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,

*The corresponding author.

†University of Southern California,
Emails: {mohammab, yanliu.cs}@usc.edu

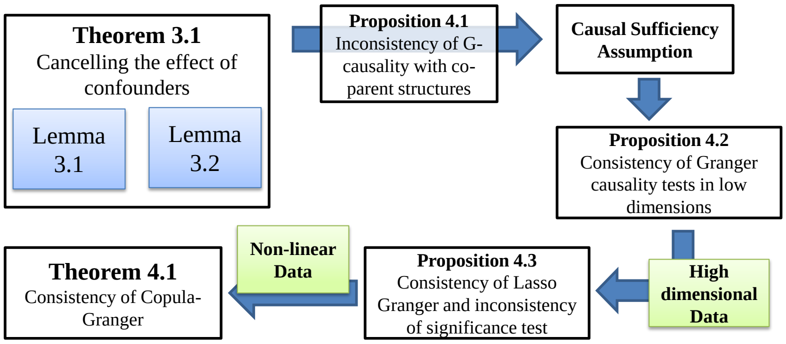

Theorem 3.1
Cancelling the effect of confounders
Lemma 3.1
Lemma 3.2
Proposition 4.1
Inconsistency of G-causality with co-parent structures
Causal Sufficiency Assumption
Proposition 4.2
Consistency of Granger causality tests in low dimensions
High dimensional Data
Proposition 4.3
Consistency of Lasso Granger and inconsistency of significance test
Non-linear Data
Theorem 4.1
Consistency of Copula-Granger

Figure 1: The sequence of the theoretical results: Theorem 3.1 utilizes path delays to find a subset of time series that via conditioning on them we are able to cancel out the spurious confounder effects. Proposition 4.1 shows that when unobserved time series are parents of multiple observed time series, there is no consistent Granger causality test. The *Causal Sufficiency* assumption excludes these structures and with this assumption both significant test and Lasso-Granger become consistent in low dimensions (Proposition 4.2). Proposition 4.3 shows that in high dimensions significant test is inconsistent but Lasso-Granger is consistent. When the data deviates from the linear model assumed in Lasso-Granger, Theorem 4.1 shows that while *Copula-Granger* is consistent in high dimensions, it can efficiently capture non-linearity in the data.

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 this 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.

In the rest of the paper, we first review Granger causality and the existing approaches to uncover Granger causality in Section 2, and then we discuss the theoretical analysis results to answer each of these two questions in Section 3 and 4, respectively. In Section 5, we show experiment results on synthetic datasets and social media application data to support our theoretical analysis, and finally summarize the paper and hint on future work.

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) $\mathcal{I}^\star(t)$, the set of all information in the universe up to time t, and (ii) $\mathcal{I}^\star_{-X}(t)$, 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 $\mathcal{I}^\star_{-X}(t)$ and $\mathcal{I}^\star(t)$ should differ. Therefore X is defined to Granger cause Y [10, 11] if

$$\mathbb{P}[Y(t+1) \in A | \mathcal{I}^\star(t)] \neq \mathbb{P}[Y(t+1) \in A | \mathcal{I}^\star_{-X}(t)], \tag{2.1}$$

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.,

$$X_j(t) = \sum_{i=1}^{V} \boldsymbol{\beta}_{j,i}^{\top} \mathbf{X}_i^{t,Lagged} + \epsilon_j(t), \tag{2.2}$$

where $\mathbf{X}_i^{t,Lagged} = [X_i(t-L), \ldots, X_i(t-1)]$ is the history of X_i up to time t, L is the maximal time lag, and $\boldsymbol{\beta}_{j,i} = [\beta_{j,i}(1), \ldots, \beta_{j,i}(L)]$ 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 $\boldsymbol{\beta}_i$ 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:

$$\min_{\{\boldsymbol{\beta}\}} \sum_{t=L+1}^{T} \left\| X_j(t) - \sum_{i=1}^{P} \boldsymbol{\beta}_{j,i}^{\top} \mathbf{X}_i^{t,Lagged} \right\|_2^2 + \lambda \left\| \boldsymbol{\beta} \right\|_1, \tag{2.3}$$

where λ is the penalty parameter, which determines the sparsity of the coefficients $\boldsymbol{\beta}$.

