University of Alberta

Nonlinear Dynamic Causality Inference in Time Series

by

Amir Reza Alizad Rahvar

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Communications

Department of Electrical and Computer Engineering

Spring 2014
Edmonton, Alberta

Dedicated to My Beloved Wife, Zahra.

Abstract

The main focus of this work is on detection of causal relationships or couplings between different processes or systems. Identification of these causal relationships has applications in many disciplines including physics, economics, biology, neuroscience, and climatology. As these couplings or causal relationships are inherently hidden in the underlying dynamics of the system and are not necessarily accessible, we develop methods to discover these interactions by some observations of the system measured in the form of a time series.

In the first part of our work, we propose a new method called the coupling spectrum (CS) for inference of the directed coupling in a deterministic system. We will observe that this method can identify the direction of coupling in sever conditions such as bidirectional couplings, nonlinear dynamics, nonidentical and multivariate systems, small sample sizes, weak couplings, as well as multi-scale and noisy data.

Later, we study a biological and a financial application of the CS method. First, we analyze the microarray data for inference of the gene regulatory networks, one of the most important biological networks that their identification has immediate applications in cancer prediction. Then, the CS method is used for detection of the temporal causality between the stock prices of two companies. The analysis of empirical data in these applications show the successful performance of the CS method in real-world problems.

In the last part of our contributions, we propose a new method for inference of the distributional causality, a kind of causality that its inference has applications in finance and econometrics. Our method provides information about the influence of the causality on the underlying distribution of the processes. The analysis of the simulated and empirical financial data shows the success of our method.

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisor Professor Masoud Ardakani for the continuous support of my Ph.D. study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my Ph.D study.

I would also like to thank my committee members, Professor Yongyi Mao, Professor Tongwen Chen, Dr. Ivor Cribben, and Dr. Majid Khabbazian for serving as my committee members. I also want to thank them for letting my defense be an enjoyable moment, and for their brilliant comments and suggestions. I would especially like to thank Dr. Ivor Cribben (Assistant Professor at the Department of Finance and Statistical Analysis, Alberta School of Business, University of Alberta) for his beneficial and friendly collaboration regarding the financial applications of my research. Moreover, I would like to thank Dr. Paul LaPointe (Assistant Professor at the Department of Cell Biology, University of Alberta) for his interest and suggestions regarding the biological application of my study.

A special thanks to my family. Words cannot express how grateful I am to my parents and parents-in-law for all of the sacrifices that they have made on my behalf. Your prayer for me was what sustained me thus far. I would also like to thank all of my friends who supported me during my Ph.D. and made an enjoyable time for me, especially, Mahdi Karami, Moslem Noori, Ali Saberali, Gayan Amarasuriya, and Hamid Moghaddas. At the end, I would like express appreciation to my beloved wife Zahra who has always been my support in every moment of my life.

Table of Contents

List of Tables

List of Figures

List of Symbols

Symbol	Definition	First Use
D	Driver system	6
R	Response system	6
$D \to R$	D causes (or drives) R	6
$c_{D \to R}$	Coupling strengths of $D \to R$ in Hénon map	6
$\{d_t\}$	Time series of the driver system	10
$\{r_t\}$	Time series of the response system	10
N	Number of the samples	10
$\mathbf{D}_{t-1}$	Time-delayed vector of the driver system	10
$\mathbf{R}_{t-1}$	Time-delayed vector of the response system	10
ϵ_t^r	Prediction error of regression model (effect of R on itself) in Granger causality test	11
$\epsilon_t^{r\|d}$	Prediction error of regression model (effect of D and R on r_t) in Granger causality test	11
$\boldsymbol{\alpha}$	Coefficient vector of regression model (effect of R on itself) in Granger causality test	11
$\boldsymbol{\beta}$	Coefficient vector of regression model (effect of D and R on r_t) in Granger causality test	11
$\sigma^2_{\epsilon_t^r}$	Variance of ϵ_t^r	11
$\sigma^2_{\epsilon_t^{r\|d}}$	Variance of $\epsilon_t^{r\|d}$	11
χ^2_L	Chi-squared distribution with L degrees of freedom	11
$f(\cdot)$	Probability density function	12
$C_\epsilon(\cdot)$	Correlation integral	12
$I(\cdot)$	Indicator function	12

$\mathrm{R}'_{\nrightarrow}$	The rule proofing $D \nrightarrow R \vert C$	34
L_d	Maximum lag value of the common driver system C	35
$\mathrm{Pa}(X_i)$	Parents of node X_i in Bayesian network	45
$\mathrm{SIG}_{\mathrm{cs}}$	Strength of coupling detected by the CS method	48
$f_y(\Theta_{y,t})$	Underlying probability density function of y_t	64
$\Theta_{y,t}$	Distribution parameter of $f_y(\Theta_{y,t})$	64
$h_{\Theta_y}(\cdot)$	The function that relates $\Theta_{y,t}$ to x_{t-1}	64
$\widehat{\Theta}_{y,t}$	The estimation of $\Theta_{y,t}$	65
t_Δ	Time indexes such that $\vert x_{t-1} - x_{t_\Delta - 1}\vert < \delta_x$ and $\vert y_{t-1} - y_{t_\Delta - 1}\vert \geq \delta_y$	65
$\mathbf{X}_{t-1}^{(q)}$	$[x_{t-1}, \cdots, x_{t-q}]$	65
$f(\widehat{\Theta}_{y,t})$	Distribution of the estimated parameter $\widehat{\Theta}_{y,t}$	66
$\widehat{\Theta}_{y,t}^p$	Estimated parameter by permuted data x_t^p	66
$f(\widehat{\Theta}_{y,t}^p)$	Distribution of $\widehat{\Theta}_{y,t}^p$	66
$f_i(\widehat{\Theta}_{y,t}^p)$	$f(\widehat{\Theta}_{y,t}^p)$ obtained by i-th permutation	66
$\bar{f}(\widehat{\Theta}_{y,t}^p)$	Average distribution of $f_i(\widehat{\Theta}_{y,t}^p)$s	66
$\mathrm{KL}(p(x) \Vert\, q(x))$	KL divergence between two distributions $p(x)$ and $q(x)$	67
$\mathrm{JS}(p(x) \Vert\, q(x))$	JS divergence between two distributions $p(x)$ and $q(x)$	67
$m(x)$	Average distribution of $p(x)$ and $q(x)$	67
JS_Θ	$\mathrm{JS}(f(\widehat{\Theta}_{y,t}) \Vert \bar{f}(\widehat{\Theta}_{y,t}^p))$	67
$\mathrm{JS}_{\Theta^p}^i$	$\mathrm{JS}(f_i(\widehat{\Theta}_{y,t}^p) \Vert \bar{f}(\widehat{\Theta}_{y,t}^p))$	67
$\mathrm{JS}_{\mathrm{THR}}$	Significance threshold of JS_Θ	67
$\sigma_{x,t}^2$	Instantaneous variance of x_t	68
$\sigma_{y,t}^2$	Instantaneous variance of y_t	68
$z_{x,t}$, $z_{y,t}$	Independent and identically distributed random variables with zero mean and unit variance, i.e., iid(0,1)	68
$f_{\max}(\widehat{\Theta}^p)$	Upper border of the non-causality area	73

$f_{\min}(\widehat{\Theta}^p)$	Lower border of the non-causality area	73
$S_{\text{min-max}}$	The area enclosed between $f(\widehat{\Theta})$ and $f_{\max}(\widehat{\Theta}^p)$ where $f(\widehat{\Theta}) > f_{\max}(\widehat{\Theta}^p)$, plus the area enclosed between $f(\widehat{\Theta})$ and $f_{\min}(\widehat{\Theta}^p)$ where $f(\widehat{\Theta}) < f_{\min}(\widehat{\Theta}^p)$	73
ν	Degrees of freedom of student's t-distribution	73

List of Abbreviations

Abbreviation	Description	First Use
ARCH	Autoregressive Conditional Heteroscedasticity	6
CS	Coupling spectrum	7
G-causality	Granger causality	11
GWT	Granger-Wald test	11
NLG-Diks	Modified nonlinear Granger causality test proposed by Diks.	13
MI	Mutual information	17
CI	Confidence interval	28
UCI	Upper bound of confidence interval	28
NSR	Noise to signal power ratio	32
GRN	Gene regulatory network	38
BN	Bayesian network	45
DAG	Directed acyclic graph	45
DBN	Dynamic Bayesian network	46
TG	Target gene	51
NLG-causality	Nonlinear Granger causality	55
AAPL	Stock prices of Apple Inc.	58
MSFT	Stock prices of Microsoft Corporation	58
S&P500	Standard and Poor's 500 index	63
KL	Kullback-Leibler divergence	67
JS	Jensen-Shannon divergence	67
t-ARCH	ARCH model by student's t-distribution	73
TDR	True detection rate	73

Chapter 1

Introduction

1.1 Causality

Causation has always been a central topic in philosophy, logic, and science. Detection of causal relationships among variables, events, or phenomena have been the fundamental question of most natural and social sciences, such as physics, finance and economics, biology, physiology, social science, and climatology. The Nobel prizes 2003 and 2011 were awarded in economic sciences for studies corresponding to cause-effect relations, a fact that reveals the importance of this field of science.

As Granger, the winner of the Nobel prize 2003, expresses in [1], there is not a universal accepted definition of causality. Granger says:

> "*Attitudes towards causality differ widely, from the defeatist one that it is impossible to define causality, let alone test for it, to the populist viewpoint that everyone has their own personal definition and so it is unlikely that a generally acceptable definition exists. It is clearly a topic in which individual tastes predominate, and it would be improper to try to force research workers to accept a definition with which they feel uneasy. My own experience is that, unlike art, causality is a concept whose definition people know what they do not like but few know what they do like. It might therefore be helpful to present a definition that some of us appear to think has some acceptable features so that it can be publicly debated and compared with alternative definitions.*"

1.2 Philosophical Causality

Throughout the history, there have been plenty of discussions in philosophy regarding the definition of causality. Two famous definitions used in philosophy are:

Necessary cause: If A is the necessary cause of B, it means provided that B presents, A has occurred necessarily beforehand (or if A does not happen, B will not happen); however, the presence of A does not imply the presence of B. For example, being a female is a necessary cause for pregnancy, i.e., if one person is pregnant, she is necessarily female, however, being a female does not mean being pregnant.

Sufficient cause: If A occurs, B must occur, i.e., occurrence of A is sufficient for occurrence of B. However, the presence of B does not reflect the presence of A (B may occur due to another cause C). For instance, missing the final exam is a sufficient cause for failing, however, failing does not mean necessarily that the student missed the final exam and it can be because of other factors.

These definitions are considering only a unique causal relationship in a deterministic situation and they are abstract and not addressing all real world situations.

1.3 Probabilistic Causality

The real-world systems are not strictly deterministic and their behavior are not predictable with certainty. Therefore, the probabilistic definitions of causality are more appealing than philosophical definitions for scientists, even though probabilistic definitions are more complicated. Hence, rather than saying "*If A occurs, B must occur*", probabilistic statements such as "*The occurrence of A increases or alters the likelihood of B*" are more realistic for scientific applications. Indeed, these kind of definitions reflect the probabilistic nature of the real-world systems or our imperfect knowledge of a deterministic system. For example, smoking increases the probability of lung cancer, but it does not mean that a smoker will necessarily get cancer. Therefore, a new kind of causality arises here that conveys uncertainty about cause and effect relationship [2]:

Contributory cause: It is a cause (among many other causes) of an effect that precedes the effect and changing it alters the effect, but is neither necessary nor

sufficient for the effect. By this definition, smoking is a contributory cause of lung cancer as it can increase the probability of getting lung cancer. However, it is unnecessary (cancer can be due to other reasons) and insufficient (not all smokers suffer from cancer) for cancer.

By the advent of this viewpoint, different probabilistic definitions were proposed such as

1. A causes B if (i) A precedes B in time; (ii) $P(A) \neq 0$; (iii) $P(B|A) > P(B)$ [3].
2. A causes B provided that $P(B|A) > P(B|$ not $A)$ [4].

Although these kind of definitions of causality are mathematically formulated and appealing, there are some serious criticisms about them (e.g., see [1] and [5]). For example, Otte in [5] claims that Suppes' definition in [3] cannot distinguish among genuine and spurious causes and direct and indirect causes.

By applying the graph theory, Markov model, and Bayesian probability, new probabilistic models of causality were vastly studied in the late 80's and 90's literature, e.g., belief networks by using the Markov models [6], graphical modeling and Bayesian networks [7–9], and influence diagram [10–13]. These models have been the center of attention in different disciplines such as computational biology, neuroscience, learning theory, and social science.

1.4 Intervention vs. Observation

The best way to study the causal effect of A on B is forcing A to change and study the effect of this change on B. For instance, to study the effect of one gene on activation/deactivation of other genes in the cell, we can activate (deactivate) that specific gene and study the effect of its presence (absence) on other genes. However, in many practical cases, this intervention or manipulation is infeasible, illegal, or unethical. For example, to study the effect of thousands of genes on each other, it is impracticable to study the effect of each gene alteration on the other genes. Also, for studying the effect of smoking on lung cancer, it is illegal and unethical to force somebody to smoke.

Accordingly, in many real-world situations, to study the causal effect of A on B we cannot find the *interventional probability* $P(B|\text{do}(A))$. Consequently, we have to use *conditional probability* $P(B|A)$ which can be estimated from observed data.

As a result, rather than finding "the probability of getting cancer for a person forced to smoke", we can find "the probability of observing cancer in smokers".

Most of the earlier research on probabilistic causality attempts to interpret the Bayesian network as the causal networks. However, these directed networks constructed by conditional probabilities does not necessarily reflect the cause-effect relationships. Indeed, we can detect the causal relationships by applying the interventional probabilities. Hence, Pearl proposes the theory of *causal calculus* to find interventional probabilities from conditional probabilities in the Bayesian network framework [10]. Accordingly, the *influence diagram* is presented based on Bayesian networks [10–13].

