Causality in Time-Series: A Short Review

Girish Keshav Palshikar, Manoj Apte, Sushodhan Vaishampayan, Akshada Shinde

TCS Research, Tata Consultancy Services Limited, Pune, India

Corresponding author: Sushodhan Vaishampayan, Email: sushodhan.sv@tcs.com

The study of causal relations has been proved to be significant in acquisition, understanding, and representation of human knowledge across physical and biological sciences, engineering, social sciences, economics. Identifying or inferring causal relations from empirical data is a crucial step in knowledge acquisition. For study of causality, time is an important factor, as the effect will always occur after its cause, and, often, within stipulated amount of time. Thus, time-series data is considered ideal for causal inference. In this chapter we review various notions of causality in time-series data. *Granger Causality* is the most widely used notion for extraction of causal relations from time-series data, and has been extended to non-linear, conditional and multivariate scenarios. We also review some of the other notions of causality like *Dynamic Bayesian network* and *Mutual Information* related causal extraction techniques. We briefly touch upon Sir Austin Bradford Hill's criteria for causality, which puts forward nine viewpoints for concluding an association relation into causality. Some applications of causality and a list of software tools are also provided. The chapter will provide sufficient information for the readers looking for a quick introduction to the field.

Keywords: Causality, Time-series data

1 Introduction

Much data in physical, biological, engineering, economics and social sciences comes in the form of time-series (TS). As one example, a blast furnace instrumented with various sensors generates a multivariate TS (MVTS) during its operation sessions. As another example, each server in a data center generates a

2023. In Satyasai Jagannath Nanda & Rajendra Prasad Yadav (eds.), *Data Science and Intelligent Computing Techniques*, 723–748. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.56155/978-81-955020-2-8-64

MVTS, as its CPU, disk and memory utilization levels are recorded by a monitoring program every second. Or, weekly advertising expenditure and sales revenue results in a TS. An important question in the analysis of TS data is: *Given two TS, is there any causal relation between them?* For instance, does increasing sales expenditure cause sales revenue to increase? To answer this question, one must formalize the notion of causality between TS in the statistical framework i.e., what it means to say that one TS "causes" another.

TS causality is an active area of research and several notions have been explored in the literature. For an excellent survey of statistical notions of causality in a non-time-series setting, see [1]. For other reviews of TS causality techniques, see [2], [3]. In this paper, we review some of the prominent approaches that have been proposed to formalize TS causality, without attempting to be comprehensive. We also briefly discuss some examples of practical applications of TS causality.

The paper is organized as follows. Section 2 explains the basic notion of Granger causality and some of the ways in which it has been extended. Section 3 explains other notions of causality in time series. Section 4 gives information of various software tools in R for computation of causality.

2 Granger Causality

2.1 Linear Granger Causality

Interdependence between two TS can be computed using cross-correlation (in time domain) or coherence (in frequency domain). However, neither captures any causal relationship between two TS. Granger devised one way to capture TS causality [4]: a TS Y causes another TS X if using Y (along with X) allows us to build a more accurate prediction model for X, than the one using only X. To use the Wald Test, for some given lag p, one can build an autoregressive (AR) model for a univariate TS X with and without another univariate TS Y, and check if the former model is more accurate than the latter: $X_t = \beta_0 + \beta_1 X_{t-1} + ... + \beta_p X_{t-p} + \epsilon_t$ $X_t = \alpha_0 + \alpha_1 X_{t-1} + ... + \alpha_p X_{t-p} + \gamma_1 Y_{t-1} + ... + \gamma_p Y_{t-p} + \delta_t$ Eq. 2.1a denotes the *restricted model* and Eq. 2.1b denotes the *full model*. The terms ϵ_t and δ_t are normally distributed white noise with mean zero and variances σ_R^2 and σ_U^2 respectively. Coefficients in both the models can be estimated using the numeric data values in the TS X and Y. The Residual Sum of Squares (RSS) values for these models, RSS_R (Eq. 1) and RSS_U, give their respective predictive accuracy. Now use the F-test to check if the full model has better accuracy than the restricted model. Compute

the F-value as follows:

$$F = \frac{(RSS_{full} - RSS_{restricted})/p}{RSS_{full}/(n-2p-1)} \tag{64.1}$$

Here, n denotes the total number of elements in TS X (we assume $|X| = |Y|$) and p is the lag value. Next, compare this computed F-value to the critical value $F^*_{\alpha,p,(n-2p-1)}$ of the F-distribution with p degrees of freedom (DoF) in numerator and $(n-2p-1)$ DoF in the denominator (this value is obtained from tables). Here, α is the level of significance; typically, $\alpha = 0.05$. If computed F-value $>$ F^* then we reject the null hypothesis H_0 (both models have the same accuracy) and accept the alternative hypothesis H_1 that adding p lag values of Y to the AR model improves prediction accuracy of X i.e., Y Granger-causes X. In general, we may have to try different lag (p) values, and we say Y *Granger-causes* X if this statement is true for at least one lag value. We have tacitly assumed that X and Y have the same sampling rate. If this is not the case, then one possibility is to re-sample both the time-series to have the same timestamps. Note that Granger causality is directed. An equivalent definition is that TS Y Granger causes TS X if the estimated variance $\hat{\sigma}^2_R$ of the residuals (i.e., prediction errors) of X in the restricted model is more than the estimated variance $\hat{\sigma}^2_U$ in the residuals of X in the full model (i.e., $ln(\frac{|\hat{\sigma}_U|}{|\hat{\sigma}_R|}) > 0$) in some well-defined statistical sense.

We need to check some conditions on TS X, Y, before we apply any Granger causality test to check if Y Granger-causes X. Informally, a TS is *integrated with order 0*, denoted $I(0)$, if its auto-covariance "quickly" decays to 0. If TS X, Y are both $I(0)$ (or at least stationary) then a Granger causality test can be applied to check if one TS Granger-causes another TS. Statistical tests, such as Augmented Dicky-Fuller (ADF) test, can check whether or not the given time-series is stationary.

If TS X is not stationary, then we can construct a new TS by using first difference: $Z_t = X_{t+1} - X_t$, for every $1 \leq t < |X|$, and then test Z for stationarity. If Z is stationary then X is called *integrated with order 1* (denoted $I(1)$). Suppose we are given two time-series X and Y. Suppose you are able to find a constant value β such that $Y_t - \beta X_t$ is relatively constant, for every $1 \leq t < |X|$, i.e., this new time-series is stationary $I(0)$. Then X and Y are said to be *co-integrated*. There are statistical tests, such as Engle-Granger test or Johansen test - which check whether the given two time-series are co-integrated. If X (or Y) is not $I(0)$ (or stationary), then we can still test them for Granger causality if X is $I(1)$ (or Y is $I(1)$), or X, Y are co-integrated.

This notion of Granger causality has several limitations. First, it is a linear notion. Second, it is seriously affected by confounding; when X and Y are both affected by a third process, it may not yield correct results. Finally, it applies only to pairs of variables; in general, we need multi-variate extensions of Granger causality.

