Temporal Lag estimation and Granger Causality on time series

Indranil Bhattacharya
Indian Institute of Science
Bangalore, Karnataka, India
indranil.bhattacharya@csa.iisc.ernet.in

Arnab Bhattacharyya
Indian Institute of Science
Bangalore, Karnataka, India
arnabb@csa.iisc.ernet.in

Praveen Pankajakshan
Shell India Markets Pvt. Ltd.
Bangalore, Karnataka, India
praveen.pankajakshan@shell.com

ABSTRACT

Inferring causal relationships among features generated as time series is an important problem with many applications. For high dimensional data, causal inference becomes a challenging task due to the *curse of dimensionality*. Graphical modeling of temporal data based on the concept of "Granger Causality" has gained much attention in this context. The blend of Granger methods along with model selection techniques such as Lasso, enables efficient discovery of a "sparse" subset of causal variables in high dimensional settings. These temporal causal inference methods use an input parameter, L, the *maximum time lag*, also called the *max lag*. This parameter is the maximum gap in time between the occurrence of the output phenomenon and the causal input stimulus. However, in many situations of interest, the maximum time lag is unknown, and indeed, finding the range of causal effects is an important problem. In this work, we propose and evaluate a data-driven and computationally efficient method for Granger causality inference in time series without foreknowledge of the maximum time lag parameter. We present two algorithms here viz. *Lasso Granger++* and *Group Lasso Granger++* which alongside inferring the hypothesis feature causal graph, also estimates a value of max-lag by balancing the trade-off between "goodness of fit" and "model complexity".

KEYWORDS

Causal Inference, Granger Causality, Time Series, Causal graphs, Temporal Lag prediction

ACM Reference format:
Indranil Bhattacharya, Arnab Bhattacharyya, and Praveen Pankajakshan. 2017. Temporal Lag estimation and Granger Causality on time series. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, Nova Scotia - Canada, August 2017 (SIGKDD'17),* 9 pages.
DOI: 10.475/123_4

1 INTRODUCTION

Discovering *causal relationships* among natural or artificial processes is a fundamental aspect of science. In current times, massive amounts of time series data have become available for analysis and mining, yet uncovering the underlying causal structure from the data itself is quite challenging since most of the data points reside in a very high dimensional space. How to develop efficient and scalable learning algorithms to uncover the temporal dependency structures between time series and reveal insights from data has become one of the key problems in machine learning and data mining community today. It is, however, well understood that mere *statistical correlation* does not imply *causation* [16, 19], but under some specific conditions, it might be possible to derive causality from correlations in the observed data [3]. Our present work is focused entirely on causal inference of *time series data*.

One of the earliest works in quantifying the causal relationship amongst temporal variables was introduced in the field of econometrics by the Nobel laureate Clive Granger (1969). The notion he proposed is now popularly known as the *Granger Causality* [8, 9] and is widely used in practice. Empirical experience shows that the notion often captures causality in an effective way. However, Granger causality does not say much about situations where there is a hidden confounding variable causally influencing two or more observed variables, or when there is an indirect chain of causal structures. We will not be addressing such issues in our current work since we assume a "causally sufficient" system of variables.

Granger causality was initially introduced for a pair of variables. Later, techniques such as randomization tests [18] and statistical tests [22] were proposed to recover the temporal structures among multiple variables but they are not very computationally efficient in the higher dimensional setting. Recently, there has been a growing interest in combining the notion of Granger Causality with model selection techniques such as Lasso [23] and Group-Lasso [2, 12, 13] for extracting the temporal causal structures from high dimensional data. One of the major advantages of using Lasso is its statistical consistency. It has been proven [17] that the probability of Lasso falsely including any of the non-neighboring variables of a given node into its neighborhood estimate vanishes exponentially fast, even if the number of non-neighboring variables may grow very rapidly with the number of observations.

Most of the methods for mining Granger Causality on time series use the Vector Autoregressive (VAR) model [14] with a *fixed* value of *maximum time lag* or *maxLag* (also called the *model order*), usually denoted by L. This parameter *maxLag* is the maximum time in past to look back to when regressing for present and subsequently future values of the target time series variable as a function of all other causal variables in the system. Incorporating a "good" estimate of L requires sufficient domain knowledge of the underlying physical system being modeled since it captures the maximum gap in time between the occurrence of the output phenomenon and the causal input stimulus. Both over and under estimate of L impacts the Mean Squared Error (MSE) of prediction. Moreover, when L is sufficiently large there is a huge blow-up in the effective number of features. Several *Statistical testing* schemes, such as the log-likelihood ratio test, F-Statistic test, exist which perform a sequence of hypothesis

SIGKDD'17, Halifax, Nova Scotia - Canada

DOI: 10.475/123_4

tests to determine the "best" estimate of model order in a VAR(L) process. However, all these testing schemes are very effective in deducing a "good" estimate of L when the number of variables is small, they do not scale well when the lag is significantly "high" and the number of time series variables is also large.

Our current work precisely focuses on this problem of "Granger Causal Inference" on *multivariate*, *high-dimensional* time series data modeled as a vector autoregressive process but with *unknown model order*. We propose a semi-automated way of estimating the *maxLag* parameter which "best" fits the Granger Causal model to the given time series data. Our algorithm not only estimates the "best" (in terms of *"goodness of fit"*) value of L but also outputs the coefficients of the causal variables in the VAR process using standard model selection mechanism. The latter is used to reconstruct the hypothesis feature causal graph. Our approach is purely data-driven and so prior domain knowledge of the underlying physical system is not mandatory. Importantly, it is scalable over high dimensions, both in terms of space and time complexity, whereas a brute force search for the same would simply not be feasible even for moderately large datasets. We attempt to characterize the performance of our method by conducting a host of experiments on both synthetic and real-world time series data.

The rest of the paper is organized as follows: In Section 2, we describe the problem formulation and state the key notions of Granger Causality. We describe some of the existing and most significant methods for modeling Granger causality on time series data in Section 3. Section 4 highlights our key idea of maxLag discovery from data. Finally, we present the results of our experimental evaluation in Section 5.

2 PRELIMINARIES

In this section we formally introduce the problem and describe the key concepts and notions used for inferring *Granger Causality* on time series data.

2.1 Problem Formulation

Given a set of features $V = \{x_1, x_2, ..., x_P\}$, where each x_i is a time series variable, we construct a directed graph over the set of features [1], called the *feature causal graph*, $G = (V, E)$. A *feature vector* at some particular point in time t is the P-tuple $(x_1^t, x_2^t, \ldots, x_P^t)$ of the features. Nodes in graph G correspond to the features and edges capture causality. An edge $e \in E$, directed from x_i to x_j is labeled with a natural number ℓ, called the *time lag* (*or temporal lag*), which essentially captures a *causal* relationship of the form $x_i^{t-\ell} \rightarrow x_j^t$ $\forall t > \ell > 0$. This basically means that the current value of x_j is affected by the value of x_i ℓ-steps back, and thus x_i becomes a causal variable for x_j. Each causal variable can have different time lags at which they influence the observations of another variable. For example, if say x_1 is causally affected by k variables $x_2, \ldots, x_{k+1}$ with corresponding time lags $\ell_2, \ldots, \ell_{k+1}$, the ℓ_i's in principle could be any arbitrary Natural number.

