Multivariate Time Series Data Causal Discovery

by

Bo Yuan Chang

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2021

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

One publication has resulted from the work presented in the thesis:

1. B. Chang, M. Naiel, S. Wardell, S. Kleinikkink, and J. Zelek, “Time-Series Causality with Missing Data”, JCVIS, vol. 6, no. 1, pp. 1-4, Jan. 2021.

Abstract

One of the goals for Artificial Intelligence is to achieve human-like intelligence. To that end, several solutions were proposed over the decades, where *causal structure discovery* was proposed as a viable tool for enabling human-like reasoning. It can be treated as two stages, first causal discovery that examines the cause-effect relationships between variables, which are then used in the second stage, referred to as *causal parameter inference*, to perform causal inference using counterfactual/logic-like reasoning similar to how human beings approach a problem. Generally speaking, there are two types of causal discovery algorithms: those that work with random variables and those that work with time series data. The focus of this thesis will be on the latter.

Performing causal studies on real world dataset is very challenging for time series data as it is prevalent to run into missing values. Currently, all existing causal algorithms require evenly-sampled time series data which unfortunately are not always available.

In this thesis I proposed a systems that can address this difficulties that is hindering causal learning on real world datasets. The proposed system performs causal discovery using time series data with missing entries (i.e., sparsely sampled data at varying intervals). The solution put forward for this task is comprised of two parts: data filling with Gaussian Process Regression, and causal learning using a the traditional Vector Autoregressive Model or Machine Learning based approach. For the first part, experiments have shown that Gaussian Process Regression outperformed all the benchmark filling techniques such as K Nearest Neighbour regression, Parametric Linear filling as well as random variable filling. The obtained Root Mean Square Error for GPR filled was the smallest under across all filling percentages, comfortably beating benchmark algorithms by margins (RMSE difference varies from 0.05 to 1.5). As for the second part, an Echo State Network for causal learning is used due to its fast running time and higher prediction capabilities when compared with other causal learning algorithms available in the industry such as algorithms like Structural Expectation Maximization (SEM), and Subsampled Linear Auto-Regression Absolute Coefficients algorithm (SLARAC). When working with a 10 percent missing

entries, the proposed system is capable of obtaining an MCC score of 0.31 on a -1 to +1 scale where +1 represents perfect prediction and -1 represents complete no usefulness of the result. The MCC score received from the proposed system significantly outperformed other methods such as SEM and SLARAC. To showcase the ability of the proposed system to adapt causal relationships on real world engineering applications, the experiment was conducted using a chemical refinery dataset called the Tennessee Eastman (TE) dataset.

Acknowledgements

I owe my deepest gratitude and appreciation to my supervisor Dr.John Zelek. Dr. Zelek has always been supportive, approachable, and helpful throughout my master study. His encouragement and understanding helped me go through the difficulties and created the space for me to develop research ideas. I thank my thesis reviewer Dr. Ning Jiang and Dr. Pan Zhao for agreeing to be on my review committee on a short notice.

I would also like to thank postdoctoral fellows Dr. Mohamed Naiel and Dr. Georges Younes for their support and advice. It is a pleasure when to work with them, and I thank them for all their mentorship.

I would like to thank the Ontario Centres of Excellence (OCE), Natural Sciences and Engineering Research Council (NSERC) and ATS Automation Tooling Systems Inc., for supporting this research work.

Dedication

This is dedicated to my mom and dad.

Table of Contents

List of Figures

List of Tables

Nomenclature

Acronym	Description
ABCNN	Attention-Based Convolutional Neural Network
AI	Artificial Intelligence
AR	Autoregressive
AUC	Area Under ROC Curve
CCM	Convergent Cross Mapping
CFIB	Canadian Federation of Independence Business
CTIR	Coarse-Grained Trans Information-Rate
EM	Expectation Maximization
ESN	Echo State Network
GC	Granger Causality
GP	Gaussian Process
GPR	Gaussian Process Regression
KNN	K Nearest Neighbour
MCC	Matthews Correlation Coefficient
MCMC	Monte Carol Markov Chain
MIME	Mutal Information on Mixed Embedding

ML	Machine Learning
PC	Peter Clark
PCMCI	Peter Clark Momentary Conditional Independence
ROC	Receiver Operating Characteristics
RNN	Recurrent Neural Network
RSS	Residual Sum of Squares
SEM	Structural Expectation Maximization
SLARAC	Subsampled Linear Auto-Regression Absolute Coefficients
TE	Tennessee Eastman
TE	Transfer Entropy
VAR	Vector Autoregressive

Chapter 1

Introduction

1.1 Background

Artificially replicating and possibly surpassing human's ability to interpret causal relationships are considered evolutional forward when compared to existing Artificial Intelligence systems [4]. Despite several decades of research, current state-of-the-art deep learning systems are still broadly interpreted as a family of highly nonlinear statistical models [5], or pattern detection regression engines that require a very large set of training samples to distill a meaningful outcome for a particular problem. In contrast, humans are capable of drawing conclusions and solving problems by identifying causal relationships between the different variables of a task using a very small amount of data. Thus, from this perspective, the processes of discovering causal relationships are fundamental tasks for researchers among all STEM disciplines. This inspired a large body of researchers to work on causal discovery algorithms in an attempt to answer various challenges in several fields, such as epidemiology [6], economy [7, 8], and medicine [9]. However, due to reasons such as complex relationships between a large number of hidden variables, learning causal relationships on real-world data can be very challenging as causality must be inferred from noisy data [10, 11]. To put things into perspective, the next section will provide us with an example of how current causal learning algorithms struggle to be applied for solving real-world

problems.

1.2 Motivation

In December 2019, a highly contagious and transmittable virus called the Coronavirus begun to spread. In just a little over one year, the virus has already caused over 120 million infections and 2.7 million deaths globally. In order to prevent the virus from further spreading, businesses are forced to shut down and billions of residences are out of jobs. According to the report from the Canadian Federation Of Independent Business (CFIB) in January 2021, 1 in 6 small businesses has already been permanently closed or will be permanently closed soon [12]. The unemployment rate, according to Statics Canada, reached 13.7 percent during the peak of Covid and it is still sitting at 9.4 percent to date. Canada's budget deficit is forecast to hit C$343.2 billion, which puts the nation at the largest shortfall since the Second World War [13]. The economy has never suffered to this extent in the history of Canada and it will take years for Canada to pay back the deficit. The impact of such a devastating virus maybe significantly reduce if there exists an algorithm that performs backtrack(i.e., based on the data presented find out the root cause and establish preventive measures in a timely manner). One of the potential approaches for root cause tracing is through the application of a causal learning algorithm.

Ideally, the researchers should be able to simply input all data for variables that may act as the potential cause for transmission into a causal learning algorithm (i.e., black box) for cause-effect learning. Using the causal relationships learned researchers should be able to pinpoint the exact root cause of the virus. Researchers can also use the black box to find the cure for such virus (i.e., substances that causes the virus to die) and start the massive vaccine production immediately. Unfortunately, that is not the case. Things are more complicated and there are a few restrictions preventing us from doing so. Some of the challenges (but not limited to this list) can be summarized as follow:

- Difficulty Obtaining Related Historic Data:

1. The data analyst can infer the correct cause-effect relationships only if the relevant dataset is provided. When there is a lack of historic data provided it is impossible for one to perform causal learning.

- Accuracy of Causal Discovery Algorithm:

  1. Although there are many algorithms proposed for the discovery of causal relationships, both for random variable and time-series data (such as Peter Clark Momentary Conditional Independence Algorithm [14], Vector Autoregressive [15], and Expectation-Maximization [16] algorithms), the accuracy is poor when dealing with real-world datasets [17] which are caused by the high degree of non-linearity in combination with noise, and as well as some of the unfeasible assumptions made by some of the learning algorithms (such as the stationary time series data assumption).

- Stochastic Elements Involved:

  1. Often in times there exists some level of stochastic elements from application to application. Such stochastic terms can make the learning process more difficult. Currently, there is a limited number of reliable open source causal learning algorithms that account for the stochastic effects. For example, the injection of a vaccine is very likely to build virus immunity but it is not guaranteed. Thus such a cause-effect relationship is stochastic rather than deterministic.

- Overfitting Issues

  1. The ability of learning algorithms to distinguish causality and correlation remains a challenge. Having a correlation among variables does not imply the variables are causality related. However, of the variables that are causally related, they are for sure correlated. While most of the causal learning packages are based on a statistical approach (i.e., equation-based), sometimes causal algorithms will classify two variables as causally related with one and another.

However, in reality, they may be completely independent of each other's existence and are only correlated.

2. When exploring the existence of causal relationships among variables, human beings can often take into consideration the information contained from the variable names themselves and reject causal relationships among variables that are very different from each other. On the other hand, a statistical-based learning algorithm will only determine the causal relationships among historical time series entries, completely neglecting the information contained from variable names themselves. This may lead up to falsely claimed causal relationships among variables that are completely unassociated with one and another.

- Missing Data Associated With Real-World Dataset:

  1. The real-world dataset is often contaminated and contains a degree of lost/missing entries. Those missing entries may be caused by different reasons such as hardware design and human mistakes.

  2. The challenge researchers faced is the ability to discover causal relationships among different variables when given the sparsely sampled dataset. Currently, all the proposed algorithms require an evenly sampled dataset as input.

I will limit the scope of my work to address the last problems mentioned previously that is, to perform causal learning on sparsely sampled time series data. I address the missing entry issue by proposing data filling with a non-parametric filling approach known as the Gaussian Process Regression (GPR) prior to causal learning. By doing so, sparsely sampled time series data is converted to evenly sample time series data which is required for all causal learning algorithms. I then compare the causal learning ability against a few benchmark filling techniques such as filling with the Nearest Neighbour, parametric linear filling, and random filling. The RMSE obtained showed that the system I propose can better preserve information and ultimately leads to a better causal prediction on both the classical learning approach as well as the machine learning algorithm.

Chapter 2

Related Work

In a real-world manufacturing process, equipment components are not only connected to each other, but the performances are also mutually dependent on each other. The concept of causality is also referred to as causation, which describes the cause-effect relationships between variables or events. To describe the causal relationships between all the variables, a network can be constructed with nodes denoting variables, and arcs denoting their causal relationships; this network is usually referred to as a causal map. Causality analysis provides an effective way to localize root cause, as well as perform bottleneck process analysis for the working production line since the causal map previously generated can clearly represent the cause-effect relationships among mechanical/electrical components. By having such a clearly labelled causal graph, engineers can perform productivity improvement in a more systematic way.

2.1 History of Causality

Although the term "causality" is still a fairly new AI concept and has only gained much attention in recent years, it has a long history. In fact, defining the cause and effect relationships among variables has been one of the very first tasks humans learned to survive in the wild jungle. The concept of causality has first been officially proposed by Athenian

philosopher Plato in 400 BC. In his words, he described causation as "if (something) that becomes or changes must do so owing to some cause; for nothing can come to be without a cause." [18]. Ever since then, different philosophers/social scientists have proposed slightly different definitions for causality through demonstrations of the different scientific experiments in the hope to generalize the definition of causality [19]. For a more in-depth review of the history of causality, please refer to [20].

2.1.1 Causation VS Correlation

Although the literature record for causality has been long engaged, the concept still seems to be fairly modern for a good portion of the general public. To this date causation is still perceived as another fancy term for correlation by many; however, correlation does not necessarily imply causality in reality. Before continuing the discussion of causality, it is important to clearly distinguish the concept of causality from correlation.

Correlation between variables refers to how the pair of variables tend to fluctuate together (e.g. when one increases, the other one will increase or decrease). Often, correlation can be represented by a mathematics equation with an equal sign. This means that when variable A increases, variable B will also increase and vice versa. On the other hand, causation indicates that the occurrence of one particular event caused another event to occur. For example, for a pair of variable A and variable B, and it is known that A causes B. This means that increasing variable A will cause variable B to increase or decrease, but increasing B may not result in an increase in A [21]. Causation cannot be expressed with an equal sign because they are not directionless [4].

Figure 2.1 demonstrates a simple example of the difference between causal and correlation. Due to the dead battery, the computer is forced to shut down, therefore the low battery percentage causes the computer to shut down and this is classified as causality. On the contrary, when the computer is forced to shut down is a good indication of the inability to run the video player and vice versa. However, neither incident is the root cause of the other one, thus this relationship is classified as correlation.

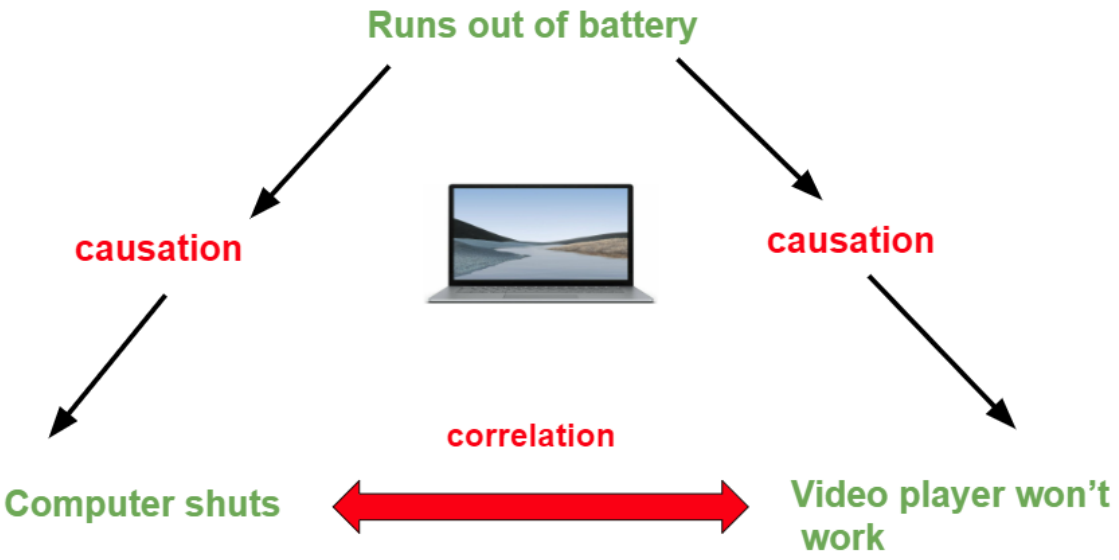


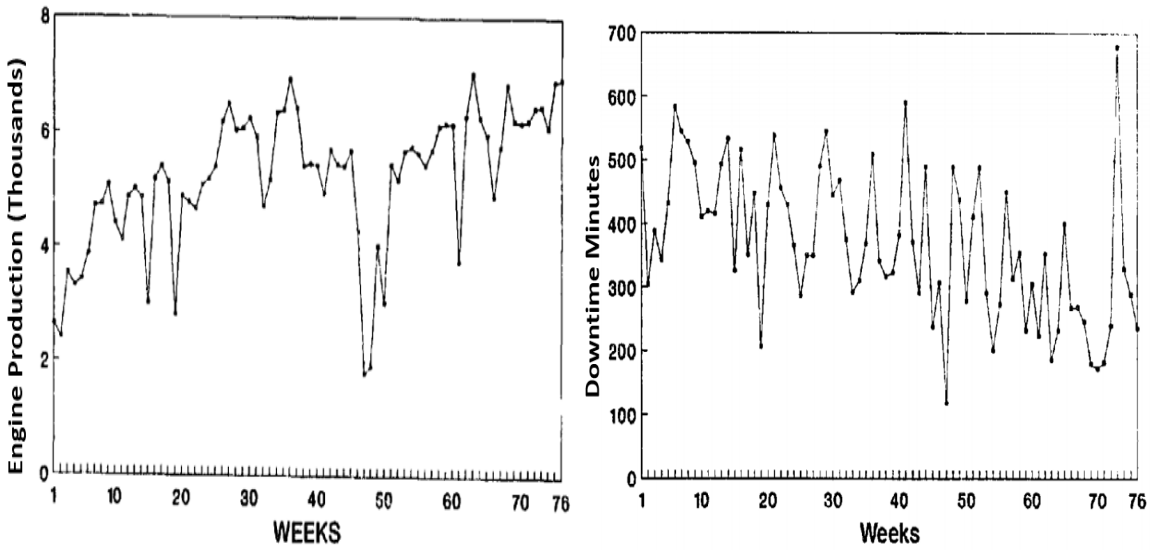
Figure 2.1 Simple illustration of the difference between causation and correlation on a laptop where the dead computer battery causes both the computer to shut down as well as the video player to not work. Whereas computer shutting down is correlated with video player not working, but one event does not cause the other.