2.2 Non-Linear Granger Causality

Linear Granger causality is unable to detect causal relations if they are non-linear; see [5] for an example. We summarize the statistical test given by Baek and Brock [6] for testing whether a TS Y *non-linearly Granger causes* another TS X. We assume that X, Y are strictly stationary, weakly dependent (i.e., $I(0)$) and satisfy some "mixing" conditions given in [7]. Let $L, M \geq 1$ denote given lag values for TS X and Y respectively, and let m denote the given *lead value* for X. For example, if $m = 3$, then the *lead vector* for X at time index t is $X_t^m = (x_{t+1}, x_{t+2}, x_{t+3})$. Similarly, the *lag vectors* are: $x_{t-L}^L = (x_{t-L}, x_{t-L+1}, \ldots, x_{t-1})$ and $y_{t-M}^M = (y_{t-M}, y_{t-M+1}, \ldots, y_{t-1})$. The null hypothesis H_0 is that Y does not non-linearly Granger cause X. The test procedure first fits a full linear model (using L lag values of X and M lag values of Y) to the data, and obtains the residuals for X i.e., difference between predicted and actual values of X. Similarly for Y. Any remaining predictive power of these residuals time-series can be considered as non-linear causal relation [6]. We now summarize the statistical hypothesis test developed by Baek and Brock [6] for this purpose.

Let $\epsilon > 0$ be a given small positive number. Let t, s be any two time indexes. Quantities $C1, C2, C3, C4$ are defined as below. $C1$ is the joint probability that (i) the distance between lag vectors X_{t-L}^{m+L} and X_{s-L}^{m+L} at time indexes t and s is $< \epsilon$; as well as (ii) the distance between lag vectorsY_{t-M}^M and Y_{s-M}^M at time indexes t and s is $< \epsilon$. $\|\cdot\|$ denotes distance, for which the *max* norm is used. $C2, C3, C4$ are understood similarly. C1(m+L, M, ⊠) = $\Pr(\|X_{t-L}^{m+L} - X_{s-L}^{m+L}\| < \epsilon, \|Y_{t-M}^M - Y_{s-M}^M\| < \epsilon)$
$C2(L, M, \epsilon) = \mathbf{Pr}(\|X_{t-L}^L - X_{s-L}^L\| < \epsilon, \|Y_{t-M}^M - Y_{s-M}^M\| < \epsilon)$
$C3(m + L, \epsilon) = \mathbf{Pr}(\|X_{t-L}^{m+L} - X_{s-L}^{m+L}\| < \epsilon)$
$C4(L, \epsilon) = \mathbf{Pr}(\|X_{t-L}^L - X_{s-L}^L\| < \epsilon)$ Then [6] gives correlation integral based estimators $\widehat{C1}, \widehat{C2}, \widehat{C3}, \widehat{C4}$ for the above quantities, which can be computed from the actual realizations of the TS X and Y. They show that under the null hypothesis and other conditions on X, Y given above, the test statistic given on the left follows the Normal distribution with mean 0 and variance σ^2 (which depends on

m, L, M, ϵ):

$$\sqrt{n}\left(\frac{\widehat{C1}}{\widehat{C2}} - \frac{\widehat{C3}}{\widehat{C4}}\right) \sim N(0, \sigma^2) \tag{64.2}$$

They give an estimator for σ^2 that can be computed from the data. Here, $n = n_0 + 1 - m - max(L, M)$ and n_0 is the length of the given realization of time-series X (and Y). The hypothesis test procedure now simply computes the value of the test statistic from given data and if it is more than the critical value obtained from the Normal distribution on the right side (for, the significance level of, say, 0.05), then H_0 is rejected and the alternative hypothesis that Y non-linear Granger causes X is accepted. A modified form of this Baek and Brock test is given by Hiemstra and Jones [8], where they have given a better estimator for σ^2. Note that the choice of values for lags L, M, lead m and ϵ are important.

2.3 Conditional Granger Causality

As discussed, the Granger causality notion detects a *direct* causal relation between *two* time-series. Suppose a univariate TS Y Granger causes a univariate TS X. In some applications, we want to know whether this is a direct causal relation or whether it is entirely due to another "mediating" TS Z. The notion of *conditional Granger causality (CGC)* helps to answer this [9]. First, we form the *restricted model* for predicting X values using only the p lag values of X and Y, omitting the lagged values of the mediating TS Z (for illustration, we use $p = 2$). Then we form the *full model* for predicting X values using p lag values of all three TS: $X_t = \beta_0 + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \beta_3 Z_{t-1} + \beta_4 Z_{t-2} + \epsilon_t$
$X_t = \alpha_0 + \alpha_1 X_{t-1} + \alpha_2 X_{t-2} + \alpha_3 Z_{t-1} + \alpha_4 Z_{t-2} + \alpha_5 Y_{t-1} + \alpha_6 Y_{t-2} + \delta_t$ The terms ϵ_t and δ_t are normally distributed white noise with mean zero and variances σ_R^2 and σ_U^2 respectively. Coefficients in both the models can be estimated using the numeric data values in the TS X, Y, Z. The R^2 values for these models, $R^2_{restricted}$ and R^2_{full}, give their respective predictive accuracy. Now, as earlier, we can use the F-test to check if the full model has better accuracy than the restricted model. If computed F-value $> F^*$ (for a given significance level α) then we reject the null hypothesis H_0 (both models have the same accuracy) and accept the alternative hypothesis H_1 that the causal relation from TS Y to TS X is entirely mediated by TS Z.

There are several issues in practice in using this notion of conditional Granger causality. For example, TS X, Y, Z may be auto-correlated and cross-correlated, which violates the assumption of independence of observations underlying the Fisher test, and which often makes the test unreliable in establishing causality; see [10].

2.4 Multi-variate Granger Causality

Several attempts have been made to extend the notion of Granger causality to deal with multivariate time-series (MVTS) as these occur frequently in practical applications. Consider an n-dimensional MVTS $\mathbf{X} = \langle \mathbf{x}_1, \ldots, \mathbf{x}_N \rangle$, where each $\mathbf{x}_t^T$ is a column vector of n entries. Consider another m-dimensional MVTS $\mathbf{Y} = \langle \mathbf{y}_1, \ldots, \mathbf{y}_N \rangle$, where each $\mathbf{y}_t^T$ is a column vector of m entries. In general, these two MVTS may be dependent on each other in some way. As earlier, we want to check if using the past values of the MVTS $\mathbf{Y}$ along with past values of $\mathbf{X}$ results in a better predictive model for $\mathbf{X}$ than the model which only uses the past values of $\mathbf{X}$.