We associate a stochastic process that generates time series data with respect to this graph. We start with a *predefined window size* L such that L [2] is at least the time lag corresponding to each edge in the graph G. This ensures that L is greater than or equal to the maximum time lag in the network. Now given this graphical model over the temporal variables, the stochastic process starts by generating an initial sequence of L feature vectors for time points $t = 0, 1, \ldots, L-1$. At each step henceforth, it generates the next feature vector according to the conditional probability distribution $P\left(\{x_i^L\} | \{x_j^t\}_{j=1,2,\ldots,P, t=0,1,\ldots,L-1}\right)$ on the graphical model where the variables x_j^t for $t = 0, 1, \ldots L-1$ are as initialized in the last L steps. This is the *"Unit Causal Graph"* generation method as stated in [5]. The conditional probabilistic model could in principle be any arbitrary statistical model [5, 10], but, in the current work, we assume that the both the conditional probabilities and the initial distribution of time series data are linear combinations of Gaussians [20]. Under these assumptions, it is easy to see that the stochastic model associated with the causal feature graph is equivalent to the VAR model [14].

The goal of a causal modeling algorithm is to infer the underlying causal structure, given as input the time series data generated by its associated stochastic process. The performance of the algorithm can be measured purely in terms of similarity between the hypothesis (output) causal graph and the original graph that gave rise to the time series data. The details about the evaluation criteria is described in the next subsection.

2.2 Evaluation Criteria

We use the metrics of Precision, Recall and F_1 measure to the problem of predicting a 0 or 1 entry in the adjacency matrix representation of the graph. Note that for any pair of features x_i and x_j, there are two entries in the adjacency matrix A, $A(i,j)$ and $A(j,i)$. The entry $A(i,j) = 1$ implies that there is a directed edge from x_j to x_i in the feature causal graph i.e. x_j causally influences x_i. Therefore, the entries marked with 1 in the i^{th} row of A (i.e. $A(i,:)$) denotes the set of variables which causally influence x_i. A bi-directional edge, $A(i,j) = A(j,i) = 1$, corresponds to causality in both directions i.e $x_i \rightarrow x_j$ and $x_j \rightarrow x_i$. Given this formulation, precision and recall are well defined. For example, predicting a bi-directional edge between x_i and x_j when there is actually a directed edge only from $x_i \rightarrow x_j$, would entail one correct prediction and one error. So precision (P) would be 0.5, recall (R) is 1, and therefore F_1 score 0.67.

Let A denote the adjacency matrix of the source (original) feature causal graph, and $\hat{A}$ denote the same for the hypothesis (output) graph, then the expressions for *precision* (P), *recall* (R) and the F_1 score which tries to balance the overall quality of prediction are given below.

$$P = \frac{|\{(i,j) \in V \times V : \hat{A}(i,j) = A(i,j)\}|}{|\{(i,j) \in V \times V : \hat{A}(i,j) = 1\}|} \tag{1}$$

$$R = \frac{|\{(i,j) \in V \times V : \hat{A}(i,j) = A(i,j)\}|}{|\{(i,j) \in V \times V : A(i,j) = 1\}|} \tag{2}$$

$$F_1 = \frac{2PR}{P+R} \tag{3}$$

[1] We call each time series variable a feature. This should be obvious once we formulate it as a feature selection problem especially using Group Lasso.

[2] We are abusing the notation L here, since we want to establish the natural connection between the model order in a VAR process and the window size.

2.3 Granger Causality

Introduced first in the field of econometrics by the Nobel laureate Clive Granger (1969), *Granger Causality* [8, 9] is one of the most popular approaches to quantify causal relationships among time series data. It is based on two major principles : (i) The cause happens prior to the effect, (instantaneous causation is ignored) and, (ii) The cause makes unique changes in the effect. A time series X is said to "Granger Cause" another time series Y, denoted by $X \rightarrow Y$ if and only if regressing with past values of both X and Y is *statistically more significant* that doing so with past values of Y alone. The original definition of Granger Causality is very general and does not assume anything about the underlying data generative model. For modeling the distributions of multivariate data, VAR models are widely used since they are simple, robust and have strong empirical performance on most of the practical applications.

2.4 Testing Granger Causality

Linear Granger Causality test was initially introduced for a pair of variables only. Later it was extended for multivariate data as well.

(a) Linear Granger Causality Test for Bivariate data : Let $X = \{x^t\}_{t=1}^T$ and $Y = \{y^t\}_{t=1}^T$ be two time series of length T. Let $\mathbf{x}_t = [x^{(t-1)}, x^{(t-2)}, \dots, x^{(t-L)}]$ and $\mathbf{y}_t = [y^{(t-1)}, y^{(t-2)}, \dots, y^{(t-L)}]$ denote the history (i.e. all L time-lagged values) of X and Y resp. up to time t, where L is the maximum time lag (the model order) of the VAR process. The *Linear Granger Causality test* to ascertain whether X *"Granger Causes"* Y is conducted as follows :

(i) First two different VAR models are fit to the data as follows :

$$y^t \approx \langle \alpha, \mathbf{y}_t \rangle + \langle \beta, \mathbf{x}_t \rangle \tag{4}$$

$$y^t \approx \langle \gamma, \mathbf{y}_t \rangle \tag{5}$$

where $\alpha = [\alpha_1, \dots, \alpha_L]$ and $\gamma = [\gamma_1, \dots, \gamma_L]$ are two different coefficient vectors of $\mathbf{y}_t$ and $\beta = [\beta_1, \dots, \beta_L]$ that of $\mathbf{x}_t$.

(ii) Then, any standard joint statistical significance test viz. F-Statistic test, χ^2-test is conducted to obtain a p-value along with the residual error which helps to determine whether model (4) is a "better fit" than model (5) with significant statistical advantage. If the first model outperforms the second, then it is concluded that X *"Granger causes"* Y. Similarly, it can be ascertained whether Y *"Granger causes"* X as well.

