FBLG: A Simple and Effective Approach for Temporal Dependence Discovery from Time Series Data

Dehua Cheng, Mohammad Taha Bahadori, Yan Liu
University of Southern California
Los Angeles, CA 90089
{dehua.cheng,mohammab,yanliu.cs}@usc.edu

ABSTRACT

Discovering temporal dependence structure from multivariate time series has established its importance in many applications. We observe that when we look in reversed order of time, the temporal dependence structure of the time series is usually preserved after switching the roles of *cause* and *effect*. Inspired by this observation, we create a *new* time series by reversing the time stamps of original time series and combine both time series to improve the performance of temporal dependence recovery. We also provide theoretical justification for the proposed algorithm for several existing time series models. We test our approach on both synthetic and real world datasets. The experimental results confirm that this surprisingly simple approach is indeed effective under various circumstances.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Time Series Analysis

Keywords

Time Series Analysis; Generalized Linear Model

1. INTRODUCTION

Discovering temporal dependence structures from multivariate time series is one of the central tasks in time series analysis. It easily finds applications in many domains. For example, in social networks, accurate identification of influence networks from users' time series activity records is of significant importance for advertising, marketing, and psychological studies. In biology, the gene regulatory networks recovered from time series microarray data reveals key information on gene functions.

Inferring dependency network structures from time series data has been extensively studied in the past. The *Granger causality* framework, which establishes temporal dependence structures based on regression techniques, has become popular due to its simplicity, robustness, and extendability [23, 7, 16, 3, 22]. Nowadays, as more and more large-scale time series data become available, traditional approaches for identifying Granger causality are confronted with a series of challenges, such as inconsistency, high computational complexity, and so on. To address these problems, penalized regression techniques (e.g. lasso or lasso-type regressions) have been applied, leading to major improvement for applications with *sparse* temporal dependence structures [29, 2, 27]. However, the overall performance of existing Granger causality techniques still leaves room for improvement. In this paper, we aim to explore a new direction by considering the procedure of reversing the time in time series data.

The inspiration for our work comes from classical mechanics where it is well-known that the basic equations of the classical physics remains valid when we look in reversed order of time, i.e., replacing time stamp t with $-t$. In a simple world, if time flows in the opposite direction, objects interact with each other under the same physical laws, and we will not notice the difference. Instead of explaining all phenomena from the underlying physical law, we usually apply simplified mathematical models to real world events. Since the underlying physics mechanism is time reversible, we would expect our model applies when the time is reversed. The question remains whether we can consolidate and enhance our estimation accuracy by utilizing the information from both directions.[1]

To fully utilize such an idea, we need to examine the effect of reversing the time on the temporal dependence structures. One important assumption of Granger causality is that *the cause occurs before the effect.* If an event A at time t causes an event B to happen at time $t+k$, we will see a correlation between events A and B with time lag k. By reversing time, the correlation between events A and B still exists, with the difference that B occurs before A. Granger causality-based algorithms should suggest that A causes B with time lag k from the original time series. Similarly, we expect that the same algorithm would also indicate that B causes A with time lag k from the reversed time series. Note that our argument is not limited to Granger causality, it also applies to other algorithms that rely on the correlation with time lags between time series, e.g., transfer entropy [26].

KDD'14, August 24–27, 2014, New York, NY, USA.
Copyright 2014 ACM 978-1-4503-2956-9/14/08 ...$15.00.
http://dx.doi.org/10.1145/2623330.2623709 .

[1] It should be noted that, for a closed complex system, the trend of entropy eliminates the ambiguity on the time direction, as suggested by the second law of thermodynamics. But since the model only addresses a particular aspect of the system, the restriction usually does not apply.

The link between the original time series and the reversed time series raises the possibility of combining these two directions for enhanced temporal dependence inference. This motivates us to propose a novel but simple approach, namely forward backward (FB) Granger causality, to infer the temporal dependence structures for multivariate time series. Firstly, we apply Granger causality-based algorithm on both the original time series and the time-reversed time series, then we combine the results by simple averaging. Note that similar approach has been applied in *Natural Language Processing*[21], where they estimate the transition kernel of the Markov chain from both directions. Performance improvement has been observed when the size of data is limited. We provide both theoretical analysis and empirical studies on the effectiveness of the proposed approach. The rest of the paper is organized as follows: we first review the preliminary and related works in Section 2. In Section 3, we describe our FB Granger causality algorithm and provide theoretical analysis on several existing models. Finally, we show experimental results in Section 4 and conclusion in Section 5.

2. PRELIMINARIES AND RELATED WORK

Notation.

We define the forward time series as the original multivariate time series $\{\mathbf{y}^{(t)}\}$, $t = \dots, 0, 1, \dots$, and the backward time series $\{\mathbf{z}^{(t)}\}$ is defined as $\mathbf{z}^{(t)} := \mathbf{y}^{(-t)}$. $\{\mathbf{y}^{(t)}\}$ and $\{\mathbf{z}^{(t)}\}$ both contain N time series; univariate time series are denoted by $\{y_i^{(t)}\}$ and $\{z_i^{(t)}\}$ for $i = 1, \dots, N$. Both $\mathbf{y}^{(t)}$ and $\mathbf{z}^{(t)}$ are vectors of the values for each time series at time t, respectively. If the ith time series at time t is caused by the jth time series at time $t-k$, we say that i is caused by j with lag k. Moreover, we represent this temporal dependence relation by the ordered *temporal dependence triplet* (i, j, k). And the inverse of temporal dependence triplet (i, j, k) is defined as (j, i, k). In addition, $\mathcal{C}_y$ denotes the set of all *temporal dependence triplets* for time series $\{\mathbf{y}^{(t)}\}$.

Related Work.

Causal inference has consistently been an important task for researchers in various fields of science. There are two main tasks in causal inference: (1) How to cancel the confounding bias, e.g., [24] and (2) How to discover the causal structures among the given variables when a set of assumptions are satisfied [28]. In this paper, the second task is our concern and we intend to improve the existing causal discovery algorithms.

The causal discovery task is challenging and may require many assumptions with weak guarantees of finding the true causal structure, see for example the theoretical discussions in [25]. *Granger causality* is one of the most popular approaches to quantify temporal dependence structures 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 [14, 15]. In practice, Granger causality tests are carried out by fitting a *Vector Auto-regression* (VAR) model. 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* [20, ch. 3.6.1]: given multiple time series $\{\mathbf{y}^{(t)}\}$, we run a VAR model as follows,

$$\mathbf{y}^{(t)} = \sum_{\ell=1}^{P} \mathbf{A}_\ell^\top \mathbf{y}^{(t-\ell)} + \boldsymbol{\epsilon}^{(t)}, \tag{1}$$

where P is the maximal time lag. We can determine that time series $\{y_j^{(t)}\}$ Granger causes $\{y_i^{(t)}\}$ if at least one value in the coefficient vector $\{\mathbf{A}_\ell\}_{ij}$ for $\ell = 1, \dots, P$ is nonzero by statistical significant tests. The second approach is the *Lasso-Granger approach* [29, 2, 27], which applies a 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. (1) can be achieved by solving the following optimization problem:

$$\min_{\mathbf{A}_\ell} \sum_{t=L+1}^{T} \left\| \mathbf{y}^{(t)} - \sum_{\ell=1}^{P} \mathbf{A}_\ell^\top \mathbf{y}^{(t-\ell)} \right\|_2^2 + \lambda \sum_{\ell=1}^{P} \|\mathbf{A}_\ell\|_1 , \tag{2}$$

where λ is the penalty parameter, which determines the sparsity of the coefficients $\mathbf{A}_\ell$.

Several approaches have been proposed for identification of Granger causality for nonlinear time series; among the notable ones, kernelized regression [22], nonparametric techniques such as [16, 23, 26], non-Gaussian structural VAR [17], generalized linear autoregressive models [19, 5], and the Copula Granger [4].