The equation corresponding to each component of the X TS is compactly written in the matrix notation as follows (this is the full vector auto-regressive (VAR) model):

$$X_t = \sum_{k=1}^{p} A^f_{xx,k} \cdot X_{t-k} + \sum_{k=1}^{p} A^f_{xy,k} \cdot Y_{t-k} + \boxtimes^f_{x,t} \tag{64.3}$$

Here X_t, Y_t are $n\times1$ and $m\times1$ column vectors respectively, $A^f_{xx,k}$, $A^f_{xy,k}$ are the $n\times n$ and $n\times m$ coefficient matrices respectively (these are p matrices each) and $\boxtimes^f_{x,t}$ is the $n \times 1$ column vector of residuals. Define the covariance matrix for the residuals of the full model as: $\Sigma^f_{xx} = cov(\boxtimes^f_{x,t})$. The restricted VAR model is formulated similarly:

$$X_t = \sum_{k=1}^{p} A^r_{xx,k} \cdot X_{t-k} + \boxtimes^r_{x,t} \tag{64.4}$$

The covariance matrix for the residuals of the restricted model is defined as: $\Sigma^r_{xx} = cov(\boxtimes^r_{x,t})$. The null hypothesis that Y does not causally affect X is: $H_0 : A^f_{xy,1} = \ldots = A^f_{xy,p} = 0$ i.e., each of the p coefficient matrices in the full model is a zero matrix. The test statistic is the ratio of the determinants of the residual covariance matrices in the two models:

$$F_{Y\rightarrow X} = \frac{|\Sigma^r_{xx}|}{|\Sigma^f_{xx}|} \tag{64.5}$$

Now the multivariate granger causality from the set of Y TS to the set of X TS can be tested using the F-test as discussed earlier. The model coefficients can be estimated using actual realizations of Y and X MVTS. MATLAB includes the MVGC toolbox for multivariate Granger causality inference [11]. This notion of unconditional multivariate Granger causality can be extended to the conditional case. We have discussed the time-domain formulation, which can also be alternatively and equivalently formulated in frequency-domain. See also [12], [13].

2.5 Temporal Causal Modeling with Graphical Granger Methods

Along with the traditional approach of using Granger causality, various extensions have been proposed and developed to overcome the limitations of canonical way of using Granger causality. In [14], authors examine some algorithms, falling loosely under the category of Granger graphical methods and compare their relative performance from multiple viewpoints. They randomly generate a large number of simulations, in which target time-series model is generated (mostly VAR), and then examine the performance of the various methods as a function of different parameters of the simulation. Similar experiments are performed on actual dataset of Standard & Poor's Compustat Data[1]. The performance of a causal modeling algorithm is measured in terms of similarity between the hypothesis graph and the actual graph which generated the input data. Few different approaches along with the canonical approach are described in [14]. The Lasso-Granger method, applies regression to neighborhood selection problem, to identify the subset of features on which the particular feature conditionally depends. The best regressor subset will be the one with least squared error, and will have non-zero coefficients. The SIN Granger method [15] is based on the observation that there is no causal relation between two variables, x_i and x_j when a subset of variables $X_s \in X \backslash \{x_i, x_j\}$ exists, such that, x_i and x_j are independent given X_s [16], [17]. SIN is based on $\rho_{xy.V}$, the partial correlation, which is the correlation among variables x and y given the remaining V variables. The VAR method [18], generalizes the univariate AR to multiple time series. The VAR estimation method is to invert the coefficient matrix and solve least squared regression problems. The paper also suggests regularized VAR method for sparse data. According to [14], the Lasso Granger method provides the advantage of consistency. They apply the frequently used metrics *precision*, *recall* and F_1*-measure* as well as graph structure for evaluating graph similarity between the generated/ hypothetical causal graph and the true causal graph. The authors emphasize that, conditions in which different approaches are most effective need to be analyzed. There is a need to explore different combinations of existing techniques.

2.6 Grouped Graphical Granger Modeling

In [19], the authors propose a novel enhancement to the graphical Granger methodology, with the use of regression methods, which are sensitive to group information, to leverage the group structure present in lagged temporal variables. They propose a new family of algorithms call *Group Boosting*. Further, the authors, make an attempt to prove that *Group Boosting* is equivalent to *Group Lasso* for the special case of linear model and orthonormal data. They try to answer the relevant variable selection question; *"whether the lagged variables for a given time-series,* ***as a group*** *(as opposed to individual variable), are to be included in regression?"*. They propose two new algorithms, *Group Boosting* and *Adaptive Group Boosting* for group variable selection. The advantages of group boosting methods are that, they are computationally inexpensive, can handle non-parametric models and,

[1]http://www.compustat.com (Accessed Date: 6^{th} September 2023)

data of mixed types. The authors suggest to further extend the research in the field of Graph Boosting algorithms, like exploring interactions with other approaches and, improving their theoretical guarantee.

2.7 Recent Advances

Granger Causality remains an active area of research (see [20] and [21] for recent reviews) and continues to be used in novel applications across diverse disciplines. While the majority of applications of Granger Causality are in economics, physics and engineering, other disciplines have also enthusiastically adopted Granger Causality analysis: neuroscience [22], politics [23], climatology [24], environment science [25] and so forth.

Despite the successful use of Granger Causality across a wide range of applications, some concerns remain: scope of its applicability, its strengths and limitations, its reliability in terms of discovering the true causal structures underlying the observed TS data and possibilities of obtaining spurious causal relations, among others. We will discuss a few papers that quantitatively demonstrate the limitations of Granger Causality formalisms.

Standard Granger Causality assumes that the underlying process that generated the data is linear and stationary. These assumptions are not valid for many real-life processes, such as neural spikes, systems with latent variables or continuous-time processes. Thus, Granger Causality should be used only after such assumptions regarding the generative process for the observed data (as underlying Granger Causality) are carefully checked. Wherever possible, the causal relations obtained using Granger Causality should be compared with the known functional relations or properties [26]. [27] show that the spurious Granger Causality may be estimated if one or both TS are non-stationary. [10] uses simulated data to demonstrate a bias-variance tradeoff in conditional Granger Causality estimates: using the true order for the AR model introduces bias and increasing the model order reduces bias but increases the variance in model parameters; they show that this leads to spurious peaks and valleys and even negative values. Standard Granger Causality requires specifying the lag in the VAR model, which can be a challenge. Penalties like hierarchical lasso have been used to automatically select the relevant lags while guarding against overfitting.

Scientists are investigating questions of what is the appropriate methodology to be followed when using a particular formulation of Granger Causality; see [28]. Philosophical foundations of Granger Causality are also being examined; see [29], which argues (using philosophical reasoning) that causal relations discovered using Granger Causality have no epistemic utility (i.e., they are meaningless) if the domain knowledge is insufficient to validate them.

Given the limited expressive power of linear models used in Granger Causality, many attempts have been made to harnesss artificial neural networks for the purpose of discovering non-linear causal structures in data. [30] proposes modified structured multilayer perceptron (MLP) and recurrent neural network (RNN) frameworks that detect non-linear causal relationships while doing automatic lag selection through sparsity-

inducing penalties on the weights in the neural network; see also [31].

Copula functions are used in quantitative finance to model dependence between random variables. By Sklar's theorem, any multivariate joint distribution can be written in terms of univariate marginal distribution functions and a copula which describes the dependence structure between the variables. This theorem allows modeling and estimating a multivariate joint distribution by estimating marginals and copulas separately. [32] proposes model the dependence between returns of two financial markets (e.g., US and Japan stock markets) using a parametric copula and infer Granger Causality by testing whether the copula function of a pair of two financial markets is the independent copula; see also [33]. Similar approaches have been used to analyze the cryptocurrency markets.

In [34], the authors take up the task of discovering non-linear directional causal relations. They introduce nonlinear Granger Causality (IsNGC), that extracts conditional granger causality between two multivariate time-series condition on a large number of confounding variables. The techniques is developed to find the causal relations in settings where the number of observations (length of the time-series) is small. To reveal the statistically significant causal relations, the interactions are modeled as nonlinear state space transformations with no apriori assumptions on functional dependencies. The functional Magnetic Resonance Imaging (fMRI) data of the brain is used for experimentation in the paper.