1.5 Predictive Causality

The first quantifiable and measurable definition of causality was proposed in 1956 by Wiener [14]:

> "*For two simultaneously measured signals, if we can predict the first signal better by using the past information from the second one than by using the information without it, then we call the second signal causal to the first one.*"

This definition presents a new kind of causality called *predictive causality*. There is a fundamental difference between the predictive and interventional causal definitions [15]. Predictive and interventional causality answer two different questions:

Predictive causality: If I know the current state of the cause, how much does it help to predict the future state of the effect?

Interventional causality: If I change the current state of the cause, to what extent does it change the future state of the effect?

In other words, the predictive causality determines the amount of information provided by the cause for prediction of the effect. In interpretation of many predictive causality results, these two concepts are not distinguished and misinterpreted interventional causal relations are drawn from predictive methods [15].

Many of causality inference methods originate from Wiener's definition. The first causality inference method developed based on predictive causality for analyzing the time series data was proposed by Granger [16], that is known as *Granger causality*.

This method applies the linear regression model to predict the future value of the effect time series. Provided that this prediction improves by including the past samples of the cause time series, we conclude the existence of Granger causality. This method has been applied in many disciplines due to its simplicity, robustness, and extendability.

1.6 Nonlinear Dynamic Causality

In the fields of physics and nonlinear dynamics, predictive causal relations are typically investigated in the sense of coupled systems where the driver system influences the response system. Identifying the couplings between the sub-systems of a complex system is essential for understanding the functionality of the interactions and controlling them. This area of study has immediate applications in various disciplines, such as physics, process and control engineering, chemical engineering, economics, biology, physiology, ecology, and climatology. Synchronized systems are one example of the coupled systems in which the coupling strength increases very much and the response system becomes synchronized with the driver system [17]. For example, many physiological signals, such as heartbeat and breath rate, are synchronized [18]. Synchronized systems are observed in physical [19], physiological [18, 20–22], and chaotic systems [23], as well as neural signals [24, 25].

The challenging issue in identifying the interactions between the coupled systems is that the couplings are inherently hidden in the underlying dynamics of the system and are not necessarily accessible. That is to say, we commonly have access to some observations of the system measured in time series. Hence, methods that based on the available observations determine whether these time series originate from coupled or decoupled sub-systems are much needed. This task will be more complicated if the goal is to determine the direction of coupling, i.e., identifying the driver and response systems. Detecting directed couplings sheds light on identifying the cause-effect relationships in causal networks.

One of the challenging issues in the task of causality inference is the analysis of time series with nonlinear dynamics. Nonlinearity complicates the inference methods as they need to utilize advanced statistical and information theoretical tools. For example, different extensions of Granger causality have been proposed for dealing with nonlinear dynamics of the systems [26–28]. We will introduce the Granger causality and its extensions and other nonlinear inference methods in Chapter 2.

1.7 Causality Strength vs. Causality Probability

In probabilistic causality, we attempt to determine with which probability we should expect the occurrence of the effect B conditioned on having the cause A. Although identification of the probabilistic behavior of the causal interactions between the subsystems can provide a deep understanding of the system, in many practical cases we need to measure the strength of the causal influences or couplings, rather than only their probabilities. For example consider a deterministic coupled system realized by the Hénon map [29], given by

$$d_t = 1.4 - d_{t-1}^2 + 0.3\ d_{t-2} \tag{1.1a}$$

$$r_t = 1.4 - r_{t-1}^2 + 0.3\ r_{t-2} + c_{D\to R}(r_{t-1}^2 - d_{t-1}^2). \tag{1.1b}$$

where the driver system D drives the response system R, denoted by $D \to R$. Here, the last term of (1.1b) reflects the coupling $D \to R$ and $c_{D\to R}$ determines the strength of this coupling. In many application, not only we want to identify the direction of coupling, but we also want to detect its strength, especially when we are dealing with bidirectional couplings. For instance, if the stock prices of two companies are influencing each other, it is desirable to know which of them has more severe effect on the other one, indeed, which of them is stronger.

1.8 Deterministic and Distributional Causality

Here, we categorize the causal relationship from different viewpoint. Generally, the causal relationship between two processes D and R can be categorized into deterministic and distributional causality. Consider two time series $\{d_t\}$ and $\{r_t\}$ observed from processes D and R, respectively, and suppose D causes R. In the absence of the noise, the deterministic causality means that the future value of $\{r_t\}$ is a deterministic function of the lagged values of $\{d_t\}$.

In the case of distributional causality, however, the lagged values of $\{d_t\}$ affect the underlying probability distribution of $\{r_t\}$. In other words, the future value of $\{r_t\}$ cannot be represented as a deterministic function of the lagged values of $\{d_t\}$. This kind of causality is typically observed in financial time series [30]. For example, in modeling of financial time series that exhibit time-varying volatility clustering, Autoregressive Conditional Heteroscedasticity (ARCH) models are commonly used [30]. In the case of a bivariate ARCH process, the variance of time series $\{r_t\}$

can be determined from the lagged values of time series $\{d_t\}$. Here, $\{d_t\}$ and $\{r_t\}$ would be uncorrelated but dependent and many existing causality inference methods, including linear Granger causality, are unable to detect the causal relationship from uncorrelated data. Hence, new methods capable of handling the uncorrelated data for inference of the distributional causality are required.

Although some methods, such as nonlinear extensions of the Granger causality test [26, 27], can detect the existence of causal relationships in the case of distributional causality, they cannot determine whether this causality is deterministic or distributional.

1.9 Summary of Contributions and Thesis Organization

In real-world problems, we usually deal with systems composed of non-identical sub-systems with fundamentally different structures as well as non-linear dynamics. Furthermore, in many cases, the available sample size is small, e.g., biological and genetic data. In addition, in many practical cases, the strength of couplings is very weak and/or asymmetric bidirectional couplings between components of the system exist. All of these scenarios are challenging for identification of causality and directed couplings. Hence, the first goal of our research is to address these challenges for the problem of identification of directed couplings. The method proposed to achieve the first goal can be categorized as a deterministic causality inference method. To infer the distributional causality and find the affected moments or distribution parameters of the underlying distribution of the effect system, we present a new method. Indeed, this method is capable of distinguishing between deterministic and distributional causality.

In Chapter 2, we introduce some existing methods for inference of directed coupling and nonlinear dynamic causality. These methods are i) linear/nonlinear Granger causality; ii) state-space reconstruction method; iii) phase synchronization method; and iv) information theoretic method.

In Chapter 3, we propose a new method to discover the direction of couplings between two or more time series in a driver-response system based on the concept of predictive causality. This method is called the coupling spectrum (CS) method. The simulation results show that the CS method can detect the direction of coupling correctly for identical and non-identical sub-systems, nonlinear dynamics, small sample sizes, as well as weak coupling strength. This method can also detect the stronger

couplings in asymmetric bidirectional couplings. Moreover, unlike some information theoretic methods, the CS method is invariant to data scaling and it is also applicable for bivariate and multi-variate couplings.

In Chapter 4, two applications of the CS method in biology and finance are presented. In the first part of this chapter, the CS method is applied for inference of biological networks. Gene-gene regulatory interactions in the cell can be considered as a driver-response coupling. The CS method can be a candidate to detect these interactions as it is capable of identifying directed couplings in severe practical conditions such as unidirectional and bidirectional coupling, nonlinear coupling, and time series with small sample sizes. The results of applying the CS method for identifying a known regulatory network from microarray data show that this method can detect these regulatory interactions with a high level of accuracy.

In the second part of Chapter 4, the CS method is applied for detection of causal relationships between financial time series. As in many financial cases the direction of causality changes with time, we combine the CS method with overlapped moving window technique to detect time-varying causality. The simulation results show the success of the windowed-CS method for detecting time-varying causality in a simulated temporal nonlinear causal system. The results are then compared to a moving window adaptation of a nonlinear extension of the Granger causality test proposed in [26]. We also apply these two methods for detecting the temporal causal relationships between the stock prices of Apple Inc. and Microsoft Corporation in more than a decade. The simulated and empirical results show that the CS method is more robust than the nonlinear Granger causality method.

In Chapter 5, a new method is proposed for inference of distributional causality between two time series. This method not only is able to detect the existence of causality, but is also capable of identifying the type of the moments or distribution parameters influenced by the distributional causality, e.g., mean, volatility, or higher order statistics. The results of our method is compared with a nonlinear extension of the Granger causality proposed in [27]. The simulation results show that with a large sample size of data, e.g., financial data, and contingent on the existence of the distributional causality, our proposed method can be superior to the nonlinear Granger causality. Identification of the causal relationships between the stock return and volume is a well-known problem in finance and econometrics. Here, we use our method to study daily S&P500 stock return and percentage change in its volume. We

will find that not only the return causes the volume change, it also affects the mean of the volume change and not its volatility. Identifying the type of the moments or distribution parameters influenced by the distributional causality was not possible based on existing nonlinear Granger causality method.

The conclusion of this dissertation and the future research directions are presented in Chapter 6.

Chapter 2

Nonlinear Dynamic Causality Inference

In this chapter, we introduce different methods proposed for inference of nonlinear dynamic coupling or causality from time series data. These methods can be divided into four major categories: (i) Granger causality [16, 26, 31]; (ii) state-space reconstruction methods [24, 32, 33]; (iii) phase synchronization approaches [34, 35]; and (iv) information theoretical methods [36, 37].

2.1 Notations

Consider a coupled system consisting of a driver system D and a response system R, denoted by $D \rightarrow R$. The samples of D are denoted by a finite time series $\{d_t\}$, consisting of N samples. Now, define $\mathbf{D}_{t-1} = (d_{t-1}, d_{t-2}, ..., d_{t-L_d})^T$ the time-delayed vector with the maximum lag value L_d. Similarly, for $\{r_t\}$ we can define $\mathbf{R}_{t-1}$ with the maximum lag value L_r. Figure 2.1 depicts two time series $\{d_t\}$ and $\{r_t\}$, and the corresponding $\mathbf{D}_{t-1}$ and $\mathbf{R}_{t-1}$, respectively, for $L_d = 3$ and $L_r = 4$.

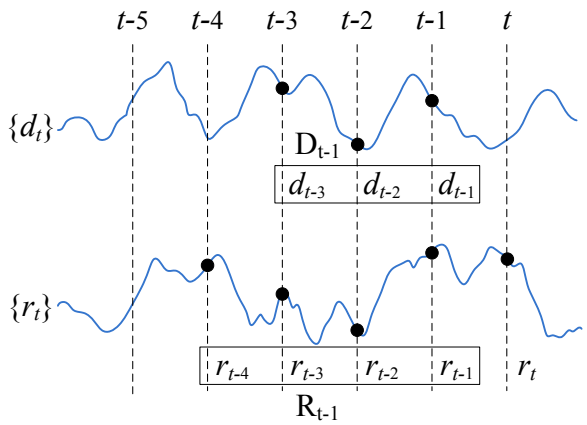


Figure 2.1: Two time series derived from systems D and R with $L_d = 3$ and $L_r = 4$.

2.2 Granger Causality

2.2.1 Linear Granger Causality

One of the first attempts to infer the causal relationships between time series is based on the concept of Granger causality (G-causality) [1, 16]. The principle idea of this method is that the cause happens prior to the effect and it contains some information about the future value of the effect. In linear G-causality method, a linear regression based model is applied. Initially, we use an autoregression model for predicting the current sample of R (i.e., r_t) by its own past samples (i.e., $\mathbf{R}_{t-1}$)

$$r_t = \alpha_0 + \boldsymbol{\alpha}^T \mathbf{R}_{t-1} + \epsilon_t^r \tag{2.1}$$

where ϵ_t^r indicates the prediction error and α_0 and $\boldsymbol{\alpha}$ are determined by autoregression to minimize ϵ_t^r. If we also consider the past samples of D to predict r_t, we have

$$r_t = \alpha_0 + \boldsymbol{\alpha}^T \mathbf{R}_{t-1} + \boldsymbol{\beta}^T \mathbf{D}_{t-1} + \epsilon_t^{r|d}. \tag{2.2}$$

If $\sigma^2_{\epsilon_t^r}$ and $\sigma^2_{\epsilon_t^{r|d}}$ denote the variance of the prediction errors ϵ_t^r and $\epsilon_t^{r|d}$, respectively, then $\sigma^2_{\epsilon_t^{r|d}} < \sigma^2_{\epsilon_t^r}$ reveals that the prediction of r_t improves by considering D. Hence, D is a Granger cause of R. To test the hypothesis of no causality, we can use the Granger-Wald test [38] defined as follows

$$\text{GWT} = N \left| \frac{\sigma^2_{\epsilon_t^r} - \sigma^2_{\epsilon_t^{r|d}}}{\sigma^2_{\epsilon_t^{r|d}}} \right|. \tag{2.3}$$

For $L_d = L_r = L$, the GWT statistic follows the chi-squared distribution with L degrees of freedom (χ^2_L) under the null hypothesis of no causality, i.e., $\sigma^2_{\epsilon_t^r} = \sigma^2_{\epsilon_t^{r|d}}$.

2.2.2 Nonlinear Granger Causality

The assumption of linearity can be violated in real applications and nonlinear causal relationships are not detectable by linear approaches [39]. Hence, different nonlinear extensions of G-causality method were proposed, e.g., a non-parametric method using the correlation integral [26] and non-parametric approaches based on Fourier or Wavelet transformation [28].