(b) Linear Granger Causality Test for Multivariate data : Given multivariate time series $X_i = \{x_i^t\}_{t=1}^T$, $\forall i = \{1, 2, \dots, P\}$, where P is the number of time series variables and T is the length of each time series. Consider X_i to be the target variable, so our objective is to find which of the time series variables $X_1, \dots, X_P$ *"Granger causes"* X_i. A VAR model of order L is fit to X_i, $\forall t = L+1$ to T as follows :

$$x_i^t = \sum_{j=1}^{P} \langle \beta_j^i, x_j^{(t,L)} \rangle + \epsilon_i^t \tag{6}$$

$$= \sum_{j=1}^{P} \sum_{\ell=1}^{L} \beta_j^i(\ell) x_j^{(t-\ell)} + \epsilon_i^t \tag{7}$$

where $x_j^{(t,L)} = [x_j^{(t-1)}, \dots, x_j^{(t-L)}]$ is the history of X_j up to time t, $\beta_j^i = [\beta_j^i(1), \dots, \beta_j^i(L)]$ is the coefficient vector modeling the effect of time series X_j on X_i, L is the maximum time lag, and ϵ_i^t is independent additive white noise. Again, by a statistical significance test [15], if *at least one* value in β_j^i is *non-zero*, we can claim that time series X_j *"Granger causes"* X_i. Thus if we conduct this test for every feature $X_i : i \in [P]$, we can find the corresponding causal variables and construct the hypothesis feature causal graph.

3 EXISTING APPROACHES

We now discuss some of the existing and popular approaches for modeling Granger Causality on multivariate time series. All of these methods use the VAR model with a *fixed* value of the model order L known a priori based on domain knowledge or intuition of the underlying data generating process.

3.1 Exhaustive Graphical Granger Method

The most trivial way of applying Granger Causality on multivariate time series data is to simply conduct the *Linear Granger causality test* (2.4(a)) for every pair of features in order to determine the presence/absence and the orientation of the corresponding edge(s) in the output feature causal graph (see [2] for details).

3.2 Vector Autoregressive Granger Method

This method parallels the underlying stochastic model of the data generation process. Here, we fix a value for maxLag L and fit a VAR model of order L to each feature as described in section 2.4(b). It is easy to see (equation (6)) that if we take all the entries of a variable, say x_i from $t = L+1$ to T and write them in a vector form of length $T - L$, and similarly the r.h.s of the same equation as a matrix X^{Lagged} of past L values of all the features, then equation (7) can also be written as :

$$Y = X^{Lagged}\beta + E \tag{8}$$

where the t^{th} entry of Y is x_i^t and the corresponding row of X^{Lagged} is a row vector $[[x_1^{t-1}, \dots, x_1^{t-L}], \dots, [x_P^{t-1}, \dots, x_P^{t-L}]]^T$ of size PL, and E is the additive Gaussian white noise vector.

This is solved by classical OLS Regression which gives a closed form representation of the coefficient vector β. The *non-zero* entries of β (if any) determines the subset of features causally affecting the target feature (x_i in this case). The above process is repeated for all the P features in the model and subsequently construct the output feature causal graph.

3.3 Lasso Granger Method

The Lasso Granger method applies Lasso-type formulation [23] to the VAR model for each feature x_i, $i = \{1, 2, \dots, P\}$ and obtains a sparse estimate of the coefficient vector. The optimization problem of Lasso Granger is as follows :

$$\min_{\beta} \sum_{t=L+1}^{T} \left(x_i^t - \sum_{j=1}^{P} \langle \beta_j^i, x_j^{(t,L)} \rangle \right)^2 + \lambda \|\beta\|_1 \tag{9}$$

where T is length of the time series, λ is the regularization parameter enforced to obtain a sparse β and the rest of the variables are same

as defined in section 2.4(b). Equation (9) can also be written in a vector form as follows :

$$\min_{\beta} \left\| Y - X^{Lagged}\beta \right\|^2 + \lambda \left\|\beta\right\|_1 \tag{10}$$

where Y is a vector of length $T-L$, the t^{th} entry of Y being x_i^t and the corresponding row of the matrix X^{Lagged} is a row vector $[[x_1^{t-1},\ldots,x_1^{t-L}],\ldots,[x_P^{t-1},\ldots,x_P^{t-L}]]^T$ of size PL. Let us emphasize the fact once again that the maxLag parameter L is fixed a priori. Like VAR method (3.2), the above process is repeated for all the P features in the model and the *non-zero* entries of the coefficient vector β are used to determine the subset of features causally affecting the target feature and hence construct the output feature causal graph.

3.4 Group Lasso Granger Method

The Lasso Granger method (3.3), although computationally efficient, has neglected one important aspect of the problem - the natural *group structure* existing among the temporal lagged variables imposed by the respective time series they belong to. The Group Lasso Granger method [12] overcomes these limitations by applying a regression method suited for high-dimensional data, and also leveraging the group structure among the temporal lagged variables according to the time series they belong to. It employs the *Group Lasso* selection [25] technique which alongside minimizing the empirical risk also performs variable selection by penalizing the intra-group and inter-group coefficients differently.

Consider a partitioning of the entire set of predictors $\{x_1,\ldots,x_P\}$ into P groups, then the group lasso formulation due to Lozano et. al. [12] is as follows :

$$\min_{\beta} \left\| Y - X^{Lagged}\beta \right\|^2 + \lambda \sum_{j=1}^{P} \sqrt{\rho_j} \left\|\beta_{\mathbb{G}_j}\right\|_2 \tag{11}$$

where Y is a vector of length $T-L$, the t^{th} entry of Y being x_i^t and the corresponding row of the matrix X^{Lagged} is the row vector $[[x_1^{t-1},\ldots,x_1^{t-L}],\ldots,[x_P^{t-1},\ldots,x_P^{t-L}]]^T$ of size PL and $\beta_{\mathbb{G}_j} = \{\beta_k : k \in \mathbb{G}_j\}$ where $\mathbb{G}_j$ denotes the set of group indices and ρ_j accounts for varying group size. When the groups are of equal size ρ_j can be set to either 1 or size of a group, $\forall j = \{1,\ldots,P\}$. Notice that using L_2-norm to penalize within a group ensures similar shrinkage of coefficient values for all intra-group predictors and the L_1-penalty on top of this group structure performs group selection. Here also the maxLag parameter L is known and fixed beforehand. The above method is repeated for each feature in the model and an edge $x_j \rightarrow x_i$ is placed in the output feature causal graph if the coefficients corresponding to x_j in $\beta_{\mathbb{G}_j}$ are non-zero implying that feature x_j is selected as a group by Group Lasso.