Given some time-series data, different causal models can be inferred for different set of samples, where each set has a different underlying causal graph. However, some relevant information is shared among this samples, that can be used to predict the effects of these causal relations as result of shared dynamics between these causal models. In [35], the authors propose Amortized Causal Discovery framework which infers causal relations from these shared dynamics. Even though the inferred causal graphs using the technique are not verifiable and the results are empirical, the paper opens up new directions for future research work, where shared dynamics could be useful to explain and identify effects that remain undiscovered using the individual causal models inferred from the data.

3 Other notions of Causality in Time-series

3.1 Dynamic Bayesian Network

A **Bayesian Network (BN)/ Bayesian Belief Network (BBN)** is a graphical model, represented as a *Directed Acyclic Graph* (*DAG*), where the nodes in the graph represent random variables and edges between these nodes represent the causal relations. Each vertex has a conditional probability table (CPT) that quantifies the effects of parents on it. Given two variables X and Y, a directed edge from X to Y indicates a causal relation between the two variables where X is the cause and Y is the effect.

A **Dynamic Bayesian Network (DBN)** is a special case of BNs, aimed at modeling temporal dependencies. The static interpretation of the system in DBNs, that is the nodes, edges, and probabilities, is identical to that of BNs. In DBN, the variables could be con-

sidered as state where each time slice could represent a different state of the system. Let us denote the random variables/ states by uppercase letters, their values by small case letters, parents of a variable X by $Pa(X)$, number of variables by N, and time boundary by T respectively. t denotes a particular time instance and X_t denotes the random variable at time slice t. In DBN it is conventional to make first order Markov assumption that the state of the system at time t depends only on its immediate past, that is the state at $t-1$.

DBN consists of probability distribution function on the sequence of T hidden state variables $X = \{x_0, x_1, ..., x_{T-1}\}$ and T observation variables $Y = \{y_0, y_1, ..., y_{T-1}\}$. Complete specification of a DBN needs defining three sets of parameters, 1)State Transition Probability Distribution Functions (PDFs) $Pr(x_t|x_{t-1})$ i.e., time dependencies between states, 2) Observation PDFs $Pr(y_t|x_t)$ i.e., dependencies of observation nodes on other nodes at time instance t, and 3) Initial state distribution $Pr(x_0)$ i.e., initial PDF at the beginning of the process [36].

Inference in DBNs

- Probabilistic inference means calculating the probability $P(X|Y = y)$, where X is the set of *query* variables (states) and Y is the set of evidence variables (observations). The joint probability distribution from time $t = 0$ to $t = T$ is given by:

$$P(X_{0:T}) = P(X_0)\prod_{t=0}^{T}\prod_{i=1}^{N}P(X_t^i|Pa(X_t^i)) \tag{64.6}$$

The above equation is the foundation of all inference algorithms in DBNs [37].

For example of inference in DBN, refer figure 1. The example has three random variables, X, Y and Z. The values of these variables at time instance t are denoted as X_t, Y_t and Z_t respectively. And the values at previous time slice $t-1$ are denoted as X_{t-1}, Y_{t-1} and Z_{t-1} respectively. As seen from the figure, inside a particular time-slice, the variable X causes Y and Y causes Z (shown by black arrows). There are some dependencies present across time-slices (shown by green arrows). The variable Y at current time-instance t denoted by Y_t is dependent on variable X at previous time-slice $t-1$, denoted by X_{t-1}. Similarly Z_t depends on Z_{t-1}. To infer the probability that a variable takes a particular value at particular time-slice is nothing but just the product of the conditional probabilities of the parents (the variables on which variable under consideration is dependent) of that variables as shown in equation 64.6. For example, the $P(Y_t) = P(X_t)P(X_{t-1})P(Y_t|X_t, X_{t-1})$. Similarly $P(Z_t) = P(X_t)P(X_{t-1})P(Y_t|X_t, X_{t-1})P(Z_{t-1}|Y_{t-1})P(Y_{t-1}|X_{t-1})P(Z_t|Z_{t-1}, Y_t)$.

For inference task, we need to learn 1) structure of the Bayesian Network and 2) probabilities (the conditional probability distributions).

1. Learning the structure

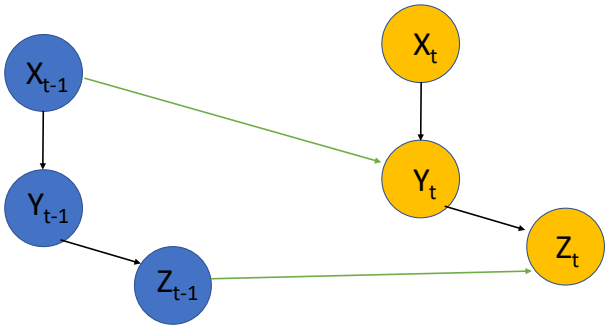

Figure 1: Dyanmic Bayesian Network Example

3.1.0.1 • The structure of a Bayesian Network can be learned by searching through the space of all possible structures and finding out the one which best describes the input data i.e., we search for the structure which maximizes the conditional probability $P(Data|\theta, M)$ where θ are the parameters of the distribution and M is the network structure. A scoring metric (e.g. Bayesian Information Criteria) is used to find the best suitable structure. Smart heuristics are used for searching through the entire space to avoid combinatorial explosion. Search algorithms like annealing, genetic algorithms. prove useful in such case.

3.1.0.2 • Another approach is to use constraint based algorithm that starts with a fully connected network and drops the edges, which show conditional independence. However, the repeated independence test by dropping the edges may result in loss of statistical power as effects of some variables will be ignored.

2. Learning the probabilities

3.1.0.3 • Estimating the probabilities for a DBN includes estimating the probabilities $P(X)$ and the conditional probability $P(X|Y)$. Non-parametric approach, uses histogram to calculate the probabilities. The other approach, which is parametric, assumes some kind of distribution for the data and estimates its parameters in a way, that best describes the probabilities from the input data. Typically

Gaussian distribution assumption is made and parameters for this assumed distribution are estimated from the data.

3.1.0.4 • The inference procedure searches through the space of all possible structures of the network, and computes the scoring metric for each, by estimating the probabilities and choosing the one with the best score. Some other tasks such as learning, decoding and prediction are also performed in DBNs [36], [38].

DBNs were introduced as a special case of singly connected Bayesian Networks that are used to represent time-series. They have been widely used to find dynamic relationships between variables in the areas such as bioinformatics, digital forensics, protein sequencing, network security and speech recognition.

Identifying dynamic relationships between infectious diseases and their related influencing factors is vital for infectious disease monitoring. [39] suggested how the DBNs can help in finding dynamic relations and improving the quality of infectious diseases surveillance. The real-world surveillance data of hand, foot, and mouth disease (HFMD) from Beijing in 2009 was used to represent dynamic relationships between weather factors (temperature(TEMP), relative humidity(RH), sunshine hours(SH).) and HFMD diseases. The weather-HFMD relationship (i.e., TEMP→ HFMD, RH→ HFMD) can be delayed because of the incubation period of infectious disease. Hence, the weekly cases of HFMD diseases and weekly average temperature and average relative humidity are considered to discover dynamic causal relations between them.