The causal relationships among variables are usually obtained in a specially designed lab environment, where researchers have full control over one variable (the independent variable) and measure this variable's impact on another variable (the dependent variable). In [22], the researcher stated that cigarette smoking may in fact causing coffee drinking and cardiovascular disease since there is plenty of research done in the field proving that one of the negative impacts of smoking is cardiovascular disease [23]. But since there is no research done on the topic of the causal effect of drinking coffee towards cardiovascular disease and vice versa, one cannot conclude the cause-effect relationship between coffee drinking and cardiovascular disease. Another example can be found in [24] where the researchers performed an experiment to study the response of the gene expression for fathead minnow when exposed to chemical wastewater that contains steroid substances. In this study, the concentration of the steroid is carefully controlled by the researcher and acts as the causing variable. Researchers concluded that the the higher concentration of wasteful content inside water the higher alternation observed in the gene [24].

2.2 Types of Data

Generally speaking, there are two types of causal discovery algorithms: the type that involves time element, which is usually referred to as the time series data algorithms, and the other type that does not involve time element. An example of non-time series data is a dataset that is collected in no particular sequential order, where the first entry is completely uncorrelated with the second entry (e.g. treated as two separate samples). An example of this type of data can be found in [25] where information such as diameter, sex, and weight are recorded for each sampled abalone. Since the focus of my work is non this type of dataset, no in-depth literature review will be conducted. Please refer to [4], [26], and [27] for a thorough literature review in this area.

In contrast, time series data is a type of dataset that is collected in a sequential order, where the data collection is a set of continuous variables collected over a period of time,



(a) Plotted Time Series Data For Daily Engine Production in thousands

(b) Plotted Time Series Data For Production Line Downtime in Minutes

Figure 2.2 Using Time Series Causal Analysis researchers in source [29] researchers showed that the production line downtime directly causes the engine production. An increase in the downtime will result in a decrease in the engine and vice versa [29].

such as the stock price or temperature over a period of time. The frequency in which the data are being collected is generally referred to as the time series frequency. See [28] for a more in-depth definition of time series data. In the following section, section 2.5, I will talk about the types of time series causal learning. An example of time series data is shown in figure 2.2 where the daily engine production and the daily production line downtime are plotted wherein [29] it is concluded that the production line downtime is directly causing the number of engines manufactured.

2.3 Evaluation Metrics

In order for researchers to evaluate the performance of causal learning the Matthews Correlation Coefficient (MCC) score is often calculated. Equation for MCC score is shown in

equation 2.1 below:

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{2.1}$$

The MCC score provides a balanced measure between the True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) cases. Its output is $\in [-1, 1]$, where a +1 score indicates a perfect prediction, 0 indicates that the prediction made is no better than random guessing, and -1 indicates complete disagreement between prediction and observation [30].

In addition to the MCC score, researchers have also used Accuracy shown in equation 2.2, and direct true positive rate (TPR), shown in equation 2.3. Receiver Operating Characteristics (ROC) curve alongside with Area under ROC curve (AUC) index are often plotted and determined for performance evaluation [17] [31].

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \tag{2.2}$$

$$\text{DTP} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2.3}$$

2.4 Data Filling

Due to the complicated nature of time-series data (e.g. seasonal, trend, stochastic term, interventions, etc.), it is often hard to predict and difficult to work with. To make the problem even more challenging missing data are often ubiquitous in many real-world data [32]. Evenly sampled time-series data is essential for causal discovery for in order to predict an outcome for a future time instant y_t, a statistical based learning algorithm all requires entries for n past instants $y_{t-1}, y_{t-2}, \cdots, y_{t-n}$ where the sampling frequency Δt is constant can be expressed as $\Delta t = (t-1) - (t-2) = \cdots = (t-n) - (t-n-1)$. But when working with real world industry data is sometimes difficult to obtain these regular samples data in

the many industry sectors due to various reasons such as the processing speed limitation of Programmable logic controller in place.

Thus, in order for me to still perform causal learning, it is required to perform proper filling to recover missing entries before preceding with causal learning. Generally speaking, there are mainly two types of approaches for time series data filling [33]. The first category is the *parametric approach*, which is to simply consider a linear (or high degree polynomial function) equation and find the line of best fit of the training dataset. This approach is simple but at the same time, the degree of the order must be defined in advance. The general equation for linear regression in 1-dimensional space can be summarized as follows:

$$f(x) = \alpha + \beta x + \varepsilon \tag{2.4}$$

Where $f(x)$ represents the value trying to predict at x, α represents the constant term and ε_i is the noise term. When generalized to a n^{th} degree polynomial, the equation can be expressed as following in equation 2.5 where $\beta_0, \cdots, \beta_n$ represents the weight terms at each degree of order respectively.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \varepsilon. \tag{2.5}$$

In real-world datasets it is often very challenging to find a single equation of best fit to represent the entire dataset, suggesting that the parametric approach is not as practical as one would hope when dealing with complicated real-world datasets. In contrast to the first category, the *non-parametric approach* is constructed according to information derived from the data rather than specifying a functional form to best represent the dataset. This technique offers much higher flexibility when dealing with more complex datasets [34].

K-Nearest Neighbour is a very popular non-parametric approach that can be applied to both classification and regression tasks [35]. The concept has been widely applied to the field of imagine classification [36] [37]. When performing regression on time series data the equation can be expressed as equation 2.6 where x_i represents the i^{th} nearest neighbour

value. Term ω represents the weight term assigned with each of the neighbour term where $\omega_1 + \cdots + \omega_K = 1$. However, the selection of the number of neighbours to include and assigning proper weight terms can be a very manual process. The filling accuracy depends heavily on the parameters chosen.

$$y = \sqrt{\sum_{i=1}^{K} \omega_i (x_i)^2} \tag{2.6}$$

Gaussian Processes (GP) is another non-parametric algorithm that can be applied to solve both complicated regression and classification problems [38]. Gaussian process refers to a time series stochastic process such that every finite collection of those the entries has a multivariate normal distribution. The GP process repetitively draw samples from the distribution to construct predicted time series entries at a given time instant. Since the distribution can result in infinite number of entries generated, it is often viewed as a process of defining a distribution over functions and generating a unique function for every single entry point. Generally speaking the GP algorithm is mainly applied in the area of supervised learning [38] while there is also some work done in areas like unsupervised learning [39] and reinforcement learning [40]. Due to the high flexibility offered by Gaussian Process Regression it is used to perform data filling prior to causal learning. More details regarding GPR will be elaborated in chapter 3.

2.5 Granger Causality

Although many of the pioneer researchers in the field acknowledges the existence of causality between different time series data, it was not until 1969 when an economist, Sir Clive Granger, proposed the concept of Granger Causality that has been first applied to the area of the economy. The concept has generalized the causality term for researchers and is later viewed as one of the fundamental theorems for time series causality [41]. Due to his unique contribution to the area, he was awarded the Nobel Memorial Prize in Economic Sciences in

2003. The term Granger Causality is a statistical hypothesis used to determine whether one time series can be used to improve prediction accuracy for another time series; the concept of Granger Causality has been used in the majority of the exciting causal algorithms today [21]. The concept of Granger Causality is intuitive and straightforward, where one time series variable Granger Causes another variable when the following two principals are satisfied:

1. Only the input/intervention from the past can Granger cause the outcome in the future. Future input/intervention cannot Granger cause any past values.

2. If having the info of variable A can improve the prediction value of variable B, then variable A Granger causes variable B.

Under the above assumptions, let a time-varying system comprised of n time-series be defined as $\S = \{T_1(t), T_2(t), \cdots, T_n(t)\}$, where $T_i(t)$ represents the i^{th} time-series component of the system; then one can state $T_i(t)$ Granger causes $T_j(t)$, for $T_i, T_j \in S$ if and only if:

$$\mathcal{P}\left[T_i(t+1) \mid \mathcal{I}(T(t))\right] \neq \mathcal{P}\left[T_i(t+1) \mid \mathcal{I}\left(T^{(-j)}(t)\right)\right] \tag{2.7}$$

Where $\mathcal{P}$ is the probability density function and $\mathcal{I}(T^{(-j)}(t))$ represents all the usable information in the universe provided by the time-series system up to time t excluding the j^{th} component. We can conclude the existence of Granger Causality between A and B if the hypothesis described in equation 2.7 is satisfied [41].

Although Granger Causality is the most popular concept in the field of causal discovery, it also received some criticisms. One of the most popular arguments against Granger Causality is that the concept lacks the ability to track down real-world causal relationships which are mainly non-linear [42]. Although there are some advancement done in exploring non-linear Granger Causality relationships, the performance is still not the greatest [43] [44] [45]. Another limiting factor for Granger Causality is the lack of ability to be rationalized

using human intuition. For example, in [46] researchers reported that the method mistakenly classified the total revenues generated by arcades as directly correlates with the number of computer science doctorates awarded in the US over the years. In other words, I can come up with a good prediction of the number of computer science doctorates graduates when knowing past years' arcades revenue. While such a claim might be mathematically correct, these two variables are certainly not causality related to each other. Similar to many other machine learning algorithms, Granger Causality lacks the intuition that humans have. As a result, it will mistakenly conclude falsely classified causal relationships similar to the one mentioned above. Due to the drawbacks mentioned previously, some researchers in the field believe it is necessary to first improve the theoretical hypotheses before increasing the specificity of GC-based models [42]. Despite all the drawbacks of GC mentioned above, the concept of Granger Causality still remains one of the hottest concepts in the field. Different variations of Granger Causality have been proposed over the years to deal with the drawbacks mentioned above.

2.6 Other Causality

2.6.1 Sims Causality

Aside from the most popular GC, there are a few other causality concepts in the field such as Sims Causality [47], and Intervention Causality [48] [49]. Most of the time series causal discovery algorithms are still based on the concept of Granger Causality.

For example, the concept of Sims Causality is defined in the year 1972, 3 years after the seminal publication from Granger, which stated that although two white noise variables might be classified as Granger Causing one and another, such relationship is not classified as Granger(Sims) Causality [47]. The concept of Sims Causality is often treated as a compliment of Granger Causality where Granger Causality implies Sims Causality but the inverse is not true. Sims stated that a pairwise Granger Causality for $X[t]$ and $Y[t]$ can

be treated as moving average along several lag terms of the two variables, expressed as following:

$$Y_t = \alpha_1 Y_{t-1} + \beta_1 X_{t-2} + \cdots + \alpha_n Y_{t-n} + \beta_n X_{t-n} + C + \epsilon \tag{2.8}$$

where ϵ represents the combined noise term from both variables. α and β terms represent the parameter values at each time lag while C is the combined constant term. Sims Causality stated that variable X does not Granger Cause Y if and only if $\beta_1, \beta_2, \cdots, \beta_n$ is being chosen identically to zero. This can be expressed as equation 2.9:

$$\beta_1 = \beta_2 = \cdots = \beta_n = 0 \tag{2.9}$$

2.6.2 Intervention Causality

The concept of Intervention causality is first proposed by Judea Pearl in 1993 [50]. The concept was first applied to work with random variables and has only been applied to time series data recently. The idea focuses on the idea of counterfactual that calculates the Average Causal Effect (ACE) which can be expressed as follows:

$$ACE_s = \mathbb{E}(Y_{t*}) - \mathbb{E}(Y_t) \tag{2.10}$$

Where $\mathbb{E}(Y_{t*})$ represents the resulting outcome for variable Y at time instant t given that the occurrence of intervention s and $\mathbb{E}(Y_t)$ represents the expected outcome for variable Y_t without the intervention [51]. While concepts such as Granger Causality and Sims Causality assume an observational framework, intervention causality requires counterfactual experiments which are not applicable in many real world applications [51]. Due to this reason, the focus of my work will be based on Granger Causality based causal learning algorithms.

2.7 Types of Causal Learning

Both time series data and non-time series data have been used by different algorithms for causal learning. Generally speaking, there are two types of learning objectives for causality: learning the causal effects and learning the causal relationships [21]. While there is certainly some overlap between the two areas, their definitions are slightly different.

Learning about the causal relationships among variables is often referred to as causal discovery or structural learning. The input for such algorithms is usually two or more time series data, where some of which are causally related while others are not. The goal of the causal discovery algorithm is to explore the causality between each pair of variables. This study usually involves learning the correct causal directions, causal lags as well as the corresponding causal index. An example of this type of causal learning can be found in [52] where researchers applied a machine learning based causal learning algorithm to discover the correct causal lag terms on the atmospheric system model.

Learning about the effect, on the other hand, is often referred to as parameter learning which describes the process of causal inference where some of the variables are known to be causally related and the drive is to study the impact of changing one variable towards the outcome of another variable. When performed incorrectly (among variables that are not causally related, to begin with), one will obtain some irrational conclusion. Thus the correct causal relationships must be correctly identified before performing inference.

The focus of my work is on learning the causal relationships among variables, thus only one technique related to causal inferences, namely Bayesian Network will be briefly discussed in section 2.9. Before the discussion on Bayesian Network, I will discuss some of the fundamental building blocks in the area of time series causal structure learning, the Granger Causality Theorem in section in section 2.8.

2.8 Time Series Causal Learning Algorithms

Ever since the seminal paper published in 1969, Granger alongside other leading researchers has begun to explore different methods to determine the existence of Granger Causality among variables, first in the field of the economy [53], and later have generalized into fields such as neuroscience [54] and climate prediction [55]. Over the years, despite some level of the modification of Granger Causality theorem for each category, the existing time-series causality algorithms can be broadly categorized into five main groups as shown in Fig. 3.1, namely, classical time-series approaches [41], chaos and dynamic systems approaches [56], information theoretic approaches [57], graphical approaches [58], and finally machine learning approaches [59]. Figure 3.1 below listed out some of the key algorithms under each category of causal learning. Readers are referred to the original publication for further details [41, 56–59].