4 CONCURRENT ESTIMATION APPROACH

As we have already pointed out before, all the existing methods for mining Granger Causality on time series data fit a VAR model, with the parameter maxLag L fixed a priori. We propose a technique which concurrently estimates the "Granger causality coefficients" in the VAR model for each time series variable and also finds the "best" value of maxLag which maximizes the "goodness-of-fit" subject to the constraint that the model is not over-parameterized - hence the name "Concurrent Estimation Approach". The maxLag (L) values for each time series could in principle be different and arbitrarily large. For example, consider the following VAR equations :

$$x(t) = a_1 * x(t-1) + a_2 * y(t-2) + \epsilon_1(t) \tag{12}$$

$$y(t) = b_1 * y(t-10) + \epsilon_2(t) \tag{13}$$

x is being causally affected by itself at lag = 1 and by y at lag = 2 time steps back. Similarly every current value of y is affected by its own value $l = 10$ time steps back in the past. Our method reports the maximum lag for each feature ($\hat{L} = 2$ and $\hat{L} = 10$ for x and y resp.) as the "best" estimate of model order in its own multivariate VAR model. It is based on the same framework as Lasso Granger and Group Lasso Granger methods and therefore enjoys the statistical consistency and similar recovery guarantees under the high-dimensional setting. We present two algorithms next - *Lasso Granger++* (4.1) and *Group Lasso Granger++* (4.2). The general approach for both the algorithms is same except for the group effect considered in the second one.

4.1 Lasso Granger++

Let $\mathbb{S}$ denote a time series dataset on P variables $\{x_1,\ldots,x_P\}$ where each x_i, $\forall i \in \{1,\ldots,P\}$ is a time series of length T. We assume an upper bound, M on the maxLag value for each feature. Note M could be as large as $T-2$. We start with an initial estimate of L, denoted by L_0 and initialize it to some positive integer ℓ. We recommend initializing $L_0 = \ell = 1$ (to avoid a two-pass run), although it is not mandatory to do so[3]. We build a multivariate VAR model of order ℓ for the target variable (let it be x_i) and use Lasso Granger method (3.3) to solve it and obtain a *sparse* coefficient vector β^0. Therefore, the first optimization problem we solve is as follows :

$$\min_{\beta} \quad \frac{1}{2n} \sum_{t=\ell+1}^{T} \left(x_i^t - \sum_{j=1}^{P} \langle \beta_j^i, x_j^{(t,\ell)} \rangle \right)^2 + \lambda \left\|\beta\right\|_1 \tag{14}$$

which is equivalent to :

$$\beta^0 = \arg\min_{\beta} \quad \frac{1}{2n} \left\| Y - X^{Lagged}\beta \right\|^2 + \lambda \left\|\beta\right\|_1 \tag{15}$$

where $n = T - \ell$ denotes the number of samples (observations), Y is the target vector of length n and X^{Lagged} is an $n \times P\ell$ matrix. The t^{th} entry of Y is x_i^t and the corresponding row of X^{Lagged} is the row vector $[[x_1^{t-1},\ldots,x_1^{t-\ell}],\ldots,[x_P^{t-1},\ldots,x_P^{t-\ell}]]^T$ of size $P\ell$. λ is the usual regularization parameter and the rest of the notations are same as described before. Each non-zero entry of the coefficient vector β^0 correspond to some time lagged version of a feature from $\{x_1,\ldots,x_P\}$, with lags from $\{1,\ldots,\ell\}$. These non zero entries form the support of β^0, i.e. $supp(\beta^0) = \{i : \beta_i^0 \neq 0\}$.

In the next iteration, we increment our maxLag estimate by ℓ and thus our new maxLag is $L_1 = 2\ell$. But now while regressing for the target variable (e.g x_i in this case) from $t = 2\ell + 1$ to T, we **do not** take into account the entire past 2ℓ values of all the features. Instead, we do the following - (a) We consider all the feature values from the "relatively older unexplored past" i.e. from time $(t-\ell-1)$ to $(t-2\ell)$, and, (b) From the "recent past" i.e. time $(t-1)$ to $(t-\ell)$ we consider only the time lagged values of features present in $supp(\beta^0)$ obtained in the previous iteration (with maxLag ℓ). Since

[3] We will come back to this issue of initializing L_0 to an integer greater than 1.

we have already considered the recent history of all features (when maxLag was ℓ) and we have the significant values in the support, hence the remaining $P\ell -$ | $supp(\beta^0)$ | variables from the recent window will have no influence in determining the present value of x_i. But the older past is still unexplored and so we consider time lagged values of all the features from that time window. This is the intuition behind our adopting the strategy of selective feature pruning. Therefore, our new optimization problem becomes:

$$\beta^1 = \arg\min_{\beta} \quad \frac{1}{2n} \left\| Y - X^{Lagged}\beta \right\|^2 + \lambda \left\|\beta\right\|_1 \tag{16}$$

where, the number of samples is $n = T - 2\ell$, $Y = [x_i^{2\ell+1}, \ldots, x_i^T]$ is the target vector of length n and X^{Lagged} is a matrix of dimensions $n \times (P\ell + k_0)$, where $k_0 =$ | $supp(\beta^0)$ |. The t^{th} row of X^{Lagged} is the row vector constructed as follows :
$\left[[\tilde{x}_i : i \in supp(\beta^0)], [x_1^{t-l-1}, \ldots, x_1^{t-2\ell}], \ldots, [x_P^{t-l-1}, \ldots, x_P^{t-2\ell}]\right]^T$ of size $P\ell + k_0$. The optimization routine (16) returns β^1 whose support ($supp(\beta^1) = \{i : \beta_i^1 \neq 0\}$) corresponds to time lagged version of features from $\{x_1, \ldots, x_P\}$, and lags from $\{1, \ldots, 2\ell\}$. This procedure is iterated for all subsequent guesses of L until the first k such that $L_{k-1} = k\ell$ exceeds the upper bound M.

Finally, we use *Akaike Information Criterion* (AIC) [1] [or AIC with bias correction (AICc) [11] when the number of samples is small] appropriately to select the "best" estimate of L. Let $\hat{L}$ denote our estimate of maximum time lag parameter for feature x_i. We choose $\hat{L}$ to be the *smallest* value for which the AIC (or AICc) is within multiplicative ϵ-bound of the *minimum* (ϵ is small, usually 0.01). This is also asymptotically equivalent to choosing the *smallest* value from our lag estimates $\{L_0, L_1, \ldots, L_{k-2}\}$ for which the MSE converges to within some ϵ-bound of the *minimum* MSE (ϵ is small, usually 0.01). The support of β corresponding to $\hat{L}$ is used to extract the features causally affecting x_i along with their coefficients (causal strengths) in the VAR model of order $\hat{L}$. The above procedure can be invoked independently for each feature x_i, $i \in [P]$ to determine the output feature causal graph.

4.2 Group Lasso Granger++

In *Group Lasso Granger++*, our approach stays the same as *Lasso Granger++* with changes only in the optimization routines. Following the same notations from 4.1, our first group lasso formulation, when $L_0 = \ell$, is :

$$\beta_G^0 = \arg\min_{\beta} \quad \frac{1}{2} \left\| Y - X^{Lagged}\beta \right\|^2 + \lambda \sum_{j=1}^{P} \sqrt{\rho_j} \left\|\beta_{\mathbb{G}_j}\right\|_2 \tag{17}$$

where Y and X^{Lagged} are same as that defined in equation 15. $\beta_{\mathbb{G}_j} = \{\beta_k : k \in \mathbb{G}_j\}$ where $\mathbb{G}_j$ denotes the set of group indices and ρ_j accounts for varying group size. Note that there are P groups here and they are of equal size (ℓ) initially.