According to [40], the DBN could be learnt from the VAR model with an effective model selection procedure. The influence of past p observations on the current observations characterized by VAR(p) model as below,

$$X_t = \mu_t + \phi_1.X_{t-1} + .. + \phi_p.X_{t-p} + \alpha_t \tag{64.7}$$

where $X_t = (X_t^0, X_t^1, ..X_t^m)^{'}$, $\mu_t = (\mu_t^0, \mu_t^1..\mu_t^m)^{'}$ is a $(m+1)$ dimension constant vector and $\alpha_t = (\alpha_t^0, \alpha_t^1..\alpha_t^m)$ is a sequence of independent and identically distributed random vectors with mean zero and constant covariance matrix. The ϕ_i^* interpreted as lag-i ($i = 0, 1, .., p$), Auto-Regressive coefficient matrix with dimensions $(m + 1) * (m + 1)$ which measures the dynamic dependencies between X_t and X_{t-i}.

The performance of DBNs was evaluated by two simulations, 1) The comparison of DBNs performance with other models such as the Granger causality test and Least Absolute Shrinkage and Selection Operator (LASSO) method 2) To assess how the DBNs could improve the forecasting ability of infectious diseases. It was observed that desirable sample size is important in identifying dynamic relations among multiple variables. The true positive rates (TPR) were higher with a large sample size than the small sample size compared to the other two models. For the second simulation, DBNs are used to identify risk factors of HFMD disease before building the forecasting model. The true positive rate for identifying risk factors was 95.48% and that of false positive rates was not more than 5%. The DBNs could reliably and efficiently detect the relationships between infectious diseases and a number of exogenous factors. This could have a real-world impact by pro-

viding the Centers for Disease Control and Prevention (CDC) with choosing prominent influencing factors of current infectious disease.

DBNs have been widely used in Bioinformatics. They are considered a promising model for inferring gene networks from time-series microarray data. DBNs can handle time delay information and generate cyclic networks since the actual gene networks feature the cyclic regulatory pathways with feedback loops. In [41], DBNs are used to infer gene networks from the real-world time-series gene data of *Saccharomyces cerevisiae*. The data contains 18, 24,17, and 14 time-points with attributes alpha, cdc15, cdc28, and elu. DBNs use time-series data to form the causal relationship between genes and construct the gene networks. Although microarray data is continuous and quite noisy, it is discretized to reduce noise. A discrete DBN model is then applied to estimate gene networks. The *Sacccharomyces cerevisiae* data is binarized into over-expressed and under-expressed genes based on whether the rate of expression was considerably more than or lower than some threshold. This model has certain flaws, such as the possibility of information loss due to discretization and the threshold value. The threshold value must be carefully chosen because the estimated networks are reliant on it. Hence the continuous DBN model is used to derive gene networks and represented as normal density function [41]. The linear DBN model is formed. However, there is no assurance that linear models will be able to represent relationships between genes. Therefore, B-splines a non-parametric regression model designed to discover non-linear causal relations is used. With another dataset KEGG, the targeted network CDC28 (YBR160w; cyclin-dependent protein kinase) forms the cyclic pathway of 45 genes. The number of false positive with DBN model is much smaller than the traditional BN. The efficient and accurate DBN model could help to infer gene networks. However, it is challenging to understand whole gene networks with only microarray data.

3.2 Bradford Hill's Criteria for Causation

The president's address [42] by Sir Austin Bradford Hill, presented nine viewpoints or criteria from which an association should be studied to conclude it as a causal relation (published in the proceeding of the Royal Society of Medicine in 1965). These criteria when examined together (rather than examining a single criteria), can help to answer the fundamental question "is there any way of explaining the set of facts together?, is there any answer equally or more likely than cause and effect?" while evaluating against an observed association. The nine criteria proposed by Hill are explained one by one as follows:

- **Strength** - The strong associations are more likely to be causal than the weaker associations, because, the strength of the affecting factor must be greater than the strength of the observed association to conclude into causality. The weaker associations could be due to bias or confounding. Today, rather than the strength of the association, the statistical significance is acceptable for causal discovery [43]. In [44], authors presented statistically significant estimates showing that employees in jobs with higher potential for flavouring chemical exposures had 2.8

times greater annual declines in forced expiratory volume (FEV) than employees in lower exposure jobs. A strong association is neither necessary nor sufficient for causality, nor is weakness necessary nor sufficient for absence of causality, and the strong association helps in only ruling out the hypothesis that the association is due to bias or confounding [45].

- **Consistency** - Consistency refers to the repeated observation of an association in different populations under different circumstances. Whether chance explains a revealed causality or not, can be answered only through repetition of the circumstances and observations. Consistency in results can justifiably infer that the association is not due to constant error or fallacy. Lack of consistency, however, does not rule out a causal association, because some effects are produced by their causes only under unusual circumstances. For example, Transfusion can cause HIV infection but it is not always the case. Consistency serves only to rule out hypotheses that the association is attributable to some factor that varies across studies.
- **Specificity** - Specificity criteria requires that a cause should have a single effect instead of multiple effects. The original criterion of specificity is widely considered weak or irrelevant. Causes of a given effect cannot be restricted to have only one effect on logical ground. It is quite natural that a cause has multiple effects. For example, smoking can affect a smoker in multiple ways. Questions have been raised on this criteria [46], [47].
- **Temporality** - Temporality refers that the cause must precede the effect. [48] referred temporality, an inarguably important criteria in terms of concluding causality. In the modern day research, there are instances where, many of effects require a much longer duration to occur, once the cause has occurred. It might also happen that the cause is gradually inducing its effect over a long period of time, making the process of causal extraction, costly, time consuming, and potentially infeasible. For example, in case of epidemiological experiments, exposures occur during specific periods of development or even in previous generation, resulting phenotype differences in offspring.
- **Biological Gradient** - It is expected that the association must reveal a biological gradient or monotone unidirectional dose-response curve suggesting that causality is more likely. For example, more smoking means more carcinogen exposure and more tissue damage, hence more carcinogenesis. Such an example may not be present always. For example, the controversial relation between alcohol consumption and mortality. Sometimes the monotone trend may be due to a causal relation between the confounding factor and the effect, rather than between the non-causal factor and the effect.
- **Plausibility** - The criteria suggests that the causality inferred should be consistent with the current body of knowledge regarding etiology. For instance, a biologically feasible causation, observed from the research, is more likely to be true.

Here the phrase "current body of knowledge is important", because the knowledge will get updated as new discoveries happen, making the currently inferred relations more likely or less likely in the future. For example, the correct cause of typhus infections, rejected on the grounds of plausibility, was accepted when the knowledge got updated. A new association observed must not be ignored, just because it is odd.

- **Coherence** - Cause and effect interpretation of the data should not seriously conflict with the generally known facts. The cause-and-effect story should make sense with all knowledge available to the researcher. Absence of coherent information should not be confused with presence of conflict with coherent information. Hill stated, histopathologic effect of smoking on bronchial epithelium (in reference to the association between smoking and lung cancer) or the difference in lung cancer incidence by sex, could reasonably be considered examples of plausibility as well as coherence.
- **Experiment** - Occasionally it is possible to appeal to the experimental or semi-experimental evidence. For example, because of some observed association, some preventive action is taken. If persons stop smoking cigarettes, is the associated event (lung cancer in this case) affected? Here the strongest support for the causation hypothesis may be revealed (through intervention or cessation).
- **Analogy** - When one causal agent is known, the standards of evidence are lowered for a second causal agent that is similar in some way [49]. For example, if there is strong evidence of a causal relationship between a particular agent and a specific disease, a weaker evidence that a similar agent may cause a similar disease, should be accepted by researchers. Analogy provides a source of more intricate causal hypothesis. Absence of analogies could indicate lack of imagination or experience rather than falsity of the hypothesis.