Hiemstra and Jones present a non-parametric and nonlinear extension of the G-causality method, called the HJ test [26]. The HJ method tests the hypothesis 'D does not cause R'. This hypothesis is tested by the following conditional

independence

$$\mathrm{H_o} : f(r_t|\ r_{t-1}, d_{t-1}) = f(r_t|\ r_{t-1}) \tag{2.4}$$

where $f(\cdot)$ denotes the probability density function. Here, we consider the maximum time lag of 1. Rejection of $\mathrm{H_o}$ means that the past value of d_t affects the future value of r_t, i.e., D Granger causes R. The HJ method uses the correlation integral for computation of the probabilities in the hypothesis $\mathrm{H_o}$ as follows

$$\mathrm{H_o} : \frac{C_\epsilon(r_t, r_{t-1}, d_{t-1})}{C_\epsilon(r_{t-1}, d_{t-1})} = \frac{C_\epsilon(r_t, r_{t-1})}{C_\epsilon(r_{t-1})}. \tag{2.5}$$

Here, the correlation integral $C_\epsilon(\cdot)$ represents the distribution $f(\cdot)$. For instance, $C_\epsilon(r_t, r_{t-1}, d_{t-1})$ is defined as the following probability

$$\begin{aligned} C_\epsilon(r_t, r_{t-1}, d_{t-1}) &= Pr\left(|r_t - r_{t'}| \le \epsilon, |r_{t-1} - r_{t'-1}| \le \epsilon, |d_{t-1} - d_{t'-1}| \le \epsilon\right) \\ &= \frac{2}{N(N-1)} \sum_{t=1}^{N-1} \sum_{t'>t}^{N} I(|r_t - r_{t'}| \le \epsilon) I(|r_{t-1} - r_{t'-1}| \le \epsilon) I(|d_{t-1} - d_{t'-1}| \le \epsilon). \end{aligned} \tag{2.6}$$

Here, $I(\cdot)$ is an indicator function. Other probabilities in equation (2.5) are represented by the correlation integral similarly.

Under the assumption that $\{d_t\}$ and $\{r_t\}$ are strictly stationary, Hiemstra and Jones introduce the following test statistic

$$\mathrm{TVAL} = \sqrt{N}\left(\frac{C_\epsilon(r_t, r_{t-1}, d_{t-1})}{C_\epsilon(r_{t-1}, d_{t-1})} - \frac{C_\epsilon(r_t, r_{t-1})}{C_\epsilon(r_{t-1})}\right) \tag{2.7}$$

where it can be shown that the asymptotic distribution of TVAL under the null hypothesis $\mathrm{H_o}$ (2.5) is $\mathrm{Normal}\big(0, \sigma^2(L_d, L_r, \epsilon)\big)$. The variance of this normal distribution is a function of L_d, L_r, and ϵ. This variance and its estimated value is presented in [26]. By using the observed value of TVAL, we can make a conclusion about $\mathrm{H_o}$. Provided that the hypothesis $\mathrm{H_o}$ is rejected, we conclude the existence of the causality $D \to R$.

Studies of Diks and Panchenko in [27] shows that the HJ test over-rejects the null hypothesis $\mathrm{H_o}$. In other words, we detect spurious Granger causal relationships by this test. To resolve the over-rejection of the HJ test, [27] proposes the following hypothesis

$$\mathrm{H_1} : E\left(\left[f(r_t|\ r_{t-1}, d_{t-1}) - f(r_t|\ r_{t-1})\right] f^2(r_t)\right) = 0 \tag{2.8}$$

where $E(\cdot)$ is the expected value. In hypothesis (2.8), $f^2(r_t)$ is a positive weight function determined by the stability criterion studied by the Monte Carlo simulations. The hypothesis H_1 in (2.8) can be simplified as

$$H_1 : E\left(f(r_t, r_{t-1}, d_{t-1})f(r_t) - f(r_t, r_{t-1})f(r_t, d_{t-1})\right) = 0 \tag{2.9}$$

Here, the local density estimator is used instead of correlation integral for estimating $f(\cdot)$. For example, for estimating $f(r_t, r_{t-1}, d_{t-1})$, we define the vector $W_t = [r_t, r_{t-1}, d_{t-1}]$ and the density estimator is as follows

$$\hat{f}(W_t) = \frac{(2\epsilon)^{-L_w}}{N-1} \sum_{t'=1, t' \neq t}^{N} I_{tt'}^{W} \tag{2.10}$$

where $I_{tt'}^{W} = I(\|W_t - W_{t'}\| \leq \epsilon)$ and L_w is the dimension of the vector W. Hence, the expected value in (2.9) can be estimated by

$$T(\epsilon) = \frac{(N-1)}{N(N-2)} \sum_{i=1}^{N} \left(\hat{f}(r_t, r_{t-1}, d_{t-1})\hat{f}(r_t) - \hat{f}(r_t, r_{t-1})\hat{f}(r_t, d_{t-1})\right). \tag{2.11}$$

Studies of [27] shows that to prevent the increasing of the false detection rate with the sample size, we have to reduce ϵ by increasing the sample size N. Hence, they introduce the following bandwidth

$$\epsilon_N = CN^{-\beta} \tag{2.12}$$

where C is a constant value and $\beta \in (\frac{1}{4}, \frac{3}{4})$. Equation (2.12) may lead to large bandwidths for small N, hence, the bandwidth is restricted by

$$\epsilon_N = \min(CN^{-\beta}, 1.5). \tag{2.13}$$

In the next chapters, we refer to this modified test by the NLG-Diks test.

2.3 State-Space Reconstruction

The principle idea of this method is that if the driver system D drives the response system R, then the closeness of the points in the driver space implies the closeness in the response space. For example, provided that $D \rightarrow R$ exists in Fig. 2.2, the closeness of the pair points $\mathbf{D}_{t_1}$ and $\mathbf{D}_{t_2}$ ($\mathbf{D}_{t_3}$ and $\mathbf{D}_{t_4}$) in the driver space results in the closeness of $\mathbf{R}_{t_1}$ and $\mathbf{R}_{t_2}$ ($\mathbf{R}_{t_3}$ and $\mathbf{R}_{t_4}$) in the response space. Different methods are proposed to quantify the dependency between the closeness in driver

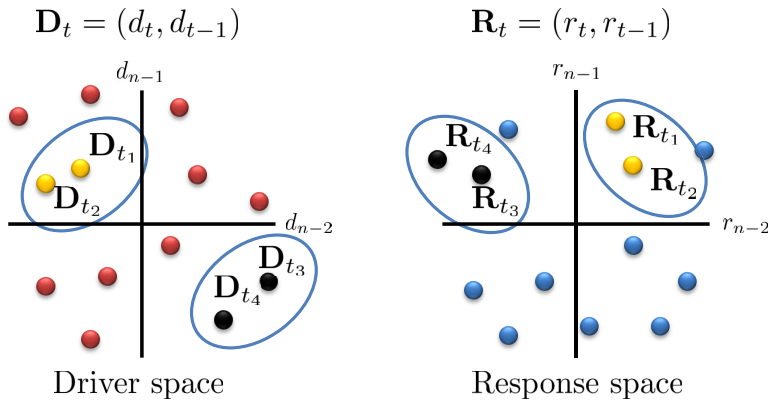


Figure 2.2: The main idea of state-space reconstruction method is shown in this figure. In the case of $D \to R$, the closeness of the points in the driver space (pair points $(\mathbf{D}_{t_1}, \mathbf{D}_{t_2})$ or $(\mathbf{D}_{t_3}, \mathbf{D}_{t_4})$) implies the closeness in the response space (pair points $(\mathbf{R}_{t_1}, \mathbf{R}_{t_2})$ or $(\mathbf{R}_{t_3}, \mathbf{R}_{t_4})$, respectively).

and response spaces [24, 32, 33]. We explain the approach of [32] that is the most famous method in the literature.

Let for $t = 1, ..., N$ and $k = 1, ..., K$, $m^d_{t,k}$ and $m^r_{t,k}$ denote the time indexes of the K nearest neighbors of $\mathbf{D}_t$ and $\mathbf{R}_t$, respectively. For each $\mathbf{R}_t$, one can define two different mean-squared Euclidean distances as follows:

$$\Delta^K_t(R) = \frac{1}{K}\sum_{k=1}^{K}\left|\mathbf{R}_t - \mathbf{R}_{m^r_{t,k}}\right|^2 \tag{2.14}$$

$$\Delta^K_t(R|D) = \frac{1}{K}\sum_{k=1}^{K}\left|\mathbf{R}_t - \mathbf{R}_{m^d_{t,k}}\right|^2 \tag{2.15}$$

where $m^r_{t,k}$ in (2.14) is replaced by $m^d_{t,k}$ in (2.15). More explicitly, (2.14) is the distance of $\mathbf{R}_t$ with its K nearest neighbors and (2.15) is its distance from the points in the response space with the time indexes derived from K nearest neighbors of $\mathbf{D}_t$ in the driver space. If $D \to R$, the closeness in the driver space implies the closeness in response space; hence, $m^d_{t,k} \approx m^r_{t,k}$ and consequently $\Delta^K_t(R|D) \approx \Delta^K_t(R)$. On the other hand, if D and R are decoupled, then there is no particular relationship between $m^d_{t,k}$ and $m^r_{t,k}$ and it follows $\Delta^K_t(R|D) \gg \Delta^K_t(R)$. Therefore, the following measure can be defined to reveal and quantify the coupling between D and R

$$S^K_t(R|D) = \frac{\Delta^K_t(R)}{\Delta^K_t(R|D)} \tag{2.16}$$

where $0 < S^K_t(R|D) \leq 1$ and the larger $S^K_t(R|D)$ means the stronger coupling. The main drawback of this method is that for weak couplings and noisy data, the detection of coupling is difficult.

2.4 Phase Synchronization

This approach is applicable for oscillatory systems that the dynamic exhibits oscillation and the instantaneous phase of the oscillations are well-defined [34, 35, 40]. Generally, the phase is not well-defined for an arbitrary signal. The Hilbert transform can be applied for phase estimation from time series [41]. In this method, the direction of coupling is determined by analyzing the relation between the phase of the sub-systems. Let $\phi_d(t)$ and $\phi_r(t)$ denote the unwrapped phase of D and R (not restricted to $[0, 2\pi]$), respectively. Provided that R is driven by D, the phase of R is influenced by the phase of D, which can be represented as

$$\phi_r(t+\tau) = \phi_r(t) + F_r(\phi_r(t), \phi_d(t)) + \xi_r(t) \tag{2.17}$$

where τ is a time delay and $\xi_r(t)$ is a zero-mean random process . F_r, representing the dependency of R to D, can have different forms such as a trigonometric polynomial form

$$F_r(\phi_r, \phi_d) = \sum_{m,n} \left[a_{m,n} \cos(m\phi_r + n\phi_d) + b_{m,n} \sin(m\phi_r + n\phi_d) \right]. \tag{2.18}$$

The strength of coupling is determined by dependence of F_r on ϕ_d, i.e., $\frac{\partial F_r}{\partial \phi_d}$.

The principle drawback of this method is that in most cases the phase of the system is not well-defined, hence, this method is not applicable.

2.5 Information Theoretical Methods

Information theory has successfully found its significance in different disciplines such as communications, physics, finance, genetics, psychology, and neuroscience. Indeed, information theory can be applied to the problems dealing with nonlinear dynamics, complex systems, and non-deterministic and probabilistic processes. Hence, information theory can be a candidate for dealing with the problem of causality inference.

A causal relationship can be understood in terms of a 'flow' between the cause and effect processes. In the case of the information theory, this flow can be considered as the flow of information from the cause system toward the effect system. However, most existing information theoretic measures, such as mutual information, do not consider cause-effect (or driver-response) relationship due to their symmetry

property. Hence, breaking this symmetry enables us to discover the causal relationships.

In this section, we begin by defining the basic information theoretic measures. Then, the transfer entropy [36] proposed for breaking the symmetry property to detect the direction of information flow will be introduced.

2.5.1 Definition of basic information theoretic measures

In this section we define the basic information theoretic measures for the continues random variable X with the probability density function $f(x)$ [42].

2.5.1.1 Self-Information

The amount of information contained in a probabilistic event X, called self-information, is defined as follows

$$I_X = \log\left(\frac{1}{f(x)}\right) \tag{2.19}$$

where log is the natural logarithm, and consequently, I_X is measured in nats. I_X has the following properties:

1. The smaller the probability of an event X, the larger its self-information. This, for example, means observing an event with small probability brings a lot of information to the observer;

2. I_X is positive;

3. I_X is additive, i.e., the self-information of a pair of independent events X and Y is $I_X + I_Y$.

2.5.1.2 Differential Entropy

Entropy is the average value of self-information I_X of the random variable X. For the continuous random variable X, differential entropy H_X is defined as follows

$$\begin{aligned} H(X) &= E[I_X] \\ &= -\int f(x)\log f(x)dx \end{aligned} \tag{2.20}$$

where $E[\cdot]$ is the expected value. Provided that the above integral exists, $H(X) \geq 0$.

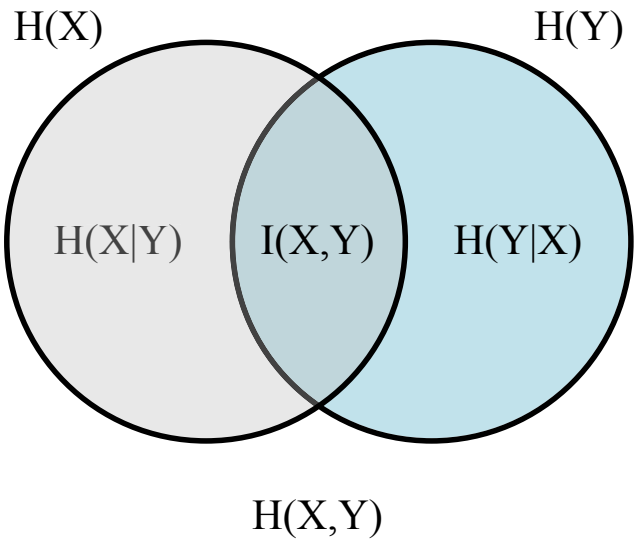


Figure 2.3: This diagram shows the relation between the entropies of two random variables X and Y and their mutual information, joint entropy and conditional entropies.

Consider two continuous random variables X and Y with a joint probability distribution function $f(x, y)$ and marginal distribution functions $f(x)$ and $f(y)$. Similar to (2.20), the average information obtained by observing X and Y can be measured by

$$H(X,Y) = -\iint f(x,y)\log f(x,y)dxdxy. \tag{2.21}$$

Hence, the conditional entropy $H(X|Y)$ can be defined by

$$H(X|Y) = -\iint f(x,y)\log f(x|y)dxdy. \tag{2.22}$$

By considering $f(x|y) = \frac{f(x,y)}{f(y)}$, we can rewrite (2.22) as

$$H(X|Y) = H(X,Y) - H(Y). \tag{2.23}$$

Indeed, the conditional entropy $H(X|Y)$ measures the average amount of new information obtained by observation of X, after observing Y.