The proposed method in this paper is similar to bootstrap aggregating (bagging) techniques [6] in the sense that it averages over the results from multiple datasets. But the fundamental difference between the two techniques stems from the way that the algorithms generate datasets: the bagging techniques sample the original dataset and create subsamples of the dataset and then average over the results of the algorithm on each new datasets. Here we do not subsample the original dataset; instead we create a new dataset by reversing time. Randomization techniques [12] constitute another wide class of dataset manipulation techniques. However, note that the our proposed method is purely deterministic.

3. MODEL ANALYSIS

In this section, we first describe our algorithms for exploiting the information in the backward time series, and then elaborate theoretical bases for the gain achieved by these algorithms.

3.1 Forward Backward Granger Causality

Given a specific temporal dependence inference algorithm, if it indicates the existence of the temporal dependence triplet (i, j, k) based on the forward time series $\{\mathbf{y}^{(t)}\}$, as argued in Section 1, intuitively we would expect the algorithm to find the triplet (j, i, k) based on the backward time series $\{\mathbf{z}^{(t)}\}$. This motivates our core design principle for utilizing this property: we can achieve more robust temporal dependence inference by appropriately combining the results from both the forward time series and the backward time series produced by the same temporal dependence inference algorithm.

The validity of such an approach depends on both the original temporal dependence inference algorithm and how we combine the results. We mainly focus on the Granger causality-based algorithms in this paper.

Algorithm 1 Naive Forward Backward Lasso Granger Causality

```
Input: Time series $\{\mathbf{y}^{(t)}\}$, lag $P$, penalty parameter $\lambda$.
Output: Coefficients $\mathbf{A}_\ell^{FB}$, $\ell = 1, 2, \ldots, P$.
Define the backward time series $\{\mathbf{z}^{(t)}\}$ by $\mathbf{z}^{(t)} = \mathbf{y}^{(-t)}$.
Get forward coefficients $\mathbf{A}_\ell$ via Lasso-Granger with $\{\mathbf{y}^{(t)}\}$, $P$, and $\lambda$.
Get backward coefficients $\mathbf{B}_\ell$ via Lasso-Granger with $\{\mathbf{z}^{(t)}\}$, $P$, and $\lambda$.
Return $\mathbf{A}_\ell^{FB} = \frac{1}{2}(\mathbf{A}_\ell + \mathbf{B}_\ell^\top)$, $\ell = 1, 2, \ldots, P$.
```

Algorithm 2 Naive Forward Backward Copula Lasso Granger Causality

```
Input: Time series $\{\mathbf{y}^{(t)}\}$, lag $P$, penalty parameter $\lambda$.
Output: Coefficients $\mathbf{A}_\ell^{FB}$, $\ell = 1, 2, \ldots, P$.
for each $i = 1, 2, \ldots, N$ do
    Transform $y_i^{(t)} \to w_i^{(t)}$ by equation 3.
end for
Get coefficients $\mathbf{A}_\ell^{FB}$ by calling algorithm 1 with $\{\mathbf{w}^{(t)}\}$, $P$, and $\lambda$.
Return $\mathbf{A}_\ell^{FB}$, $\ell = 1, 2, \ldots, P$.
```

In general, suppose that the assumptions for correctness of Granger causality are satisfied such that the coefficients estimated by Granger causality indicate the existence of temporal dependence relationships; such assumptions have been studied in [13, 4]. The simplest way to combine the results is to add the coefficients for (i, j, k) in the forward time series and (j, i, k) in the backward time series, which yields the *Naive Forward Backward Lasso Granger Causality Algorithm*, shown in Algorithm 1. It is important to note that since we only flipped the temporal order of the original dataset, the results from forward and backward time series are expected to be correlated. But the coefficients are not fully correlated for time series with finite length, which is supported by our experimental results.

However, Granger causality is designed for linear time series, which is not always the case for the data of interest. Given a time series $\{\mathbf{x}^{(t)}\}$, we can map the data using the empirical marginal distribution of time series to the Gaussian distribution by

$$y_i^{(t)} = s_i \Phi^{-1}(\hat{F}(x_i^{(t)})), \quad \text{for } i = 1, \ldots, N, \tag{3}$$

where $\hat{F}$ is the empirical *cumulative distribution function* (CDF) of the ith time series, Φ is the CDF for standard Gaussian distribution and s_i is the standard derivative of the ith time series, which helps to retain original information. $\{\mathbf{y}^{(t)}\}$ will be treated as a linear representation for the original time series, to which we can apply Granger causality based algorithms, e.g., the *Copula Lasso Granger Causality* [4], and similarly, *Naive Forward Backward Copula Lasso Granger Causality* as described in algorithm 2.

3.2 Analysis of Continuous Time Series

In this section[2], we show that for the time series generated from *Vector Autoregressive Model* (VAR), the backward time series is also a VAR under some conditions. To do so, we assume $\boldsymbol{\epsilon}^{(t)} \sim \mathcal{N}(0, \gamma I)$, where γ is a constant which governs the level of noise, and $\mathcal{N}(\mu, \sigma^2)$ denotes the Gaussian (normal) distribution. The VAR model with the Gaussian noise uniquely defines a multivariate Gaussian distribution on $(\mathbf{y}^{(t)}, \mathbf{y}^{(t-1)}, \ldots)$. This provides the foundation for us to derive the conditional distribution of the same form as equation (1) for *backward time series* $\{\mathbf{z}^{(t)} | \mathbf{z}^{(t)} \equiv \mathbf{y}^{(-t)}\}$. We show that the backward time series is also a VAR model only with different set of coefficients $\mathbf{B}_i$ and noise. However, for arbitrary VAR, the causation defined on the forward time series $\mathbf{y}^{(t)}$ is not the same as the inverse of the causation defined on the backward time series $\{\mathbf{z}^{(t)}\}$, i.e., the set of causation triplets of the backward time series $\mathcal{C}_z \neq \{(j, i, k) | (i, j, k) \in \mathcal{C}_y\}$ where $\mathcal{C}_y = \{(i, j, k)\}$ is the set of causation triplets of the forward time series. In the following theorem, we show that the temporal dependence triplets on the backward time series $\{\mathbf{z}^{(t)}\}$ is closely related to the inverse of the triplets on $\{\mathbf{y}^{(t)}\}$.

Theorem 3.1. *If the forward time series $\{\mathbf{y}^{(t)}\}$ is stable and there exists $\delta > 0$ and a matrix norm $|||\cdot|||$, e.g., the Frobenius norm, so that $|||\mathbf{A}_i||| < \delta, \forall i = 1, 2, \ldots, P$. Then the backward time series $\{\mathbf{z}^{(t)}\}$ is also a VAR, defined by*

$$\mathbf{z}^{(t)} = \sum_{i=1}^{P} \mathbf{B}_i \mathbf{z}^{(t-i)} + \boldsymbol{\omega}_t,$$

where $\boldsymbol{\omega}_t \sim \mathcal{N}(0, \gamma[\mathbf{I} + \Theta(\delta)])$ and $\mathbf{B}_i = \mathbf{A}_i^\top + o(\delta)$, $\forall$ i.