2.8.1 Classical Time Series Approach

The very first type of approach to examine causality falls under the category of the classical time series approach. *Classical time-series approaches* are widely adopted for dealing with time-series causality and are built upon the principle of Granger Causality using a statistical-based model. This category is later defined as structural equation modelling by Judea Pearl [60]. The most basic/fundamental approach to measure GC is done by implementing a linear bi-variate Vector Autoregression model. The VAR model is built upon the concept of the Autoregressive(AR) model, in which a n degree AR can be expressed in equation 2.11:

$$y_t = \alpha_1 y_{t-1} + \cdots + \alpha_n y_{t-n} + C_1 + \epsilon \tag{2.11}$$

where C_1 and ϵ are the constant values and a noise term, respectively. Term y_t is the time-series value of dependent variable at time t, x_{t-i} is the time-series of the independent variable x at time $t - i$, and α terms are corresponding parameter values. The equation can then be generalized into equation 2.12 shown below:

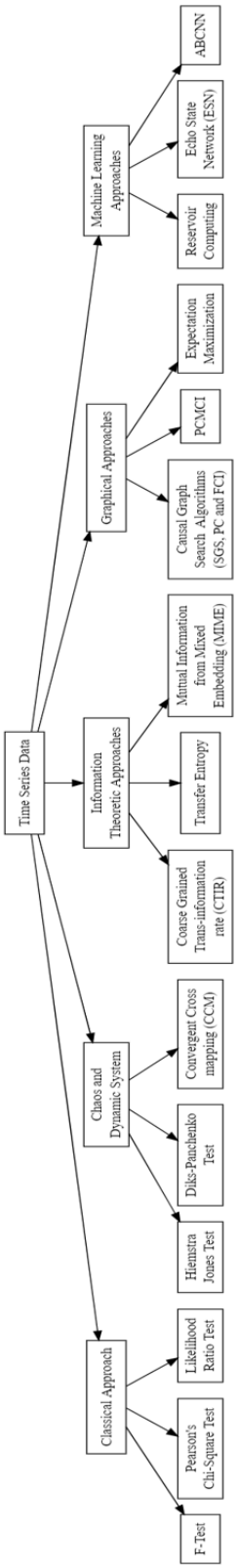


Figure 2.3 Time series causal learning algorithms can be categorized into 5 groups, namely the classical methods, chaos and dynamic system, information theory, graphical approach, and machine learning based algorithms.

$$y_t = \sum_{\tau=1}^{L_1} \alpha_\tau y_{t-\tau} + C_1 + \epsilon(t) \tag{2.12}$$

A VAR model incorporated another variable into the equation, a general pairwise VAR equation is shown in below in equation 2.13:

$$y_t^* = \sum_{\tau=1}^{L_1} \alpha_\tau y_{t-\tau} + \sum_{\tau=1}^{L_2} \beta_\tau x_{t-\tau} + C + \epsilon(t) \tag{2.13}$$

Term y_t^* is the updated prediction for variable y using the new variable x introduced to the equation. $\beta_1, \beta_2, \cdots, \beta_{L_1}$ are the parameter values at each time lag for variable x. If equation 2.13 contains additional information than equation 2.12, then by definition, variable x contains some unique information of variable y that can allow improvement in the predictability towards y at future time instants, hence the reason why variable x Granger causes y. One of the most popular ways to verify the impact is to use the F-test, which can be calculated using the Residual Sum of Squares (RSS) for equation 2.12 and 2.13 along with their lag counts, equation can be expressed in equation 2.14. A higher F-test value indicates a stronger causality between the pair of variables.

$$F = \frac{(\mathrm{RSS_y} - \mathrm{RSS_{y^*}})/(\mathrm{L_2\text{-}L_1})}{(\mathrm{RSS_{y^*}})/(T - L_2)} \tag{2.14}$$

Where term T represent the entire length of the time series and term RSS_y and RSS_{y^*} are denoting the total sum of the prediction error terms for equation 2.12 and 2.13, expressed in equation 2.15 and 2.16 respectively:

$$\mathrm{RSS_y} = \sum_{t=1}^{n} (y_t - f(y_{t-1}, \cdots, y_{t-L_1}))^2 \tag{2.15}$$

$$\mathrm{RSS_{y^*}} = \sum_{t=1}^{T} (y_t - f(y_{t-1}, \cdots, y_{t-L_1}, x_{t-1}, \cdots, x_{t-L_2}))^2 \tag{2.16}$$

In addition to the VAR model, one can also verify Granger Causality using Pearson Correlations [61], Equations are shown in 2.17 where N represents the total number of

sample entries selected whereas term x_i and y_i are the individual sample point picked from the x and y respectively. $\bar{x}$ and $\bar{y}$ represents the mean value.

$$r = \frac{\sum_{t=1}^{n}(x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^{N}(x_t - \bar{x})^2}\sqrt{\sum_{t=1}^{n}(y_t - \bar{y})^2}} \tag{2.17}$$

While these methods are generally intuitive and easy to implement, they struggle to identify complicated non-linear causal relationships despite some variants being proposed to work with nonlinear data. Due to these issues, sometimes in working with complicated nonlinear multivariate datasets, not only are the true causal relationships are missed, but this approach may also lead to a significant number of falsely defined causal relationships, narrowing the usefulness of such method in complex applications [62] [63].

2.8.2 Information Theory Approach

Inspired by the Transfer Entropy (TE) principle from physics literature in 2000, a computer scientist researcher, Schreiber Thomas, originally motivated TE as an alternative to lagged mutual information that takes into consideration of the shared information due to history input signals [64]. The TE measurement from one variable to another is the information-theoretic distance measured between the transition probability that includes the causing variable, and the one that excludes the causing variable [64]. This transfer of entropy between variable x and y is expressed in equation 2.18:

$$TE_{X\to Y} = H\left(y_t \mid y_{t-1:t-L}\right) - H\left(y_t \mid y_{t-1:t-L}, x_{t-1:t-L}\right) \tag{2.18}$$

Where $H(\cdots)$ represents Shannon's entropy at the given condition and $TE_{X\to Y}$ can be viewed as the amount of information transfer from X to Y at t instant. This principle is surprisingly similar to the Granger Causality theorem described previously. Although there is no direct reference made from Schreiber, many researchers in the field sometimes treat the information theory approach as a method that shares a lot of fundamental similarities with

the foundation of Granger Causality, but later has emerged from the discovery of Granger Causality [65].

The performance of the information theory approach, in general, is more accurate when working with non-linear datasets compared to classical time series approaches [66] [67] [68]. It is sometimes regarded as a nonlinear generalization of Granger Causality [69]. However, despite its ability to deal with non-linear data, most of the TE approaches have mostly been applied to bivariate experiments (rather than multi-variable) as it is difficult to determine the information flow when working with higher dimension datasets [63]. Thus, it is usually recommended to perform dimensional reduction when working with TE-based methods [63]. When working with real-world datasets, sometimes it is very difficult to isolate and perform causal experiments on two variables only. Other information theory methods under the information theory approach include techniques, such as Mutual Information on Mixed Embedding (MIME) [70], and Coarse-Grained Trans Information-Rate (CTIR) [71]. Since this is a fairly new research area, there are currently limited open source packages available to the public for validation.

2.8.3 Chaos and Dynamic Approaches

Chaos and Dynamic approaches share many similarities with *Information-Theoretic approaches*; thus, it is often viewed as a complementary approach to the information theory approach [72]. The methods under this category are built upon the idea where oscillations in dynamical systems can be excited by other dynamical systems, necessitate that a phase can be extracted from a time series. Therefore, the signal must "circulate" in phase space to form cycles. This characteristics can be represented using a set of state vectors, shown in equation 2.19:

$$\vec{x}_t = (x_t, x_{t-d}, \cdots, x_{t-(m-1)d}) \tag{2.19}$$

Where m is the embedding dimension and d is the embedding delay. Similar to the Granger Causality principle, if another time series can be used to improve the prediction of the dynamics in the reconstructed phase space, one can conclude a causal relationship among the two variables [73].

The transfer entropy values can instead be measured in a different way to analyze causality. For example, causal discovery is performed in the field of assessment of cardiovascular regulatory sees [74]. Methods that can be classified under these frameworks are used to examine linear and non-linear Granger Causality including the Hiemstra-Jones test [75], and Convergent cross mapping (CCM) [72]. When working with a time series dataset that is predominantly stochastic (such as an industrial process), such a method failed to generalize [63] [76].

2.8.4 Graphical Approaches

Graphical Approaches models causality in a multivariate setting by representing each variable as a node on a graph with Granger Causality represented as directed edges between the nodes. The direction arrows on the graph denote the causal link among the pair of variables. A simple example is shown in figure 2.4 where the researcher proposed the cause-effect relationship between “race” to “age”, "age" to "Death from COVID" and "race" to "Death from COVID" based on the observed data provided by Centers for Disease Control website [1].

Such a format of visualization has later been adopted by other types of causal learning algorithms for a clear and concise schematic representation. Most of the methods under the graphical approach take into consideration of the probabilistic element into causal relationships, where the causal effect is a probability rather than a certainty. The focus of graphical approach algorithms evolves around the application of the Causal Markov Condition which stated that every variable is conditionally independent of its non-descendants (variables that are not the causing variable), given its parents [77]. The existence of conditional independence can be categorized as causality and vice versa. While there are a few

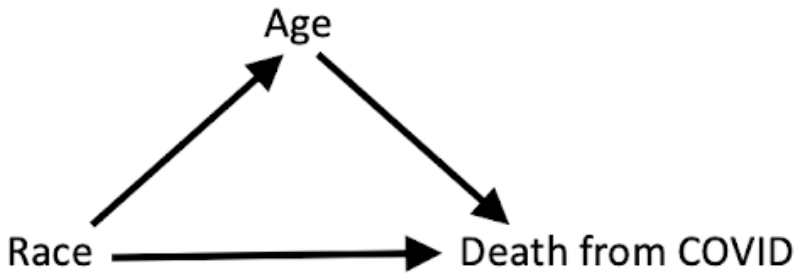


Figure 2.4 Researcher in [1] proposed both age and race are directly causing the number of death from COVID-19 based on data provided.

variations of graphical algorithms, they can be summarized into three main steps. First, the graphical algorithm will assume complete dependency among every pair of variables by connecting all variables on the graph together. Then the conditional independence test will be performed on each pair of variables and causal edges will be eliminated on variable pairs that are independent of each other. Lastly, the direction of cause-effect relationships is determined based on the statistical approach proposed by each algorithm.

Some of the methods under graphical approaches include Spirtes Glymour Scheines Algorithm (SGS) [77], Fast Causal Inference Algorithm (FCI) [77], Peter Clark Momentary Conditional Independence Algorithm (PCMCI) [14] Expectation-Maximization (EM) [16], and Structural Expectation Maximization (SEM) [78]. The graphical method is ideal for applications that involve stochastic elements and the running speed is very fast. However, in general, the causal prediction accuracy of graphical approaches is somewhat inconsistent for the performance may depend heavily on the manual fine-tuning of hyper-parameters as well as how much prior information feed-in(i.e., predefined causal relationships before learning). In addition, although the graphical methods can be performed using no additional prior information, a recent case study done in neuro-imagining reported that the accuracy obtained is very poor (high false-positive and extremely low true positive when working with no prior knowledge). It is often recommended to include as much prior knowledge as possible [17].

2.8.5 Machine Learning

Machine Learning (ML) is a new area that researchers are currently exploring for causal discovery. Unlike the traditional time series methods in which regression models are used and the inference accuracy of these methods depend greatly on whether or not the model can be well fitted to the data, the machine learning approaches use a different variation of neural networks that result in better prediction when working with nonlinear datasets. Unlike other methods mentioned previously that have been around for decades, machine learning algorithms have only gained popularity in the past decade or so. Although the potential for machine learning-based algorithms in time series causal learning is promising, there is limited research done in this area at the moment.

Some of the recent machine learning methods include the Attention-Based Convolutional Neural Network (ABCNN) approach [79], Reservoir computing [52], and Echo State Network approach [3]. There are many advantages favouring ML-based algorithms over the other approaches. For example, they can handle relatively large and complicated data and require less human intervention for tuning; they can be adaptable to a wide range of applications and are backed by a largely motivated support and development community [80]. While there are several drawbacks to ML methods such as the expensive computational cost and the relatively long training time, the advantages outweigh the disadvantages and hence it is adopted to the proposed system for causal discovery. Similar to other applications such as imagine classification or natural language processing, it is foreseeable to expect more and more deep learning based algorithms in the field of time series causality. An ESN based causal learning algorithm is adapted into the system proposed in chapter 4 due to the advantages mentioned above. Unlike a standard Convolutional Neural Network, the ESN contains a sparsely connected hidden layer typically with no more than 10% connectivity in between neurons which results in a fast training time of the model. The time series data is fed directly into the model in a sequential order where the last layer of the model will calculate the estimated value (using weight terms defined previously) and compare it to the expected value. More elaboration of the ESN model will be discussed in section 4.2.

2.9 Causal Inference

After identifying the causal relationship between variables using one of the previously mentioned approaches, causal inference can be performed to answer meaningful questions, such as given the number of people without a mask on during a protest last Friday, estimate the probability of daily COVID count surpass 4,000 tomorrow. This can be achieved using Bayesian Network [81].

In the simplest language a Bayesian Network model is a directed graphical model for representing conditional in-dependencies between multiple variables (can be both random variables or time series variables). A Bayesian Network for a specific application consists of n random variables $X = \{X_1, \cdots, X_n\}$; a set of values $D = \{a_1, \cdots, a_d\}$, where each $X_i \in X$ has an associated finite domain $(X_i) \subseteq D$ of possible values; and a joint probability distribution $P(X_1, \cdots, X_n)$ over the possible assignments to the variables in X. The joint probability distribution for any entry in the joint probability distribution can be expressed in the following expression:

$$P(X_1, \cdots, X_n) = \prod_{i=1}^{n} P(X_i | parents(X_i))) \tag{2.20}$$

Where any of the variables in the system are independent of its non-descendants and non-parents given its parents. The parent-child nodes in the Bayesian Network represent the cause-effect relationships obtained through learning algorithms mentioned in the previous section or through lab experiments. Bayesian Network is first proposed to work with random variables but later Dynamic Bayesian Network or DBN in short, was proposed to work with time series variables [82]. Inside DBN models each variable's log at each time instant is treated as a separate variable. Techniques such as Monte Carol Markov Chain (MCMC) [83] and variable elimination [84] can then be used to perform causal inference on the causal graph one generated. Prior to performing Bayesian inference, however, if the cause-effect relationship provided is invalid, then the inference algorithm may conclude misleading fatal

conclusions. Since the focus of my work is on causal learning, I will not go in-depth about the Bayesian Inference. Please see [85] for more details.

2.10 Conclusion

To summarize, causal inference is an area of research that has great potential and has gained a notable amount of attention recently. However, this is still an ongoing research area with several unsolved issues. One of the unexplored challenges being the incompetence of existing algorithms to perform causal learning on sparsely sampled datasets. When working with real world applications it is inevitable for one to encounter some degree of missing entries. In this research, I present a novel approach (Chapter 3) that applies data filling with Gaussian Process Regression before performing causal discovery. The filling ability for GPR is compared against both the parametric approach (linear regression) as well as the non-parametric approach (K Nearest Neighbour) for performance evaluation. Furthermore, I have adapted a machine learning based algorithm, namely the Echo State Network and performed causal learning on the GPR-filled data (Chapter 4). The causal prediction result is compared against methods under the graphical approach (Structural Expectation Maximization), and the classical approach (Multivariate Granger Causality).