Figure 2.3 depicts the Venn diagram showing $H(X), H(Y), H(X,Y), H(X|Y)$, and $H(Y|X)$.

2.5.1.3 Mutual Information

To measure the mutual information (MI) between the random variables X and Y, denoted by $I(X,Y)$, (i.e., the intersection of $H(X)$ and $H(Y)$ in the Venn diagram

shown in Fig. 2.3) we can use the following formulas

$$I(X,Y) = H(X) + H(Y) - H(X,Y); \tag{2.24}$$

$$= H(X) - H(X|Y); \tag{2.25}$$

$$= H(Y) - H(Y|X). \tag{2.26}$$

Indeed, the MI can be interpreted as the amount of information about X gained from Y. By using the joint probability density function $f(x,y)$ and the marginal distributions $f(x)$ and $f(y)$, we have

$$I(X,Y) = -\iint f(x,y)\log\left(\frac{f(x,y)}{f(x)f(y)}\right)dxdy. \tag{2.27}$$

The mutual information $I(X,Y)$ has the following properties:

1. $I(X,Y) \geq 0$;
2. Symmetry property: $I(X,Y) = I(Y,X)$.

2.5.2 Measurement of Information Flow

As it is mentioned before, the causal relationship (coupling) $X \to Y$ can be considered as the flow of information from cause (driver) X to effect (response) Y. However, it is seen in Sec. 2.5.1.3 that the mutual information between the random variables X and Y is symmetric, i.e., $I(X,Y) = I(Y,X)$. Therefore, the mutual information itself cannot show the flow of information between X and Y. Hence, to detect the flow of information between two coupled systems we have to use asymmetric measures. Here, we introduce some asymmetric information theoretic measures for detection of the flow of information.

2.5.2.1 Directed Information

One way to break the symmetry property of the mutual information is using the conditional MI. Massey used this idea and proposed the *directed information* as follows [43]

$$I(\mathbf{D^N} \to \mathbf{R^N}) = \sum_{t=2}^{N} I(\mathbf{D^t}, r_t|\mathbf{R^t}) \tag{2.28}$$

where $\mathbf{D^t} = [d_1, d_2, \cdots, d_t]$ and $\mathbf{R^t} = [r_1, r_2, \cdots, r_t]$. As it is obvious, $I(\mathbf{D^N} \to \mathbf{R^N}) \neq I(\mathbf{R^N} \to \mathbf{D^N})$ and the flow of information is measurable by directed information.

2.5.2.2 Transfer Entropy

The most popular asymmetric measure to identify the direction of coupling is *transfer entropy* (TE) [36], which determines the direction of information flow between two random variables. Generally, TE can be defined based on two information [36, 37]:

1. $I_{r|R,D}$: Information about the current sample r_t gained from past samples of R and D;

2. $I_{r|R}$: Information about the current sample r_t gained from past samples of R.

Accordingly, TE is defined by

$$\mathrm{TE}(D \to R) = I_{r|R,D} - I_{r|R} \tag{2.29}$$

that is the information flow from D to R. According to this definition, if there is no causality between D and R, we cannot get any information about r_t from D. Consequently, $I_{r|R,D}$ and $I_{r|R}$ would be the same, and therefore, $\mathrm{TE}(D \to R) = 0$. On the other hand, provided that $D \to R$, we obtain more information about the current sample r_t by considering the past samples of both R and D rather than only R. Hence, $I_{r|R,D}$ is greater than $I_{r|R}$ and consequently $\mathrm{TE}(D \to R) > 0$.
The mathematical form of TE is [44]

$$\begin{aligned}\mathrm{TE}(D \to R) &= H(r_t|\mathbf{R}_{t-1}) - H(r_t|\mathbf{R}_{t-1}, \mathbf{D}_{t-1}) \\ &= H(r_t, \mathbf{R}_{t-1}) - H(\mathbf{R}_{t-1}) - H(r_t, \mathbf{R}_{t-1}, \mathbf{D}_{t-1}) + H(\mathbf{R}_{t-1}, \mathbf{D}_{t-1})\end{aligned} \tag{2.30}$$

which can be written as

$$\mathrm{TE}(D \to R) = \iiint f(r_t, \mathbf{R}_{t-1}, \mathbf{D}_{t-1}) \log \frac{f(r_t|\mathbf{R}_{t-1}, \mathbf{D}_{t-1})}{f(r_t|\mathbf{R}_{t-1})}\, dr_t\, d\mathbf{R}_{t-1}\, d\mathbf{D}_{t-1}. \tag{2.31}$$

TE can be estimated by [44]

$$\mathrm{TE}(D \to R) = \sum_{t=1}^{N} \log \frac{g(r_t|\mathbf{R}_{t-1}, \mathbf{D}_{t-1})}{g(r_t|\mathbf{R}_{t-1})} \tag{2.32}$$

where $g(\cdot)$ is the density function which can be estimated from observed data by using the kernel estimator [45]

$$\hat{g}(\mathbf{R}_{t-1}, \mathbf{D}_{t-1}) = \frac{1}{N^*} \sum_{t'} \frac{1}{h_r^{L_r} h_d^{L_d}} \prod_{j=1}^{L_r} K\left(\frac{r_{t-j} - r_{t'-j}}{h_r}\right) \prod_{j=1}^{L_d} K\left(\frac{d_{t-j} - d_{t'-j}}{h_d}\right). \tag{2.33}$$

Here, $N^* = N - \max\{L_r, L_d\}$, $K(\cdot)$ is the kernel function, and h_d and h_r are the bandwidth of the kernel function. We use the Gaussian kernel function defined by

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}}. \tag{2.34}$$

The optimal value of the bandwidths h_d and h_r in (2.33) to estimate a normal distribution with the standard deviation σ is $\sigma N^{-1/(L+1)}$ where L is the maximum lag value [46].

The main problem with TE is that the direction of detected coupling is affected by data scaling, i.e., the detected direction may reverse by scaling the data [44]. This problem, in particular, is more severe for multi-nature data that are physically different and are not comparable. For example, for a set of physiological data consisting of the 'heart rate' (H) and 'breath rate' (B) time series, different scalings alter the inferred direction of coupling (Fig. 2.4) [44].

It is noteworthy to mention that the studies of [47] demonstrates the relationship between the Granger causality, directed information, and transfer entropy.

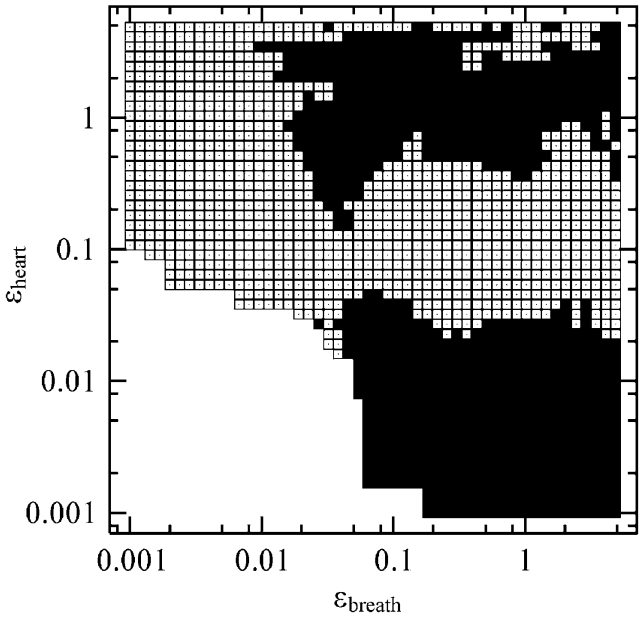


Figure 2.4: Direction of coupling derived by TE for different scalings between 'heart rate' (H) and 'breath rate' (B) time series (ϵ_{heart} and ϵ_{breath} represent the scaling factors of H and B, respectively): ■ H→B □ B→H [44].

Chapter 3

Coupling Spectrum Method

3.1 Introduction

In the fields of nonlinear dynamics and time series analysis, identifying the couplings between the sub-systems of a complex system is essential for understanding the functionality of the interactions and controlling them. This area of study has immediate applications in various disciplines, such as physics, economics, biology, physiology, ecology, and climatology.

The challenging issue in identifying the interactions is that the couplings are inherently hidden in the underlying dynamics of the system and are not necessarily accessible. That is to say, we commonly have access to some observations of the system measured in time series. Hence, methods that based on the available observations determine whether these time series originate from coupled or decoupled sub-systems are much needed. This task will be more complicated if the goal is to determine the direction of coupling.

Methods that are proposed to detect the directed coupling can be divided into four categories: (i) regression-based methods (Granger causality) [16, 31]; (ii) state-space reconstruction methods [24,32]; (iii) phase synchronization approaches [34,35]; (iv) information theoretic methods [36, 37, 48]. These methods are compared in [40, 49,50]. Comparisons in [40] between state-space and phase-synchronization methods show that the phase synchronization methods rely on a meaningful definition of the phase and this is not always accessible in real data. Hence, between these two methods, state-space approaches are generally preferred. Also, studies of [49] suggest that transfer entropy (TE) [36], which is an information theoretic method, is generally more reliable than regression-based and state-space methods.

In real-world problems, we usually deal with systems composed of non-identical

sub-systems with fundamentally different structures. The data derived from these systems may have different dynamics with non-comparable scales. Some methods need to scale data sets to make them comparable; for example, the data are normalized to unit variance. However, in some cases, the scaling reverses the direction of the inferred coupling. For instance, [44] shows that the direction of coupling obtained by TE may alter by data scaling. Furthermore, in many cases, the available sample size is small and the strength of coupling is very weak, making most of the proposed methods erroneous.

To resolve the mentioned difficulties of the real-world data, inspired by [51], we propose a new method to identify the direction of coupling[1]. The work of [51] proposes a method to identify the dependencies between the samples of a single time series. Here, we generalize this method to discover the direction of couplings between two or more time series. According to different kinds of causal relationships and definitions described in Chapter 1, this method is developed for inference of deterministic causality based on the definition of predictive causality. Moreover, the simulation results show that it is not applicable for detection of distributional causality. Generally, this method can be considered as a member of the state-space reconstruction methods which are working based on the closeness of the points in the driver and response spaces.

The simulation results show that our method can detect the direction of coupling correctly for identical and non-identical sub-systems, nonlinear dynamics, small sample sizes, as well as weak coupling strength. The proposed method can also detect the stronger couplings in asymmetric bidirectional couplings. Moreover, our method is invariant to data scaling. Same as other kinds of state-space reconstruction methods, the proposed method is sensitive to noise.

In this chapter, we first propose our method for a simple bivariate coupling system and show its successful performance. Then we generalize our method for multivariate systems.

3.2 Bivariate Coupling

Consider a finite time series $\{r_t\}$, consisting of N samples and define the time-delayed vector $\mathbf{R}_{t-1} = (r_{t-1}, r_{t-2}, ..., r_{t-L_r})$ with the maximum lag value L_r. We use

[1]The results of this work were published as a journal paper entitled '*Finding weak directional coupling in multi-scale time Series*' in Physical Review E, vol. 86, no. 1, pp. 16215, Jul. 2012.

the maximum norm to measure the distance between $\mathbf{R}_{t-1}$ and $\mathbf{R}_{t'-1}$, i.e.,

$$\rho^R_{tt'} = \|\mathbf{R}_{t-1} - \mathbf{R}_{t'-1}\| = \max_{1\le k\le L_r} \{|r_{t-k} - r_{t'-k}|\}. \tag{3.1}$$

Accordingly, we define

$$\rho^r_{tt'} = |r_t - r_{t'}|. \tag{3.2}$$

Similarly, for another time series $\{d_t\}$, we can define $\mathbf{D}_{t-1}$ with the maximum lag value L_d, $\rho^D_{tt'}$, and $\mathbf{D}_{t'-1}$ and $\rho^d_{tt'}$.

Consider a coupling system consisting of the driver system D and the response system R with corresponding samples $\{d_t\}$ and $\{r_t\}$, respectively. The goal is to identify the existing directed coupling between these two systems from observed samples. If a coupling exists from D to R, denoted by $D \to R$, the current sample of R should be predictable by the past samples of R and D, i.e.,

$$r_t = f(\mathbf{R}_{t-1}, \mathbf{D}_{t-1}) + \eta_t. \tag{3.3}$$

where $f(\cdot)$ is a continuous function and $\partial f/\partial \mathbf{D}_{t-1} \neq 0$. Here, η_t denotes the indeterminable part which originates from the real noise or insufficient considered dimension of the system [51]. To characterize the dimensions of the system correctly, it is sufficient that L_d and L_r be greater than the minimum lag values of D and R, respectively, and in this case, η_t vanishes for a noiseless scenario. Hereafter, we assume that there is no noise.

If the distance between $\mathbf{R}_{t-1}$ and $\mathbf{R}_{t'-1}$ is smaller than $\delta^r > 0$, i.e., $\rho^R_{tt'} < \delta^r$, and provided that the distance of $\mathbf{D}_{t-1}$ from $\mathbf{D}_{t'-1}$ is smaller than $\delta^d > 0$, i.e., $\rho^D_{tt'} < \delta^d$, then the probability that the distance between the corresponding outputs of (3.3), i.e., r_t and $r_{t'}$, is smaller than a fixed value $\epsilon^r_o > 0$ is denoted by

$$P(\epsilon^r_o|\delta^r, \delta^d) = P(\rho^r_{tt'} < \epsilon^r_o | \rho^R_{tt'} < \delta^r, \rho^D_{tt'} < \delta^d). \tag{3.4}$$

Studying the behavior of $P(\epsilon^r_o|\delta^r, \delta^d)$ as a function of δ^r and δ^d sheds light on developing a method for identifying the directed coupling $D \to R$.