Proof. By the definition of $\{\mathbf{y}^{(t)}\}$, we have

$$\mathbf{y}^{(t)} | \mathbf{y}^{(t-1)}, \ldots, \mathbf{y}^{(t-P)} \sim \mathcal{N}(\sum_{i=1}^{P} \mathbf{A}_i \mathbf{y}^{(t-i)}, \gamma \mathbf{I}).$$

Because time series $\{\mathbf{y}^{(t)}\}$ is stable, so it is also strictly stationary [20, Ch. 2.1.3] and the marginal distribution of the P consecutive time stamps has the following representation:

$$\mathbf{Y}^{(t)} = (\mathbf{y}^{(t+1)\top}, \mathbf{y}^{(t+2)\top}, \ldots, \mathbf{y}^{(t+P)\top})^\top \sim N(0, \gamma\{\Lambda_{ij}\}^{-1}),$$

where the $\mathbf{Y}^{(t)}$ is an $NP \times 1$ vector, and the $\{\Lambda_{ij}\}$ is the precision matrix (represented in blocks), each block Λ_{ij} is an $N \times N$ matrix. Given the stationarity of the time series, we can set $t = 0$ without the loss of generality. The probability density function (PDF) of the marginal distribution of the $P + 1$ consecutive time stamps are proportional to

$$\begin{aligned}
&\exp[-\frac{1}{2\gamma}(\sum_{i,j=1}^{P} \mathbf{y}^{(i)\top} \Lambda_{ij} \mathbf{y}^{(j)} \\
&+ (\mathbf{y}^{(P+1)} - \sum_{i=1}^{P} \mathbf{A}_i \mathbf{y}^{(i)})^\top (\mathbf{y}^{(P+1)} - \sum_{i=1}^{P} \mathbf{A}_i \mathbf{y}^{(i)}))] \\
&= \exp[-\frac{1}{2\gamma}(\sum_{i,j=1}^{P+1} \mathbf{y}_i^\top (\Lambda_{ij} + \mathbf{A}_i^\top \mathbf{A}_j) y_j)],
\end{aligned}$$

where $\mathbf{A}_{P+1} = -\mathbf{I}$, $\Lambda_{P+1,i} = \Lambda_{j,P+1} = \mathbf{0}$. The PDF of the conditional distribution of $\mathbf{y}^{(1)}$ given $\mathbf{y}^{(2)}, \ldots, \mathbf{y}^{(P+1)}$ is proportional to

$$\exp[-\frac{1}{2\gamma}(\mathbf{y}^{(1)\top}(\Lambda_{11}+\mathbf{A}_1^\top \mathbf{A}_1)\mathbf{y}^{(1)} - 2\sum_{i=2}^{P+1} \mathbf{y}^{(i)\top}(\Lambda_{i1}+\mathbf{A}_i^\top \mathbf{A}_1)\mathbf{y}^{(1)})$$

$$\sim \mathcal{N}((\Lambda_{11}+\mathbf{A}_1^\top \mathbf{A}_1)^{-1}(-\sum_{i=2}^{P+1}(\Lambda_{1i}+\mathbf{A}_1^\top \mathbf{A}_i)\mathbf{y}^{(i)}), \gamma(\Lambda_{11}+\mathbf{A}_1^\top \mathbf{A}_1)^{-1}).$$

[2] Note that by continuous or discrete time series, we refer to whether the variables take on continuous or discrete values.

To study the structure of the covariance matrix $\{\Lambda_{ij}\}^{-1}$, we recall that the *Moving Average* representation of VAR model, which is

$$\mathbf{Y}^{(t)} \sim \sum_{i=1}^{+\infty} \mathbf{A}^i \mathbf{U}^{(t-i)},$$

where $\mathbf{U}^{(t)} = (0^\top, 0^\top, \ldots, \epsilon^{(t)\top})^\top$, and

$$\mathbf{A} = \begin{bmatrix} 0 & \mathbf{I} & 0 & 0 \\ 0 & 0 & \mathbf{I} & \vdots \\ 0 & 0 & \ddots & \mathbf{I} \\ \mathbf{A}_P & \mathbf{A}_{P-1} & \ldots & \mathbf{A}_1 \end{bmatrix}.$$

So we have that

$$\gamma\{\Lambda_{ij}\}^{-1} = \gamma \sum_{i=1}^{+\infty} \mathbf{A}^i \Sigma_{t-i} (\mathbf{A}^i)^T,$$

where the lower right block of $\mathbf{\Sigma}_{t-i}$ is $\mathbf{I}$, otherwise is $\mathbf{0}$.

By some derivations, we can show that

$$\{\Lambda_{ij}\}^{-1} = \mathbf{I} + \Theta(\delta),$$

where the diagonal blocks are $\mathbf{I}+\Theta(\delta)$ by setting all $\mathbf{A}_i = \mathbf{0}$. And the first row of blocks are $(\mathbf{I}+\Theta(\delta), \mathbf{A}_P^\top + o(\delta), \ldots, \mathbf{A}_2^\top + o(\delta))$, which can be derived by studying the first $P+1$ terms in the series. So by inversion, we have that

$$\Lambda_{1i} = -\mathbf{A}_{P+2-i}^T + o(\delta), \forall i = \{2, \ldots P\},$$

$$\Lambda_{11} = \mathbf{I} + o(1).$$

Together with $\mathbf{A}_1^T \mathbf{A}_i = o(\delta), \forall i = 1, 2, \ldots, P$, we have

$$-(\Lambda_{1i} + \mathbf{A}_1^T \mathbf{A}_i) = \mathbf{A}_{P+2-i}^T + o(\delta),$$

$$(\Lambda_{11} + \mathbf{A}_1^T \mathbf{A}_1) = \mathbf{I} + o(1) = \mathbf{I} + \Theta(\delta).$$

By replacing $\mathbf{y}^{(t)}$ with $\mathbf{z}^{(-t)}$, we have

$$\mathbf{z}^{(t)} = \sum_{i=1}^{P} \mathbf{B}_i \mathbf{z}^{(t-i)} + \boldsymbol{\omega}_t = \sum_{i=1}^{P} [\mathbf{A}_i^\top + o(\delta)] \mathbf{z}^{(t-i)} + \boldsymbol{\omega}_t,$$

where $\boldsymbol{\omega}_t \sim \mathcal{N}(0, \gamma[\mathbf{I} + \Theta(\delta)])$. □

The assumption in Theorem 3.1 implies that the strength of the influence in time series $\{\mathbf{y}^{(t)}\}$ has an upper bound, which is sufficient but not necessary. It can be relaxed, since the proof only requires that the higher order product between $\mathbf{A}_i$ is negligible comparing with $\mathbf{A}_i$ itself. This assumption can be easily satisfied when the temporal dependence structure is sparse, which is usually true in real world applications.

Theorem 3.1 shows that the first order components of the influence on the backward VAR is exactly the inverse of the forward VAR, i.e., not only $(i, j, k) \in \mathcal{C}_y \Leftrightarrow (j, i, k) \in \mathcal{C}_z$, but they also share the same strength $\{\mathbf{A}_k\}_{ij}$, which indicates a much stronger link between the forward time series and the backward time series. The link also results in a simple form, which justifies our approach of combining the results from both directions, i.e., averaging on the corresponding coefficients. Moreover, Granger causality provides an unbiased estimation for both directions, and the results in [27] indicate that they have the same variance. Therefore the average of both directions is also unbiased with smaller variance, when the correlation of two estimations are strictly less than 1. Additionally, when we have sufficiently long time series, i.e., $T \gg N$, we would expect that the forward backward approach provides an estimation similar to the original Lasso Granger causality, since both forward and backward should provide accurate coefficients estimation, as suggested by the consistency of the penalized maximal likelihood estimation. This phenomenon has been observed in our experiments on synthetic datasets.

For nonlinear time series, we apply the copula transformation before Granger causality. In order to show similar theoretical results for this approach, we need the data to be generated from the *Granger Non-paranormal (G-NPN) model* as follows:

Definition.