Chapter 3

Missing Data Impact In Causal Discovery Using Linear VAR

Currently, all the causal learning techniques mentioned in chapter 2 require the input to be uniformly dense time series data where the entries must be collected at the same frequency. When working with real industrial processes it is not uncommon to encounter missing values. Some of which were caused by human error while others are by design (e.g. A process that is designed to only collect randomly sampled entries in order to save disc memory). In order to still perform causal learning on sparsely sampled data, I proposed a system that allows us to perform causal learning under situations where missing data is unavoidable.

3.1 Abstract

Over the past years, researchers have proposed various methods to discover causal relationships among time-series data [68] [4] [86] as well as algorithms to fill in missing entries in time-series data [87] [88]. However little to no work has been done in combining the two strategies for the purpose of learning causal relationships using unevenly sampled multivariate time-series data. In this chapter, I examine how the causal parameters learned from

unevenly sampled data (with missing entries) deviate from the parameters learned using the evenly sampled data (without missing entries). However, obtaining the causal relationship from a given time series requires evenly sampled data, which suggests filling the missing data values before obtaining the causal parameters. Therefore, the proposed method is based on applying a Gaussian Process Regression (GPR) model for missing data recovery, followed by several pairwise Granger causality equations in Vector Autoregressive form to fit the recovered data and obtain the causal parameters. Experimental results show that the causal parameters generated by using GPR data filling offer much lower RMSE when compared against benchmark filling techniques such as parametric linear filling, filling with Nearest Neighbour, and random filling. The result I obtained is suggesting that GPR data filling can better preserve the causal relationships than all other benchmark techniques listed. Thus this method should be considered when dealing with unevenly sampled time-series causality learning.

In this chapter, I performed missing data recovery using the Gaussian Process Regression technique for filling missing values in time-series data to obtain pairwise Granger Causality parameters. In addition, I compared the quality of filling the missed data by comparing the Granger causality parameters estimated using original time-series data against its GPR filled version where the RMSE values under each filling percentage are calculated. The same procedure is repeated with benchmark filling techniques such as filling with Nearest Neighbour, parametric linear filling, and random number filling where their RMSE values are compared for evaluation of the performance.

3.2 Proposed Method

Figure 3.1 illustrates the proposed pipeline in order to study the performance of causal discovery with irregularly sampled data [89]. Given multivariate time-series data, the proposed method randomly drops X% of the original data entry and the missing values are then filled using either (a) Gaussian Process Regression (Section 3.2.1), (b) Filling with

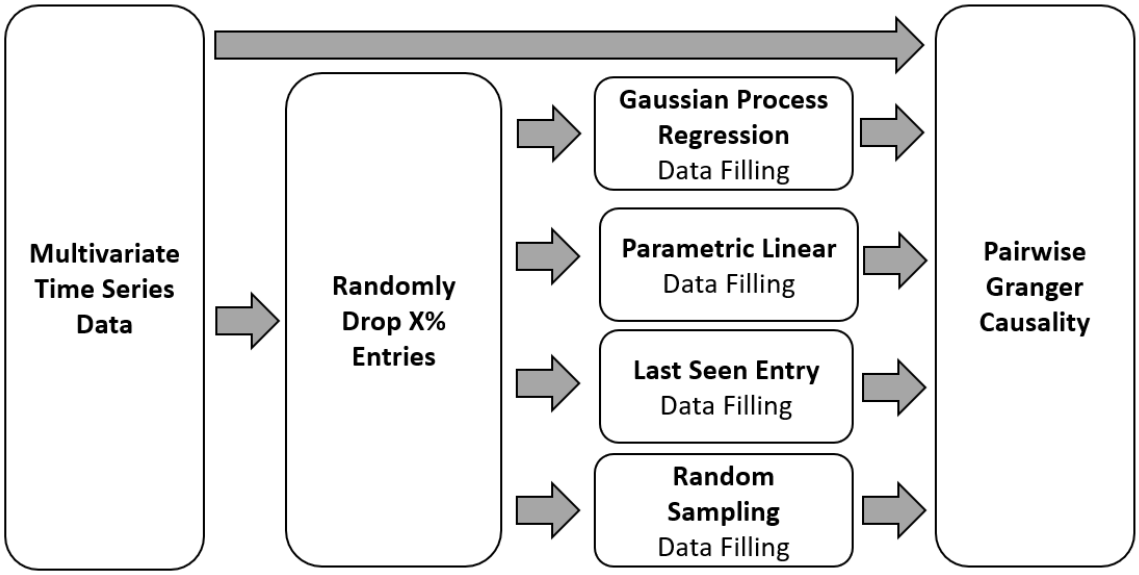


Figure 3.1 The proposed pipeline for Granger causality from irregularly sampled data. The original time series data is first contaminated by randomly dropping X% of its original entries. One of the filling techniques is then used to recover the dataset before performing pairwise causal learning. The causal parameter learned (using the filling technique) are then compared to the parameters learned using the original data and RMSE is calculated for performance comparison.

Nearest Neighbour (c) Parametric Linear Filling or (d) Random Filling approach. Next, the two recovered datasets are then used to obtain the parameters of the pairwise Granger Causality with Vector Autoregression functions (Section 2.5). Finally, the root means square error (RMSE) for each filling technique is calculated with respect to the causal parameters obtained from the original dataset.

3.2.1 Gaussian Process Regression for Data Filling

Although the GP requires an entire training set to perform prediction and lose efficiency with higher dimensions [90], it offers probabilistic predictions and allows the incorporation of different kernels which leads to flexibility in implementation. In [38], it was stated that the GP process can be interpreted with two views: weight-space view and function-space view. A quick discussion regarding GP's hyper-parameters as well as GP sampling function

is also included. However, for more details about GPR, the reader is referred to [38].

Weight-Space View: The equation from the Bayesian analysis of the standard linear regression model can be written as [38]:

$$f(x_t) = {x_t}^T w \tag{3.1}$$

$$y = f(x_t) + \epsilon_i \tag{3.2}$$

where w is the weight vector, ϵ_i is the noise term and it follows a normal distribution in such $\epsilon_t \sim \mathcal{N}(0, {\sigma_\epsilon}^2)$ with 0 mean and ${\sigma_\epsilon}^2$ as the variance from the Bayes' rule. Term x_t represents the new input vector at t instant. The posterior distribution can be obtained as [91]:

$$p(w|y_t, X_t) \; \alpha \; p(y_t|X_t, w)p(w) \tag{3.3}$$

$$p(w|y_t, X) = \mathcal{N}({\sigma_\epsilon}^{-2} A_t^{-1} X y_t, A_t^{-1}) \tag{3.4}$$

where $A_t = \sum^{-1} + {\sigma_\epsilon}^{-2} X X^T$, $\sum$ is the covariance matrix and I want to predict y_t for a new input point x_t with the information obtained prior to instant t. The term X is the matrix that contains the aggregated predicted y_t and the real values for each time instants in the time series [92].

The form shown above is often referred to as the weight space view of regression [91]. In order to predict the y^* at new point x^*, I can average over all the possible parameter values that are provided by the function f, predicting $f(x^*) = y^* + \epsilon_*$. Without going into the actual derivation, the predictive distribution with respect to the Gaussian posterior can be written as [91]:

$$p(f(x^*)|x^*, X_t, y_t) = \int p(f(x^*)|x^*, w)p(w|X, y)dw \tag{3.5}$$

After performing integration equation (3.5) can be expressed as [91]:

$$p(f(x^*)|x^*, X, y_t) = \mathcal{N}({\sigma_\epsilon}^{-2} x^{*T} A^{-1} X, y, X^{*T} A^{-1} x^*) \tag{3.6}$$

Weights are first generated from this posterior distribution and the final predictions are generated using the weight generated previously. The term can be generalized from 1-dimensional spacing to higher-dimensional space [38]. The model now becomes :

$$f(x) = \phi(x)^T w \tag{3.7}$$

where $\phi(x) = (1, x, x^2, x^3, \cdots, x^n)$. The predictive distribution then becomes [38]:

$$p(f(x^*)|x^*, X, y) \sim \mathcal{N}({\sigma_n}^{-2} \phi(x^*)^T A^{-1} \phi_y, \phi(x^*)^T A^{-1} \phi(x^*)) \tag{3.8}$$

Function-Space View: Another way to understand the GP algorithm is to focus directly on its distribution over functions [38]. As stated previously, the GP algorithm defines a distribution over several functions: if I pick any two (or more) points inside a function, our observations at the selected points follow a joint multivariate Gaussian distribution [93]. In [38], the Gaussian process is defined as a collection of random variables, any finite number of which have a joint Gaussian distribution. Similar to the assumption made in linear regression, I can write the Gaussian Process regression equation as:

$$y = f(x) + \epsilon \tag{3.9}$$

where the noise term $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$, reflects the randomness or uncertainty of our observation. The selection process for the noise term σ_ϵ^2 are fine-tuned to a value in order to best represent the data in hand [38]. Based on the definition provided, I can specify a Gaussian Process by its mean function and covariance matrix function, thus Gaussian process can be expressed as follow :

$$f(x) \sim GP(m(x), k(x, x')) \tag{3.10}$$

where $m(x)$ is the mean function and the $k(x, x')$ is the covariance function (also known as the kernel function) for the randomly selected two points x and x'. Equations 3.10 can be

expressed as following :

$$m(x) = \mathbb{E}[f(x)] \tag{3.11}$$

$$k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x')] \tag{3.12}$$

The covariance function, k, is more commonly referred to as the kernel of the GP algorithm [94]. There are many kernel functions available and the choice of which kernel function to use is based on the prior knowledge of the data (e.g. information such as will variable b be affected when variable a is larger, if so to what degree etc.). The choice of the kernel function is also based on factors such as the smoothness and the cycle patterns of the observed values. Although there exists some software that allows automatic selection of the kernel function, it is often required to manually test out different kernel functions on the training data in hand [95].

Hyper-Parameter-Based Kernels: The hyper-parameters refer to the pre-defined constant terms inside the kernel functions. Since there are many possible kernel functions, there will be different hyper-parameters for each kernel function. Readers are encouraged to explore more kernel functions if interested. A very popular kernel function is the radial basis function kernel or RBF kernel in short [96]. The kernel function can be expressed as the following:

$$k(x, x') = \sigma_f^2 \exp(-\frac{\|x - x'\|^2}{2\lambda^2}) \tag{3.13}$$

where $\|\cdot\|$ denotes the euclidean distance. Term x and x' are the two points passed into the kernel function. There are two hyper-parameters inside the radial basis kernel function: λ and σ_f^2. The term λ refers to the length scale while the term σ_f^2 is the data variance of the kernel function. These two hyper-parameters can be increased or decreased to better fit the working dataset. Usually, this is an iterative process for users should test out values before finding the most optimal hyper-parameter values for the working dataset. The GP can then be used to draw prior functions once the mean function and the kernel functions are selected.

Sampling From GP: Let X^* be a matrix that contains all the new input points where $x_i^*, i = 1, 2, \cdots, n$. The kernel function in (3.13) is constructed for all the pairs between the input points. The expression can be displayed in a matrix form as follow [91]:

$$K(X^*, X^*) = \begin{pmatrix} k(x_1^*, x_1^*) & k(x_1^*, x_2^*) & \cdots & k(x_1^*, x_n^*) \\ k(x_2^*, x_1^*) & k(x_2^*, x_2^*) & \cdots & k(x_2^*, x_n^*) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n^*, x_1^*) & k(x_n^*, x_2^*) & \cdots & k(x_n^*, x_n^*) \end{pmatrix} \tag{3.14}$$

Where $k(x_1^*, x_2^*)$ is the kernel function constructed using point x_1^* and x_2^* selected from X^* which contains all input values. To simplify the equation obtained from (3.10), the mean function $m(x)$ is set to 0 and (3.14) is substituted. The following term for a normal distribution is obtained as [91]:

$$f(x^*) \sim \mathcal{N}(0, K(X^*, X^*)) \tag{3.15}$$

Where the notation $f(x^*)$ represents the samples from the defined function. Our observed values defined in previous section is $D_t = \{(x_i, y_i) | i = 1, 2, \cdots, n\}$ and I would like to draw new entry X^*'s predictions from function $f(x^*)$ using the posterior distribution. Let x_t (value at instant t) be the value drawn from X^*. Then the matrix form of the distribution can be expressed as follows [91]:

$$\begin{bmatrix} y_t \\ f(x^*) \end{bmatrix} = \mathcal{N}\Big(0, \begin{bmatrix} K(x_t, x_t) + \sigma_\epsilon^2 I & K(x_t, x^*) \\ K(x^*, x_t) & K(x^*, x^*) \end{bmatrix}\Big) \tag{3.16}$$

where σ is the noise level term and I is the identity matrix. By implementing the Gaussian Identities Theorem for conditional distribution $p(f(x^*)|X_t, y_t, X^*)$ provided in [38], I can rewrite equations (3.16) and (3.8) as the following expression:

$$p(f(x^*)|X^*, x_t, y_t) \sim \mathcal{N}\big(m_t(x), k_t(x, x')\big) \tag{3.17}$$

where the mean function and the kernel function in equations (3.11) and (3.12). The sample functions $f(x^*)$ can now be sampled using (3.18) and (3.19) stated below. [91, 97]:

$$m_t(x) = K(X^*, x_t)[K(x_t, x_t) + \sigma_\epsilon^2 I]^{-1} y_t \tag{3.18}$$

$$k_t(x, x') = K(X^*, X^*) - K(X^*, x_t)K(X^*, X^*)^{-1}K(x_t, X^*) \tag{3.19}$$

3.3 Experimental Results

Data Description: In order to validate the performance of recovery of the proposed method, I use a public dataset called the Prognostics and Health Management (PHM08) system dataset [2, 98]. The PHM08 [2] is a turbofan engine degradation simulation dataset created by NASA using the Commercial Modular Aero Propulsion System Simulation Tool (C-MAPSS). The engine is simulated to failure point and the average sensor/operational measurements are recorded for each cycle. Engines inside the training set lasted anywhere from 130 cycles to 362 cycles before the failure point.

Although the ground truth data for causality is not available for this PHM08 dataset, it is safe to make the assumption that causality relationships did exist between these sensor measurements. In a real-world scenario, it is often rare to spot the breakdown of a complicated system caused by all intermediate components fails in one instant. It is more common to have a breakdown of one component (sensor) which leads to failure of surrounding components and ultimately leads to the malfunctioning of the system. Figure 4.1 is an illustration of a simplified jet engine diagram. The first 11 engines in the first training set of the PHM08 [2] dataset are selected for this experiment. There are 9 constant sensor readings (with little to no fluctuation) out of the given 24 time-series, thus are neglected for this experiment and the remainder 15 sensor data are used.