1. If $\delta^d \to \infty$, then $P(\epsilon^r_o|\delta^r, \delta^d \to \infty) = P(\epsilon^r_o|\delta^r)$.

2. Provided that $D \to R$, for a fixed $\delta^r = \delta^r_o$, by increasing the distance of $\mathbf{D}_{t-1}$ from $\mathbf{D}_{t'-1}$, the probability that r_t stays in the ϵ^r_o neighborhood of $r_{t'}$ reduces. Hence, $P(\epsilon^r_o|\delta^r_o, \delta^d)$ decreases monotonically as δ^d increases.

3. Lack of coupling from D to R, denoted by $D \nrightarrow R$, yields $P(\epsilon_o^r|\delta^r, \delta^d) = P(\epsilon_o^r|\delta^r)$.

According to the above statements 2 and 3, the existence of coupling from D to R can be verified by the following rules for a fixed $\delta^r = \delta_o^r$.

$\mathrm{R}_{\rightarrow}$: $P(\epsilon_o^r|\delta_o^r, \delta^d)$ is a decreasing function of $\delta^d \Rightarrow\ D \rightarrow R$

$\mathrm{R}_{\nrightarrow}$: $P(\epsilon_o^r|\delta_o^r, \delta^d)$ does not vary by $\delta^d \Rightarrow\ D \nrightarrow R$

We can visualize $P(\epsilon_o^r|\delta^r, \delta^d)$ by a two-dimensional representation, here referred to as coupling spectrum (CS), denoted by $\mathrm{CS}(D \rightarrow R)$ (see Figs. 3.1(a) and 3.1(b)). In CS, the value of the conditional probability is mapped to a color for each pair of (δ^r, δ^d). The horizontal and vertical axes of $\mathrm{CS}(D \rightarrow R)$ correspond to δ^r and δ^d, respectively. Therefore, if we observe a change of color in each column of the CS, meaning that $\mathrm{R}_{\rightarrow}$ is satisfied, we conclude that the coupling $D \rightarrow R$ exists. Otherwise, if all the columns of the CS lack the color change, the rule $\mathrm{R}_{\nrightarrow}$ is satisfied meaning that the coupling does not exist.

We now explain how to determine the fixed value of ϵ_o^r in $P(\epsilon_o^r|\delta^r, \delta^d)$. Obviously, for specific values of δ^r and δ^d, $P(\epsilon_o^r|\delta^r, \delta^d)$ increases with ϵ_o^r and this probability saturates to 1 for large value of ϵ_o^r and small values of δ^r and δ^d. By generalizing the discussion of [51], in presence of noise, $P(\epsilon_o^r|\delta^r, \delta^d)$ will not saturate to 1 as ϵ_o^r drops below $\Delta\eta_{\max}$ where $\Delta\eta_{\max}$ is the maximum distance of the noise samples, i.e., $\max\{|\eta_t - \eta_{t'}|\}$. Indeed, $\epsilon_o^r > \Delta\eta_{\max}$ provides more space for the fluctuation of the neighbor points due to the noise. Hence, to determine the value of ϵ_o^r, we try to have the saturation level of 1 in the CS to be able to tolerate the noise (or we try to find a value of ϵ_o^r that results the closest saturation level to 1). On the other hand, for very large values of ϵ_o^r, $P(\epsilon_o^r|\delta^r, \delta^d)$ saturates approximately to 1 for all values of δ^r, δ^d. Hence, ϵ_o^r should be neither too larger nor too small. Thus, we first consider the smallest value of ϵ_o^r for which the maximum of $P(\epsilon_o^r|\delta^r, \delta^d)$ equals 1. However, in that case, it is possible $P(\epsilon_o^r|\delta^r, \delta^d)$ to be close to 1 for all values of δ^r and δ^d. Hence, no color change will be visible in $\mathrm{CS}(D \rightarrow R)$. To prevent this situation, we further reduce ϵ_o^r such that the minimum of $P(\epsilon_o^r|\delta^r, \delta^d)$ is smaller than a threshold, e.g., 0.75. This way, the change of color in each column becomes more visible in the presence of any existing coupling. It is noteworthy to mention that in presence of noise, generally, the value of ϵ_o^r becomes larger that that of the noiseless scenario.

Moreover, $\rho^D \leq \max\{\rho^d\}$ and $\rho^R \leq \max\{\rho^r\}$; hence, there is no need to consider $\delta^d > \max\{\rho^d\}$ and $\delta^r > \max\{\rho^r\}$.

Now, let us provide numerical results. Firstly, counting is used to calculate $P(\epsilon_o^r|\delta^r, \delta^d)$ from the time series, i.e.,

$$P(\epsilon_o^r|\delta^r, \delta^d) = \frac{n(\rho_{tt'}^r < \epsilon_o^r, \rho_{tt'}^R < \delta^r, \rho_{tt'}^D < \delta^d)}{t(\rho_{tt'}^R < \delta^r, \rho_{tt'}^D < \delta^d)} \tag{3.5}$$

where $n(\cdot)$ is the number of pairs satisfying the distance constraints.

Secondly, to avoid statistical fluctuations, we disregard values of $P(\epsilon_o^r|\delta^r, \delta^d)$ derived by $n(\cdot) < n_{\min}$ in (3.5). Finally, as mentioned earlier, TE generally outperforms other existing methods. Hence, we compare CS against TE. For estimation of the TE, we use the kernel density estimator (2.33). If a specific value is not mentioned for h_r and h_d in simulation results, the optimal bandwidth of normal distribution is used.

As an example, consider an identically coupled system realized by the Hénon map [29], given by:

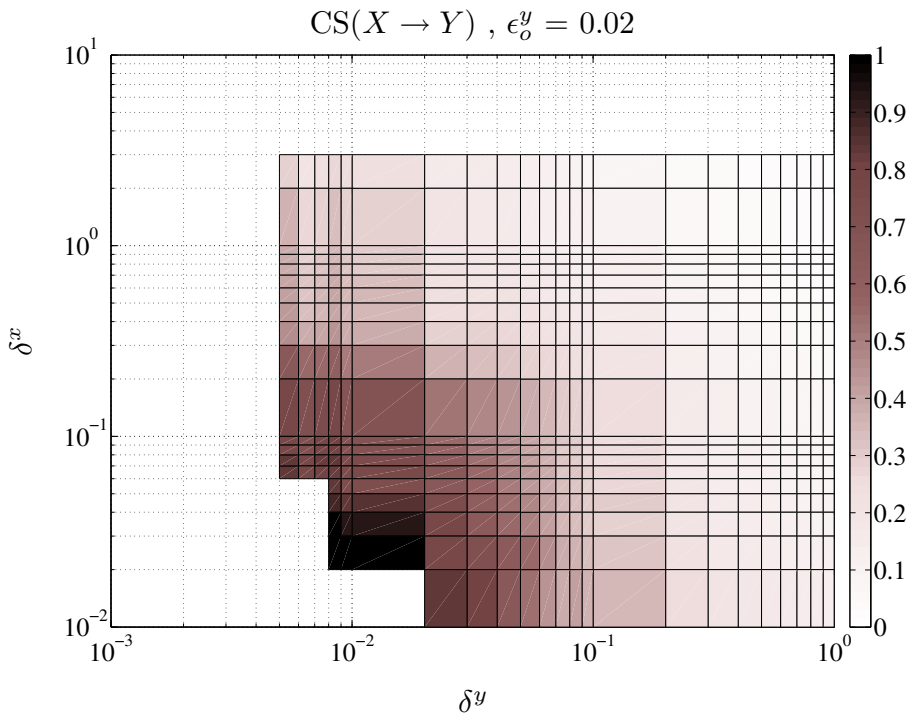
$$x_t = a - x_{t-1}^2 + bx_{t-2} + c_{y\to x}(x_{t-1}^2 - y_{t-1}^2) \tag{3.6a}$$

$$y_t = a - y_{t-1}^2 + by_{t-2} + c_{x\to y}(y_{t-1}^2 - x_{t-1}^2) \tag{3.6b}$$

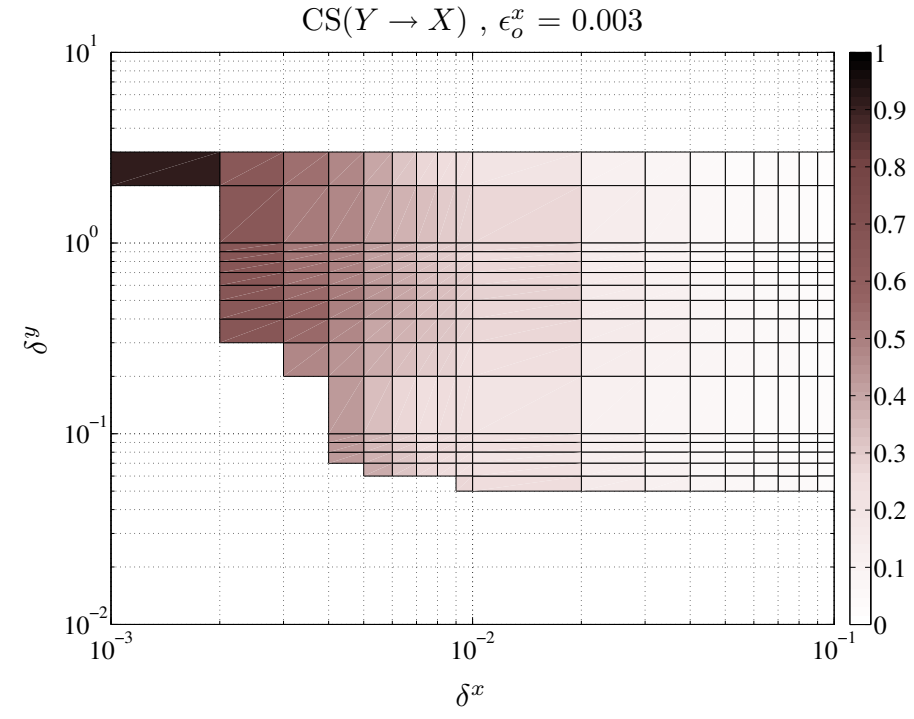
where $a = 1.4$, $b = 0.3$, $N = 1000$, $L_x = L_y = 2$, and $n_{\min} = 10$. The initial values of x_0 and y_0 are uniformly distributed in $[0, 0.5]$. To have a unidirectional coupling $X \to Y$, we set the coupling strengths $c_{y\to x} = 0$ and $c_{x\to y} = 0.2$. As Fig. 3.1(a) shows, for $\delta^y < 0.2$, the color of each column in $\mathrm{CS}(X \to Y)$ changes drastically by variation of δ^x, which means the dependency of $P(\epsilon_o^y|\delta_o^y, \delta^x)$ to δ^x. Hence, $\mathrm{CS}(X \to Y)$ confirms the existence of coupling from X to Y. On the contrary, as Fig. 3.1(b) reveals, the color of no column in $\mathrm{CS}(Y \to X)$ changes with δ^y, which indicates $Y \nrightarrow X$.

Since variation of color over the columns of the CS is a qualitative measure, we suggest using a quantitative measure of variability, e.g., standard deviation. In fact, we calculate the standard deviation of the values of $P(\epsilon_o^r|\delta_o^r, \delta^d)$ in each column for different values of δ^d. This measure is denoted by σ_{cs}. The corresponding σ_{cs} of Figs. 3.1(a) and 3.1(b) are plotted in Fig. 3.2 where $\sigma_{\mathrm{cs}}(X \to Y)$ is considerably greater than $\sigma_{\mathrm{cs}}(Y \to X)$. Hence, one concludes that $X \to Y$ exists.

For non-identical and structurally different systems, a direct comparison of $\sigma_{\mathrm{cs}}(D \to R)$ and $\sigma_{\mathrm{cs}}(R \to D)$ maybe become meaningless. This means that we need a way to measure the significance of each standard deviation individually. Hence,



(a) The change of color over each column for $\delta^y < 0.2$ shows the existence of $X \to Y$.



(b) The color of each column is fixed, indicating the lack of coupling from Y to X.