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:

$$\begin{aligned} X_1(t) &= \alpha X_4(t-2) + \varepsilon_1(t), \; X_3(t) = \varepsilon_3(t), \\ X_2(t) &= \beta X_4(t-1) + \gamma X_3(t-1) + \varepsilon_2(t), \; X_4(t) = \varepsilon_4(t), \end{aligned} \tag{3.4}$$

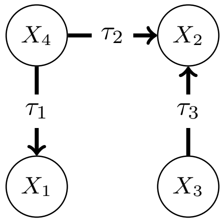


Figure 2: In this toy Granger graphical model, according to *m-separation* criteria, when X_4 is unobserved, a spurious edge $X_1 \leftarrow X_3$ is detected. However, the spurious edge is not detected when $\tau_3 - \tau_2 + \tau_1 \leq 1$, where L is the maximum lag in Granger causality test.

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 $Z \to X$ must have smaller delay than $Z \to Y$.

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 $X_1 \leftarrow X_4 \to X_2 \leftarrow X_3$ which is connected under the "m-separation" criteria. However, a quick inspection shows that when $\tau_1 \leq 1$ the spurious edge $X_3 \to X_1$ 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 τ_1 and τ_2, respectively. A simple inspection shows that the identified direction of causality between X and Y depends on the relative value of τ_1 and τ_2. In particular,

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

$$\begin{aligned} \tau_1 < \tau_2 \quad &\Rightarrow \quad X \to Y \\ \tau_1 > \tau_2 \quad &\Rightarrow \quad Y \to X \\ \tau_1 = \tau_2 \quad &\Rightarrow \quad \textit{No Causality} \end{aligned} \tag{3.5}$$

In other words, the path from X to Y, when Z is unobserved, is blocked if $\tau_1 \geq \tau_2$ while the path $X \leftarrow Z \to Y$ 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:

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

where the noise term ε_Z is $\mathcal{N}(0, \sigma)$. The causal relationships between X, Y and Z include $X \to Z$ and $Y \to Z$, where Z is called the *Collider Node*. Fig. 3b shows an example of the collider structure where the effects of X and Y reach to Z with τ_1 and τ_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 $X \to Y$ 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 $T_{j,i}(P) = \sum_{k=1}^{p-1} \alpha_{(k),(k+1)} \tau_{(k),(k+1)}$ where $\alpha_{(k),(k+1)} = +1$ if the edge between $X_{(k)}$ and $X_{(k+1)}$ is oriented as $X_{(k)} \to X_{(k+1)}$ 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.

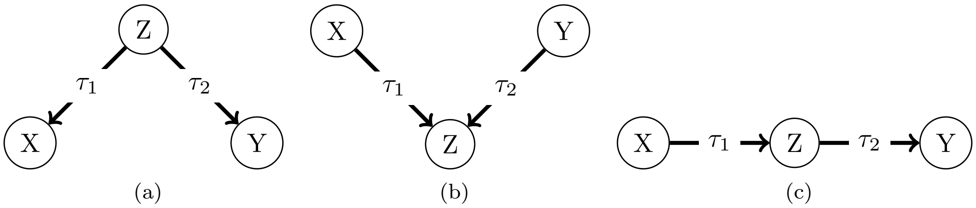


Figure 3: Three of four possible directed graphs created by three nodes (a) the coparent, (b) the collider and (c) the chain structures. The fourth structure is the chain with reversed edge directions.

Theorem 3.1. *Consider a Granger network $G(V, E)$ with set of nodes $V = \{X_i\}$ for $i = 1, \dots, n$, set of directed edges E and the edge delays $\tau_{i,j} \in \mathbb{Z}^+$ for every edge $X_i \to X_j \in E$. 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 $X_1 \to X_5 \to X_3$, however the $X_1 \leftarrow X_4 \to X_2 \leftarrow X_3$ 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 $X_1 \leftarrow X_4 \to X_2 \leftarrow X_3$, unidentifiability only occurs when the path delay $T_{3,1} > 0$ and we have higher possibility of successful causal inference.

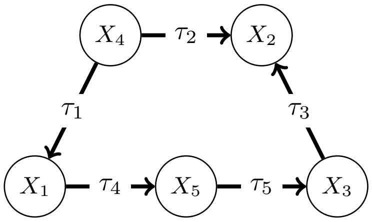


Figure 4: An example of canceling spurious causation. Time series X_1, X_2 and X_3 are observed while X_4 and X_5 are unobserved.

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 the presence of hidden variables.