These above mentioned 9 criteria could guide researchers in distinguishing causality from association and perform complete and thorough checks from significant views before concluding an association into causality. These criteria prove as mere guidelines and assist the researchers to provide conclusions with higher confidence. However, the criteria do not guarantee the truthfulness of the conclusions.

In [50], authors describe the general causation approach provided by Hill's criteria as an assessment tool for specific causation with regards to post traumatic headache (PTH) and sexual assault. They emphasize that, fact finders are left at an evidential impasse, rather than a standard process to find the valid opinion among the conflicting opinions. They try to answer two major forensic questions *"could the exposure have caused the disease or injury outcome in this case?"* and those that answer the question *"did the exposure cause the disease or injury outcome in this case?"*. Hill's criteria are grouped into three causal milestones : *"Biological Plausibility"*, *"Temporality"*, and *"Strength of causal association"*. *"Biological Plausibility"* is meant to demonstrate whether or not the exposure *could* have caused the disease or injury outcome, instead of *how often. "Temporality"*

suggests that the headaches must start after trauma and within a reasonable amount of time. *"Strength of causal association"* is computed by comparing the risk of the condition relative to the suspected exposure, to the competing risk of the condition had the exposure not occurred, given the time frame of the exposure. Relative risk is used as a quantitative metric for specific causation. They conclude that the relative risk and the role of time between the trauma and onset of symptoms are the most critical concepts for the forensic examiner in dictating the strength of a causal relationship.

3.3 Causality detection using Information Theoretic approaches

Mutual Information (MI) of two random variables is the measure of mutual dependence between the two variables. MI can be expressed as

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y) \tag{64.8}$$

where $H(X), H(Y)$ are marginal entropies and $H(X|Y), H(Y|X)$ are conditional entropies and $H(X,Y)$ is the joint entropy of X and Y.

Similarly, Conditional Mutual Information (CMI) is defined as the expected value of the mutual information of two random variables given the value of third variable.

$$\begin{aligned} I(X;Y|Z) &= H(X,Z) + H(Y,Z) - H(X,Y,Z) - H(Z) \\ &= H(X|Z) - H(X|Y,Z) = H(X|Z) + H(Y|Z) - H(X,Y|Z) \end{aligned} \tag{64.9}$$

[51] compare the method of CMI for detection of causality with different methods. They compute the CMI between two variables of interest and propose that a significantly high value of CMI indicates the presence of a causal link between the investigated variables. They consider $x(t)$, $y(t)$ as two time series that represent the coordinates of two coupled dynamic systems X and Y respectively. They indicate information flow between $x(t)$ and $y(t+\tau)$ conditioned on $y(t)$ and its history. If this holds for various forward lags τ after averaging then there exists a causal link X to Y.

[52] first construct five examples of nonlinear systems where the Granger causality approach fails and then they propose Time Delayed Mutual Information (TDMI) to capture the relation between the time series with time delay τ.

$$I(X,Y,\tau) = \sum_{x_t} \sum_{y_t-\tau} p(x_t, y_{t-\tau}) log \frac{p(x_t, y_{t-\tau})}{p(x_t)p(y_{t-\tau})} \tag{64.10}$$

where $p(x_t, y_{t-\tau})$ is the joint probability distribution of two stationary signals $X = x_t$ and $Y = y_t$, $p(x_t)$ and $p(y_t)$ are their marginal probability distributions.

They propose that non-zero amplitude of mutual information indicates existence of interactions between the two signals and the sign of the time lag where $I(X, Y, \tau)$ reaches its peak magnitude is used to infer the causal direction of the interaction.

The five scenarios considered where Granger causality fails are a)unidirection is misinferred as no interaction, b)unidirection is misinferred as bidirection, c) unidirection misinfered as reversed unidirection, d)bidirection misinfered as no interaction and e)bidirection misinfered as unidirection. In first case of unidirection misinfered as no interaction they start with a linear dynamical system $x_t = \epsilon_t$ and $y_t = -0.1x_{t-1} + \eta_t$ where ϵ_t and η_t are independent and identically distributed standard Gaussian random variables. Realization of this system generates time series x_t and y_t with data length 10^7. They found $F_{x\to y} \approx 1.0 \times 10^{-2}$ and $F_{y\to x} \approx 6.0\times10^{-8}$ for the significance threshold $F_{thr} \approx 1.1\times10^{-6}$. So the linear Granger causality identifies the direction of causal interaction as $x \to y$. Next they construct a new signal $\tilde{X}$ with its realization $\tilde{X} = x^2$. Here Granger causality results give no causal interaction which is not correct. In contrast TDMI analysis identified the direction correctly in both linear as well as nonlinear systems.

They show that the TDMI analysis is applicable in high dimensional complex systems and have applied it to neural data to show the existence of θ-driving neuron in rat hippocampus that was reported in mouse hippocampus previously.

[53] investigate causal relationship between sentiment about a company in social media and it's stock price. They use transfer entropy to detect strength and direction of transfer of information between the sentiment and prices. This information transfer between two variables X and Y in terms of conditional mutual information for a given lag k is given as

$$TE^{(k)}_{(X\to Y)} = I(Y_t, X_{t-k}|Y_{t-k}) = H(Y_t|Y_{t-k}) - H(Y_t|X_{t-k}, Y_{t-k}) \tag{64.11}$$

Transfer entropy returns a non-negative real value and the magnitude of the number representing the amount of information measured. To give a benchmark they compare it with a null hypothesis from the dataset where any causal information is removed. Such null hypothesis data is obtained from the original data by randomly shuffling the time sequence of observations. They compare the statistical significance of the Transfer entropy results with Z as below:

$$Z = \frac{TE - \mu_{shuffle}}{\sigma_{shuffle}} \tag{64.12}$$

where $\mu_{shuffle}$ and $\sigma_{shuffle}$ are the mean and standard deviation of the shuffled transfer entropy. Larger Z score implies value of transfer entropy that is more significantly deviating from the expected values implying higher causality. They applied this to a set of top 50 companies of S&P for stock price provided by Yahoo Finance and the sentiment index provided by Brain [2] from November 2018 to November 2020. The study revealed significant causal relationships between companies price and sentiment.

[54] apply several methods based on Information theory to observational data in ecohydrologic and other systems. They construct an example based on observed 1 minute weather data consisting of air temperature T_a, shortwave solar radiation R_g, wind speed WS and relative humidity RH collected over 12 hour period from 6 AM to 6 PM on 30th Aug 2014 at the Sangamon Forest Preserve site in Central Illinois. They consider each of the four variables as 'target' of information and apply Transfer Entropy, Information decomposition and causal history analysis to study time dependent interactions.