Figure 3.1: The coupling spectrum (CS) for the coupling $X \to Y$ realized by the unidirectional Hénon map

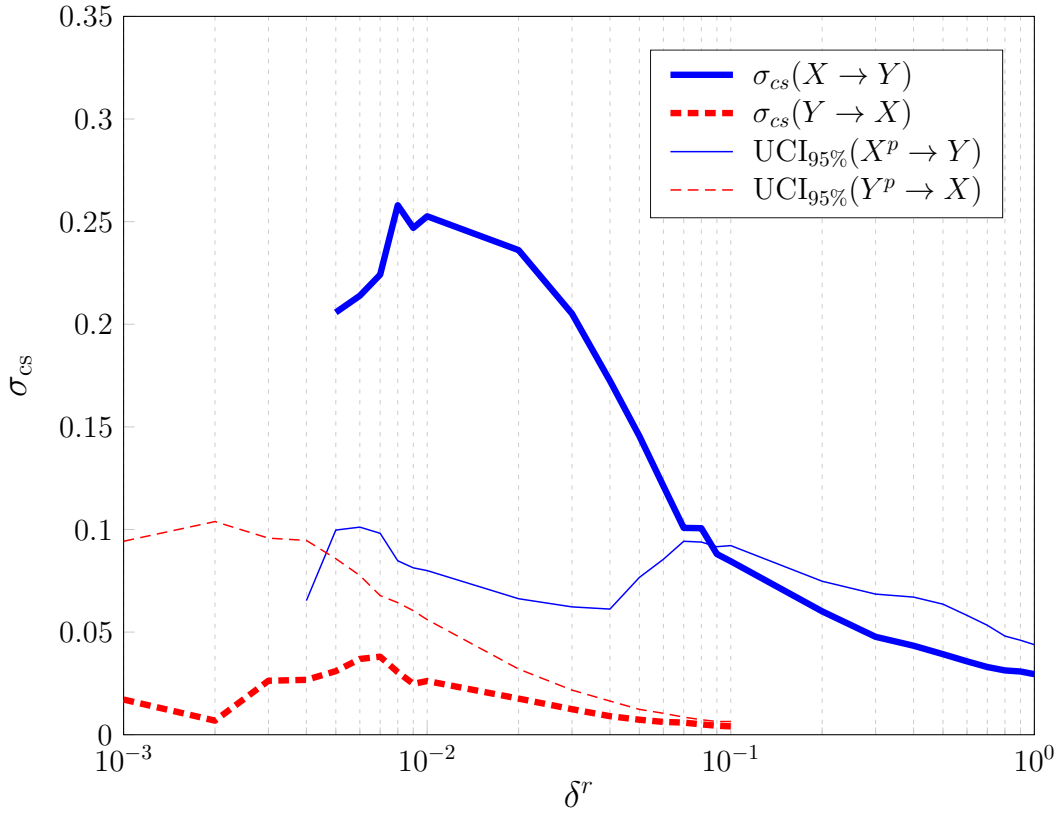


Figure 3.2: The standard deviation of each column of CS (σ_{cs}) is shown for both directions of coupling between X and Y for the coupling $X \to Y$ realized by the Hénon map. The significance of σ_{cs} is assessed by the 95% confidence interval derived by permuting the driver time series.

we use the permuted $\{d_t\}$, represented by $\{d_t^p\}$, to destroy any inherent coupling hidden in the dynamics of the system. To assess the significance of $\sigma_{\text{cs}}(D \to R)$, we obtain the 95% confidence interval ($\text{CI}_{95\%}$) of $\sigma_{\text{cs}}(D^p \to R)$ based on the percentile bootstrap method[1] [52]. Provided that $\sigma_{\text{cs}}(D \to R)$ lies outside the $\text{CI}_{95\%}$, we confirm that $\sigma_{\text{cs}}(D \to R)$ is significant, i.e., $D \to R$; otherwise, that coupling does not exist. Let us return to the unidirectional Hénon map explained above and illustrate the upper bound of the $\text{CI}_{95\%}$ ($\text{UCI}_{95\%}$) with the permuted time series of the driver in Fig. 3.2. Obviously, $\sigma_{\text{cs}}(X \to Y)$ for $\delta^y < 0.09$ is higher than $\text{UCI}_{95\%}(X^p \to Y)$; consequently, the hypothesis $X \to Y$ is accepted. Conversely, for all values of δ^x, $\sigma_{\text{cs}}(Y \to X)$ lies lower than $\text{UCI}_{95\%}(Y^p \to X)$; hence, $Y \nrightarrow X$.

To illustrate the capability of the CS method to detect the asymmetric bidirectional coupling, consider $c_{y\to x} = 0.13$ and $c_{x\to y} = 0.07$ in (3.6). Thus, the coupling $Y \to X$ is stronger than $X \to Y$. Fig. 3.3 shows that both $\sigma_{\text{cs}}(X \to Y)$ and $\sigma_{\text{cs}}(Y \to X)$ are greater than their corresponding $\text{UCI}_{95\%}$; however, the maximum of $\sigma_{\text{cs}}(Y \to X)$ is larger than that of $\sigma_{\text{cs}}(X \to Y)$, which indicates that the coupling $Y \to X$ is stronger than $X \to Y$.

Let us investigate the effect of the sample size on the CS method for the unidirectional Hénon map described above. For each value of N, $\sigma_{\text{cs}}(X \to Y)$ and $\sigma_{\text{cs}}(Y \to X)$ are computed for 1000 different initial values of x_0 and y_0 in (3.6). The mean of $\max\{\sigma_{\text{cs}}\}$ over 1000 trials is plotted in Fig. 3.4, where the error bars indicate the corresponding standard deviations. Fig. 3.4 reveals that the error bars are distinctly separated for $N \geq 80$. Repeating this simulation with TE reveals that $N \geq 80$ is required again for error bars to be separated. Thus, CS is as strong as TE in handling small sample size.

One important feature of the CS method is that it suits non-identical processes with fundamentally different structures. Let us consider a non-identical coupling system that the driver is a three-dimensional discrete-time Rössler hyperchaotic system [53] defined by

[1] In percentile bootstrap method, the time series $\{d_t\}$ is permuted N_p times . For each permuted time series we calculate the corresponding $\sigma_{\text{cs}}(D^p \to R)$, and then, these values are sorted. The upper bound of $\alpha\%$ confidence interval is the $(1 - \frac{\alpha}{2})$ percentile of $\sigma_{\text{cs}}(D^p \to R)$ values, i.e., the value below which $100 \times (1 - \frac{\alpha}{2})$ percentage of $\sigma_{\text{cs}}(D^p \to R)$ values fall which is the value of the $\lceil (1 - \frac{\alpha}{2})N_p \rceil$-th sorted $\sigma_{\text{cs}}(D^p \to R)$.

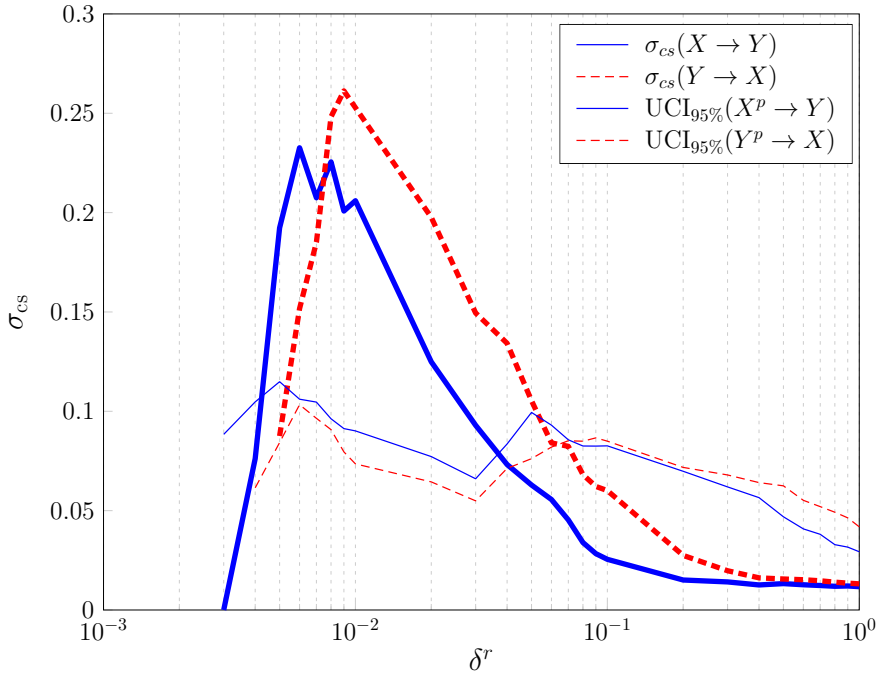


Figure 3.3: The CS method is applied to a bidirectional coupling between X and Y in which the coupling from Y to X is stronger. The CS method shows a significant coupling for both directions as well as the larger maximum of σ_{cs} for the stronger coupling, i.e., $Y \to X$.

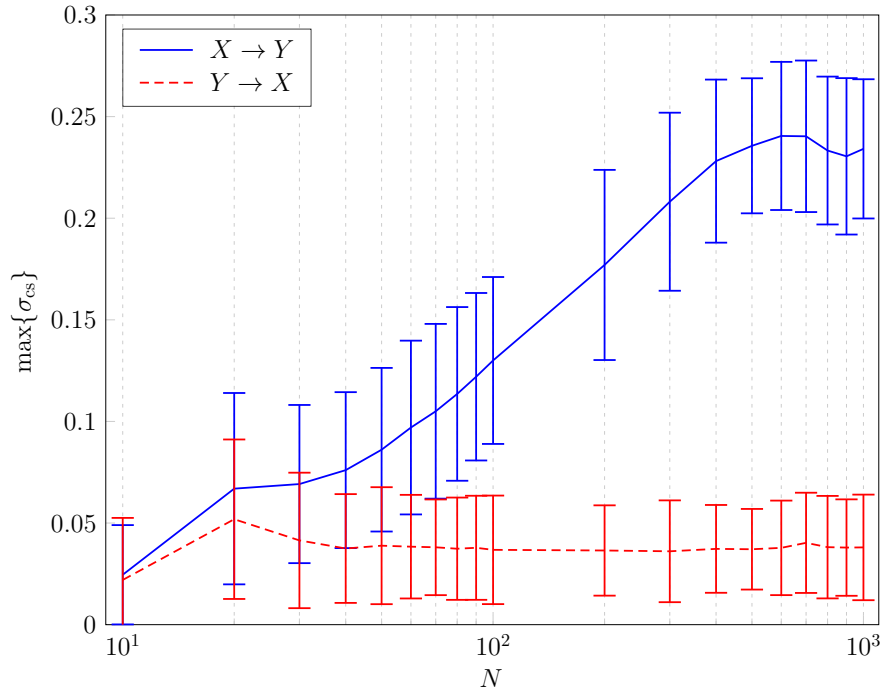


Figure 3.4: The maximum of $\sigma_{\mathrm{cs}}(X \to Y)$ and $\sigma_{\mathrm{cs}}(Y \to X)$ are depicted against the number of the samples for a unidirectional Hénon coupling $X \to Y$. The CS method can identify the direction of coupling correctly for $N \geq 80$.

$$x_t = \alpha x_{t-1}(1 - x_{t-1}) - \beta(z_{t-1} + \gamma)(1 - 2y_{t-1}) \tag{3.7a}$$

$$y_t = \delta y_{t-1}(1 - y_{t-1}) + \zeta z_{t-1} \tag{3.7b}$$

$$z_t = \eta\left[(z_{t-1} + \gamma)(1 - 2y_{t-1}) - 1\right](1 - \theta x_{t-1}) \tag{3.7c}$$

and the response system is the two-dimensional discrete Lorenz system [53]

$$x'_t = (1 + ab)x'_{t-1} - b\ x'_{t-1}y'_{t-1} \tag{3.8a}$$

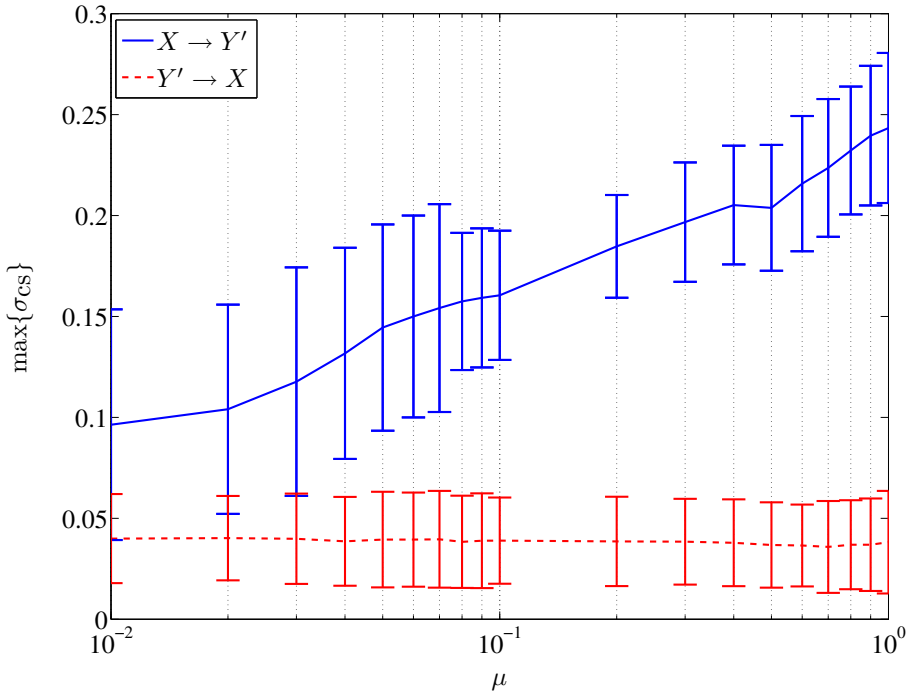
$$y'_t = (1 - b)y'_{t-1} + b\ {x'_{t-1}}^2 + \mu x_{t-1}. \tag{3.8b}$$

The driving term μx_{t-1} is applied to (3.8b) of the response system where μ is the coupling strength. The parameters of the Rössler system are set to $\alpha = 1.25$, $\beta = 0.75$, $\gamma = 0.35$, $\delta = 3.78$, $\zeta = 0.2$, $\eta = 0.1$, and $\theta = 1.9$ and the parameters of the Lorenz system are $a = 1.25$, and $b = 0.75$. To have a stable coupling, the initial conditions of (3.7) and (3.8) are uniformly distributed as x_0 and $y_0 \in [0.5, 1]$, $z_0 \in [0, 0.25]$, and x'_0 and $y'_0 \in [0, 0.5]$. Two variables X and Y' are observed and both L_x and $L_{y'}$ are set to 2. Fig. 3.5(a) depicts the maximum of $\sigma_{\text{cs}}(X \to Y')$ and $\sigma_{\text{cs}}(Y' \to X)$ versus μ over 500 distinct realizations with $N = 1000$. The error bars show the standard deviation of $\max\{\sigma_{\text{cs}}\}$ for each value of μ. It is obvious in Fig. 3.5(a) that the error bars are separated for $\mu > 0.03$. For the similar scenario of simulation for TE, Fig. 3.5(b) shows that the minimum coupling strength that is detectable by TE is 10 times greater than that of CS. Hence, the CS method is more applicable for weak directional couplings.