PROPOSITION 4.1. *In the presence of hidden variables, no test for Granger causality can be consistent.*

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

$$\mathbb{P}[\text{Error}] = \mathbb{P}[\{(i,j) : \widehat{\boldsymbol{\beta}}_{i,j} \neq \mathbf{0}\} \neq \{(i,j) : \boldsymbol{\beta}_{i,j} \neq \mathbf{0}\}].$$

where $\widehat{\boldsymbol{\beta}}$ 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 tests 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 $\mathbb{P}[Error] \leq cL\sqrt{T-L}\exp\left(-\frac{c^2}{2}(T-L)\right)$ 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:*

$$\mathbb{E}_{\mathbf{X},\boldsymbol{\varepsilon}}[\widehat{\boldsymbol{\beta}}_{\lambda}] = \begin{cases} \boldsymbol{\beta} & \textit{if } (n+1)L \leq T \\ (\frac{T/L-1}{n})\boldsymbol{\beta} & \textit{if } (n+1)L > T. \end{cases}$$

as the penalization parameter in the ridge regression $\lambda \to 0$. The expectation is over outcomes of the data $\mathbf{X}$ and noise $\boldsymbol{\varepsilon}$. 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 $B = \{\boldsymbol{\beta}_{i,j}\}$. More specifically, the joint distribution for the transformed random variables $Z_j \triangleq F_j(X_j)$*

can be factorized as following

$$p_{\mathbf{Z}}(\mathbf{z}) = \mathcal{N}(\mathbf{z}(1,\dots,L)) \times \prod_{j=1}^{n}\prod_{t=L+1}^{T} p_{\mathcal{N}}(z_j(t)| \sum_{i=1}^{n} \boldsymbol{\beta}_{i,j}^{\top} z_i^{t,Lagged}, \sigma_j),$$

where $p_{\mathcal{N}}(z|\mu,\sigma)$ is the Gaussian density function with mean μ and variance σ^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 $\widehat{F}_i$. (ii) Map the observations into the copula space as $\widehat{f}_i(X_i^t) = \widehat{\mu}_i + \widehat{\sigma}_i \Phi^{-1}\left(\widehat{F}_i(X_i^t)\right)$. (iii) Find the Granger causality among $\widehat{f}_i(X_i^t)$. In practice we have to use the Winsorized estimator of the distribution function to avoid the large numbers $\Phi^{-1}(0^+)$ and $\Phi^{-1}(1^-)$:

$$\tilde{F}_j = \begin{cases} \delta_n & \text{if } \widehat{F}_j(X_j) < \delta_n \\ \widehat{F}_j(X_j) & \text{if } \delta_n \leq \widehat{F}_j(X_j) \leq 1-\delta_n \\ (1-\delta_n) & \text{if } \widehat{F}_j(X_j) > 1-\delta_n \end{cases}$$

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 $X \perp\!\!\!\perp Y$ if and only if $g(X) \perp\!\!\!\perp h(Y)$ for any arbitrary random variables X and Y and deterministic one-to-one transformation functions $g(.)$ and $h(.)$, 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 $i = 1,\dots,n$ and $t = 1,\dots,T$ generated according to $G - NPN(X,B,F)$. Select $\delta_{T-L} = \left(4(T-L)^{1/4}\sqrt{\pi\log(T-L)}\right)^{-1}$ and $\lambda_{T-L} \propto \sqrt{(T-L)\log(nL)}$. Suppose the incoherent design condition in [20] holds for both covariance matrices $C \triangleq \mathbb{E}[X_i(t)X_j(t')]$ and $\tilde{C} \triangleq \mathbb{E}[\tilde{F}_i(X_i(t))\tilde{F}_j(X_j(t'))]$ for $i,j = 1,\dots,N$ and $t,s = t-L,\dots,t-1$. The Copula-Granger estimate of the B is asymptotically consistent as $T \to \infty$*

$$\left\|\widehat{\boldsymbol{\beta}}_{i,j} - \boldsymbol{\beta}_{i,j}\right\|_2 = \mathcal{O}_P\left(K_{T-L}\sqrt{\frac{s\log(nL)}{T-L}}\right), \tag{4.6}$$

where $\widehat{\boldsymbol{\beta}}_{i,j}$ are estimates of $\boldsymbol{\beta}_{i,j}$ using Copula-Granger, s is the number of non-zero coefficients among nL coefficients under analysis and K_{T-L} is proportional to $\frac{\phi_{max}}{\phi^2_{min}(se_n^2)}$ where ϕ_{max} and $\phi_{min}(m)$ are maximum and m-sparse minimum eigenvalue of the matrix $\tilde{C}$ and e_n is a saparisty multiplier sequence as defined in [20]. The subindex P in $\mathcal{O}_P$ denotes convergence in probability.