Similarly the next optimization routine with $L_1 = 2\ell$ is :

$$\beta_G^1 = \arg\min_{\beta} \quad \frac{1}{2} \left\| Y - X^{Lagged}\beta \right\|^2 + \lambda \sum_{j=1}^{P} \sqrt{\rho_j} \left\|\beta_{\mathbb{G}_j}\right\|_2 \tag{18}$$

where, n and Y are same as that defined in equation 16. X^{Lagged} is a $n \times (P\ell + k_0)$ matrix, $k_0 =$ | $supp(\beta_G^0)$ |. The t^{th} row of X^{Lagged} is the row vector constructed as follows :
$\left[[\tilde{x}_i : i \in supp(\beta_G^0)], [x_1^{t-l-1}, \ldots, x_1^{t-2\ell}], \ldots, [x_P^{t-l-1}, \ldots, x_P^{t-2\ell}]\right]^T$ of size $P\ell + k_0$. Here ρ_j might be different for each group and is set to be the dimension of each group. The rest of the algorithm is same as *Lasso Granger++*. Also running *Group Lasso Granger++* independently for each feature x_i results in construction of the output feature causal graph.

Few crucial aspects : (a) The choice of λ, the regularization parameter plays in key role in both Lasso and Group Lasso routines discussed above. We tune λ by varying it in different step sizes within an interval $[0.001, 20]$ and for every choice of lag L_i, we choose that λ for which the AIC (or AICc in case of small samples) is *minimum*. Notice that very high value of λ will result in under-fit since it will over sparsify the model and very small value of λ will unnecessarily over-parameterize the model and nullify the sparsity of the coefficient vector β. Since AIC (or AICc) is the most widely used "Model Selection" criterion and in our case there is in fact a direct correspondence between a choice of λ and the degree of freedom captured by AIC, we appeal to it's use.

(b) The initial value of maxLag L_0 can be any positive integer greater than 1. But to get the "best" estimate for maxLag, we can adopt the following two pass approach. For example, let $L_0 = 5$. So, in the first pass, our subsequent guesses for L are $L_1 = 10$, $L_2 = 15$, $L_3 = 20$ and so on. Let the maxLag value $\hat{L}$ chosen as per our selection criteria be say 55. From the selection condition of our method itself we know that "true" maxLag cannot exceed 55. Therefore, we can run a second pass of our algorithm but now with the upper bound $M = 55$ and $L_0 = 1$. This will refine our estimate $\hat{L}$ if possible.

(c) In *Group Lasso Granger++* method, the construction of column indices for each group is crucial for all lag estimates greater than L_0. Every row of X^{Lagged} ($\forall L_i : i > 0$) has two components - (a) all time lagged feature values from older past, and, (b) selective feature values from recent past. So the assignment of column indices to every group from both these parts must be taken care of accordingly.

4.3 Complexity Analysis

The time complexity of Lasso Granger method, for a fixed L (with the efficient Lasso LARS [7] implementation) is $O(n(PL)^2)$, where $n = T - L$ is the number of samples. If instead of doing feature pruning at each estimate, we simply increment L in steps of ℓ and consider the L time-lagged values of all the P features, then the problem will become intractable soon. Let us see why is that so. When the maxLag estimate is $L_{k-1} = k\ell$, the effective number of features in the regression is precisely $PL_{k-1} = Pk\ell$. Under the high dimensional setting (i.e. P is large) and when the "true" maxLag itself is large enough, the product PL_{k-1} will be very large. Then the time complexity of this brute force approach would be $\tau_b = O\left(\sum_{j=1}^{\frac{T-2}{\ell}} (T - j\ell)P^2 j^2 \ell^2\right) = O(\frac{P^2T^4}{\ell})$. Since the effective number of features at each step governs the amount of memory space consumed, hence the space complexity of this brute force approach would be $\kappa_b = O\left(\sum_{j=1}^{\frac{T-2}{\ell}} Pj\ell\right) = O(\frac{PT^2}{\ell})$.

But with our method of selective feature pruning, we are able to bring down both the time and space complexity dramatically. With

the initial lag estimate $L_0 = \ell$, the time complexity of Lasso Granger is $O((T-\ell)(P\ell + s_0)^2)$, where $s_0 = 0$. When the lag estimate is $L_1 = 2\ell$, the running time is $O((T-2\ell)(P\ell + s_1)^2)$, where $s_1 = |\ supp(\beta^0)\ |$ is the cardinality of the support of the coefficient vector from previous estimate of L i.e. L_0. Note that $0 \leq s_1 \ll P\ell$ because of the sparsity inducing property of the L_1 norm operator. Thus when the maxLag estimate is L_{j-1}, the time complexity of Lasso Granger is $O((T-j\ell)(P\ell + s_{j-1})^2)$, where $s_{j-1} = |\ supp(\beta^{j-2})\ |$ is the cardinality of the support of the coefficient vector from previous estimate L_{j-2}. Therefore, the overall time complexity (τ) of *Lasso Granger++* is given by the expression :

$$\tau = O\left(\sum_{j=1}^{\frac{T-2}{\ell}} (T-j\ell)(P\ell + s_{j-1})^2\right), \quad 0 \leq s_{j-1} \ll P(j-1)\ell \tag{19}$$

$$= O\left(T^2 \max\left(P^2\ell, \frac{s^2}{\ell}, Ps\right)\right), \quad where \quad s = \max_j s_{j-1} \tag{20}$$

It is easy to see that each term in the r.h.s of τ (equation 19) is less than that of τ_b and therefore τ as a whole is much smaller than τ_b.

Similarly, the space complexity (κ) of *Lasso Granger++* is :

$$\kappa = O\left(\sum_{j=1}^{\frac{T-2}{\ell}} (P\ell + s_{j-1})\right), \quad 0 \leq s_{j-1} \ll P(j-1)\ell \tag{21}$$

$$= O\left(T \max\left(P, \frac{s}{\ell}\right)\right), \quad where \quad s = \max_j s_{j-1} \tag{22}$$

which is significantly smaller than $\frac{PT^2}{\ell}$.

Thus the effective number of features **does not** grow linearly with L and this is what makes our algorithm efficient as well as scalable both in terms of space and time complexity. We demonstrate this space and time complexity savings with one of our experiments on synthetic data in section 5.

5 EXPERIMENTS AND RESULTS

In this section, we present the results of our experimental evaluation on both synthetic data where the ground truth is known and a real world gene expression data of the human cancer cell cycle (HeLa S3) which is also validated against a de-facto database. The source code is available at https://github.com/MessianNil/GrangerCausality.