TE quantifies the information transferred to a target Y_t, from a sequence of historical states of another variable $X_{t-1:t-\tau} = \{X_{t-1}, X_{t-2}, ..., X_{t-\tau}\}$, given the knowledge of its own past $Y_{t-1:t-\tau} = \{Y_{t-1}, Y_{t-2}, ..., Y_{t-\tau}\}$ and is given by $TE_{X\to Y}(\tau) = I(Y_t; X_{t-1:t-\tau}|Y_{t-1:t-\tau})$.

In the weather station example RH and Rg provide information to air temperature Ta at very short time lags and Ta and RH provide information to Rg at longer delays when transfer entropy is used. Their framework serves to analyze the system at different levels of pairwise, joint and multivariate causal interactions. They state that Information flow analyses to infer causal dependencies provides novel insights into system or network level behavior.

4 Software tools for time-series causality

Causal inference from time-series data is a widely researched and applied field in many real world domains. The traces of the field being studied can be traced some centuries before. Some applications of extracting causal relations, using different notions, are given in the section 3. It is obvious for such a widely studied area, to have already developed powerful software tools, by the experts of the field. In Table 1, we provide in brief, information about the various packages[3]

[2] https://braincompany.co (Accessed Date: 6^{th} September 2023)

[3] Mentioned tools and software packages are the ones available and widely used at the time of writing this chapter. Some of them may become obsolete or get replaced with new packages in the future.

available for causal inference and extraction, in R (one of the most widely used language for statistical research). There are equally powerful tools developed in other languages like Python, Matlab, Java. Here we provide the information about the tools available in R only.

Table 1: Software tools for time-series causality

Sr. No.	Package Name	Related Notion	Description	Important Functions
1	dbnlearn	Dynamic Bayesian Network	Allows to learn structure of univariate time-series, learn parameters, and perform forecasting.	fit(), learn(), predict(), preprocessing()
2	dbnR	Dynamic Bayesian Network	Provides learning and inference in DBNs of arbitrary Markovian order. Learns network from data (by offering 3 structure learning algorithms), performs exact inference, and provides forecasts of arbitrary lengths.	learn_dbn_struc(), plot_dynamic_network()
3	lmtest	Granger Causality	Performs Granger causality test in bivariate time-series.	grangertest()
4	grangers	Granger Causality	Inference on Granger causality in the frequency domain. Provides functions for calculation of unconditional and conditional Granger-causality spectra.	bc_test_cond(), bc_test_uncod(), Granger.conditional(), Granger.unconditional()
5	vars	Granger Causality	Estimation, lag selection, diagnostic testing, forecasting, causality analysis, forecast error variance decomposition and impulse response functions of VAR models and estimation of SVAR and SVEC models.	causality()
6	infotheo	Information Theory	The package implements various measures of information theory based on several entropy estimators	mutinformation(), condinformation()
7	praznik	Information Theory	A toolbox of fast, native and parallel implementations of various information-based importance criteria estimators and feature selection filters	cmiMatrix(), cmiscores()

5 Conclusion

A glance through the area of causal inference in time-series data, gives an impression that Granger causality is the most widely used and studied notion of causality. We have presented some important concepts related to Granger causality and its extensions in the section 2. However, Granger causality, too, has some limitations. It needs constraints of stationarity to be fulfilled. Also, it fails to capture all the aspects of causality. It cannot address the causal question where two features have a hidden common cause [14]. We describe some alternatives to Granger causality in the section 3. Dynamic Bayesian Networks, are considered equally powerful to the notion of Granger causality. The Hill's criteria could also be used to have a theoretical base while looking for causality. The Hill's criteria allows us to have a higher confidence on the results of the inference. In another

notion, sufficiently high value of Conditional Mutual Information / Relative Entropy is considered to indicate presence of a causal link between the variables under consideration. Readers can use the tools mentioned in section 4, to get a quick start in practicing detection of causality in time-series.

From the multiple notions of causality, the best fitting notion needs to be used considering the application and the objective of the research as well as the statistical nature of the data. There are some studies that showcase the suitable settings where these notions should be considered. For example in [55] the author compared the notion of Granger's causality to the that of DBN and came up with a critical point related to the length of the time-series, which suggests that DBN should be used when the time-series length is shorter than the critical point, otherwise, Granger's notion should be used. Similarly [51] gives comparative study of six methods of causality in bivariate time-series. Overall, the field of causality in time series is extremely vast and there are many more notions of causality than the ones presented in the chapter. For example, the use of Temporal logic for causal inference is another promising notion [56], [57]. The chapter contains a very small portion of the field, and tries to capture the most widely utilized and relevant notions, which could serve as a good introduction to the field.

References

[1] J. Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

[2] Samantha Kleinberg and George Hripcsak. A review of causal inference for biomedical informatics. *Journal of Biomedical Informatics*, pages 1102–1112, Dec. 2011.

[3] Rothman K.J. and Greenland S. Causation and causal inference in epidemiology. *American Journal of Public Health*, 95:Suppl 1:S144–50, 2005.

[4] J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Acta Physica Polonica B*, 37:424–438, 1969.

[5] W. Brock. Causality, chaos, explanation and prediction in economics and finance. In *Beyond belief: randomness, prediction and explanation in science*, pages 230–279, 1991.

[6] E. Baek and W. Brock. A nonparametric test for independence of a multivariate time series. In *Statistica Sinica 2*, pages 137–156, 1992.

[7] M. Denker and G. Keller. On u-statistics and von-mises statistics for weakly dependent processes. *Zeitschrift fur Wahrscheinlichkeitstheorie und Vervandte Gebiete*, 64:505–522, 1983.

[8] C. Hiemstra and J.D. Jones. Testing for linear and nonlinear granger causality in the stock price-volume relation. In *Journal of Finance*, volume 49, pages 1639–1664, 1994.

[9] J. F. Geweke. Measures of conditional linear dependence and feedback between time series. *Journal of the American Statistical Association*, 79(388):907–915, 1984.

[10] Patrick A. Stokesa and Patrick L. Purdon. A study of problems encountered in granger causality analysis from a neuroscience perspective. *Proc. of the National Academy of Sciences*, 114(34):E7063–E7072, 2017.

2023. In Satyasai Jagannath Nanda & Rajendra Prasad Yadav (eds.), *Data Science and Intelligent Computing Techniques*, 723–748. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.56155/978-81-955020-2-8-64

[11] Lionel Barnett and Anil K. Seth. The mvgc multivariate granger causality toolbox: A new approach to granger-causal inference. *Journal of Neuroscience Methods*, 223:50–68, Feb. 2014.

[12] Xiaotong Wen, Govindan Rangarajan, and Mingzhou Ding. Multivariate granger causality: an estimation framework based on factorization of the spectral density matrix. *Philosophical Transactions of the Royal Society A*, 371:20110610, 2013.

[13] Elsa Siggiridou and Dimitris Kugiumtzis. Granger causality in multivariate time series using a time-ordered restricted vector autoregressive model. *IEEE Transactions on Signal Processing*, 64(7):1759–1773, 2016.

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

[15] Mathias Drton and Michael D Perlman. A sinful approach to gaussian graphical model selection. *Journal of Statistical Planning and Inference*, 138(4):1179–1200, 2008.

[16] Judea Pearl. *Causality*. Cambridge university press, 2009.

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

[18] Paul D Gilbert. Combining var estimation and state space model reduction for simple good predictions. *Journal of Forecasting*, 14(3):229–250, 1995.