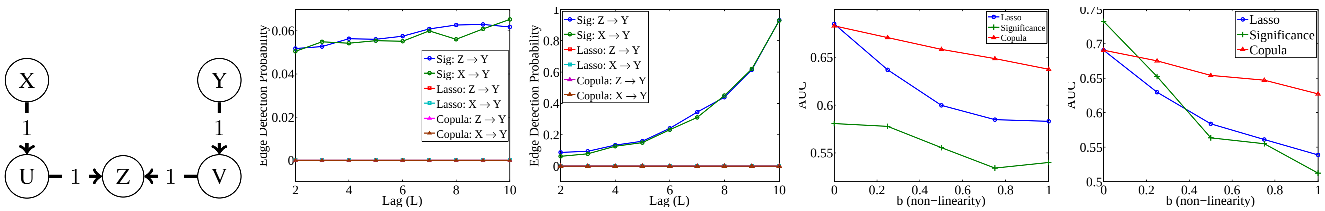
Proof. The proof provided in the supplementary materials relies on the result of [16] which shows that the covariance matrix of the samples transformed by the non-parametric Winsorized distribution estimator is concentrated around the true covariance matrix. Using this concentration bound, we can bound 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 $X \to Y$



(a) The model (b) $n = 500$ (c) $n = 50$ (d) $n = 20$ (e) $n = 10$

Figure 5: (a-c) Edge detection probabilities in the collider scenario. (d-e) Accuracy of Lasso-Granger, Significance test and Copula-Granger methods in the non-linear settings. The time series are selected to be short ($T = 30$).

Table 1: The top ten influential users detected by the algorithms.

Significance Test		Lasso-Granger		Granger-Copula	
Name	# of Tweets	Name	# of Tweets	Name	# of Tweets
imzadi1	43	monsterroxanne	13	prayer_network	85
thinkingofrob	49	untoothershaiti	265	contactolatino	77
wyclef	41	trendsbyminute	122	woodringstpreux	158
bduguay	31	lustlove	85	epiccolorado	75
gregdominica	24	clarlune	51	lumicelestial	74
margofranssen	35	1upmaria	49	viequesbound	114
joannasimkin	27	haiti_tweets	33	shirley1376	55
porque2012	31	srgryph	30	kareenaristide	101
mekaemanuel	22	hope_for_haiti	26	nancy19087	49
catweazle1961	30	mrlyphe	24	alaingabriel	97
	333		698		**885**

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-linearity. Using this function we define the following set of time series: $X_1(t) = \sum_{i=1}^{p} \left[X_i(t-1) + bX_i^3(t-1)\right] + \varepsilon_1(t)$ and $X_j(t) = \varepsilon_j(t), \quad j = 1, \ldots, n$, where $\varepsilon_j(t)$ for $j = 1, \ldots, n$ are white $\mathcal{N}(0, 0.1)$ noises. Fig. 5 shows the effect of non-linearity on the performance of the three algorithms for high dimensional ($n = 20$ for Fig. 5d) and low dimensional ($n = 10$ for Fig. 5e) cases. Note the robustness of the copula approach with respect to nonlinearity. 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.

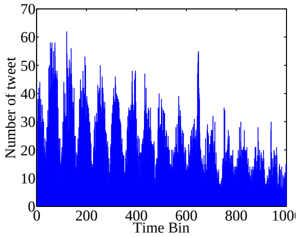


Figure 6: Aggregate number of tweets about Haiti in 17 days after the earthquake. The time axis is divided to 1000 intervals.

Table 2: The RMS prediction error.

Method	RMSE
Significance Test	5.2×10^{-3}
Lasso-Granger	3.8×10^{-3}
Granger-Copula	3.9×10^{-3}

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 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

[1] H. Akaike. A new look at the statistical model identification. *IEEE Trans. on Automatic Control*, 1974.

[2] A. Arnold, Y. Liu, and N. Abe. Temporal causal modeling with graphical granger methods. In *KDD*, 2007.

[3] I. Asimakopoulos, D. Ayling, and W. M. Mahmood. Non-linear granger causality in the currency futures returns. *Economics Letters*, 2000.

[4] A. Brovelli, M. Ding, A. Ledberg, Y. Chen, R. Nakamura, and S. L. Bressler. Beta oscillations in a large-scale sensorimotor cortical network: directional influences revealed by Granger causality. *PNAS*, 2004.

[5] M. Eichler. A graphical approach for evaluating effective connectivity in neural systems. *Phil. Trans. R. Soc. B*, 2005.