Discussion: To evaluate the proposed scheme, as I indicated in Figure 3.1, the selected data is dropped by 10 20%,$\cdots$, or 80% of its' original entries to simulate an unevenly sparsely

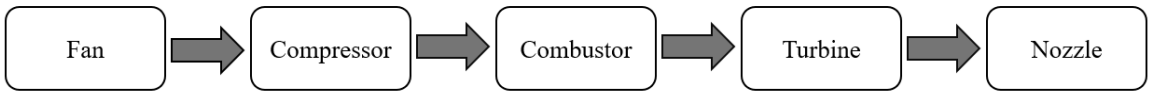


Figure 3.2 Simplified diagram for engine components in the PHM08 dataset [2].

sampled time-series data. Gaussian Process Regression[1] is then used to recover those missing values and finally the recovered multivariate time-series data is feed into the VARs model[2] to calculate the causality parameters. Those parameter values are then compared against the parameter values obtained from the original dataset and the average RMSE, for all the considered engines, values are recorded under each missing value percentage. The same test is repeated using the other benchmark filling methods.

Figure 4.2 shows the comparison between the GP Regression prediction, parametric linear filling, Filling with the last seen approach, random filling approach, and the ground truth values for engine 1 sensor 7 with 50% missing data. It is clear that GPR is able to follow the changes in the time-series data better than other methods. In addition, GPR filling is able to provide a smoothing effect to reduce the noise level. The RMSE values in predicting the causal parameters for the proposed method and other benchmark techniques under different filling percentages are also summarized and plotted in Figure 3.4. As shown in this figure, the GPR-filled data can better preserve the pairwise causal relationships in the original data when compared against other benchmark approaches.

[1]GPR function in pymc3 is used https://github.com/pymc-devs/pymc3

[2]VARs package in R is used https://cran.r-project.org/web/packages/vars/vars.pdf

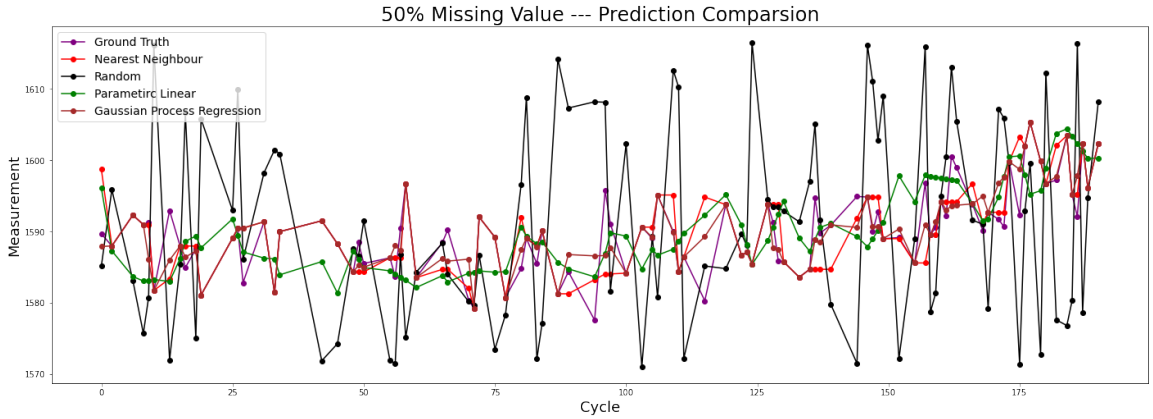


Figure 3.3 Comparison between GP Regression prediction, filling with Nearest Neighbour, parametric linear filling, random filling and ground truth value for engine 1 sensor 7 that contains 50% missing values, where the PHM08 dataset is used.

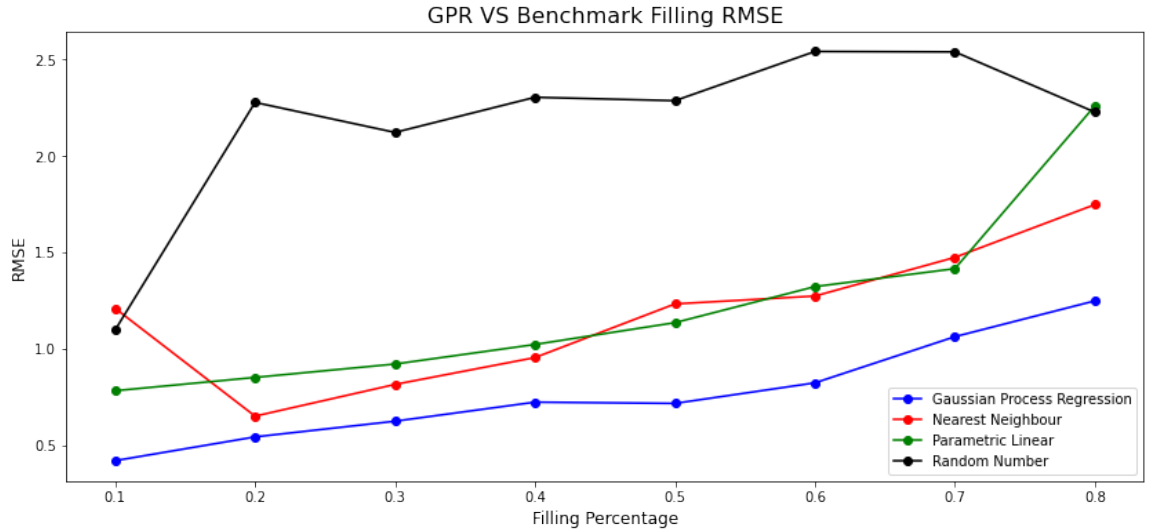


Figure 3.4 RMSE Comparison Between GPR Filling and all other benchmark filling techniques w.r.t. Pairwise Granger Causality Parameters learned. The result showed that the performance for GPR filling obtains a much smaller RMSE when compared with all other techniques.

3.4 Summary

In this chapter [99], I have studied the ability of Gaussian Process Regression to recover missing time-series data values for the purpose of determining the pairwise Granger causality. The proposed method has been tested by using the PHM08 dataset subjected to different missing value percentages that can affect the causal parameter values obtained from pairwise Granger causality. The results show that the Gaussian Process recovered data is better preserved for the pairwise Granger causality relations when compared to those obtained by other benchmark filling approaches. In the next chapter, I will apply the same system to a machine learning based causal learning algorithm to study the ability to recover data and pick up causal relationships among variables.

Chapter 4

Machine Learning Based Causal Algorithm on Sparsely Sampled Data

4.1 Architecture

Due to all the advantages associated with the machine learning based causal learning algorithm mentioned in chapter 2 I've decided to test out the learning ability on sparsely sampled data using an Echo State Network (ESN) algorithm. The ESN causal learning algorithm is selected due to its fast computation time and its ability to handle nonlinear datasets [100]. The proposed approach is summarized in Fig. 4.1 where T is a time-series input with some missing data, T^* represents the multivariate time-series that was data-filled by GPR, and $GC(x_i, x_j)$ is the causal relation between x_i and x_j. The input is a $N \times M$ multivariate time-series where N is the total number of entries for each variable (feature) and M is the number of variables. The GPR filled time-series is then processed through the GC-ESN estimator (dynamic reservoir), to generate an $M \times M$ causality matrix, which can then be displayed as a heat map of causal relations amongst variables. The performance is evaluated using the MCC score and is then compared to benchmark algorithms such as Structural Expectation Maximization (SEM), Subsampled Linear Auto-Regression Absolute Coefficients (SLARAC), and Multivariate Granger Causality (MVGC).

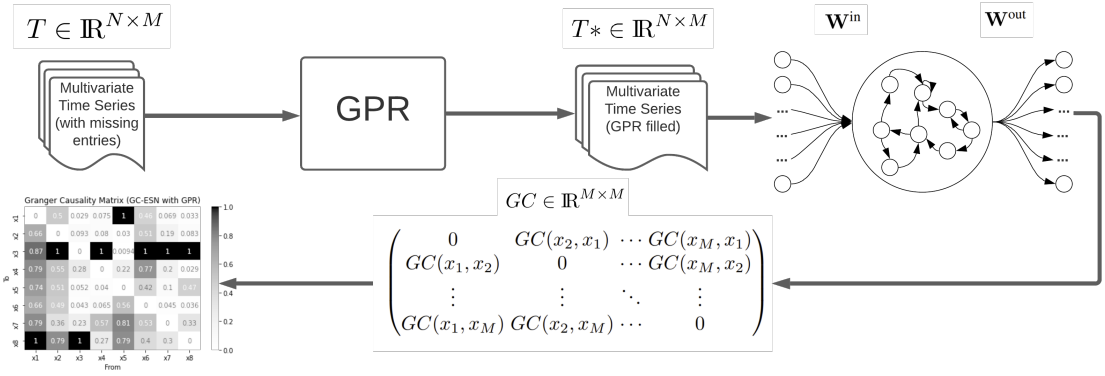


Figure 4.1 Workflow of the proposed solution; N time-series each with M of entries that contains x% missing entries is first filled using GPR before feeding into a GC-ESN learning algorithm to determine causal relations among variables. The output of the algorithm is a N by N causal matrix that outlines the predicted causal index among each pair of variables. The index is between 0 and 1 with 0 being completely causally independent while 1 indicates complete causal-effect connection among the pair of variables.

4.2 Granger Causality-Based Echo State Network

In this work, I opt to use a special type of Recurrent Neural Networks (RNNs) known as the Echo State Network (ESN) [3] for causal discovery. Compared to traditional RNNs, ESNs train much faster as the weights of an ESN are randomly initialized and fixed during both training and inference. The ESN's output layer acts as a linear regressor thus providing much more flexibility to the ESN than the general RNN architecture. However, due to the non-linear nature of each unit's activation function, the model is able to capture non-linear causality with a much faster speed than the traditional RNN. The basic structure of an ESN is shown in Fig. 4.2, where the inputs are propagated through the input layer and into the reservoir (the Internal Units). The Internal Units are randomly connected and their weights are fixed after kernel initialization [3]. Let $x(t)$ be the n^{th} internal unit, then during training, an update is computed according to equation 4.1:

$$\mathbf{x}(t+1) = \mathbf{f}^{\text{out}}(\mathbf{W}^{\text{in}}\mathbf{u}(t+1) + \mathbf{W}\mathbf{x}(t) + \mathbf{W}^{\text{out}}\mathbf{y}(t)), \tag{4.1}$$

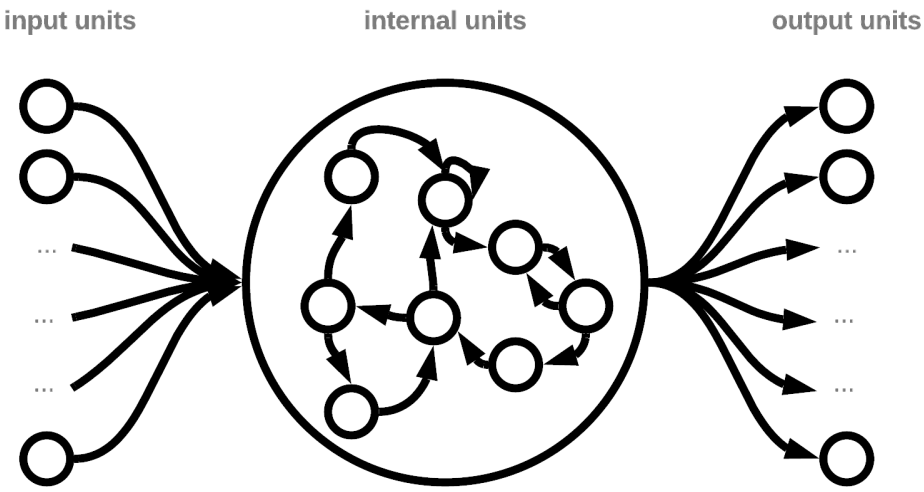


Figure 4.2 Echo State Network Neurons Structure [3]

where $\mathbf{u}(t)$, $\mathbf{x}(t)$, and $\mathbf{y}(t)$ represent the input, internal and output units at a time step t respectively. $\mathbf{W}^{\text{in}}$, $\mathbf{W}$ and $\mathbf{W}^{\text{out}}$ are their corresponding weight matrices. During inference, the weights $\mathbf{W}^{\text{in}}$ are fixed and the output is computed using equation 4.2:

$$\mathbf{y}(t+1) = \mathbf{f}^{\text{out}}\left(\mathbf{W}^{\text{out}}\ [\mathbf{u}(t+1);\quad \mathbf{x}(t+1);\quad \mathbf{y}(t)]\right), \tag{4.2}$$

Where $\mathbf{f}^{\text{out}}$ is the output function and $\mathbf{W}^{\text{out}} \in \mathbb{R}^{1\times(1+\text{M}+\text{L})}$ given a reservoir size M and input size L.

In this work, I followed the approach presented in [101], that is I model the ESN's reservoir units with an often used short-term memory term α and a variation term of the internal unit function [101, 102] $\tilde{\mathbf{x}}$ as following:

$$\mathbf{x}(t) = (1-\alpha)\mathbf{x}(t-1) + \alpha\tilde{\mathbf{x}}(t). \tag{4.3}$$

To model non-linear data, $\tilde{\mathbf{x}}$ is used since it is a linear combination of the input units $\mathbf{W}^{\text{in}}u(t)$ and the previous reservoir states $\mathbf{W}\mathbf{x}(t-1)$, and is computed using:

$$\tilde{\mathbf{x}}(t) = f(\mathbf{W}^{\text{in}}[1;\mathbf{u}(t)] + \mathbf{W}\mathbf{x}(t-1)). \tag{4.4}$$

Term $\alpha\epsilon[0,1]$ controls the level of memorization. $\alpha \to 0$ indicates a pure memorization characterization, recalling all $t-2, t-3, \cdots, 0$ input signals while $\alpha \to 1$ will only consider

the input on the $t-1$ instants. Such variable can have a large impact on the system's performance [3]. It is often required to run the model with different α values to find the best fit for the data in hand [103]. I then add an extended state $\mathbf{z}(t)$ which contains $\mathbf{x}$ and $\mathbf{u}$ [101]:

$$\mathbf{z}(t) = [1; \quad \mathbf{x}(t); \quad \mathbf{u}(t)] \tag{4.5}$$

Given a reservoir size of M, I am then able to find the expected value of $u_i(t+1)|\mathcal{I}(\mathbf{u}(t))$:

$$E[u_i(t+1)|\,\mathcal{I}(\mathbf{u}(t))] = \mathbf{W}^{\text{out}}\mathbf{z}(t) \tag{4.6}$$

Eventually, I can find the ES-GC strength of j Granger causes i using the i-th squared residuals, $\varepsilon_i(t)$:

$$\varepsilon_i(t) = u_i(t+1) - \mathbf{W}^{\text{out}}\mathbf{z}(t) \tag{4.7}$$

The ES-GC strength of feature j causes i, $\text{GC}_{j\to i}$ is described as follow:

$$\text{GC}_{j\to i} = \log(\varepsilon^{-j}/\varepsilon_i) \tag{4.8}$$

4.3 Experiments and Evaluation

4.3.1 Dataset Description

I evaluated the proposed system on the Tennessee Eastman's Process Dataset (or TE for short) [104]. TE is a dataset that simulates industrial chemical processes and has been widely applied in the study of fault diagnosis and root cause analysis [81] [105]. The process flow consists of 5 major physical components: the reactor, condenser, vapor-liquid separator, compressor, and the product stripper. There are several sensor measurements available (such as flow rate, temperature, pressure, and feed rate) for each component. Overall, the TE dataset contains 500 operation cycles' measurements from 52 sensor readings. Each operation cycle consists of 500 entries being recorded every 3 minutes for a total duration of 1500 minutes. For further information about the TE dataset, the interested reader is referred to [106].

