{a}(1, \ldots, L)) \times \prod_{j=1}^n \prod_{t=L+1}^T p_N(z_j(t) | \sum_{i=1}^n \beta_{i,j} z_i(t-\text{Lagged}), \sigma_j),
\]
where \( p_N(z|\mu, \sigma) \) is the Gaussian density function with mean \( \mu \) and variance \( \sigma^2 \).

Based on the copula technique [16], the G-NPN model aims to separate the marginal properties of the data from its dependency structure. The marginal distribution of the data can be efficiently estimated using the non-parametric techniques with exponential convergence rate [30, ch. 2]. The estimation of the dependency structure requires more effort; because there are at least \( \mathcal{O}(n^2) \) pairwise dependency relationships. Thus, we resort to \( L_1 \) regularized techniques for efficient estimation of the dependency structure in high dimensional settings.

Learning Granger Non-paranormal models consists of three steps: (i) Find the empirical marginal distribution for each time series \( \hat{F}_i \). (ii) Map the observations into the copula space as \( \hat{F}_i(X_i^t) = \hat{\mu}_i + \sigma_i \Phi^{-1}(\hat{F}(X_i^t)) \). (iii) Find the Granger causality among \( \hat{F}_i(X_i^t) \). In practice we have to use the Winsorized estimator of the distribution function to avoid the large numbers \( \Phi^{-1}(0^+) \) and \( \Phi^{-1}(1^-) \):

\[
\hat{F}_j = \begin{cases} 
\delta_n & \text{if } \hat{F}_j(X_j^t) < \delta_n \\
\hat{F}_j(X_j^t) & \text{if } \delta_n \leq \hat{F}_j(X_j^t) \leq 1 - \delta_n \\
(1 - \delta_n) & \text{if } \hat{F}_j(X_j^t) > 1 - \delta_n 
\end{cases}
\]

First we have the following proposition that connects the Granger causality results identified by the Copula-Granger method to the true Granger causality values:

**Proposition 4.4.** The independence relationships in the copula space are the same as the independence relationships among original time series.

**Proof.** Since \( X \perp Y \) if and only if \( g(X) \perp h(Y) \) for any arbitrary random variables \( X \) and \( Y \) and deterministic one-to-one transformation functions \( g(\cdot) \) and \( h(\cdot) \), the proposition is established.

The next theorem establishes the consistency rate of the Copula-Granger method.

**Theorem 4.1.** Consider the time series \( X_i(t) \) for \( i = 1, \ldots, n \) and \( t = 1, \ldots, T \) generated according to \( G = \text{NPN}(X, B, F) \). Select \( \delta_{T-L} = \left( 4(T-L)^{1/4} \sqrt{\log(T-L)} \right)^{-1} \) and \( \lambda_{T-L} \propto \sqrt{(T-L) \log(nL)} \). Suppose the incoherent design condition in [20] holds for both covariance matrices \( C = \mathbb{E}[X_i(t)X_j(t')] \) and \( \tilde{C} = \mathbb{E}[\tilde{F}_i(X_i(t))\tilde{F}_j(X_j(t))] \) for \( i, j = 1, \ldots, N \) and \( t, s = t-L, \ldots, t-1 \). The Copula-Granger estimate of the \( B \) is asymptotically consistent as \( T \to \infty \).

\[
(\tilde{\beta}_{i,j} - \beta_{i,j})_2 = \mathcal{O}_P \left( K_{T-L} \sqrt{\frac{s \log(nL)}{T-L}} \right),
\]

where \( \tilde{\beta}_{i,j} \) are estimates of \( \beta_{i,j} \) using Copula-Granger, \( s \) is the number of non-zero coefficients under analysis and \( K_{T-L} \) is proportional to \( \frac{\phi_{\max}(\sigma^2)}{\phi_{\min}(\sigma^2)} \). The subindex \( P \) in \( \mathcal{O}_P \) denotes convergence in probability.

**Proof.** The proof provided in the supplementary materials relies on the result of [16] which shows that the copulation matrix of the samples transformed by the non-parametric Winsorized distribution estimator is concentrated around the true covariance matrix. Using this concentration bound, we can bound the maximum eigenvalue of the matrix \( \tilde{C} - C \). Repeating the steps of [20] gives the rate above.