[6] M. Eichler. Granger causality and path diagrams for multivariate time series. *Journal of Econometrics*, 2007.

[7] M. Eichler. Graphical modelling of multivariate time series. *Probability Theory and Related Fields*, 2012.

[8] M. Eichler and V. Didelez. On granger causality and the effect of interventions in time series. *Lifetime Data Analysis*, 2010.

[9] J. H. Friedman, T. Hastie, and R. Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 2010.

[10] C. W. J. Granger. Investigating Causal Relations by Econometric Models and Cross-spectral Methods. 1969.

[11] C. W. J. Granger. Testing for causality: A personal viewpoint. *J. of Econ. Dyn. and Cont.*, 1980.

[12] C. Hiemstra and J. D. Jones. Testing for Linear and Nonlinear Granger Causality in the Stock Price- Volume Relation. *The Journal of Finance*, 1994.

[13] P. O. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *NIPS*, 2008.

[14] A. Hyvärinen, K. Zhang, S. Shimizu, P. O. Hoyer, and P. Dayan. Estimation of a Structural Vector Autoregression Model Using Non-Gaussianity. *JMLR*, 2010.

[15] Y.-H. Kim, H. H. Permuter, and T. Weissman. Directed information, causal estimation, and communication in continuous time. 2009.

[16] H. Liu, J. D. Lafferty, and L. A. Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *JMLR*, 2009.

[17] H. Lütkepohl. *New Introduction to Multiple Time Series Analysis*. Springer, 2005.

[18] D. Marinazzo, M. Pellicoro, and S. Stramaglia. Kernel Granger causality and the analysis of dynamical networks. *Physical review. E*, 2008.

[19] N. Meinshausen and P. Bühlmann. High-Dimensional Graphs and Variable Selection with the Lasso. *The Annals of Statistics*, 2006.

[20] N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics*, 2009.

[21] C. D. Panchenko and Valentyn. Modified hiemstra-jones test for granger non-causality. Technical report, Society for Computational Economics, 2004.

[22] J. Pearl. *Causality: Models, Reasning and Inference*. Cambridge University Press, 2009.

[23] C. Quinn, N. Kiyavash, and T. P. Coleman. Directed Information Graphs. 2012.

[24] J. M. Robins, R. Scheines, P. Spirtes, and L. Wasserman. Uniform consistency in causal inference. *Biometrika*, 2003.

[25] T. Schreiber. Measuring Information Transfer. *Physical Review Letters*, 2000.

[26] S. Song and P. J. Bickel. Large Vector Auto Regressions. 2011.

[27] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search, Second Edition*. The MIT Press, 2001.

[28] P. A. Valdés-Sosa, J. M. Sánchez-Bornot, A. Lage-Castellanos, M. Vega-Hernández, J. Bosch-Bayard, L. Melie-García, and E. Canales-Rodríguez. Estimating brain functional connectivity with sparse multivariate autoregression. *Phil. Trans. R. Soc. B*, 2005.

[29] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity. In *Allerton*, 2006.

[30] L. Wasserman. *All of Nonparametric Statistics (Springer Texts in Statistics)*. Springer, 2005.

[31] P. Zhao and B. Yu. On Model Selection Consistency of Lasso. *JMLR*, 2006.

7 Supplementary Materials

7.1 Proofs of Section 3

7.1.1 Proof of Lemma 3.1 Without loss of generality let's assume $\tau_1 < \tau_2$. We need to show that there is an spurious edge from X to Y and there is no spurious edge from Y to X. Formally, we need to show that $(Y(t) \not\perp\!\!\!\perp X(t-\tau))|Y(t-1,\dots,t-L)$ for $\tau = 1,\dots,L$. Using the SEM notation, as shown in Fig. 7a, we can see that there is always a directional path from $Y(t)$ via $Z(t-\tau_2)$ to $X(t-\tau_2+\tau_1)$. However, all paths from $Y(t-\tau)$, $\tau = 1,\dots,L$ to $X(t)$ are blocked. Note that if $Z(t)$ is not white, there exists at least a spurious path which goes through history of $Z(t)$; however in practice the attenuation of this path is significant enough that makes the Lemma approximately hold for non-white unobserved variables.

7.1.2 Proof of Lemma 3.2 An argument similar to the previous proof can be used here to show that $(Y(t) \perp\!\!\!\perp X(t-\tau))|Y(t-1,\dots,t-L)$ for $\tau = 1,\dots,L$. Fig. 7b shows the scenario corresponding to $\tau_1 = 1$ and $\tau_2 = 2$. We can see that the observations at $t-1$ block all the directed paths from past to $X(t)$ and $Y(t)$ which concludes the proof.