I selected 8 sensor readings from the first 200 operation cycle's entry inside the training fault-free file for our experiment. The description for the 8 selected variables is shown in Table 4.1. In the experimental procedure, I aim to recover these causal relationships and compare them against their ground truth values.

Table 4.1 Selected Sensor's Descriptions

Variable ID	Header in data	Description	Units
1	xmeas_5	Recycle Flow	km^3/h
2	xmeas_6	Reactor feed rate	km^3/h
3	xmeas_7	Reactor Pressure	kPa
4	xmeas_8	Reactor Level	%
5	xmeas_9	Reactor Temperature	°C
6	xmeas_12	Separator Level	%
7	xmeas_20	Compress Work	KW
8	xmeas_21	Reactor cooling water Outlet Temperature	°C

4.3.2 Experimental Setup

The selected TE data is first contaminated by removing 10% of its original entries. This provides ground truth data to validate the success of our proposed data filling process. Figure 4.3 shows the original *vs.* filled data for sensor 5 (recycling flow rate measurement). The contaminated data is then processed through the proposed GPR method to fill out the missing entire, and subsequently fed into the Echo State Network to discover causal relationships among the variables.

The process is repeated on three other off-the-shelf causal discovery algorithms, namely:

- Structural Expectation Maximization (SEM) [78]: a graphical approach that is capable of learning causal relationships from sparsely sampled time series data; as such, the GPR process is not used for this algorithm, and the missing data are directly fed

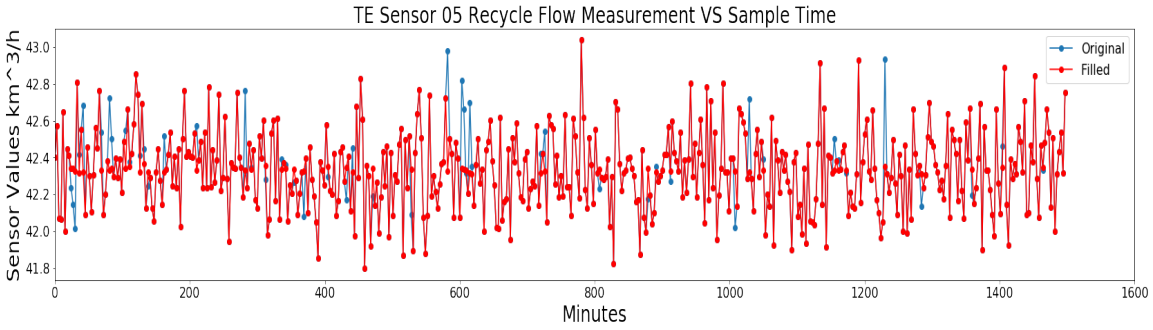


Figure 4.3 Original Data VS 10% GPR Filled Data

into the SEM algorithm for causal learning. I report on the results of *bnstruct* [107], an R implementation of SEM.

- Subsampled Linear Auto-Regression Absolute Coefficients algorithm (SLARAC): a classical approach for causal discovery written in python, and can be found at [108].
- Multivariate Granger Causality (MVGC): a classical approach is written in Matlab and can be found at [109].

The GPR filling is applied to both classical approaches before performing causal discovery. To validate the impact of the data filling process, I also included our results obtained from then ESN with the original data (no missing values). The experimental comparison set up for the various systems is summarized in Fig. 4.4.

4.3.3 Evaluation Metrics

I validated the performance of the various systems using the Matthews Correlation Coefficient (MCC) score. mentioned previously in section 2.3. In addition to the MCC index comparison, I will also report on the ROC (Receiver Operating Characteristic) curves for the various causal discovery systems. The ROC curves offer valuable insights into the specificity and sensitivity of each model at different cutoff thresholds in the causal matrices. I also report on the AUC (Area Under Curve) score for all ROC curves. Note that due to the binary matrix output nature of the SEM algorithm, I was unable to plot its ROC curve.

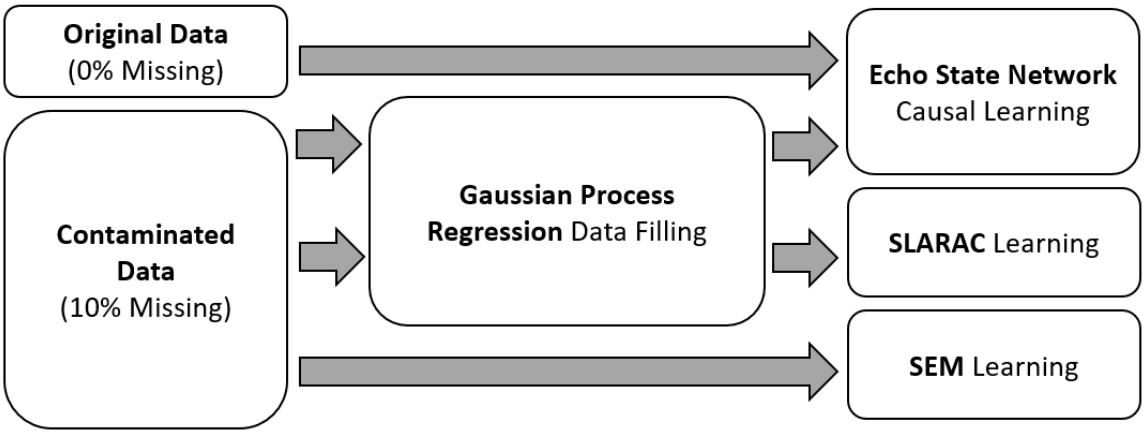


Figure 4.4 The Proposed Experiment Setup for Performance Comparison.

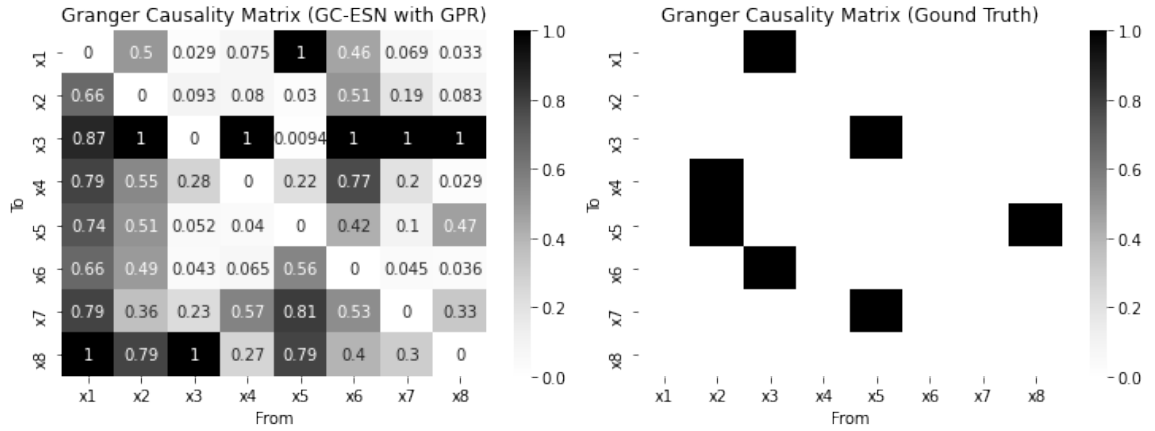
4.3.4 Results and Discussion

Figure 4.5 shows the causal relations recovered from our proposed GC-ESN system (a) side-by-side with its corresponding ground-truth causality matrix (b). On the other hand, table 4.2 summarizes the MCC index of the proposed system (GPR + ESN 10% Missing) against the other algorithms.

Table 4.2 MCC score results shows that the proposed system (GPR filling and ESN causal learning) is able to obtain an MCC score of 0.31 which is better than the result obtained from three benchmark algorithms by a margin.

	GC-ESN 0% Missing	GPR + GC-ESN 10% Missing	GPR + SLARAC 10% Missing	MVGC 10% Missing	SEM 10% Missing
TP	5	5	7	7	0
FP	13	15	47	48	13
TN	44	42	8	9	44
FN	2	2	0	0	7
MCC	0.34	0.31	0.15	0.14	-0.18

As shown in Table. 4.2, our system was capable of recovering five of the seven causal relations by applying thresholds that yield the highest F1 scores, and despite the 10 %



(a) Causal Matrix of GC-ESN with 10% GPR

(b) The Ground Truth Causal Matrix

Figure 4.5 Causal Matrix Comparison between GC-ESN with 10% Missing Data Filling using GPR and the Ground Truth.

missing entries, our proposed system still achieved an MCC index (Table 4.2) of 0.31 which is slightly lower (by 0.03) than the MCC index obtained with the original data (using ESN causal learning). This indicates that the GPR filling process I adapted inside our system is very effective and did restore a satisfactory amount of information comparable to that of the original data. Furthermore, compared to the remaining algorithms on missing data, the proposed GC-ESN estimator achieved the highest MCC score. This indicates that our proposed system is capable of offering more precise and reliable causal link suggestions than the other systems.

The results of the ROC curves (shown in Fig. 5.1) further solidify the claims as the proposed GC-ESN system reports a significantly higher AUC score than other benchmark algorithms (other than ESN on the original data). Moreover, by varying the threshold values, GC-ESN has the highest true positive rate, proving that among the three estimators, GC-ESN exhibits the best performance.

Figure 4.7 shows the results from Figure 4.5 using a causal diagram. Different colours represent different causal features and different thicknesses of connections represent different weights. It is consistent with the comparison results from Figure 5.1.

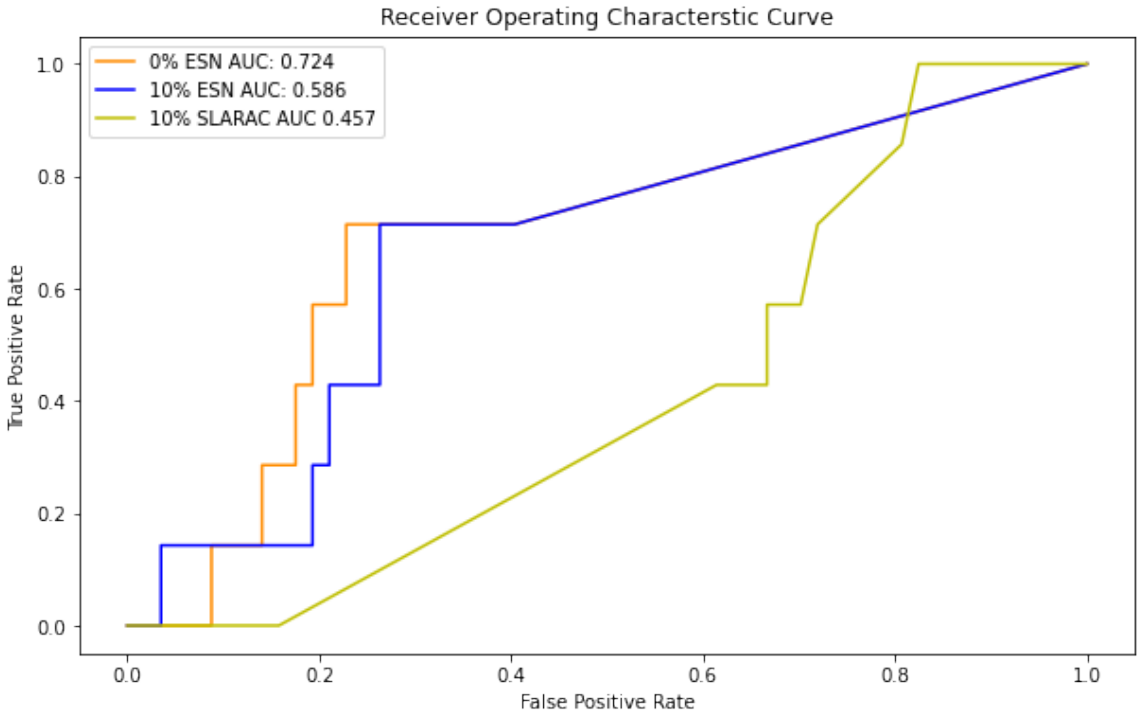
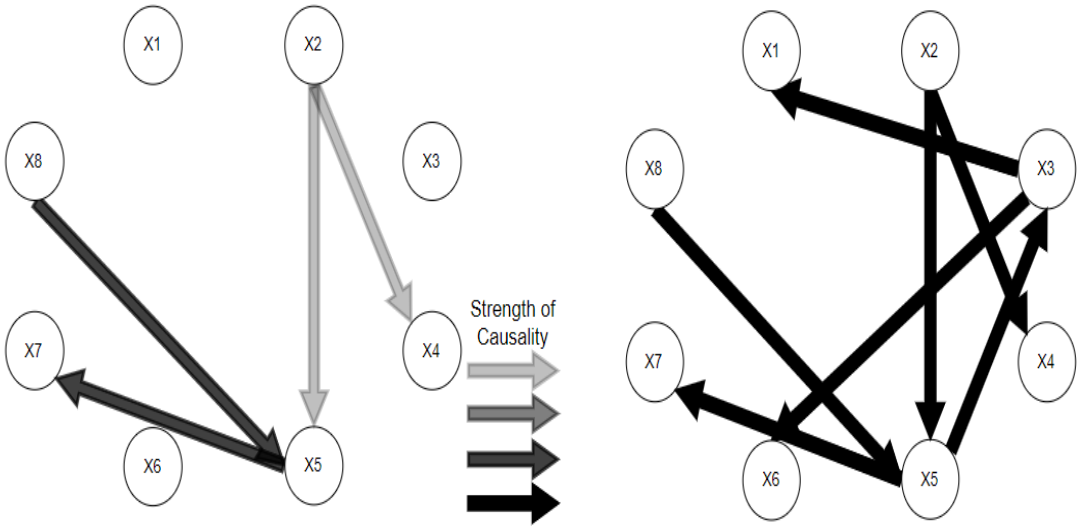


Figure 4.6 ROC Curve for Different Missing Percentage GC-ESN and other estimators. The proposed system is able to outperform the rest of the estimators by a noticeable margin in their AUC scores.



(a) Causal Diagram of GC-ESN with 10% GPR

(b) The Ground Truth Causal Diagram

Figure 4.7 Results Comparison between GC-ESN (true positive) with 10% missing data by applying a 0.5 threshold value filling using GPR and the Ground Truth, I am able to obtain 4 true positive causal relations out of 7 which is better than the result obtained from other methods.

4.4 Conclusion

In this chapter I have proposed a system capable of performing causal discovery for sparsely sampled multivariate time series data. The system consists of two parts: (1) Data filling with Gaussian Process Regression, and (2) causal learning with an Echo State Network. The proposed system is evaluated on the Tennessee Eastman (TE) process dataset with 10 percent missing entries. In order for us to evaluate the performance, the proposed system is compared and shown to outperform several other methods including Structural Expectation Maximization (SEM), Subsampled Linear Auto-Regression Absolute Coefficients (SLARAC), and Multivariate Granger Causality (MVGC). I also perform an ablation study to evaluate the effectiveness of the GPR in recovering causal information by comparing its results to that of causal discovery using the original (uncontaminated) data, and found that the proposed data filling process is capable of recovering causal relationships reliably and performed only marginally worse had the full original data was used. This work shows promise in recovering causal relationships from imperfect data better than current SOTA (State Of The Art) methods. The obtained results show great potential in applying the proposed system in more complicated real-world scenarios as it outperforms all other methods by a comfortable margin in both AUC scores and MCC indices. That being said, the proposed system still falls short in comparison to a human subject expert in identifying causal relationships.