5.1 Synthetic data

We used VAR model throughout as the generative model for synthetic data. We compare the true maxlag for each time series variable in the original model with the ones predicted by our method. We also compare the similarity between the original feature causal graph and the hypothesis graph, in terms of the evaluation criteria already mentioned before. The results presented are in fact averages over 10 independent simulations of each experiment.

Experiment 1 : We consider the 3 basic building structures in a Granger Causal Network viz. the Co-parent structure, the Collider structure and the Chain structure. Let X, Y, Z be three time series variables which causally affect each other through different path delays (lags) labeled as τ_1 and τ_2 in figure 1. In the Co-parent structure (a) Z is the common cause for X and Y, in the Collider structure (b) Z is causally affected by both X and Y, and, lastly the

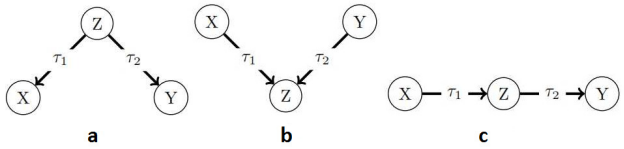


Figure 1: Three basic structures in a Granger Causal Network

Chain structure (c) where X is the indirect cause of Y through Z. The initial distribution of X, Y, Z are Gaussian with mean=0 and variance=1 and the lag values along the edges τ_1, τ_2 are integers chosen uniformly at random from the interval $[1, 10]$. The coefficients of the VAR model are drawn from the uniform distribution $\mathcal{U}(0,1)$ and the additive independent white noise processes η_i's are simulated as $\mathcal{N}(0, \sigma^2)$, $\sigma = 0.3$. The results of our method on these three models is presented in table 1. We also included for comparison the normal (standard) Lasso Granger method with L known and fixed beforehand. The evaluation scores are comparatively lower than our method. The relatively lower values of precision (P) is

Lasso Granger++				
Structure	P	R	F_1	Lag prediction accuracy
Co-Parent	0.824	1.000	0.884	1.000
Collider	0.764	0.950	0.808	0.967
Chain	0.867	1.000	0.914	1.000
Group Lasso Granger++				
Structure	P	R	F_1	Lag prediction accuracy
Co-Parent	0.884	1.000	0.927	1.000
Collider	0.791	0.950	0.831	1.000
Chain	0.867	1.000	0.914	1.000
Normal Lasso Granger (with fixed L)				
Structure	P	R	F_1	Lag prediction accuracy
Co-Parent	0.727	1.000	0.808	N.A
Collider	0.757	0.950	0.824	N.A
Chain	0.768	1.000	0.854	N.A

Table 1: Results on Experiment 1 (averaged over 10 sims.)

due to some random instances where the choice of lags τ_1 and τ_2 results in violation of the *m-separation criterion* ([4]). But in all cases our maxlag estimation is highly accurate as shown in the Lag prediction accuracy column (table 1). For the Co-parent structure, the maxlag value reported for X and Y are τ_1 and τ_2 respectively. Similarly, for the Collider structure, the maxlag value reported for Z is $\max(\tau_1, \tau_2)$ and so on for the Chain structure.

Experiment 2 : We consider the 3-variable VAR model used in [6]. This model gives rise to spurious causality due to smaller time lags. The VAR equations are :

$$x(t) = 0.8 * x(t-1) - 0.5 * x(t-2) + 0.4 * z(t-1) + \eta_1(t)$$
$$y(t) = 0.9 * y(t-1) - 0.8 * y(t-2) + \eta_2(t)$$
$$z(t) = 0.5 * z(t-1) - 0.2 * z(t-2) + 0.5 * y(t-1) + \eta_3(t)$$

where the initial distributions of x, y and z are Gaussian with zero mean and unit variance. The noise variables η_i's are distributed as $\mathcal{N}(0, \sigma^2)$ with $\sigma = 0.3$. Figure 2(a) is the ground truth representa-

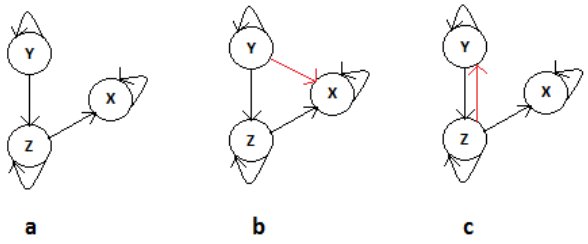


Figure 2: Ground Truth and Spurious edges

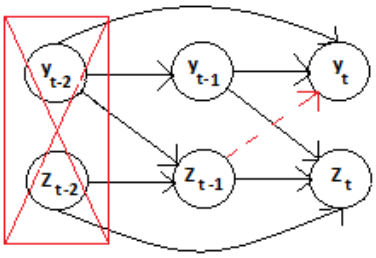


Figure 3: Spurious causality when model order is 1

tion of the feature causal graph and parts (b) and (c) highlights the spurious edges in red. The spurious edge from Y to X is detected when the maxlag (model order) is 2, because of causal influence of Y on Z at lag 1 and that of Z to X again at lag 1. Similarly when the model order is 1, the edge from Z to Y is inferred because of the causal edge from y_{t-2} to z_{t-1} as shown in figure 3. The performance of both the algorithms is summarized in table 2. In Lasso Granger++, since there is no group-wise penalization, the spurious edges are more frequent there resulting in relatively lower values of precision (P) and F_1 score, but the same is not true in case of Group Lasso Granger++. The maxlag prediction accuracy here also is 100% ($\hat{L}_x = \hat{L}_y = \hat{L}_z = 2$) for both the algorithms. We also included for comparison the normal (standard) Lasso Granger method with maxlag known and fixed beforehand. The evaluation scores are comparatively lower than our method. Also if the lag value chosen is incorrect in the standard methods, then it gives very wrong inferences.

Lasso Granger++			
P	R	F_1	Lag prediction accuracy
0.675	1.000	0.803	1.000
Group Lasso Granger++			
P	R	F_1	Lag prediction accuracy
0.921	1.000	0.956	1.000
Standard Lasso Granger (L fixed)			
P	R	F_1	Lag prediction accuracy
0.648	1.000	0.784	N.A

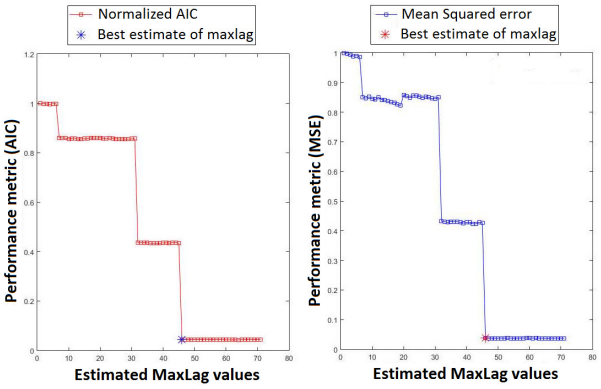
Table 2: Results on Experiment 2 (averaged over 10 sims.)