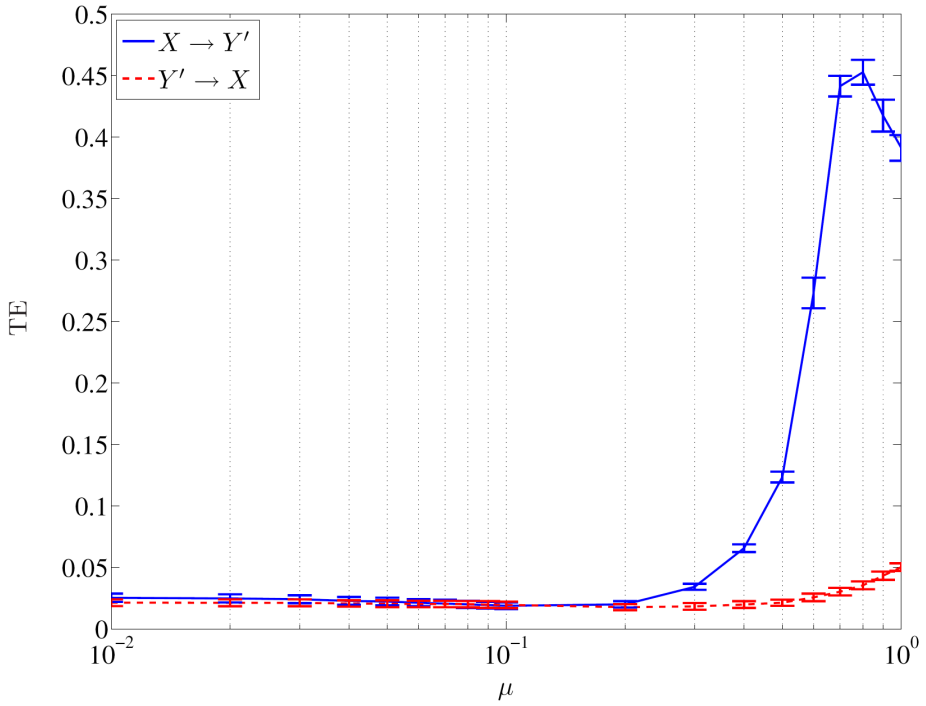
Invariance to scaling is an important feature when working with multi-scale data, specially when the data is derived from structurally different systems. Interestingly, σ_{cs} is independent of scaling. If two time series $\{d_t\}$ and $\{r_t\}$ are scaled as $d'_t = \alpha d_t$ and $r'_t = \beta r_t$, then we have $\rho^{D'}_{tt'} = \alpha\rho^{D}_{tt'}$, $\rho^{R'}_{tt'} = \beta\rho^{R}_{tt'}$, and $\rho^{r'}_{tt'} = \beta\rho^{r}_{tt'}$. Consequently, it can be shown that

$$P(\rho^{r'}_{tt'} < \epsilon^{r'}_o | \rho^{R'}_{tt'} < \delta^{r'}, \rho^{D'}_{tt'} < \delta^{d'}) = P(\epsilon^{r}_o | \delta^{r}, \delta^{d}) \tag{3.9}$$

where $\epsilon^{r'}_o = \beta\epsilon^{r}_o$, $\delta^{r'} = \beta\delta^{r}$, and $\delta^{d'} = \alpha\delta^{d}$. Therefore, $\sigma_{\text{cs}}(D' \to R')$ obtained from $P(\epsilon^{r'}_o | \delta^{r'}, \delta^{d'})$ is equal to $\sigma_{\text{cs}}(D \to R)$. Here, the CS method is compared with TE for different relative scalings of data. As it is mentioned in [44], the relative scaling of two samples $\{d_t\}$ and $\{r_t\}$ is equivalent to varying the corresponding bandwidth of each data set in the kernel function, i.e., h_d and h_r in (2.33). Similarly, $\{d_t\}$ and



(a) CS



(b) TE

Figure 3.5: The performances of the CS and TE methods are depicted for the non-identical discrete R?ssler-Lorenz system against the coupling strength.

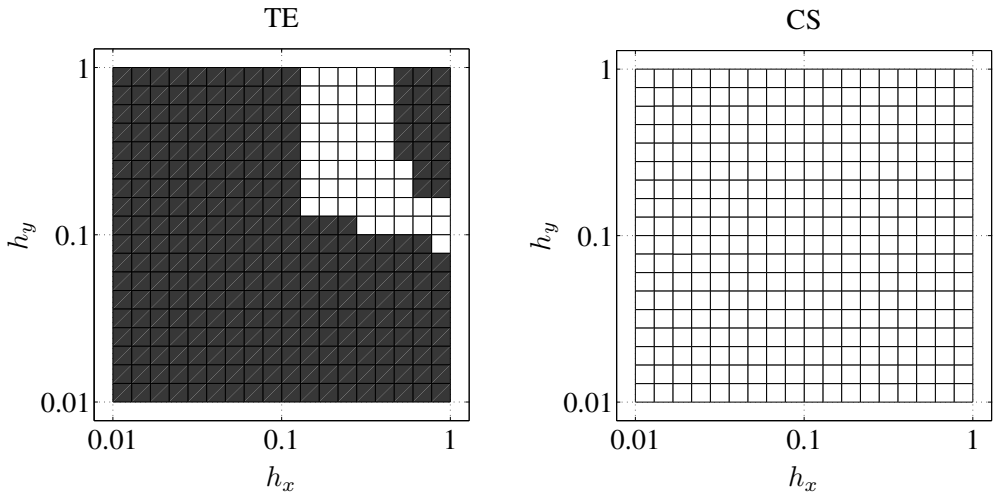


Figure 3.6: TE and CS methods are compared for different scalings of unidirectional Hénon map data. □: $X \to Y$ and $Y \nrightarrow X$ are deduced; ■: wrong directions of couplings are detected.

$\{r_t\}$ are scaled with h_d and h_r, respectively, to be used in the CS method. The unidirectional Hénon map $X \to Y$ corresponding to Fig. 3.1 with different scalings h_x and h_y is considered here. The permuting method is performed to evaluate the significance of the results. In Fig. 3.6, white boxes denote regions where a correct coupling direction is detected, i.e., $X \to Y$ and $Y \nrightarrow X$ are deduced. Black boxes represent regions over which the direction of couplings between X and Y is wrongly detected. Fig. 3.6 reveals that the TE method finds the correct direction of coupling just in a specific range of scaling, however, the CS method performs successfully for the whole range.

In the presence of noise, $\sigma_{\rm cs}(X \to Y)$ corresponding to the coupling $X \to Y$ reduces and becomes closer to $\sigma_{\rm cs}(Y \to X)$, and consequently, they are not distinguishable for strong noise. To examine the effect of noise on the CS and TE methods, the unidirectional Hénon map explained above with power σ_s contaminated by additive white Gaussian noise with power σ_η is considered. In Fig. 3.7, success rate represents the probability of $\max\{\sigma_{\rm cs}(X \to Y)\} > \max\{\sigma_{\rm cs}(Y \to X)\}$ for CS method and probability of ${\rm TE}(X \to Y) > {\rm TE}(Y \to X)$ for TE method. Fig. 3.7 depicts the success rate for each noise to signal power ratio (NSR), i.e. σ_η/σ_s, where 500 different noisy signals are generated for each value of NSR. For CS method, the direction of coupling is detected with a probability greater than 0.9 for NSR smaller than 0.12 and 0.06 with 1000 and 100 samples, respectively. For the same level of

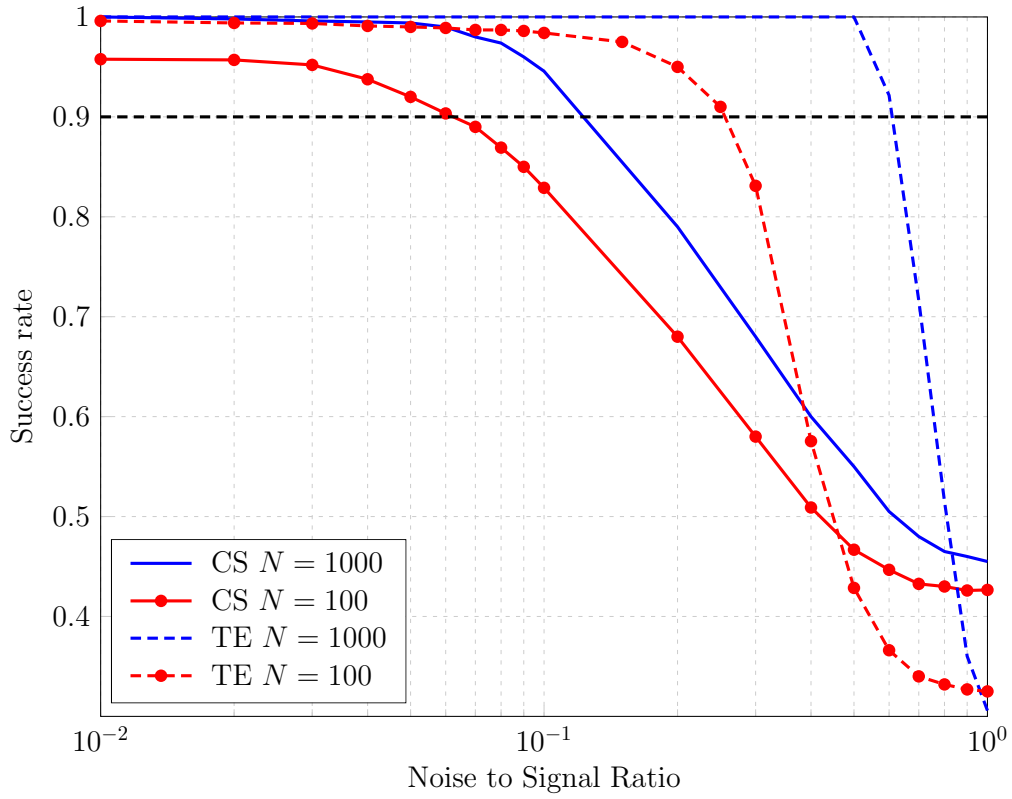


Figure 3.7: The effect of noise on the CS and TE methods are depicted for $X \to Y$ realized by the Hénon map with 100 and 1000 samples. The success rate shows the probability of $\max\{\sigma_{\text{cs}}(X \to Y)\} > \max\{\sigma_{\text{cs}}(Y \to X)\}$ for CS method and probability of $\text{TE}(X \to Y) > \text{TE}(Y \to X)$ for TE method.

success rate, the maximum NSR of the TE method is 4-5 times greater than that of the CS method.

It is noteworthy that for calculation of bivariate CS, we should calculate $P(\epsilon_o^r|\delta^r, \delta^d)$ for a mesh grid of δ^r and δ^d values. Therefore, if we have N_δ values for each of δ^r and δ^d, $P(\epsilon_o^r|\delta^r, \delta^d)$ is calculated N_δ^2 times. However, it is possible to optimize the implementation of the code to prevent a whole calculation for each grid point. In our implementation, the execution time is in the order of seconds on a 2.8 GHz CPU with C-MEX implementation in MATLAB® without any optimization in the code. For example, for 1000 samples, the calculation of CS and σ_{cs} takes around 35 seconds. In comparison, the TE method calculates the TE just for specific values of the bandwidths of the kernel estimator and it is faster. However, as we showed in Fig. 3.6, we do not know whether the selected bandwidths result in the correct direction of coupling. Thus, despite the low computational cost of TE, the results are not reliable for multi-scale data.

3.3 Multivariate Coupling

Now, we modify the CS method to discover the couplings in a multivariate system. The proposed CS method can be used for each pair of variables in a multivariate system. Indirect coupling commonly appears when a pairwise measure is used to detect the couplings between more than two processes. For example, consider a coupled system as $R_1 \leftarrow D \rightarrow R_2$. Aside from the obvious couplings $D \rightarrow R_1$ and $D \rightarrow R_2$, two indirect couplings represented by $R_1 \dashrightarrow R_2$ and $R_2 \dashrightarrow R_1$ are also observed. Indeed, this indirect couplings are due to the common driver D hidden in the underlying dynamics of both R_1 and R_2. As another example, the indirect coupling $D \dashrightarrow R_2$ may be seen in $D \rightarrow R_1 \rightarrow R_2$.

To deal with indirect couplings, we generalize the bivariate measure of the CS method to a triple-variate measure. Consider a coupling system as $D \leftarrow C \rightarrow R$ that the CS method detects C and D as the drivers of R. The goal is to examine whether the couplings $C \rightarrow R$ and $D \rightarrow R$ are direct or indirect. We can exclude the influence of the common driver C in $P(\epsilon_o^r|\delta_o^r, \delta^d)$ by conditioning on C, i.e., $P(\epsilon_o^r|\delta_o^r, \delta_o^c, \delta^d)$. Hence, in the presence of C in the conditional probability, all the potential couplings from D to R are covered by C, and consequently, D becomes ineffective and we have $P(\epsilon_o^r|\delta_o^r, \delta_o^c, \delta^d) = P(\epsilon_o^r|\delta_o^r, \delta_o^c)$. In contrast, if the coupling $D \rightarrow R$ really exists, $P(\epsilon_o^r|\delta_o^r, \delta_o^c, \delta^d)$ depends on δ^d and varies by it. Therefore, we can modify the rules $\mathrm{R}_{\rightarrow}$ and $\mathrm{R}_{\nrightarrow}$ for a triple-variate system for the fixed values of $\delta^r = \delta_o^r$ and $\delta^c = \delta_o^c$ as

$\mathrm{R}'_{\rightarrow}$**:** $P(\epsilon_o^r|\delta_o^r, \delta_o^c, \delta^d)$ is a decreasing function of $\delta^d \Rightarrow\ D \rightarrow R|C_o$

$\mathrm{R}'_{\nrightarrow}$**:** $P(\epsilon_o^r|\delta_o^r, \delta_o^c, \delta^d)$ does not vary by $\delta^d \Rightarrow\ D \nrightarrow R|C_o$

where conditioning on C_o represents the conditioning on C for a specific $\delta^c = \delta_o^c$.