Granger Non-paranormal (G-NPN) model We say a time series $\mathbf{x}^{(t)} = (x_1^{(t)}, \ldots, x_N^{(t)})$ has Granger-Nonparanormal distribution $G - NPN(\mathbf{x}, \mathbf{A}, F)$ if there exist monotonically increasing functions $\{F_i\}_{i=1}^N$ such that $F_i(x_i^{(t)})$ for $i = 1, \ldots, N$ are jointly Gaussian and can be factorized according to the VAR model with coefficients $\mathbf{A} = \{\mathbf{A}_k\}_{k=1}^P$. More specifically, the joint distribution for the transformed random variables $y_i^{(t)} \triangleq F_i(x_i^{(t)})$ can be factorized as follows

$$p_{\mathbf{y}^{(t)}}(\mathbf{y^{(t)}}) = \mathcal{N}(\mathbf{y}^{(1:P)}) \prod_{i=1}^{N} \prod_{t=P+1}^{T} p_{\mathcal{N}}(y_j^{(t)} | \sum_{k=1}^{P} \mathbf{A}_k \mathbf{y}^{(t-k)}, \sigma_j),$$

where $p_{\mathcal{N}}(y|\mu, \sigma)$ is the Gaussian density function with mean μ and variance σ^2.

Proposition 3.1. *Using the copula transformation on the data generated from G-NPN model, the forward and backward relationships in Theorem 3.1 hold for the transformed random processes.*

Proof. Our proof is mainly to show that the copula transformation recovers the original vector auto-regressive process. The key step to prove this result is to show that if $X \sim \mathcal{N}(0, 1)$ and $Y = F(X)$, where $F(\cdot)$ is a monotonically increasing function, then $\Phi^{-1}(F_Y(Y)) \sim \mathcal{N}(0, 1)$. Furthermore, the independence relationships will be preserved a the copula mapping, as the transformation $\Phi^{-1}(F_Y(\cdot))$ is a deterministic transformation. This is because $X \perp\!\!\!\perp Y$ 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 $g(\cdot)$ and $h(\cdot)$.

After applying the above result to each variable $x_i^{(t)}$, we can show [4] that by using the copula transformation we obtain $y_i^{(t)}$ which are multivariate Gaussian as defined in the definition of G-NPN. □

The definition of the Granger Non-paranormal (G-NPN) model indicates that the underlying mechanism of time series $\{\mathbf{x}^{(t)}\}$ is a linear time series, which subsumes numerous circumstances. When $\{\mathbf{x}^{(t)}\}$ is an observation of $\{\mathbf{y}^{(t)}\}$ with deterministic bias, the copula transformation helps to restore the original information as stated in Proposition 3.1. And then we can apply our argument for the VAR model.

3.3 Analysis of Discrete Time Series

Nowadays, social media provides a rich source for time series analysis because the interactions among individuals are naturally reflected in the time series of action logs. One way to analyze the social influence among users is to create time series of user activity by assigning 1 to a user at a particular time interval if she has at least one activity in that time interval and 0 otherwise [1]. For example, tweeting (i.e., posting on Twitter) activity naturally defines a time series, where $y_i^{(t)} = 1$ if user i posts at time interval t, and $y_i^{(t)} = 0$ otherwise. We can also recover influence relationship among users based on retweeting. For example, if user i retweets user j, we say that i has been influenced by j. Such social network time series pose a unique challenge for the temporal dependence inference algorithms.

In this section, we first define a general type of binary time series, which includes many existing models. Then, with additional Assumption 3.2, we prove that applying Granger causality to binary time series provides consistent temporal dependence structure recovery. Furthermore, by Assumption 3.2 and Lemma 3.4, we build the connection between the forward times series and the backward time series, which leads to the consistency results on temporal dependence structure recovery for the backward time series. By consistency we refer to that with appropriate thresholding on the estimated coefficients, we can correctly recover all temporal dependence triplets from the time series. Note that we are applying Granger causality on a misspecified model (i.e., the binary time series is not generated from VAR), the consistency results also justify our approach which applies Granger causality on certain non-VAR time series.

Given a binary multivariate time series $\{\mathbf{y}^{(t)}\}$, we say the ith series is activated at time t if and only if $y_i^{(t)} = 1$. In addition to the previous notations, we denote $\mathbf{y}^{(t-1:t-P)} = (\mathbf{y}^{(t-1)}, \mathbf{y}^{(t-2)}, \dots, \mathbf{y}^{(t-P)})$. $\Omega_{\mathbf{y}^{(t-1:t-P)}}$ is the set of activated variables in $\mathbf{y}^{(t-1:t-P)}$. We have the following assumptions:

Assumption 3.1.
Markov Assumption

The probability of activation for any variable at time t only depends on the states of the most recent P times $(t-1, t-2, \dots, t-P)$, which is

$$\begin{aligned} P(y_i^{(t)} = 1|\mathbf{y}^{(t-1)}, \mathbf{y}^{(t-2)}, \dots) &= P(y_i^{(t)} = 1|\mathbf{y}^{(t-1:t-P)}) \\ &= P(y_i^{(t)} = 1|\Omega_{\mathbf{y}^{(t-1:t-P)}}). \end{aligned}$$

The last equality is because of variables only take binary value, so the status of $\mathbf{y}^{(t-1:t-P)}$ is uniquely defined by $\Omega_{\mathbf{y}^{(t-1:t-P)}}$.

Activation Rate Monotonicity

$y_j^{(t-k)} \notin \Omega_{\mathbf{y}^{(t-1:t-P)}}$ implies

$$P(y_i^{(t)} = 1|\{y_j^{(t-k)}\} \cup \Omega_{\mathbf{y}^{(t-1:t-P)}}) \geq P(y_i^{(t)} = 1|\Omega_{\mathbf{y}^{(t-1:t-P)}}),$$

which means that there is no negative influence on the activation rate if more variables from the histories become activated.

Influcence Significance

If there exist a set $\Omega_{\mathbf{y}^{(t-1:t-P)}}$, such that $y_j^{(t-k)} \notin \Omega_{\mathbf{y}^{(t-1:t-P)}}$ and $P(y_i^{(t)} = 1|\Omega_{\mathbf{y}^{(t-1:t-P)}} \cup \{y_j^{(t-k)}\}) > P(y_i^{(t)} = 1|\Omega_{\mathbf{y}^{(t-1:t-P)}})$, then there exist $\delta_{ijk} > 0$, for any $\Omega_{\mathbf{y}^{(t-1:t-P)}}$, $y_j^{(t-k)} \notin \Omega_{\mathbf{y}^{(t-1:t-P)}}$ implies

$$P(y_i^{(t)} = 1|\Omega_{\mathbf{y}^{(t-1:t-P)}} \cup \{y_j^{(t-k)}\}) > P(y_i^{(t)} = 1|\Omega_{\mathbf{y}^{(t-1:t-P)}}) + \delta_{ijk}.$$

And we say j is a cause for i with lag k

The last term in Assumption 3.1 actually implies that the causation is significant under any circumstances, i.e. the activation of time series j at time $t-k$ will increase the activation rate of time series i at time t by at least δ_{ijk}, regardless of other variables in $\mathbf{y}^{(t-1:t-P)}$, or the status of $y_i^{(t)}$ does not depends on the status of $y_j^{(t-k)}$ at all. We stress that Assumption 3.1 can be easily satisfied in practical applications and many existing models fall in this category.

Example 3.1. ***Independent Cascade**[18] (IC) model is originally proposed for modeling the diffusion process in social networks. We modify it to model the activity on networks over time. IC model defined on a weighted directed graph $\{\mathcal{V}, \mathcal{E}\}$, with each vertex represents an individual in the network. If vertex v is activated at time t, it attempts to activate its neighbor s with probability $p_{v\to s}$ at time $t+1$ independently. If any neighbor of s activates s successfully, s will be marked as activated at time $t+1$. It can also activate itself by probability μ_s. The activation rate has the following representation*

$$P(s^{(t+1)} = 1) = 1 - (1-\mu_s) \prod_{\substack{v^{(t)}=1 \\ (v\to s)\in\mathcal{E}}} (1 - p_{v\to s}).$$

And the IC model satisfies Assumption 3.1.