Experiment 3 : This experiment demonstrates the discovery of maxlag when the lag values of individual features are significantly different from each other and also the maxlag is somewhat large. Consider a star graph with $P = 5$ time series variables $\{x_1, \ldots, x_5\}$. Features x_2 through x_5 are noise variables simulated from $\mathcal{N}(0,1)$ independently. The only target variable is x_1 which is causally affected by all the other variables $\{x_2, x_3, x_4, x_5\}$. The lag values (path delays) for $\{x_2, x_3, x_4, x_5\}$ are integers drawn uniformly at random from the interval $[1, 50]$, and their coefficients are also drawn from $\mathcal{U}(0,1)$. The error curve (for x_1) of both the algorithms is shown in figure 4. In both parts (a) and (b) of this figure, the curve in red is the AIC curve against different guesses of maxlag and the blue curve is the MSE against different guesses of L. It is clear from this figure that the maxlag selection due to MSE-ϵ convergence criterion and the AIC criterion are asymptotically similar as the number of samples $n = T - L$ grow. In one of the simulations, the lag values for $\{x_2, x_3, x_4, x_5\}$ were chosen randomly to be $\{46, 7, 46, 32\}$. The maxlag value chosen by both *Lasso Granger++* and *Group Lasso Granger++* is 46 (marked with blue asterisk in figure 4). The stepping nature of the error curve reflects the significant difference in lag values of the individual features causally affecting x_1. The first step is observed at $L = 7$, the second one at $L = 32$ and finally the MSE converges (AIC is minimized) at $L = 46$ which is what we predict as $\hat{L}$. Given this information, the set of features which have causal influences at significantly different time lags, can also be extracted by simply looking up at the set of active features computed by our method at these different lag points ($L = 7$ and $L = 32$ in this case).

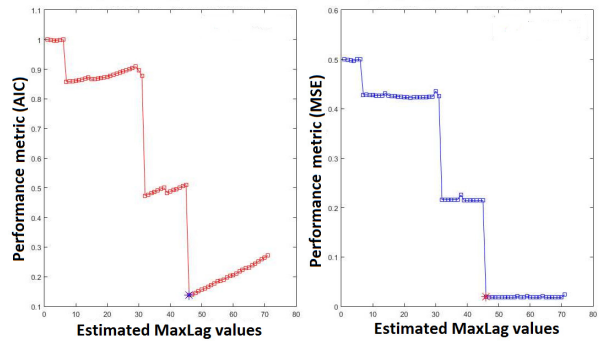
Experiment 4 : Here we demonstrate the efficiency, in terms of both space usage and time complexity, of our method (in particular Lasso Granger++) when compared against the brute force approach (without enforcing sparsity selection at each step) of Lag prediction and the usual Lasso Granger method with L fixed a priori. Figure 5 demonstrates the running time comparison of these algorithms. Of course, standard Lasso Granger (the curve in black) performs better than the other two algorithms since L is provided to it. But we claim that the running time behavior of our algorithm Lasso Granger++ (the red curve) compares favorably to it and is much superior to the brute force Lasso Granger (the blue curve) approach. Figure 6 demonstrates the efficiency of Lasso Granger++ with respect to the amount of memory space required as a function of maxlag. Brute force Lasso Granger (the blue curve) shows a linear growth rate whereas Lasso Granger++ (the red curve) shows a sub-linear growth rate since the effective number of features is much smaller than PL. We emphasize that if the number of features is in the thousands and the maximum lag is reasonably large, our algorithm still can be run on a standard laptop whereas this is not possible for the brute force Lasso Granger algorithm because of time and space constraints.

5.2 Real World data

We applied our concurrent estimation method to the gene expression data of the human cancer cell cycle (HeLa S3) [24]. We focus on the first four experiments from [12] only, having 12, 27, 48 and 19 data points resp. recording the expression levels of a handful of pre-identified genes well-studied by Whitfield et. al. [24] and Sambo et. al. [21]. For all the experiments, we follow the same



(a) Lasso Granger++ error curves



(b) Group Lasso Granger++ error curves

Figure 4: Step wise error curve highlighting $\hat{L}$ and the lag values of other causal features

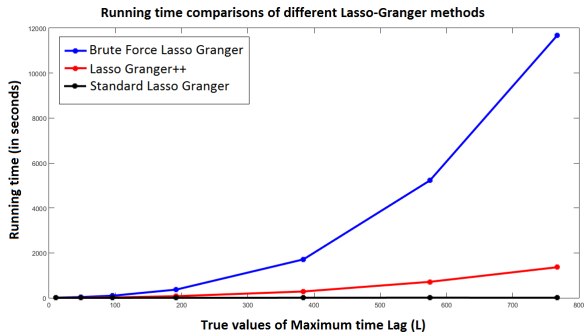


Figure 5: Running time as a function of L

data pre-processing steps as detailed in [12] (sec. 4). We cannot evaluate the performance of our method by comparing the "discovered" feature causal graph to the "true" graph since the latter is simply not known. Hence we focus on the particular subset of 9 genes as selected in [21] and compare the discovered gene-to-gene interactions to those reported in the BioGRID[4] database. However,

[4]https://thebiogrid.org/

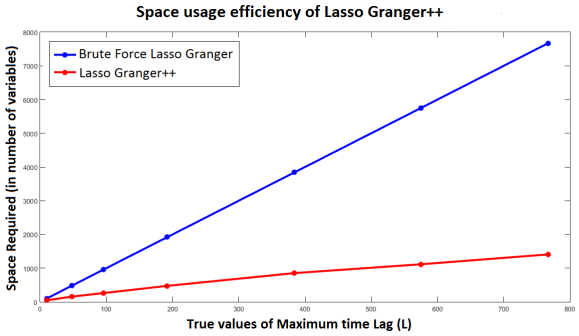


Figure 6: Space usage as a function of L

it is important to note that the list of interactions reported in BioGRID is not exhaustive. Also some of the interactions may not be direct. So caution must be taken when interpreting the results of such comparison : *false positives* in the output causal network with respect to BioGRID may not necessarily be *false*, and may contain actual links that are unknown till date, or known but have not yet been incorporated into the database. Figure 7 is the known network of interactions among these 9 genes as reported in th BioGRID database. We applied our method on data from both experiment 3 and 4. In experiment 3, the measurements are taken 1 hour apart, whereas for experiment 4 each observation is taken 2 hours apart. The networks discovered by our method is shown in figure 8. The recall (R) is 0.57 and 0.62 for *Lasso Granger++* and *Group Lasso Granger++* resp. For the extra edges reported by our method, we indeed found in literature [12, 24] that most of these genes are "regulated" genes (i.e. in-degree is high) except E2F1 which is also a "traffic/hub" gene. Also the maxlag values reported for this set of 9 genes are significantly large. For example. in experiment 4, the maxlag values estimated by our method based on AIC score is shown in table 3. The genes are labeled using alphabets A to I in figure 7, 8 and their correspondence is shown in table 3.