7.1.3 Proof of Theorem 3.1 Proof can be established by induction. Informally, the proof is a sequence of reduction of the fundamental structures to there equivalent spurious edge. The final path will be one of the three fundamental structures for which the equation $T_{j,i} > 0$ is satisfied.

7.2 Proof of Proposition 4.2 The proof is done in two steps: (i) finding the convergence rate of Maximum Likelihood estimation of VAR models and (ii) Proposition 7.1 to complete the proof.

Proof. **Convergence rate of Maximum Likelihood estimation of VAR models** The main idea of the proof is that learning edges for every node is a variable selection problem isolated from other nodes. Define the consistency by introducing the probability of errors for VAR-type models as follows:

$$\begin{aligned}\mathbb{P}[\text{Error}] = \mathbb{P}[\exists\ell:\ |\widehat{\beta}_{i,j}(\ell)| > \alpha_0|\boldsymbol{\beta}_{i,j} = \mathbf{0}]\mathbb{P}[\boldsymbol{\beta}_{i,j} = \mathbf{0}]\\ + \mathbb{P}[\forall\ell:\ |\widehat{\beta}_{i,j}(\ell)| < \alpha_0|\boldsymbol{\beta}_{i,j} \neq \mathbf{0}]\mathbb{P}[\boldsymbol{\beta}_{i,j} \neq \mathbf{0}],\end{aligned}$$

where ℓ can take values in the $1,\dots,L$ range. In the significance test method, first perform an ordinary least squares to obtain $\beta_{i,j}$. Then construct tests as following:

$$\mathcal{H}_0^1 : \beta_{i,j}(1) = 0, \tag{7.7}$$

$$\mathcal{H}_0^2 : \beta_{i,j}(2) = 0, \tag{7.8}$$

$$\dots, \tag{7.9}$$

$$\mathcal{H}_0^L : \beta_{i,j}(L) = 0. \tag{7.10}$$

We report an edge if any of the hypotheses in Eq. (7.10) is rejected. The consistency of the test above can be established using asymptotic normality of the ordinary least squares estimate. For significance tests in the form of $\beta_{i,j}(\ell) < \alpha_0$ for some $\alpha_0 < \alpha$ the error can be bounded as following:

$$\begin{aligned}\mathbb{P}[\text{Error}] &\leq \mathbb{P}[\text{Identify } x_i \to x_j|x_i \not\to x_j]\\ &= \sum_{\ell=1}^{L}\mathbb{P}[|\widehat{\beta}_{i,j}(\ell)| > \alpha_0|\boldsymbol{\beta}_{i,j} = 0]\\ &= 2LQ\left(\frac{\alpha_0}{\sqrt{T-Lv}}\right).\end{aligned}$$

where $Q(t)$ is the tail probability of the Gaussian distribution and v is the variance of individual $\widehat{\beta}_{i,j}$ (without loss of generality they are assumed to be equal.). Since $Q(t) < t^{-1}\exp(-\frac{1}{2}t^2)$ for $t > 0$, the above probability can be bounded as follows

$$\mathbb{P}[\text{Error}] \leq 2cL\sqrt{T-L}\exp\left(-\frac{c^2}{2}(T-L)\right), \tag{7.11}$$

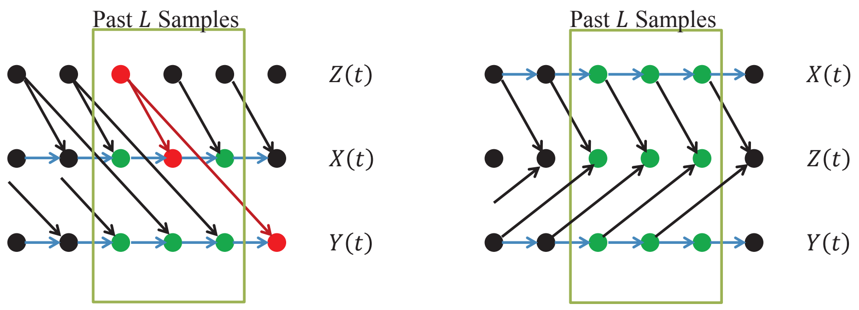
where $c = \frac{v}{\alpha_0}$. Uniform consistency is established due to existence of uniformly consistent tests for association [24].