Chapter 5

Discussion and Conclusion

Learning the cause-effect relationships among variables is a fundamental task for researchers in various disciplines of science. Many believed that causal inference may be the next breakthrough for the creation of a human-like artificial intelligence algorithm. However, due to the unsolved issues mentioned in chapter 2, it is still an ongoing research area where causal elements have limited participation in real-world applications to date. When working with real-world industry datasets it is very common for us to encounter missing entries. While there are several types of causal learning algorithms, such as classical time series algorithms, information theory algorithms, dynamics system algorithms, graphical approach methods, and machine learning algorithms, none can be used to work with sparsely sampled data. To address this deficiency, I have proposed to perform data filling with a non-parametric approach called Gaussian Process Regression (GPR) prior to causal learning. When compared with other benchmark techniques such as parametric linear filling and filling with Nearest Neighbour, the GPR filling method achieved much lower RMSE values than all other benchmark filling techniques. When GPR filling is used prior to Echo State Network causal learning at a 10% filling rate, we achieved an MCC score of 0.31 which is significantly higher than the result obtained from other benchmark learning algorithms such as Structural Expectation Maximization (SEM), Subsampled Linear Auto-Regression Absolute Coefficients (SLARAC) and Multivariate Granger Causality (MVGC). The result

is showing great potential for the system which I proposed.

In addition to the learning with sparsely sampled data concern, another unsolved challenge is the overfitting concerns. One of the most widely accepted criticisms for Granger Causality stated that Granger Causality does not necessarily imply true causality. Granger Causality is a statistical-based concept that says that if the past entry for one variable can be fitted into an equation to better predict another variable's future outcome, then Granger Causality exists between the two. Based on the definition, it is possible for researchers to conclude the existence of Granger Causality among two variables that are completely unassociated. Due to this drawback, prior knowledge from subject experts is required to categorize variables into groups where causal learning is only applied to between variables that belong to the same group. Although prior knowledge does not translate directly into fully defined causal relationships, the knowledge contains partial information towards the data and should be incorporated to improve the overall usefulness of the prediction.

While there are many causal learning packages out there [79] [110] [108], there isn't a single package that can be used as a direct replacement for human expert's knowledge. One of the deficiencies with exciting algorithms lies in the lack of ability for algorithms to perform common sense rejection. While such concept of Granger Causality may be mathematically valid, it neglects the fact that the two variables may be completely not causally associated, to begin with. Causal learning should not be blindly applied among variables. Instead, causal learning should only be studied among variables that maybe causally related and this decision should be made using the prior knowledge from the subject expert.

To elaborate on the previous claim, here is an example, Figure 5.1 shown below contains the number of annual Ph.D. computer science graduates in the US alongside the annual arcade revenue in the US. By considering the prior understanding towards these two variables I know that these two variables are not directly causally related regardless of the conclusion received from the learning package. If such prior information is neglected and the two time series data were blindly fed into a causal discovery algorithm, such as the MVGC algorithm,

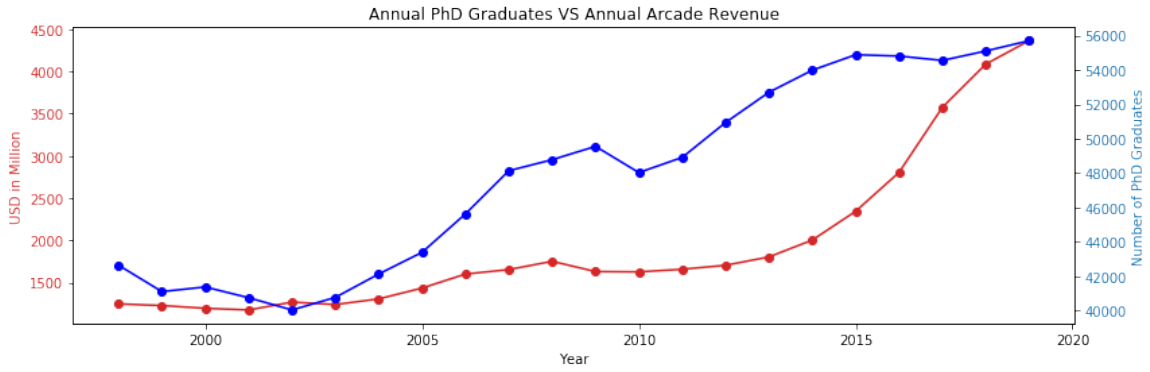


Figure 5.1 Arcade Revenue vs Computer Science PhD Graduates

the result indicates that the total revenue of the casino is directly causing the number of doctorate graduates by an F score of 0.2. By definition, the Arcade revenue Granger Causes the number of CS Ph.D. graduates. If this is the reality then the government should simply build more casinos across the country in order to obtain more highly educated Ph.D. computer science graduates from school, which judging from intuition I know that this is a falsely claimed statement. However, according to the definition of Granger Causality, the past information of arcade revenue can be used to better predict the future Ph.D. graduates and thus annual arcade revenue Granger Causes CS Ph.D. graduates.

When working with datasets that are more complicated, it is more challenging for researchers to decide whether causal learning should be studied among the pair or not. Although many researchers are fully aware that the existence of Granger Causality may not imply true causality, little to no work has been done to articulate on the idea that by properly leverage the prior information researchers can significantly reduce the number of falsely defined causal relationships and ultimately obtain more reliable conclusions from GC learning.

In order for us to address this concern, it is necessary for researchers to clearly specify when and when not to perform causal learning. More specifically, researchers should categorize different variables into groups based on prior knowledge of each variable. Then, instead of performing causal learning between one variable against all other variables, one should only perform a causal learning algorithm against all the other nodes that belong to

the same cluster. The process of us grouping variables can be achieved using the prior understanding towards each variable. This process can be viewed as classifying whether two variables may appear to be causally related or not based on their prior understanding of the variables which is usually done through the comparison of their features.

Say we have a multivariate dataset that consists of M time series variables. With no consideration of prior knowledge, one may study causal relationships among every pair of variables. The $M \times M$ dimension causal matrix obtained can be summarized as follow:

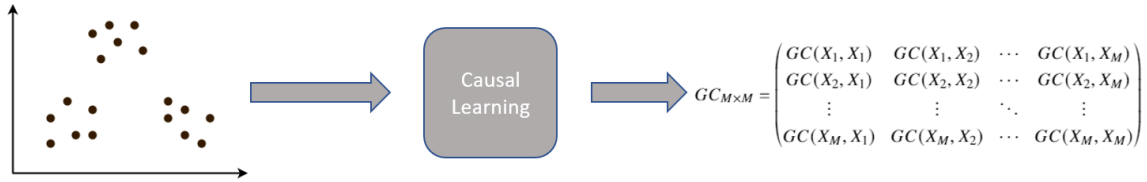
$$GC_{M\times M} = \begin{pmatrix} GC(X_1, X_1) & GC(X_1, X_2) & \cdots & GC(X_1, X_M) \\ GC(X_2, X_1) & GC(X_2, X_2) & \cdots & GC(X_2, X_M) \\ \vdots & \vdots & \ddots & \vdots \\ GC(X_M, X_1) & GC(X_M, X_2) & \cdots & GC(X_M, X_M) \end{pmatrix} \tag{5.1}$$

Where $X_1, \cdots, X_M$ each represents a time series in the dataset. The term $GC(X_1, X_2)$ $\epsilon [0, 1]$ represents the the causal index between variable X_1 and X_2 and the causal relationship is accepted if it is above a certain pre-determined threshold value. The pipeline can be summarized in figure 5.2a.

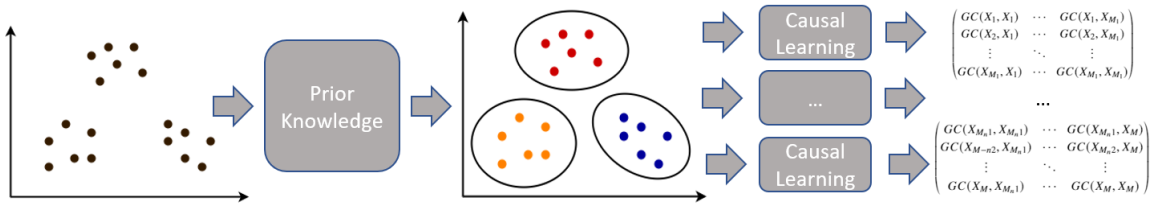
However, like mentioned previously it is possible for us to falsely conclude causal relationship between two variables that are completely independent from each other in matrix 5.1. To reduce these falsely defined relationships, prior knowledge should be used to categorize variables into subgroups based on their relevance. Matrix 5.1 can be divided into n sub-matrices, shown in equation 5.2:

$$\begin{pmatrix} GC(X_1, X_1) & \cdots & GC(X_1, X_{M_1}) \\ GC(X_2, X_1) & \cdots & GC(X_2, X_{M_1}) \\ \vdots & \ddots & \vdots \\ GC(X_{M_1}, X_1) & \cdots & GC(X_{M_1}, X_{M_1}) \end{pmatrix}, \cdots, \begin{pmatrix} GC(X_{M_n1}, X_{M_n1}) & \cdots & GC(X_{M_n1}, X_M) \\ GC(X_{M-n2}, X_{M_n1}) & \cdots & GC(X_{M_n2}, X_M) \\ \vdots & \ddots & \vdots \\ GC(X_M, X_{M_n1}) & \cdots & GC(X_M, X_M) \end{pmatrix} \tag{5.2}$$

Where the M variables are being classified into n groups where $M_1 + M_2 + \cdots + M_n = M$.



(a) Original System Flow Chart Causal Learning on the entire dataset



(b) Proposed System Flow Chart For Reducing Falsely Classified Causal Relationships

Figure 5.2 Comparison between the original pipeline and the proposed pipeline for causal learning. The original approach shown in figure 5.2a did not consider the prior knowledge hence is at risk for overfitting whereas the proposed pipeline, shown in figure 5.2b incorporated the subject expert's knowledge and grouped the variables before applying causal learning only if the pair belongs to the same group. By doing so we can avoid falsely defined causal relationships.

Variable $X_1, \cdots, X_{M_1}$ belongs to group 1 and variable $X_{M_n}, \cdots, X_M$ belongs to group n. The causal learning will then only be performed between variables that belong to the same group to avoid undesired overfitting issues. In other words, any variable that belongs to M_1 will not perform causal learning with any variables that belong to $M_2, M_3, \cdots, M_n$, hence they are not causally associated. By doing so I have successfully eliminated the overfitting issue for Granger Causality(statistical-based) by leveraging the prior knowledge. The proposed pipeline can be summarized in figure 5.2b.

To conclude, in this thesis I have successfully explored and proposed a systems that can be used to address the existing challenges with causal learning, namely performing causal learning with sparsely sampled data. The work that I did can be generalized to applications outside of manufacturing. While the work showed great potentials for generalization, they also come with some limitations. For example, although the performance of the proposed system in Chapter 4 is better than all the benchmark methods, it is still not

accurate enough to be directly used as a replacement of the subject's knowledge for causal discovery. Currently, the ESN algorithm I used in the system is an off-the-shelf package. With the recent advancements in Recurrent neural networks, the accuracy gap with learning algorithms could possibly be bridged. For future works, I would like to come up with my own causal algorithm using deep learning based architecture such as a transformer or Long Short Term Memory (LSTM) in order to further improve our system's performance.

References

[1] D. Mackenzie, Race, covid mortality, and simpson's paradox (2020).

[2] A. Saxena, K. Goebel, D. Simon, N. Eklund, Damage propagation modeling for aircraft engine run-to-failure simulation, International Conference on Prognostics and Health Management (2008). doi:10.1109/PHM.2008.4711414.

[3] J. Herbert, The "echo state" approach to analysing and training recurrent neural networks, GMD-Report 148, German National Research Institute for Computer Science (2001).

[4] J. Pearl, Theoretical impediments to machine learning with seven sparks from the causal revolution, CoRR abs/1801.04016 (2018). arXiv:1801.04016.

[5] P. L. Bartlett, A. Montanari, A. Rakhlin, Deep learning: a statistical viewpoint, arXiv preprint arXiv:2103.09177 (2021).

[6] M. Hernán, B. Brumback, J. Robins, Marginal structural models to estimate the causal effect of zidovudine on the survival of hiv-positive men., Epidemiology 11 5 (2000) 561–70.

[7] J. Hicks, et al., Causality in economics, Australian National University Press, 1980.

[8] J. J. Heckman, Econometric causality, International statistical review 76 (2008) 1–27.

[9] R. P. Thompson, Causality, mathematical models and statistical association: dismantling evidence-based medicine, Journal of Evaluation in Clinical Practice 16 (2010) 267–275.

[10] R. Guo, L. Cheng, J. Li, P. R. Hahn, H. Liu, A survey of learning causality with data, ACM Computing Surveys 53 (2020) 1–37. doi:10.1145/3397269.

[11] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, A. Zhang, A survey on causal inference, 2020. arXiv:2002.02770.

[12] L. Jones, One year of covid-19: 7 ways the world has changed for small business (2021).

[13] Labour force survey, january 2021 (2021).

[14] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, D. Sejdinovic, Detecting and quantifying

causal associations in large nonlinear time series datasets, Science Advances 5 (2019) eaau4996. doi:10.1126/sciadv.aau4996.

[15] H. Y. Toda, P. C. B. Phillips, Vector autoregressions and causality, Econometrica 61 (1993) 1367–1393. URL: http://www.jstor.org/stable/2951647.

[16] M. Gong, K. Zhang, B. Schoelkopf, D. Tao, P. Geiger, Discovering temporal causal relations from subsampled data, in: International Conference on Machine Learning, PMLR, 2015, pp. 1898–1906.

[17] P. V. Xinpeng Shen, Sisi Ma, G. Simon, Challenges and opportunities with causal discovery algorithms: Application to alzheimer's pathophysiology, 2020. doi:https://doi.org/10.1038/s41598-020-59669-x.

[18] A. Taylor, Plato: Timaeus and Critias (RLE: Plato), Routledge, 2013.

[19] S. Kern, A cultural history of causality: Science, murder novels, and systems of thought, Princeton University Press, 2009.

[20] S. Mumford, R. L. Anjum, Causation: a very short introduction, OUP Oxford, 2013.

[21] R. Guo, L. Cheng, J. Li, P. R. Hahn, H. Liu, A survey of learning causality with data, ACM Computing Surveys 53 (2020) 1–37. URL: http://dx.doi.org/10.1145/3397269. doi:10.1145/3397269.

[22] F. Russo, Causation and correlation in medical science: Theoretical problems, in: Handbook of the Philosophy of Medicine, Springer Science+ Business Media Dordrecht, Netherlands, 2017, pp. 839–849.

[23] I. S. Ockene, N. H. Miller, Cigarette smoking, cardiovascular disease, and stroke: a statement for healthcare professionals from the american heart association, Circulation 96 (1997) 3243–3247.