Name of the Gene	$\hat{L}$ predicted by Lasso Granger++
(A) CCNA2	8
(B) E2F1	9
(C) CDC6	9
(D) CDC2	9
(E) CCNB1	8
(F) PCNA	9
(G) CCNE1	10
(H) RFC4	9
(I) CDKN3	9

Table 3: Maxlag predicted for each gene (expt. 4)

6 CONCLUSIONS AND FUTURE WORK

In this paper, we proposed and evaluated a pure data-driven and computationally efficient method of estimating the maximum time lag value while modeling Granger Causality on time series data. The two algorithms we presented here viz. *Lasso Granger++* and *Group*

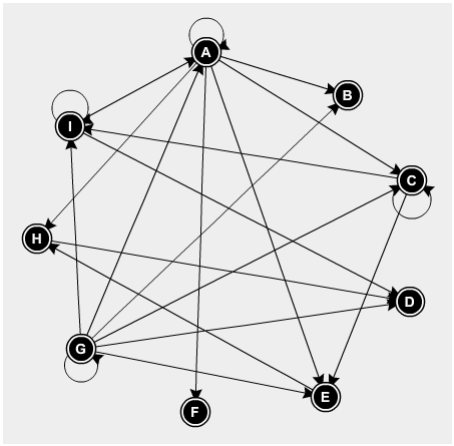


Figure 7: Network from the BioGRID database

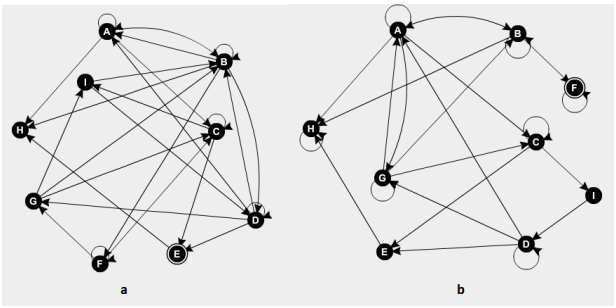


Figure 8: Networks discovered by our method - (a) Lasso Granger++ and (b) Group Lasso Granger++

Lasso Granger++ not only infers the hypothesis feature causal graph, but also estimates a value of max-lag by balancing the trade-off between "goodness of fit" and "model complexity" using standard model selection tools such as AIC or AICc. This enabled us to report a "sparse" subset of causal variables in high dimensional settings. Our empirical evaluations have demonstrated that the proposed approach has high predictive accuracy and low space requirements. Future efforts will be directed towards further investigating and exploring the range of real-world problems where our proposed method could add significant value.

REFERENCES

[1] Hirotogu Akaike. 1998. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike*. Springer, 199–213.

[2] Andrew Arnold, Yan Liu, and Naoki Abe. 2007. Temporal causal modeling with graphical granger methods. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 66–75.

[3] Frank Arntzenius. 1999. Reichenbach's common cause principle. (1999).

[4] Mohammad Taha Bahadori and Yan Liu. 2013. An examination of practical granger causality inference. In *Proceedings of the 2013 SIAM International Conference on data Mining*. SIAM, 467–475.

[5] T Chu and C Glymour. 2006. Semi-parametric Causal Inference for Nonlinear Time Series Data. *J. of Machine Learning Res., submitted* (2006).

[6] Mingzhou Ding, Yonghong Chen, and Steven L Bressler. 2006. 17 Granger causality: basic theory and application to neuroscience. *Handbook of time series analysis: recent theoretical developments and applications* 437 (2006).

[7] Bradley Efron, Trevor Hastie, Iain Johnstone, Robert Tibshirani, and others. 2004. Least angle regression. *The Annals of statistics* 32, 2 (2004), 407–499.

[8] Clive WJ Granger. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society* (1969), 424–438.

[9] Clive WJ Granger. 1980. Testing for causality: a personal viewpoint. *Journal of Economic Dynamics and control* 2 (1980), 329–352.

[10] Patrik O Hoyer, Shohei Shimizu, and Antti J Kerminen. 2006. Estimation of linear, non-gaussian causal models in the presence of confounding latent variables. *arXiv preprint cs/0603038* (2006).

[11] Clifford M Hurvich and Chih-Ling Tsai. 1989. Regression and time series model selection in small samples. *Biometrika* 76, 2 (1989), 297–307.

[12] Aurélie C Lozano, Naoki Abe, Yan Liu, and Saharon Rosset. 2009. Grouped graphical Granger modeling for gene expression regulatory networks discovery. *Bioinformatics* 25, 12 (2009), i110–i118.

[13] Aurelie C Lozano, Hongfei Li, Alexandru Niculescu-Mizil, Yan Liu, Claudia Perlich, Jonathan Hosking, and Naoki Abe. 2009. Spatial-temporal causal modeling for climate change attribution. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 587–596.

[14] Helmut Lütkepohl. 2011. *Vector autoregressive models*. Springer.

[15] Daniele Marinazzo, Mario Pellicoro, and Sebastiano Stramaglia. 2008. Kernel-Granger causality and the analysis of dynamical networks. *Physical review E* 77, 5 (2008), 056215.

[16] Robert Matthews. 2000. Storks deliver babies (p= 0.008). *Teaching Statistics* 22, 2 (2000), 36–38.

[17] Nicolai Meinshausen and Peter Bühlmann. 2006. High-dimensional graphs and variable selection with the lasso. *The annals of statistics* (2006), 1436–1462.

[18] Patrick Onghena and Eugene S Edgington. 1994. Randomization tests for restricted alternating treatments designs. *Behaviour research and therapy* 32, 7 (1994), 783–786.

[19] Judea Pearl. 2009. *Causality*. Cambridge university press.

[20] Sam Roweis and Zoubin Ghahramani. 1999. A unifying review of linear Gaussian models. *Neural computation* 11, 2 (1999), 305–345.

[21] Francesco Sambo, Barbara Di Camillo, and Gianna Toffolo. 2008. CNET: an algorithm for reverse engineering of causal gene networks. *NETTAB2008, Varenna, Italy* (2008).

[22] Peter Spirtes, Clark N Glymour, and Richard Scheines. 2000. *Causation, prediction, and search*. MIT press.

[23] Robert Tibshirani. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* (1996), 267–288.

[24] Michael L Whitfield, Gavin Sherlock, Alok J Saldanha, John I Murray, Catherine A Ball, Karen E Alexander, John C Matese, Charles M Perou, Myra M Hurt, Patrick O Brown, and others. 2002. Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Molecular biology of the cell* 13, 6 (2002), 1977–2000.

[25] Ming Yuan and Yi Lin. 2006. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 68, 1 (2006), 49–67.