Obtaining the probability of error for Lasso requires further assumptions in [31]; it can be shown that the model selection error is linearly proportional to L and diminishes with rate $o(\exp(-T^{\nu}))$, for some $0 \leq \nu < 1$.[1]

Proposition 7.1. *Suppose in a VAR model with some unobserved time series. If Assumption 4.1 holds. The graph $\widehat{G}$ learned Significance test and any consistent L_1 variable selection method does not have any spurious causal edge. It may be different with the true Granger graph G in identification of direct causality structures.*

Proof. The proof follows from the fact that given causal sufficiency (Assumption 4.1) an unobserved variable can have two possibilities: (i) it is not a descendant of another observed variable. In this case it can be merged

[1] Note that as long as a method selects individual elements of $\boldsymbol{\beta}_{i,j}$ separately, the error will scale linearly with L for $L \ll T$. The Group-Lasso which selects all elements of $\boldsymbol{\beta}_{i,j}$ at once improves this linear dependence.



(a)

(b)

Figure 7: The diagrams for proving (a) Lemma 3.1 and (b)Lemma 3.2. The green circles are observed variables and the red path shows a d-connected path.

with noise term in the VAR model or (ii) it has an observed parent X. In this case all the causal paths such as $X \to Z \to Y$ where Z unobserved can be converted to a direct causality $X \to Y$ structures. The resulting VAR model can be consistently inferred as shown in Proposition 4.2.

7.3 Proof of Lemma 7.1 First Lets establish another lemma.

Lemma 7.1. *Suppose* $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ *where* $\mathbf{X}$ *is a* $n \times p$ *design matrix with iid Gaussian random elements,* $\boldsymbol{\beta}$ *is a* $p \times 1$ *constant vector and* $\boldsymbol{\varepsilon}$ *is a iid Gaussian random vector. Let* $\widehat{\boldsymbol{\beta}}_\lambda$ *denote the solution of ridge regression with regularization parameter* λ. *We have the following result for* $\lambda \to 0$:

$$\mathbb{E}_{\mathbf{X},\boldsymbol{\varepsilon}}[\widehat{\boldsymbol{\beta}}_\lambda] = \begin{cases} \boldsymbol{\beta} & \text{if } p \le n \\ \frac{n}{p}\boldsymbol{\beta} & \text{if } p > n. \end{cases}$$

Proof. The Ridge regression solution can be written as $\widehat{\boldsymbol{\beta}}_\lambda = (\mathbf{X}^\top\mathbf{X} + \lambda I)^{-1}\mathbf{X}^\top\mathbf{y}$. Suppose the Singular Value Decomposition (SVD) of $\mathbf{X}$ is in the form of $\mathbf{X} = UDV^\top$. Substitution of the decomposition in the regression yields $\widehat{\boldsymbol{\beta}}_\lambda = V(D^\top D + \lambda I)^{-1}D^\top U^\top\mathbf{y}$. Defining $\overline{D}_\lambda = (D^\top D + \lambda I)^{-1}D^\top$ and using the assumption that $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ yields:

$$(7.12) \qquad \widehat{\boldsymbol{\beta}}_\lambda = V\overline{D}_\lambda DV^\top\boldsymbol{\beta} + V\overline{D}_\lambda U^\top\boldsymbol{\varepsilon}.$$

Taking expectation with respect to $\boldsymbol{\varepsilon}$ eliminates the second term because of independence of $\boldsymbol{\varepsilon}$ and $\mathbf{X}$. The matrix $\overline{D}_\lambda D$ is a $p \times p$ matrix in the form of $\mathrm{diag}(\frac{d_1^2}{d_1^2+\lambda}, \ldots, \frac{d_n^2}{d_n^2+\lambda}, 0, \ldots, 0)$. The first term in the right hand side of Eq. (7.12) can be interpreted as (i) rotation via $V^\top$, (ii) setting $p - n$ elements to zero via multiplication with $\overline{D}_\lambda D$ and finally (iii) rotating back to original space via V. Because of randomness of $\mathbf{X}$, V will be uniformly rotating random matrix. Using a geometrical argument we can complete the proof. Fig. 8 shows the three steps: (1) rotate clockwise by θ (blue arrow), (2) Project into one dimension (green on vertical axis) and (3) rotate counterclockwise by θ (the inclined green arrow). Note that there is always another arrow from the process for $-\theta$ rotations (shown in red) which neutralizes the components in the vertical direction.