[24] A. L. Filby, E. M. Santos, K. L. Thorpe, G. Maack, C. R. Tyler, Gene expression profiling for understanding chemical causation of biological effects for complex mixtures: a case study on estrogens, Environmental science & technology 41 (2007) 8187–8194.

[25] D. Dua, C. Graff, UCI machine learning repository, 2017. URL: http://archive.ics.uci.edu/ml.

[26] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, A. Zhang, A survey on causal inference, arXiv preprint arXiv:2002.02770 (2020).

[27] R. Guo, L. Cheng, J. Li, P. R. Hahn, H. Liu, A survey of learning causality with data: Problems and methods, ACM Computing Surveys (CSUR) 53 (2020) 1–37.

[28] D. R. Brillinger, Time series: data analysis and theory, SIAM, 2001.

[29] T. M. Somers, Y. P. Gupta, Using multiple time-series analysis, of assembly-line production of automobile engines—a case study, Engineering Costs and Production Economics 21 (1991) 243–258.

[30] B. Matthews, Comparison of the predicted and observed secondary structure of t4 phage lysozyme, Biochimica et Biophysica Acta (BBA) - Protein Structure 405 (1975) 442–451.

[31] I. Guyon, et al., Practical feature selection: from correlation to causality, Mining massive data sets for security: advances in data mining, search, social networks and text mining, and their applications to security (2008) 27–43.

[32] I. Pratama, A. Permanasari, I. Ardiyanto, R. Indrayani, A review of missing values handling methods on time-series data, 2016, pp. 1–6. doi:10.1109/ICITSI.2016.7858189.

[33] T. M. Mitchell, Machine Learning, 1 ed., McGraw-Hill, Inc., USA, 1997.

[34] J. S. Racine, Nonparametric econometrics: A primer, volume 4, Now Publishers Inc, 2008.

[35] E. Fix, J. L. Hodges, Discriminatory analysis. nonparametric discrimination: Consistency properties, International Statistical Review/Revue Internationale de Statistique 57 (1989) 238–247.

[36] M.-L. Zhang, Z.-H. Zhou, Ml-knn: A lazy learning approach to multi-label learning, Pattern recognition 40 (2007) 2038–2048.

[37] H. Zhang, A. C. Berg, M. Maire, J. Malik, Svm-knn: Discriminative nearest neighbor classification for visual category recognition, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), volume 2, IEEE, 2006, pp. 2126–2136.

[38] C. E. Rasmussen, C. K. I. Williams, Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning), The MIT Press, 2005.

[39] M. Kac, A. J. F. Siegert, An explicit representation of a stationary gaussian process, Ann. Math. Statist. 18 (1947) 438–442.

[40] M. P. Deisenroth, Efficient reinforcement learning using gaussian processes, 2010.

[41] M. Farné, A. Montanari, A bootstrap test to detect prominent granger-causalities across frequencies, 2018. arXiv:1803.00374.

[42] S. Grosche, Limitations of granger causality analysis to assess the price effects from the financialization of agricultural commodity markets under bounded rationality (2012).

[43] H. Lütkepohl, Non-causality due to omitted variables, Journal of Econometrics 19 (1982) 367–378.

[44] C. Hiemstra, J. D. Jones, Testing for linear and nonlinear granger causality in the stock price-volume relation, The Journal of Finance 49 (1994) 1639–1664.

[45] D. Marinazzo, M. Pellicoro, S. Stramaglia, Kernel method for nonlinear granger causality, Physical review letters 100 (2008) 144103.

[46] E. Kiciman, A. Sharma, Causal inference and counterfactual reasoning (3hr tutorial), in: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, 2019, pp. 828–829.

[47] C. A. Sims, Money, income, and causality, The American economic review 62 (1972) 540–552.

[48] H. White, Time-series estimation of the effects of natural experiments, Journal of Econometrics 135 (2006) 527–566.

[49] E. Eells, Probabilistic causality, volume 1, Cambridge University Press, 1991.

[50] J. Pearl, [bayesian analysis in expert systems]: comment: graphical models, causality and intervention, Statistical Science 8 (1993) 266–269.

[51] S. Palachy, Inferring causality in time series data (2019). URL: https://towardsdatascience.com/inferring-causality-in-time-series-data-b8b75fe52c46#586a.

[52] Y. Huang, Z. Fu, C. L. Franzke, Detecting causality from time series in a machine learning framework, Chaos: An Interdisciplinary Journal of Nonlinear Science 30 (2020) 063116.

[53] J. Geweke, Inference and causality in economic time series models, Handbook of econometrics 2 (1984) 1101–1144.

[54] M. Ding, Y. Chen, S. Bressler, Granger causality: Basic theory and application to neuroscience. 2006, Handbook of Time Series Analysis [Internet]. Wiley, Wienheim (2020).

[55] R. K. Kaufmann, D. I. Stern, Evidence for human influence on climate from hemispheric temperature relations, Nature 388 (1997) 39–44.

[56] J. Goldstein, Causality and emergence in chaos and complexity theories, in: Nonlinear dynamics in human behavior, World Scientific, 1996, pp. 159–190.

[57] P.-O. Amblard, O. J. Michel, The relation between granger causality and directed information theory: A review, Entropy 15 (2013) 113–143.

[58] R. Dahlhaus, M. Eichler, Causality and graphical models in time series analysis, Oxford Statistical Science Series (2003) 115–137.

[59] B. Schölkopf, Causality for machine learning, arXiv preprint arXiv:1911.10500 (2019).

[60] J. Pearl, Causality: Models, reasoning and inference cambridge university press, Cambridge, MA, USA, 9 (2000) 10–11.

[61] K. Mainali, S. Bewick, B. Vecchio-Pagan, D. Karig, W. F. Fagan, Detecting interaction networks in the human microbiome with conditional granger causality, PLoS computational biology 15 (2019) e1007037.

[62] J. Peters, D. Janzing, B. Schölkopf, Causal inference on time series using restricted structural equation models, in: Advances in Neural Information Processing Systems, 2013, pp. 154–162.

[63] J. Runge, Detecting and quantifying causality from time series of complex systems (2014).

[64] T. Schreiber, Measuring information transfer, Physical review letters 85 (2000) 461.

[65] P.-O. Amblard, O. J. Michel, The relation between granger causality and directed information theory: A review, Entropy 15 (2013) 113–143.

[66] A. Papana, C. Kyrtsou, D. Kugiumtzis, C. Diks, et al., Identifying causal relationships in case of non-stationary time series, Department of Economics of the University of Macedonia, Thessaloniki (2014).

[67] P. Duan, F. Yang, T. Chen, S. L. Shah, Direct causality detection via the transfer entropy approach, IEEE transactions on control systems technology 21 (2013) 2052–2066.

[68] F. Yang, P. Duan, S. L. Shah, T. Chen, Capturing connectivity and causality in complex industrial processes, Springer Science & Business Media, 2014.

[69] A. García-Medina, G. González Farías, Transfer entropy as a variable selection methodology of cryptocurrencies in the framework of a high dimensional predictive model, PloS one 15 (2020) e0227269.

[70] I. Vlachos, D. Kugiumtzis, Nonuniform state-space reconstruction and coupling detection, Physical Review E 82 (2010) 016207.

[71] M. Palus, V. Komarek, Z. Hrncír, K. Sterbová, Synchronization as adjustment of information rates: Detection from bivariate time series, Physical review. E, Statistical, nonlinear, and soft matter physics 63 (2001) 046211. doi:10.1103/PhysRevE.63.046211.

[72] G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, S. Munch, Detecting causality in complex ecosystems, science 338 (2012) 496–500.

[73] Y. Chen, G. Rangarajan, J. Feng, M. Ding, Analyzing multiple nonlinear time series with extended granger causality, Physics letters A 324 (2004) 26–35.

[74] L. Faes, G. Nollo, A. Porta, Information-based detection of nonlinear granger causality in multivariate processes via a nonuniform embedding technique, Phys. Rev. E 83 (2011) 051112. URL: https://link.aps.org/doi/10.1103/PhysRevE.83.051112. doi:10.1103/PhysRevE.83.051112.

[75] E. G. Baek, W. A. Brock, A nonparametric test for independence of a multivariate time series, Statistica Sinica 2 (1992) 137–156.

[76] D. Maraun, J. Kurths, Epochs of phase coherence between el nino/southern oscillation and indian monsoon, Geophysical Research Letters 32 (2005).

[77] P. Spirtes, C. Glymour, R. Scheines, Causation, Prediction, and Search, volume 81, 1993. doi:10.1007/978-1-4612-2748-9.

[78] N. Friedman, The bayesian structural em algorithm, 2013. arXiv:1301.7373.

[79] M. Nauta, D. Bucur, C. Seifert, Causal discovery with attention-based convolutional neural networks, Machine Learning and Knowledge Extraction 1 (2019) 312–340. doi:10.3390/make1010019.

[80] O. Biran, C. Cotton, Explanation and justification in machine learning: A survey, in: IJCAI-17 workshop on explainable AI (XAI), volume 8, 2017, pp. 8–13.

[81] X. Chen, J. Wang, J. Zhou, Probability density estimation and bayesian causal analysis based fault detection and root identification, Industrial & Engineering Chemistry Research 57 (2018) 14656–14664. doi:10.1021/acs.iecr.8b03009.

[82] P. Dagum, A. Galper, E. Horvitz, Dynamic network models for forecasting, in: Uncertainty in artificial intelligence, Elsevier, 1992, pp. 41–48.

[83] B. Baesens, M. Egmont-Petersen, R. Castelo, J. Vanthienen, Learning bayesian network classifiers for credit scoring using markov chain monte carlo search, in: Object recognition supported by user interaction for service robots, volume 3, IEEE, 2002, pp. 49–52.

[84] N. L. Zhang, D. Poole, A simple approach to bayesian network computations, in: Proc. of the Tenth Canadian Conference on Artificial Intelligence, 1994.

[85] J. Pearl, Bayesian networks (2011).

[86] M. Nauta, D. Bucur, C. Seifert, Causal discovery with attention-based convolutional neural networks, Machine Learning and Knowledge Extraction 1 (2019) 312–340. URL: http://dx.doi.org/10.3390/make1010019. doi:10.3390/make1010019.

[87] W. F. Velicer, S. M. Colby, A comparison of missing-data procedures for arima time-series analysis, Educational and Psychological Measurement 65 (2005) 596–615.

[88] D. MacKay, Introduction to gaussian processes, 1998.

[89] B. Chang, M. Naiel, S. Wardell, S. Kleinikkink, J. Zelek, Time-series causality with missing data, Journal of Computational Vision and Imaging Systems 6 (2021) 1–4. URL: https://openjournals.uwaterloo.ca/index.php/vsl/article/view/3552. doi:10.15353/jcvis.v6i1.3552.

[90] D. Duvenaud, Automatic model construction with Gaussian processes, Ph.D. thesis, 2014.

[91] E. Schulz, M. Speekenbrink, A. Krause, A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions, Journal of Mathematical Psychology 85 (2018) 1 – 16.

[92] C. K. I. Williams, Prediction with Gaussian Processes: From Linear Regression to Linear Prediction and Beyond, Springer Netherlands, Dordrecht, 1998, pp. 599–621. URL: https://doi.org/10.1007/978-94-011-5014-9_23. doi:10.1007/978-94-011-5014-9_23.

[93] J. Bernardo, J. Berger, A. Dawid, A. Smith, et al., Regression and classification using gaussian process priors, Bayesian statistics 6 (1998) 475.

[94] F. Jäkel, B. Schölkopf, F. Wichmann, A tutorial on kernel methods for categorization, Journal of Mathematical Psychology 51 (2007) 343–358.

[95] D. Duvenaud, Automatic model construction with Gaussian processes, Ph.D. thesis, University of Cambridge, 2014.

[96] J. Vert, K. Tsuda, B. Schölkopf, A primer on kernel methods, Kernel Methods in Computational Biology, 35-70 (2004) (2004).

[97] C. E. Rasmussen, H. Nickisch, Gaussian processes for machine learning (gpml) toolbox, Journal of Machine Learning Research 11 (2010) 3011–3015. URL: http://jmlr.org/papers/v11/rasmussen10a.html.

[98] D. Frederick, J. DeCastro, J. Litt, User's guide for the commercial modular aero-propulsion system simulation (c-mapss), NASA Technical Manuscript 2007–215026 (2007).

[99] B. Chang, M. Naiel, S. Wardell, S. Kleinikkink, J. Zelek, Time-series causality with missing data, Journal of Computational Vision and Imaging Systems 6 (2021) 1–4. URL: https://

openjournals.uwaterloo.ca/index.php/vsl/article/view/3552. doi:10.15353/jcvis.v6i1.3552.

[100] H. Jaeger, H. Haas, Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication, science 304 (2004) 78–80.

[101] A. Duggento, M. Guerrisi, N. Toschi, Echo state network models for nonlinear granger causality, bioRxiv (2019). URL: https://www.biorxiv.org/content/early/2019/05/27/651679. doi:10.1101/651679. arXiv:https://www.biorxiv.org/content/early/2019/05/27/651679.full.pdf.

[102] H. Jaeger, Short term memory in echo state networks, volume 5, GMD-Forschungszentrum Informationstechnik, 2001.

[103] E. Di Gregorio, C. Gallicchio, A. Micheli, Combining memory and non-linearity in echo state networks, in: International Conference on Artificial Neural Networks, Springer, 2018, pp. 556–566.

[104] X. Chen, Tennessee eastman simulation dataset, 2019. URL: https://dx.doi.org/10.21227/4519-z502. doi:10.21227/4519-z502.

[105] H. Gharahbagheri, S. A. Imtiaz, F. Khan, Root cause diagnosis of process fault using kpca and bayesian network, Industrial & Engineering Chemistry Research 56 (2017) 2054–2070. doi:10.1021/acs.iecr.6b01916.

[106] C. A. Rieth, B. D. Amsel, R. Tran, M. B. Cook, Issues and advances in anomaly detection evaluation for joint human-automated systems, in: J. Chen (Ed.), Advances in Human Factors in Robots and Unmanned Systems, Springer International Publishing, Cham, 2018, pp. 52–63.

[107] A. Franzin, F. Sambo, B. di Camillo, bnstruct: an r package for bayesian network structure learning in the presence of missing data, Bioinformatics 33 (2017) 1250–1252. doi:10.1093/bioinformatics/btw807.

[108] S. Weichwald, M. E. Jakobsen, P. B. Mogensen, L. Petersen, N. Thams, G. Varando, Causal structure learning from time series: Large regression coefficients may predict causal links better in practice than small p-values, in: H. J. Escalante, R. Hadsell (Eds.), Proceedings of the NeurIPS 2019 Competition and Demonstration Track, volume 123 of *Proceedings of Machine Learning Research*, PMLR, 2020, pp. 27–36. URL: http://proceedings.mlr.press/v123/weichwald20a.html.

[109] L. Barnett, A. K. Seth, The mvgc multivariate granger causality toolbox: A new approach to

granger-causal inference, Journal of Neuroscience Methods 223 (2014) 50–68. doi:https://doi.org/10.1016/j.jneumeth.2013.10.018.

[110] A. Duggento, M. Guerrisi, N. Toschi, Echo state network models for nonlinear granger causality, bioRxiv (2019). doi:10.1101/651679.