Note that because we are applying Granger causality as a misspecified model, we need one more assumption to support the consistency results on binary time series.

Assumption 3.2. ***Diminishing Influence** Let a binary time series $\{y_i^{(t)}\}i = 1, 2, \dots, N$ satisfy Assumption 3.1. Let's denote $\mathbf{y}^{(t-1:t-P)} - \{y_j^{(t-k)}\}$ by $\mathcal{S}_{jk}$, then we assume*

$$\frac{P(\mathcal{S}_{jk}|y_j^{(t-k)} = 1) - P(\mathcal{S}_{jk}|y_j^{(t-k)} = 0)}{P(\mathcal{S}_{jk}|y_j^{(t-k)} = 1) + P(\mathcal{S}_{jk}|y_j^{(t-k)} = 0)} = O(\frac{1}{N}),$$

for all possible values of $\mathcal{S}_{jk}$ as $N \to +\infty$.

And for any vector β_{jk} of the same size as $\mathcal{S}_{jk}$, we assume that $||\beta_{jk}||_\infty = O(1)$ and

$$E[\beta_{jk} \cdot \mathcal{S}_{jk}|y_j^{(t-k)} = 1] - E[\beta_{jk} \cdot \mathcal{S}_{jk}|y_j^{(t-k)} = 0] = O(\frac{1}{N}),$$

for all possible value of $\mathcal{S}_{jk}$ as $N \to +\infty$.

Assumption 3.2 shows that the influence of an individual on the entire network is diminishing as the size of the network increases. For example, the difference of joint distribution of $\mathcal{S}_{jk}$ given $y_j^{(t-k)} = 0$ or $y_j^{(t-k)} = 1$ represents the influence of $y_j^{(t-k)}$ on $\mathcal{S}_{jk}$, e.g., when there is no difference, $y_j^{(t-k)}$ and $\mathcal{S}_{jk}$ are independent. Moreover, note that, if Assumption 3.2 holds for $\{\mathbf{y}^{(t)}\}$, then $\{\mathbf{z}^{(t)}\}$ also satisfy the same assumption, since $\mathbf{z}^{(t-1:t-P)} = \mathbf{y}^{(1-t:P-t)}$.

Assumption 3.3. *Given a binary time series $\{y_i^{(t)}\}$, $i = 1, 2, \dots, N$, we have*

$$P(y_i^{(t)}|y_j^{(t-k)} = 1) - P(y_i^{(t)}|y_j^{(t-k)} = 0) \geq \Theta(\delta_{ijk}) > 0,$$

if j is a cause for i with lag k. Otherwise,

$$P(y_i^{(t)}|y_j^{(t-k)} = 1) - P(y_i^{(t)}|y_j^{(t-k)} = 0) = O(\frac{1}{N}),$$

as $N \to +\infty$.

Assumption 3.3 helps us to establish the connection between the forward time series and the backward time series, and it is important for the consistency result in Theorem 3.3. Furthermore, we have an important lemma connecting the Assumption 3.3 with 3.1 and 3.2.

Lemma 3.2. *Given a binary time series $\{y_i^{(t)}\}, i = 1, 2, \dots, N$ satisfy Assumption 3.1 and 3.2, then it satisfies Assumption 3.3.*

Proof. Given the binary time series $\{y_i^{(t)}\}$, $i = 1, 2, \dots, N$, we denote $y^{(t-1:t-P)} - \{y_j^{(t-k)}\}$ by $\mathcal{S}_{jk}$, and we have

$$P(y_i^{(t)}|y_j^{(t-k)}) = \sum_{\mathcal{S}_{jk}} P(y_i^{(t)}|y^{(t-1:t-P)})P(\mathcal{S}_{jk}|y_j^{(t-k)}).$$

By Assumption 3.1 and 3.2, if j is a *cause* for i with lag k, we have

$$\begin{aligned}P(y_i^{(t)}|y_j^{(t-k)} = 1) - &P(y_i^{(t)}|y_j^{(t-k)} = 0)\\ &\geq \sum_{\mathcal{S}_{jk}} (\Theta(\delta_{ijk}) + O(1/N))P(\mathcal{S}_{jk}|y_j^{(t-k)} = 0)\\ &= \Theta(\delta_{ijk}).\end{aligned}$$

Otherwise, we have

$$\begin{aligned}P(y_i^{(t)}|y_j^{(t-k)} = 1) - &P(y_i^{(t)}|y_j^{(t-k)} = 0)\\ &= \sum_{\mathcal{S}_{jk}} O(1/N)P(\mathcal{S}_{jk}|y_j^{(t-k)} = 0)\\ &= O(1/N),\end{aligned}$$

as $N \to +\infty$. □

We now state our main theorem in this section, which suggests the consistency of Granger causality on binary time series with appropriate thresholding.

Theorem 3.3. *Let a binary time series $\{y_i^{(t)}\}, i = 1, 2, \dots, N$ satisfy Assumption 3.2 and 3.3. We denote the coefficients by $\mathbf{A}_k, k = 1, 2, \dots, P$ as in VAR model, which are estimated by applying Granger causality on time series $\{\mathbf{y}^{(t)}\}$, we have*

$$\{\mathbf{A}_k\}_{ij} \geq \Theta(\delta_{ijk}) > 0,$$

if j is a cause *for i with lag k. Otherwise,*

$$\{\mathbf{A}_k\}_{ij} = O(\frac{1}{N}),$$

as $N \to +\infty$.

Proof. As $T \to +\infty$, we have the objective function of the regression as follows:

$$\sum_{\mathbf{y}^{(t-1:t-k)}} \sum_{i} \sum_{y_i^{(t)}} P(y_i^{(t)}, \mathbf{y}^{(t-1:t-k)})(y_i^{(t)} - b_i - \sum_{k=1}^{P} \mathbf{A}_k \mathbf{y}^{(t-k)})^2.$$

Note that we do not sum over t, since we weight the loss term by its own marginal distribution. Without loss of generality, we study the coefficient $\{\mathbf{A}_k\}_{ij}$ (denoted by β_{ijk}), which connecting $y_i^{(t)}$ and $y_j^{(t-k)}$. Let us denote $\mathbf{y}^{(t-1:t-P)} - \{y_j^{(t-1:t-P)}\}$ by $\mathcal{S}_{jk}$ and the associated coefficients by β_{jk}. By absorbing the constant into b, we can shift the value of $y_j^{(t-k)}$ from $\{0, 1\}$ to $\{-0.5, 0.5\}$. The related objective is as follows:

$$\sum_{\substack{\mathcal{S}_{jk}\\ \mathbf{y}^{(t-1:t-k)}}} P(y_i^{(t)}, \mathbf{y}^{(t-1:t-k)})(y_i^{(t)} - b_i - y_j^{(t-k)}\beta_{ijk} - \mathcal{S}_{jk} \cdot \beta_{jk})^2.$$

By taking the derivative w.r.t. β_{ijk}, we have

$$\begin{aligned}&P(y_i^{(t)} = 1|y_j^{(t-k)} = -0.5) - P(y_i^{(t)} = 1|y_j^{(t-k)} = 0.5) + \beta_{ijk}\\ &+E[\mathcal{S}_{jk} \cdot \beta_{jk}|y_j^{(t-k)} = 0.5] - E[\mathcal{S}_{jk} \cdot \beta_{jk}|y_j^{(t-k)} = -0.5] = 0.\end{aligned}$$

Recall that $\{\mathbf{y}^{(t)}\}$ satisfy Assumption 3.2, which indicates that

$$\beta_{ijk} \geq \Theta(\delta_{ijk}) > 0,$$

if j is a *cause* for i with lag k. Otherwise,

$$\beta_{ijk} = O(\frac{1}{N}),$$

as $N \to +\infty$. This proves the theorem. □

Theorem 3.3 indicates that by setting an appropriate threshold (on the order of $\min_{\{i,j,k\}}\{\delta_{ijk}\}$) on the coefficients, we can reconstruct the correct temporal dependence structure. We now explain the connection between the forward time series $\{\mathbf{y}^{(t)}\}$ and the backward time series $\{\mathbf{z}^{(t)}\}$.