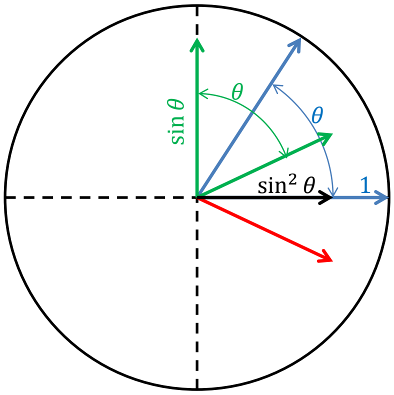


Figure 8: The illustration of the three step process rotation-projection-rotation in two dimensions.

Note that since we assume the rotation is uniformly random, for each rotation $+\theta$ (shown in green) there is neutralizing vector from $-\theta$. Thus for $\boldsymbol{\beta} = (1,0)^\top$ in Fig. 8 we have:

$$\mathbb{E}_{\mathbf{X},\varepsilon}[\widehat{\beta}_\lambda] = \mathbb{E}_{\theta,\varepsilon}[\sin^2\theta] = \frac{1}{2\pi}\int_0^{2\pi}\sin^2\theta d\theta = \frac{1}{2}.$$

The proof holds for any arbitrary standard basis. Since the ridge regression solution is a linear function of $\boldsymbol{\beta}$, we can decompose any $\boldsymbol{\beta}$ as $\beta_1(1,0)^\top + \beta_2(0,1)^\top$ and use the above proof for each basis. This concludes the proof.

Now Lemma 7.1 can be easily shown:

Proof. The proof is based on a direct application of Lemma 7.1 which implies all the edge values decay as the dimensionality increases. Note that in the the Granger causality test we perform a regression with $T - L$ observations and nL features which yields $\mathbb{E}_{\mathbf{X},\boldsymbol{\varepsilon}}[\widehat{\boldsymbol{\beta}}_\lambda] = (\frac{T/L-1}{n})\boldsymbol{\beta}$ as $\lambda \to 0$ which highlights the effect of large L choice.

7.4 Proof of Theorem 4.1

Proof. In order to prove Theorem 4.1 we need to show that the corresponding regression problem with Winzorized mapped version of variables is consistent. Here we show the proof for the bias term; the proof for the variance term follows the same lines.

Consider the following linear model:

$$y = \boldsymbol{\beta}^\top \mathbf{x} + \varepsilon,$$

where $\mathbf{x}$ is a $p \times 1$ zero mean Gaussian random vector, $\boldsymbol{\beta}$ is the coefficient vector and ε is a zero-mean Gaussian noise. Suppose in observation of n samples $\mathbf{x}_i$ for $i = 1, \dots, n$, we have access to noisy versions of them $\tilde{\mathbf{x}}_i$ and $\tilde{y}_i$. We know that the estimation of covariance based on $\tilde{x}_i$ is consistent with the following rate [16]

$$\max_{j,k} \left| \tilde{S}^n_{jk} - \widehat{S}^n_{jk} \right| = \mathcal{O}_P\left(\sqrt{\frac{\log p \log^2 n}{n^{1/2}}} \right),$$

where $\widehat{S}^n_{jk} = (\mathbf{X}^\top \mathbf{X})_{jk}$ and $\tilde{S}^n_{jk} = (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})_{jk}$ is our estimate of covariance using the tilted samples $\tilde{\mathbf{x}}_i$.

We assume that the matrix $\Delta = \tilde{C} - C$ is positive semi-definite. We can relax this assumption, but we leave it as future work. Modify the bound in Eq. 22 of [20] as following:

$$\gamma^\top \tilde{C} \gamma \leq \lambda \sqrt{s} \left\| \gamma \right\|_2 + \varphi_{max}(\Delta) \tag{7.13}$$

Bounding $\varphi_{max}(\Delta) \leq K_2 \max \left| \tilde{S}^n_{jk} - \widehat{S}^n_{jk} \right|$ for some constant K_2 and deriving the lower bound in Eq. 26 using the fact that $\varphi_{min}(\Delta) \geq 0$ yields the following equation:

$$K \phi_{min} \left\| \gamma \right\|_2^2 \leq \frac{\frac{\lambda}{n}\sqrt{s}}{K \phi_{min} + \varphi_{\max}} \tag{7.14}$$

Since $\varphi_{\max}(\Delta)$ diminishes with respect to $\phi_{min}(\tilde{C})$ according to results from [16] and having the incoherent design assumption [20] for lower bound of $\phi_{min}(\tilde{C})$ the proof establishes by following the steps in [20].