Theorem 4.1 states that the convergence rate for Copula-Granger is the same as the one for Lasso which suggests efficient Granger graph learning in high dimensions via Copula-Granger.

**5 Experiments**

In this section, we conduct experiments on synthetic datasets and a Twitter application dataset to study the properties of significance tests, Lasso-Granger and the semi-parametric approach for Granger causality analysis and verify our theoretical results. In all the experiments, we use implementation of Lasso in GLMnet package [9] and tune the penalization parameter of Lasso via AIC [1]. In the prediction task, we train the algorithm on the 90% of the data and test it on the rest.

**Verification of the Theoretical Consistency Results**

We generated multiple synthetic datasets to verify the claim in Theorem 3.1. We provide an example of such experiments. Fig. 5a shows the graph of a synthetic dataset generated to verify the claim in Lemma 3.2. In this dataset, \( X, Y \) and \( Z \) are observed, but \( U \) and \( V \) are unobserved. Fig. 5b shows the causality relationships identified by three algorithms when we set the length of the time series to 500. As we can see none of the edges \( Z \to Y \) and \( X \to Y \)
are detected by algorithms. In Fig. 5c the length of time series is reduced to 50. Neither Lasso-Granger nor Copula-Granger identify any edge, while the significance test approach over-rejects the null hypothesis. These results also confirm the loss due to large lags when the length of the time series is short.

**The Effect of Non-Linearity** Similar to [13], we design a non-linear system with a parameter to control the amount of non-linearity. We choose the nonlinear function \( g(x) = x + bx^3 \) and \( b \in [0, 1] \) where \( b \) is used to control it non-linearly. Using this function we define the following set of time series: 

\[
X_1(t) = \sum_{i=1}^{p} \left[ X_i(t-1) + bX_i^3(t-1) \right] + \varepsilon_1(t) \quad \text{and} \quad X_j(t) = \varepsilon_j(t), \quad j = 1, \ldots, n, \text{ where } \varepsilon_j(t) \text{ for } j = 1, \ldots, n \text{ are white } \mathcal{N}(0, 0.1) \text{ noises.}
\]

Fig. 5 shows the effect of non-linearity on the performance of the three algorithms for high dimensional \( (n = 20 \text{ for Fig. 5d}) \) and low dimensional \( (n = 10 \text{ for Fig. 5e}) \) cases. Note the robustness of the copula approach with respect to non-linearity. Fig. 5e points out the fact that in low dimensional settings the ridge regression with small penalization terms has lower bias and is more accurate.

**Social Networking Dataset** We used a complete Twitter dataset to analyze the tweets about “Haiti earthquake” by applying different Granger causality analysis methods to identify the potential top influential on this topic (i.e. those Twitter accounts with the highest number of effect to the others). We divided the 17 days after the Haiti Earthquake on Jan. 12, 2010 into 1000 interval and generated a multivariate time series dataset by counting the number of tweets on this topic for the top 1000 users who tweeted most about it. Fig. 6 shows the aggregate number of tweets about Haiti in the dataset. Table 2 compares the prediction performance of the algorithms. Associating the number

![Image](image-url)
of outgoing edges with the social influence of a node, we find the most influential users identified by each algorithm by counting the number of outgoing edges for each user. The top ten most influential users identified by each algorithm are listed in Table 1. For each user, we also count the number of tweets by the user about the topic in the interval of study. The top ten influential users identified by Copula-Granger technique have significantly more tweets which confirms the superior performance of the Copula-Granger approach.

6 Conclusion

In this paper, we studies the theoretical properties of large-scale Granger causality inference algorithms. We utilized the confounder path delays to find a subset of time series that via conditioning on them we are able to cancel out the spurious confounder effects. After study of consistency of different Granger causality techniques, we propose Copula-Granger and show that while it is consistent in high dimensions and scalable for large data, it can efficiently capture non-linearity in the data. For future work we are interested in theoretical analysis of properties of a wider class of algorithms. Investigation of different techniques for high-dimensional non-linear Granger causality inference is another line of future work.

Acknowledgment

We thank the anonymous reviewers for their valuable comments. This research was supported by the NSF research grants IIS-1134990 and DARPA award No. W911NF1210034. The views and conclusions contained herein are those of the authors and should not be interpreted as representing the official policies or endorsements, either expressed or implied, of any of the above organizations or any person connected with them.

References