Lemma 3.4. *Given a binary time series $\{y_i^{(t)}\} i = 1, 2, \dots, N$ satisfy Assumption 3.3, we have*

$$P(y_j^{(t-k)}|y_i^{(t)} = 1) - P(y_j^{(t-k)}|y_i^{(t)} = 0) \geq \Theta(\delta_{ijk}) > 0,$$

if j is a cause *for i with lag k. Otherwise,*

$$P(y_j^{(t-k)}|y_i^{(t)} = 1) - P(y_j^{(t-k)}|y_i^{(t)} = 0) = O(\frac{1}{N}).$$

as $N \to +\infty$, if $P(y_i^{(t)}) = \Theta(1), \forall i, t$.

Proof. Let A and B be two binary random variables with $P(A, B) = \Theta(1)$. And for simplicity, we denote $P(A = i, B = j)$ by p_{ij}. We have

$$\begin{aligned}&P(A = 1|B = 1) > P(A = 1|B = 0) + \delta\\ &\Leftrightarrow \frac{p_{11}}{p_{01} + p_{11}} > \frac{p_{10}}{p_{00} + p_{10}} + \delta\\ &\Leftrightarrow \frac{p_{00}}{p_{10}} > \frac{p_{01}}{p_{11}} + \Theta(\delta)\\ &\Leftrightarrow \frac{p_{00}}{p_{01}} > \frac{p_{10}}{p_{11}} + \Theta(\delta)\\ &\Leftrightarrow \frac{p_{11}}{p_{10} + p_{11}} > \frac{p_{01}}{p_{00} + p_{01}} + \Theta(\delta).\\ &\Leftrightarrow P(B = 1|A = 1) > P(B = 1|A = 0) + \Theta(\delta).\end{aligned}$$

Replace $y_j^{(t-k)}$ and $y_i^{(t)}$ by A, B, we proved the Lemma. □

Note that $y_j^{(t-k)}$ and $y_i^{(t)}$ are switched, compared with Assumption 3.3. Simply combining Lemma 3.4 with Theorem 3.3, we have similar consistency results for the backward time series:

Theorem 3.5. *Let a binary time series $\{y_i^{(t)}\}, i = 1, 2, \dots, N$ satisfy Assumptions 3.1 and 3.2. We define the backward time series by $\{z_i^{(t)}|z_i^{(t)} = y_i^{(-t)}\}$ and denote the coefficients* *by* $\mathbf{B}_k, k = 1, 2, \ldots, P$ *as in VAR model, which are estimated by applying Granger causality on time series* $\{z^{(t)}\}$*. We have*

$$\{\mathbf{B}_k\}_{ji} \geq \Theta(\delta_{ijk}) > 0,$$

if j *is a* cause *for* i *with lag* k *w.r.t.* $\{y_i^{(t)}\}$*. Otherwise,*

$$\{\mathbf{B}_k\}_{ji} = O(\frac{1}{N}),$$

as $N \to +\infty$*, if* $P(y_i^{(t)}) = \Theta(1), \forall i, t$*.*

Proof. We prove this theorem based on existing theorems and lemmas. Forward time series $\{y_i^{(t)}\}$, $i = 1, 2, \ldots, N$ satisfy Assumption 3.2, so does the backward time series since Assumption 3.2 is symmetric in time. Forward time series $\{y_i^{(t)}\}$, $i = 1, 2, \ldots, N$ satisfy Assumption 3.1 and 3.2, by Lemma 3.2, it also satisfies Assumption 3.3. Then by Lemma 3.4, the backward time series also satisfy Assumption 3.3. Then by Theorem 3.3, we prove Theorem 3.5. □

Theorem 3.5 indicates that applying Granger causality on the backward time series also provides consistent temporal dependence inference results. Note that we only assume 3.1 and 3.2 for the forward time series. In fact, the backward time series might not satisfy Assumption 3.1 at all.

3.4 Summary and Discussion

In Section 3.2 and 3.3, we investigate several well-known time series models and establish the connection between forward and backward time series in terms of temporal dependence structures. For VAR, the strength $\{\mathbf{A}_k\}_{ij}$ of triplet (i, j, k) in forward time series is approximately the same as that for $\{\mathbf{B}_k\}_{ji}$ of triplet (j, i, k) in backward time series. For binary time series models that satisfy Assumptions 3.1 and 3.2, the strength of triplets (i, j, k) and (j, i, k) for forward and backward time series, respectively, share the same order of magnitude. These connections justify our approach to combine the results from the forward and the backward time series for better temporal dependence inference.

Note that information inferred from the backward time series is not exactly the same as that inferred from the forward time series, but they indeed share considerable similarities, which can be utilized as suggested. Moreover, when applied to real world data, model misspecification should be considered, which is absent in our current analysis. In addition, one should be aware of the data preprocessing procedure, to make sure it is compatible with our assumptions, especially when the preprocessing relies on a specific order of time.

4. EXPERIMENT RESULTS

In the experiments, we evaluate the effectiveness of our proposed approach on several synthetic datasets and two real world datasets. Next, we describe the data collections, baseline methods, evaluation metric and experimental results.

4.1 Datasets

Synthetic Datasets Since we do not have the access to the true underlying temporal dependence structure in most applications, we generate synthetic datasets to evaluate the performance of temporal dependence structure recovery.

For discrete time series, we generate two synthetic datasets: one is generated from the IC model (as discussed in example 3.1), and the other is an instance of the generalized linear models (later referred to as LOG model), with a Bernoulli distribution and the link function $\sigma(\cdot)^{-1}$ as a logistic sigmoid function. Specifically, the distribution of $\mathbf{y}^{(t)}$ is a Bernoulli random vector with parameter

$$P(y_i^{(t)} = 1|\mathbf{y}^{(t-1:t-P)}) = E[y_i^{(t)}|\mathbf{y}^{(t-1:t-P)}], i = 1, \ldots, N,$$

defined as follows:

$$\sigma^{-1}(E[\mathbf{y}^{(t)}|\mathbf{y}^{(t-1:t-P)}]) = \mu + \sum_{k=1}^{P} \mathbf{A}_k \mathbf{y}^{(t-k)}.$$

If $\mathbf{A}_k$ are all nonnegative matrices, it satisfies Assumption 3.1.

For continuous time series, we also generate two synthetic datasets: one is linear time series generated according to VAR, and the other is nonlinear time series generated from generalized linear model with polynomial link function and Gaussian noises (POLY). Specifically, the distribution of $\mathbf{y}^{(t)}$ is a multivariate Gaussian, i.e.,

$$\mathbf{y}^{(t)}|\mathbf{y}^{t-1:t-P} \sim \mathcal{N}(f(\sum_{k=1}^{P} \mathbf{A}_k \mathbf{y}^{(t-k)}), \epsilon I),$$

where $f(\cdot)$ is defined by $f(x) = x + bx^3$. We vary b to change the level of nonlinearity.

For each model, we set the lag to 1 and generate a sparse temporal dependence structure $\mathbf{A}$ with 5% nonzero entries. For the IC, each nonzero entry in $\mathbf{A}$ is drawn from a uniform distribution Uniform$(0, 1)$. For LOG, μ_i is drawn from $\mathcal{N}(0, 1)$, and each nonzero entry in $\mathbf{A}$ is drawn from $\mathcal{N}(0, 1)$. For VAR and POLY, each nonzero entry in $\mathbf{A}$ is drawn from $\mathcal{N}(0, 1)$, and then we normalize $\mathbf{A}$ by its Frobenius norm to ensure the stability of the time series. Moreover, for each model, we generate time series $\{y_i^{(t)}\}$, $i = 1, \ldots, N$, $t = 1, \ldots, T$ by setting N and T with two scenarios: (1) $N = 30, T = 2000$, which corresponds to the low-dimensional case, and (2) $N = 100, T = 150$, which mimics the high-dimensional case.