[19] Aurelie C Lozano, Naoki Abe, Yan Liu, and Saharon Rosset. Grouped graphical granger modeling methods for temporal causal modeling. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 577–586, 2009.

[20] Michael Eichler. Causal inference in time series analysis. In Carlo Berzuini, editor, *Causality : statistical perspectives and applications*, pages 327–352. Wiley, 3 edition, 2012.

[21] Ali Shojaie and Emily B Fox. Granger causality: A review and recent advances. *arXiv preprint arXiv:2105.02675*, 2021.

[22] Seth A.K., Barrett A.B., and Barnett L. Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience*, 35:3293–3287, 2015.

[23] John R. Freeman. Granger causality and the times series analysis of political relationships. *American Journal of Political Science*, 27(2):327–358, 1983.

[24] Dmitry A Smirnov and Igor I Mokhov. From granger causality to long-term causality: Application to climatic data. *Physical Review E*, 80(1):016208, 2009.

[25] Cathy W.S. Chen, Ying-Hen Hsieh, Hung-Chieh Su, and Jia Jing Wu. Causality test of ambient fine particles and human influenza in taiwan: Age group-specific disparity and geographic heterogeneity. *Environment International*, 111:354–361, 2018.

[26] M. Apte, S. Vaishampayan, and G.K. Palshikar. Detection of causally anomalous time-series. *International Journal of Data Science and Analytics*, 11:141–153, 2021.

[27] Zonglu He and Koichi Maekawa. On spurious granger causality. *Economics Letters*, 73(3):307–313, 2001.

[28] Cees Diks and Valentyn Panchenko. A new statistic and practical guidelines for nonparametric granger causality testing. *Journal of Economic Dynamics and Control*, 30(9-10):1647–1669, 2006.

[29] Mariusz Maziarz. A review of the granger-causality fallacy. *The journal of philosophical economics: Reflections on economic and social issues*, 8(2):86–105, 2015.

[30] Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily Fox. Neural granger causality. *arXiv preprint arXiv:1802.05842*, 2018.

[31] Maciej Rosoł, Marcel Młyńczak, and Gerard Cybulski. Granger causality test with nonlinear neural-network-based methods: Python package and simulation study. *Computer Methods and Programs in Biomedicine*, 216:106669, 2022.

[32] T–H. Lee and W. Yang. Granger-causality in quantiles between financial markets: using copula approach. *International Review of Financial Analysis*, 33:70–78, 2014.

[33] Hyuna Jang, Jong-Min Kim, and Hohsuk Noh. Vine copula granger causality in mean. *Economic Modelling*, 109:105798, 2022.

[34] Axel Wismüller, Adora M Dsouza, M Ali Vosoughi, and Anas Abidin. Large-scale nonlinear granger causality for inferring directed dependence from short multivariate time-series data. *Scientific reports*, 11(1):7817, 2021.

[35] Sindy Löwe, David Madras, Richard Zemel, and Max Welling. Amortized causal discovery: Learning to infer causal graphs from time-series data. In *Conference on Causal Learning and Reasoning*, pages 509–525. PMLR, 2022.

[36] V Mihajlovic and Milan Petkovic. Dynamic bayesian networks: A state of the art. *University of Twente Document Repository*, 2001.

[37] Xiao-Guang Gao, Jun-Feng Mei, Hai-Yang Chen, and Da-Qing Chen. Approximate inference for dynamic bayesian networks: sliding window approach. *Applied intelligence*, 40(4):575–591, 2014.

[38] Kevin Patrick Murphy. *Dynamic bayesian networks: representation, inference and learning*. University of California, Berkeley, 2002.

[39] Tao Zhang, Yue Ma, Xiong Xiao, Yun Lin, Xingyu Zhang, Fei Yin, and Xiaosong Li. Dynamic bayesian network in infectious diseases surveillance: a simulation study. *Scientific reports*, 9(1):1–12, 2019.

[40] Rainer Opgen-Rhein and Korbinian Strimmer. Learning causal networks from systems biology time course data: an effective model selection procedure for the vector autoregressive process. *BMC bioinformatics*, 8(2):1–8, 2007.

[41] Sun Yong Kim, Seiya Imoto, and Satoru Miyano. Inferring gene networks from time series microarray data using dynamic bayesian networks. *Briefings in bioinformatics*, 4(3):228–235, 2003.

[42] Austin Bradford Hill. The environment and disease: association or causation?, 1965.

[43] Kristen M Fedak, Autumn Bernal, Zachary A Capshaw, and Sherilyn Gross. Applying the bradford hill criteria in the 21st century: how data integration has changed causal inference in molecular epidemiology. *Emerging themes in epidemiology*, 12(1):1–9, 2015.

[44] Kathleen Kreiss, Chris A Piacitelli, and Jean Cox-Ganser. *Lung Function (spirometry) Testing in Employees at Flavorings Manufacturing Plant-Indiana*. US Department of Health and Human Services, Public Health Service, Centers ..., 2011.

[45] Kenneth J Rothman and Sander Greenland. H ill's criteria for causality. *Encyclopedia of biostatistics*, 4, 2005.

[46] Kenneth J Rothman, Sander Greenland, Timothy L Lash, et al. *Modern epidemiology*, volume 3. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia, 2008.

[47] Philip E Sartwell. "on the methodology of investigations of etiologic factors in chronic diseases"—further comments. *Journal of chronic diseases*, 11(1):61–63, 1960.

[48] Kenneth J Rothman and Sander Greenland. Causation and causal inference in epidemiology. *American journal of public health*, 95(S1):S144–S150, 2005.

[49] Mervyn Susser. What is a cause and how do we know one? a grammar for pragmatic epidemiology. *American Journal of Epidemiology*, 133(7):635–648, 1991.

[50] Michael D Freeman and Sean S Kohles. Application of the hill criteria to the causal association between post-traumatic headache and assault. *Egyptian Journal of Forensic Sciences*, 1(1):35–40, 2011.

[51] Anna Krakovská, Jozef Jakubík, Martina Chvosteková, David Coufal, Nikola Jajcay, and Milan Paluš. Comparison of six methods for the detection of causality in a bivariate time series. *Physical Review E*, 97(4):042207, 2018.

[52] Songting Li, Yanyang Xiao, Douglas Zhou, and David Cai. Causal inference in nonlinear systems: Granger causality versus time-delayed mutual information. *Physical Review E*, 97(5):052216, 2018.

[53] Roberta Scaramozzino, Paola Cerchiello, and Tomaso Aste. Information theoretic causality detection between financial and sentiment data. *Entropy*, 23(5):621, 2021.

[54] Allison E Goodwell, Peishi Jiang, Benjamin L Ruddell, and Praveen Kumar. Debates—does information theory provide a new paradigm for earth science? causality, interaction, and feedback. *Water Resources Research*, 56(2):e2019WR024940, 2020.

[55] Cunlu Zou and Jianfeng Feng. Granger causality vs. dynamic bayesian network inference: a comparative study. *BMC bioinformatics*, 10(1):1–17, 2009.

[56] James F Allen and George Ferguson. Actions and events in interval temporal logic. *Journal of logic and computation*, 4(5):531–579, 1994.

[57] Samantha Kleinberg and Bud Mishra. The temporal logic of causal structures. *arXiv preprint arXiv:1205.2634*, 2012.