Twitter Datasets We collect the *Haiti* dataset[5] with all the tweets published between Oct 2009 and Jan 2010 on "Haiti earthquake". We choose this topic because it was one of the hot topics during that time period and many tweets have been generated around the event.

For the *Haiti* dataset, we collect the tweets by searching the keyword "Haiti" from Jan. 12, 2010 for 17 days. We then generate multivariate time series datasets by counting the number of tweets from the top 1000 users (who tweet most on the topics) over these 1000 intervals. For accurate modeling, we remove the users that are highly correlated with each other, most of which are operated by the same persons and tweet exactly the same contents. We also remove robot-like user-accounts who tweet on very regular intervals. Finally, we select a set of users with at least one interaction with another user, which results in a subset of 274 users.

Microarray Dataset Most multicellular organisms rely on their immune systems to defend against the infection from a multitude of pathogens. We collect the time series microarray data on macrophages from human immune cells from the supporting website of [9, 11]. It consists of 1651 genes with 9 time series observations. We apply the proposed model to this dataset in order to infer the temporal dependence networks for immune system genes. Due to the

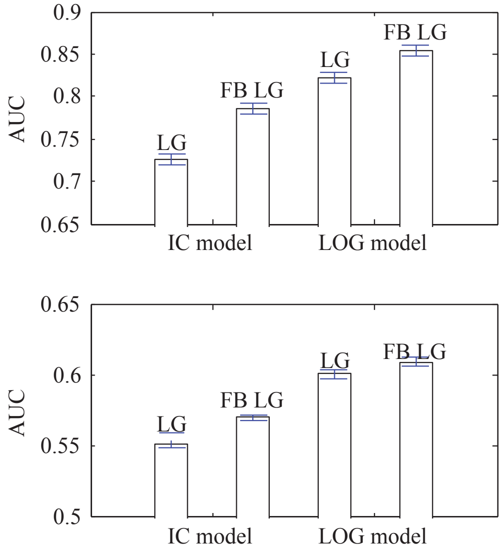

Figure 1: The performance of temporal dependence recovery on IC and LOG datasets. Top: $N = 30, T = 2000$**; Down:** $N = 100, T = 150$**. Results suggest that LG benefits from the forward backward approach.**

space limit, we only show the results of a subset of 6 genes, whose interactions have been well studied.

4.2 Baseline Algorithms

We use the following baselines for comparison analysis on the synthetic datasets:

- *Lasso Granger Causality* (LG) and *Naive Forward Backward Lasso Granger Causality*[1] (FB LG) as described in section 3.1.

- *Copula Lasso Granger Causality* (CLG) and *Naive Forward Backward Copula Lasso Granger Causality*[2] (FB CLG). For nonlinear time series, we apply the copula transformation for each time series first, and then apply the *Granger causality* based algorithms.

For *Lasso* based algorithm, we choose the penalty parameter λ to minimize the prediction error on the validation dataset.

For the Haiti dataset, we also test the *Transfer Entropy* (TE) algorithm [26]. Transfer entropy is another related technique which identifies the temporal dependence structure between two time series by measuring the decrease in uncertainty of one time series in the future, given the past information of the other time series. Namely, the transfer entropy is defined as

$$T_{\mathbf{y}_j \to \mathbf{y}_i} = \mathcal{H}(y_i^{(t)}|\mathbf{y}_i^{t-1:t-P}) - \mathcal{H}(y_i^{(t)}|\mathbf{y}_i^{t-1:t-P}, \mathbf{y}_j^{t-1:t-P}),$$

where $\mathcal{H}(x)$ is the Shannon entropy of the random variable x. We also test the *Naive Forward Backward Transfer Entropy* (FB TE), where we measure the influence from j to i by $T_{\mathbf{y}_j \to \mathbf{y}_i} + T_{\mathbf{z}_i \to \mathbf{z}_j}$.

4.3 Evaluation Measures

To evaluate the performance of different methods in recovering temporal dependence structures, we choose the Area Under the Curve (AUC) measure as it is a good performance measure for the ground truth with unbalanced ratio of positive and negative labels. The value of AUC is the probability that the algorithm will assign a higher value to a randomly chosen positive (existing) edge than a randomly chosen negative (non-existing) edge in the graph [10].

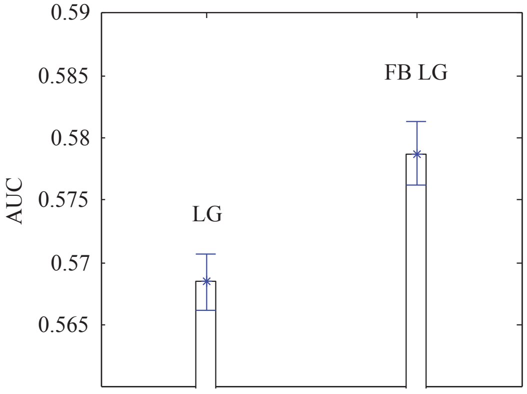

Figure 2: The performance of temporal dependence recovery on VAR dataset with $N = 100, T = 150$**. Results suggest that LG benefits from the forward backward approach.**

For synthetic datasets, we calculate AUC against the ground truth, i.e., the temporal dependence structure defined by $\mathbf{A}$. The reported results are averaged over 20 randomly generated datasets.

For the Twitter dataset, since we do not have access to the true underlying influence graph in the social network, we use the retweet information as indirect evaluation. It has been argued that the retweet graph in the future time can reflect the influence in social networks to a certain extent [8]. We first represent the retweet information by a weighted graph $\mathcal{G}_{RT}$, where the weight of an edge $(s \to t)$ denotes the number of tweets from user s retweeted by user t. The retweet graph $\mathcal{G}_{RT}$ on Haiti earthquake has 867 edges.

For the microarray dataset, we do not have the complete ground truth as well. We therefore compare our results with those reported interactions in the BioGRID database[3], a curated biological database of protein-protein and genetic interactions.

4.4 Experiment Results

In this section, we present the result of our experiments. We focus on the difference of performance between the original version of algorithm and the forward backward version.

Results on Synthetic Datasets We aim to test whether the forward and backward approach can improve the performance of LG on IC and LOG datasets. We report the AUC scores of LG and FB LG in Figure 1. We can see that FB LG consistently outperforms LG, which suggests that FB LG indeed benefits from combining estimates from forward and backward time series.

On the VAR dataset, we focus on the performance for linear continuous time series. We report the AUC scores of LG and FB LG in Figure 2. For high-dimensional time series, FB LG outperforms LG significantly, which indicates that combining two directions helps to improve the performance.

[3] www.thebiogrid.org

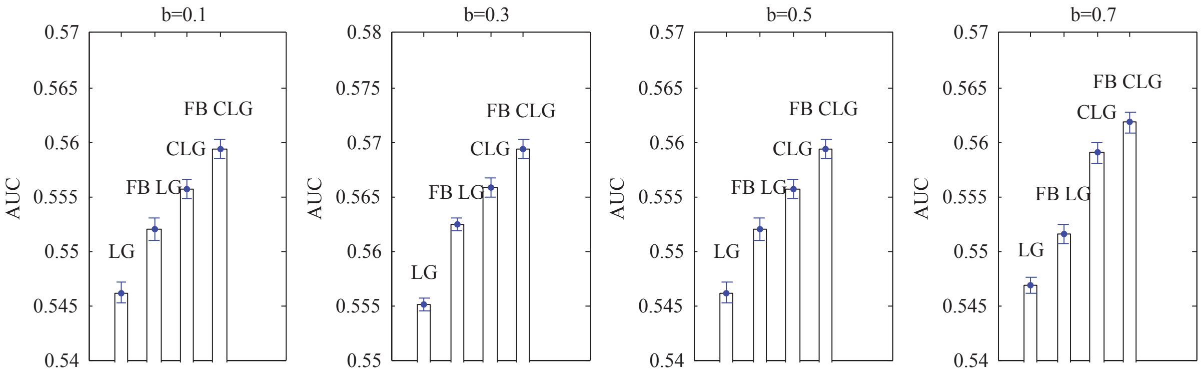

Figure 3: The performance of temporal dependence recovery on POLY dataset with different b and $N = 100, T = 150$. Results suggest that both LG and CLG benefit from the forward backward approach.

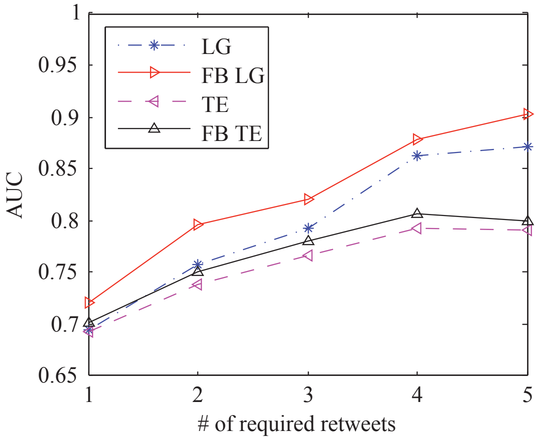

Figure 4: The performance of temporal dependence recovery on Haiti dataset. The performance of FB LG and FB TE outperform the LG and TE, respectively.

On the POLY dataset, we aim to test whether the forward and backward approach can further improve the performance of LG and CLG for nonlinear time series. We report the AUC scores of LG, FB LG, CLG, FB CLG on the POLY datset in Figure 3. As we can see CLG can improve LG thanks to the copula transformation. And FB CLG can further improve CLG for high-dimensional case. For both VAR and POLY datasets, when $N = 30, T = 2000$, we find that the performance of LG and FB LG are similar, as suggested in Section 3.2.

Results on Twitter Dataset We test LG, FB LG, TE, and FB TE on the Haiti dataset (see Figure 4 for results). We set the lag $P = 15$ for all algorithms for fair comparison, which corresponds to 6 hours approximately. The AUC is calculated against the retweet graph $\mathcal{G}_{RT}$, and we vary the required number of retweets, so that only if the retweets from j by i passes the required number n, we establish an edge from i to j. Intuitively, n screens the weak influence between users. As we can see, all algorithms perform better as we increase n. In addition, the forward backward approach improves the performance for both baseline algorithms.

Table 1: Top 20 predictions for gene interaction by LG and FB LG. Bold terms are ground truth suggested by BioGRID database.

Lasso Granger	F/B Lasso Granger
PCNA→CCNA2	CDC2→E2F1
E2F1→CCNA2	RFC4→E2F1
CDKN3→CDC2	CDKN3→CDC2
CDC2→RFC4	**PCNA→CCNA2**
CCNA2→RFC4	RFC4→CDKN3
CCNA2→CDC2	CCNA2→RFC4
RFC4→CDKN3	CCNA2→CDC2
CDC2→E2F1	**RFC4→PCNA**
CCNA2→CDKN3	PCNA→E2F1
E2F1→CDC2	**E2F1→CCNA2**
PCNA→RFC4	PCNA→RFC4
RFC4→PCNA	CDC2→RFC4
CDKN3→CCNA2	RFC4→CDC2
CDKN3→RFC4	CCNA2→CDKN3
RFC4→CCNA2	CDKN3→RFC4
RFC4→CDC2	**CCNA2→E2F1**
CDC2→CDKN3	CDC2→PCNA
E2F1→PCNA	CDKN3→E2F1
CCNA2→PCNA	E2F1→CDC2
RFC4→E2F1	CDKN3→PCNA

Results on Microarray Dataset We test LG and FB LG on the time series microarray dataset, and achieve AUCs of **0.6923** and **0.7308**, respectively. Moreover, we list the top edges identified by both algorithms in Table 1. The bold ones are the ground truth suggested by BioGRID database. LG correctly identified 3 interactions while FB LG identified all 4 known interactions.

5. CONCLUSION AND FUTURE WORK

Inspired by time-reversibility of physical laws and its effect on temporal dependency structure, we proposed the forward backward approach to improve the performance of Granger causality. We developed the forward and backward Lasso Granger causality algorithm, which combines the coefficients estimated from the forward time series and backward time series to provide better performance of temporal dependency structure recovery. We show that with the copula transformation, we can extend our algorithm for non-linear time series. Theoretical analysis on several existing times series models including VAR and IC model confirms our intuition. Our empirical results on both synthetic and real world datasets demonstrate that the forward backward approach can improve the performance of temporal dependence inference using forward time series only.

For future work, we will investigate other combination strategies for forward backward approach, other than simply averaging the coefficients. the performance. We will also examine more general types of time series models where the forward backward approach is applicable.

6. ACKNOWLEDGMENT

We thank Aram Galstyan, Greg Ver Steeg for discussions, and David C. Kale for helpful suggestions. This research was sponsored by the NSF research grants IIS-1254206, Okawa Foundation Research Award, and U.S. Defense Advanced Research Projects Agency (DARPA) under Social Media in Strategic Communication (SMISC) program, Agreement Number W911NF-12-1-0034. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agency or the U.S. Government.

References

[1] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *KDD*, 2008.

[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. *Econ. Letters*, 2000.

[4] M. T. Bahadori and Y. Liu. An examination of practical granger causality inference. In *SDM*, 2013.

[5] M. T. Bahadori, Y. Liu, and E. P. Xing. Fast structure learning in generalized stochastic processes with latent factors. In *KDD*, 2013.

[6] L. Breiman. Bagging predictors. *Mach. Learning*, 1996.

[7] 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.

[8] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy. *ICWSM*, 2010.

[9] D. Chaussabel, R. T. Semnani, M. A. McDowell, D. Sacks, A. Sher, and T. B. Nutman. Unique gene expression profiles of human macrophages and dendritic cells to phylogenetically distinct parasites. *Blood*, 2003.

[10] C. Cortes and M. Mohri. Confidence intervals for the area under the roc curve. In *NIPS*, 2005.

[11] C. S. Detweiler, D. B. Cunanan, and S. Falkow. Host microarray analysis reveals a role for the salmonella response regulator phop in human macrophage cell death. *PNAS*, 2001.

[12] E. S. Edgington and P. Onghena. *Randomization Tests*. Chapman & Hall/CRC, 2007.

[13] M. Eichler. Graphical modelling of multivariate time series. *Probab. Theory Related Fields*, 2012.

[14] C. W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 1969.

[15] C. W. Granger. Testing for causality: A personal viewpoint. *J. Econ. Dynam. Control*, 1980.

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

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

[18] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *KDD*. ACM, 2003.

[19] Y.-H. Kim, H. H. Permuter, and T. Weissman. Directed information, causal estimation, and communication in continuous time. *IEEE Trans Inf Theory*, 2009.

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

[21] C. D. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT press, 1999.

[22] D. Marinazzo, M. Pellicoro, and S. Stramaglia. Kernel Granger causality and the analysis of dynamical networks. *Phys. Rev. Lett. E*, 2008.

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

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

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

[26] T. Schreiber. Measuring Information Transfer. *Phys. Rev. Lett.*, 2000.

[27] S. Song and P. J. Bickel. Large Vector Auto Regressions. *arxiv:1106.3915*, 2011.

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

[29] 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.