Likewise $P(\epsilon_o^r|\delta^r, \delta^d)$, $P(\epsilon_o^r|\delta^r, \delta^c, \delta^d)$ can be represented by a three-dimensional coupling spectrum, denoted by $\mathrm{CS}(D \rightarrow R|C)$, where the x-axis, y-axis, and z-axis correspond to δ^r, δ^d, and δ^c, respectively. In fact, on each xy-plane of $\mathrm{CS}(D \rightarrow R|C)$ a two-dimensional CS for $D \rightarrow R$ with corresponding $\delta^c = \delta_o^c$ can be seen, which is denoted by $\mathrm{CS}(D \rightarrow R|C_o)$. Hence, the changes of the color in each column on the plane, which can be quantified by the concept of σ_{cs} denoted by $\sigma_{\mathrm{cs}}(D \rightarrow R|C_o)$, determines whether $\mathrm{R}'_{\rightarrow}$ or $\mathrm{R}'_{\nrightarrow}$ is satisfied. To accept the hypothesis $D \rightarrow R$, $\sigma_{\mathrm{cs}}(D \rightarrow R|C_o)$ should be independent of C_o for all the planes, i.e., $\mathrm{R}'_{\rightarrow}$ should be met for all

values of δ_o^c. Therefore, it is required to check the satisfaction of $\mathrm{R}_{\rightarrow}^{'}$ in all the xy-planes of $\mathrm{CS}(D \to R|C)$. Provided that $\mathrm{R}_{\rightarrow}^{'}$ is satisfied in all the planes, existence of coupling from D to R is accepted, otherwise the observed coupling is indirect. As an illustration, consider a coupling system $X \leftarrow Z \rightarrow Y$ achieved by a three-variable Hénon map as follows

$$z_t = a - z_{t-1}^2 + bz_{t-2} \tag{3.10a}$$

$$x_t = a - x_{t-1}^2 + bx_{t-2} + c_{zx}(x_{t-1}^2 - z_{t-1}^2) \tag{3.10b}$$

$$y_t = a - y_{t-1}^2 + by_{t-2} + c_{zy}(y_{t-1}^2 - z_{t-2}^2) \tag{3.10c}$$

where Z affects Y by one delay more than X. Here, $a = 1.4$, $b = 0.3$, $c_{zx} = c_{zy} = 0.2$, $L_x = L_y = L_z = 2$, and $N = 5000$. As Fig. 3.8(a) represents, $\sigma_{\mathrm{cs}}(Z \to Y)$ is large, indicating a coupling from Z to Y. Moreover, for all values of δ^x, $\sigma_{\mathrm{cs}}(Z \to Y|X_o)$ fluctuates in close proximity of $\sigma_{\mathrm{cs}}(Z \to Y)$, which means that $\mathrm{R}_{\rightarrow}^{'}$ is satisfied for all the planes. In other words, the coupling from Z to Y is independent of X, and consequently, $Z \to Y$ is a direct coupling. Similarly, Fig. 3.8(b) shows that $\sigma_{\mathrm{cs}}(X \to Y)$ represents a significant coupling from X to Y in the lack of conditioning on Z. Albeit, for $\delta^z \to \infty$, $\sigma_{\mathrm{cs}}(X \to Y|Z_o)$ equals $\sigma_{\mathrm{cs}}(X \to Y)$. The effect of Z is more severe when δ^z reduces and as Fig. 3.8(b) shows, smaller δ^z makes $\sigma_{\mathrm{cs}}(X \to Y|Z_o)$ less significant. Hence, Z affects $X \to Y$, and thus, this coupling is indirect.

3.4 Relationship Between CS and HJ

Before finishing this chapter, it is noteworthy to mention the relationship between the CS and HJ methods. In Section 2.2.2 we introduced the HJ test as a nonlinear extension of the G-causality method. Here, we show that the HJ test is a specific case of the CS method and explain the common underpinnings of both the HJ test and the CS method. The CS method is developed based on the following probability

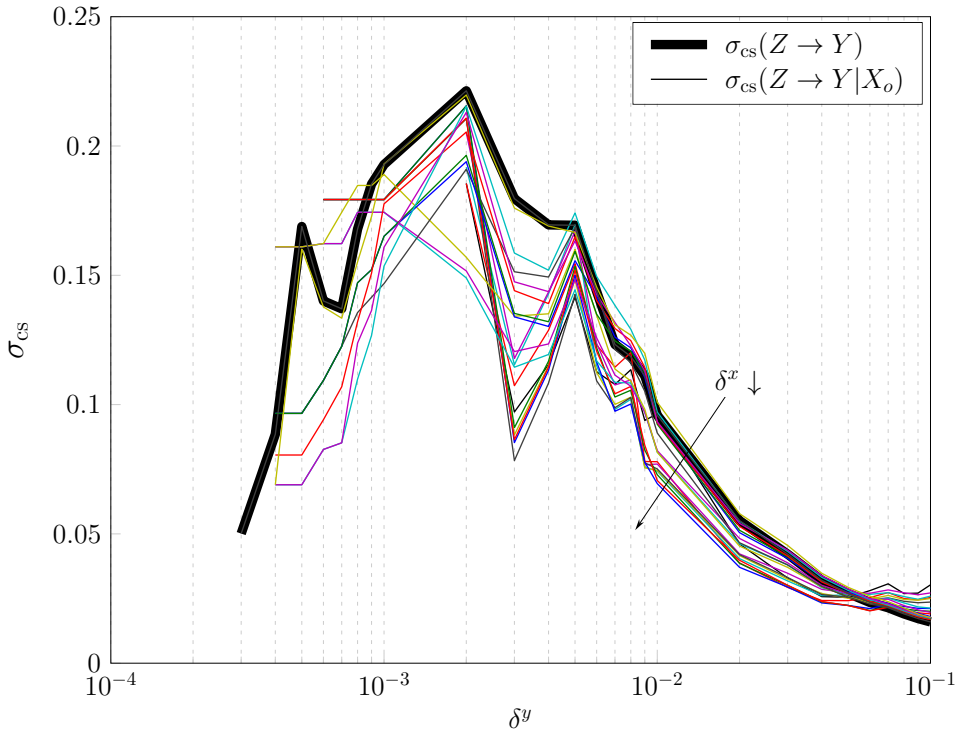
$$P(\epsilon_o^r|\delta^r, \delta^d) = P(\rho_{nn'}^r < \epsilon_o^r|\rho_{nn'}^R < \delta^r, \rho_{nn'}^D < \delta^d). \tag{3.11}$$

On the other hand, the HJ test is a hypothesis test for the following hypothesis

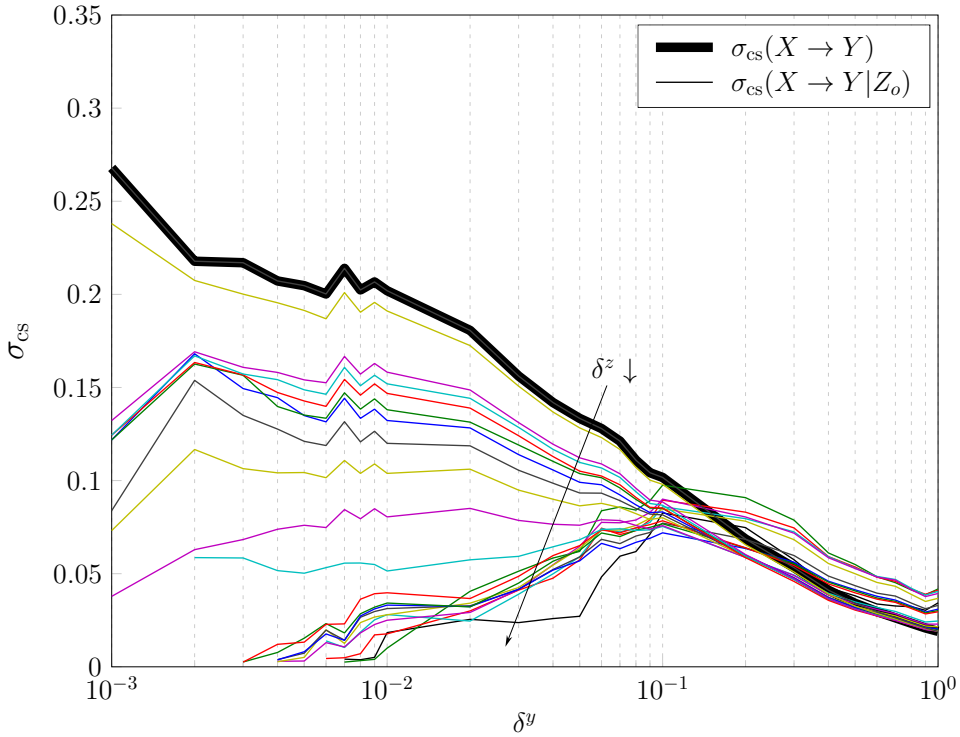
$$H_o : D \text{ does not Granger cause } R. \tag{3.12}$$

If we define $P_\epsilon(\epsilon_o^r|\delta^r, \delta^d) = P(\epsilon_o^r = \epsilon|\delta^r = \epsilon, \delta^d = \epsilon)$ for all $\epsilon > 0$, the hypothesis H_o of the HJ test presented in (2.5) can be rewritten as follows

$$P_\epsilon(\epsilon_o^r|\delta^r, \delta^d) = P_\epsilon(\epsilon_o^r|\delta^r). \tag{3.13}$$



(a) $\sigma_{\text{cs}}(Z \to Y|X_o)$ stays close to $\sigma_{\text{cs}}(Z \to Y)$ for all values of δ^x, which indicates $Z \to Y$ is a direct coupling.



(b) $\sigma_{\text{cs}}(X \to Y|Z_o)$ becomes less significant by decreasing δ^z revealing that $X \to Y$ is an indirect coupling.

Figure 3.8: Conditional σ_{cs} is depicted for a triple-variate coupling $X \leftarrow Z \to Y$.

Therefore, the TVAL value introduced in (2.7) can be presented as

$$\text{TVAL} = \sqrt{N}\left(P_\epsilon(\epsilon_o^r|\delta^r,\delta^d) - P_\epsilon(\epsilon_o^r|\delta^r)\right) \overset{a}{\sim} \text{Normal}\left(0, \sigma^2(L_d, L_r, \epsilon)\right). \tag{3.14}$$

It is noteworthy that $P_\epsilon(\epsilon_o^r|\delta^r,\delta^d)$ used in the HJ test is a specific case of $P(\epsilon_o^r|\delta^r,\delta^d)$ in the CS method where $\epsilon_o^r = \delta^r = \delta^d = \epsilon$. In other words, the HJ method considers the coupling spectrums in Fig. 3.1 only for one pair of $(\delta^r,\delta^d) = (\epsilon,\epsilon)$ and ϵ_o^r should also be equal to ϵ. As we see in Section 4.2.2, the value of ϵ has a severe impact on the results of the HJ method. However, in the CS method, we investigate $P(\epsilon_o^r|\delta^r,\delta^d)$ for the whole range of δ^r and δ^d values and ϵ_o^r is determined independently of δ^r and δ^d. In Section 4.2.2, we illustrate how the flexibility and generality of the parameters in the CS method makes it more robust than the HJ test.

3.5 Conclusion

We proposed a new method for identifying the directed couplings between time series. It was observed that this method identifies the direction of coupling in different scenarios such as unidirectional and bidirectional couplings, nonlinear dynamics, identical and non-identical sub-systems, multivariate systems, small and large sample sizes, weak and strong couplings, and in the presence of the noise. Moreover, this method is invariant to scaling of the data. These features of our method make it suitable for practical applications where we deal with multi-structure systems as well as small sample sizes and weak couplings. Comparing our method against TE, it was revealed that our method better detects weak couplings. Unlike TE, it is invariant to scaling. In terms of handling small sample size, it is as strong as TE, but for noisy data TE performs better.

Chapter 4

Applications of Coupling Spectrum Method

In this chapter, we focus on some of the applications of the coupling spectrum (CS) method, proposed in Chapter 3. In particular, we study applications in biology and finance[1]. In biological applications that provide time series data, the CS method can be used for the inference of the biological networks. In other words, we can apply the CS method to detect the interactions between biological components of the cell. The main challenges in identification of biological networks are the small sample size of data and noise. Here, we use the CS method for inference of gene regulatory networks (GRN), one of the most important biological networks in the cell. We try to detect a GRN known by the biological studies. The results of biological data analysis show the successful performance of the CS method for inference of GRNs.

Identifying dynamic causal relationships between financial data has many applications in finance and econometrics. As most of financial data are available in time series form, we can use the CS method for identification of the causal relationships between these data, e.g., the existing causality between the return and volume of a stock price, inflation and unemployment rate, or the effect of the stock price of a company on that of another company. Since the causal relationships are usually studied during a long time, for example over a decade, the direction of causality may changes over time. Hence, we combine the CS method with moving window techniques to deal with temporal causality. Here, we apply the CS method to detect the

[1]The results of biological application of the CS method were accepted in GlobalSIP 2013 Symposium on: Bioinformatics and Systems Biology (IEEE), Austin, Texas, U.S.A., Dec. 2013; however, it was withdrawn for publishing in a journal. Moreover, the financial application of the CS method is published as a conference paper entitled '*The Coupling Spectrum: A new method for detecting temporal non-linear causality in financial time series*' in proceeding of the 7th International Days of Statistics and Economics, Prague, Czech, Sep. 2013.

temporal causal effects of the stock prices of Apple Inc. and Microsoft Corporation on each other in more than a decade.

4.1 Biological Application

4.1.1 Introduction

Traditionally, the research in molecular biology has been restricted to studying single components of the cells one at a time. However, the biological entities of the cell interact with each other and work as a network rather than a collection of single biological components. To gain a thorough understanding of biological phenomena and diseases, e.g. cancer and diabetes, and the underlying processes, we need to see the cell as a whole and to unravel the interaction of molecular components involved in cellular networks. This idea yields a newly emerging multidisciplinary field termed 'Systems Biology', which provides a system-wide view of the cell and alternative solutions in medicine and biotechnology.

During the last decade, owing to the development of high throughput genomic and proteomic measurement technologies, e.g., DNA microarray and ChIP-on-chip technique, molecular biology is rapidly evolving into a quantitative science and it increasingly relies on mathematics, physics, engineering, and computing science to model and infer the biological networks. Consequently, a new discipline called 'Computational Systems Biology' has recently emerged that its aim is to discover the cellular networks through computational methods. To infer these biological networks from quantitative data, we need promising computational tools such as information theory and Bayesian networks.

Gene regulatory networks (GRNs) are one of the most important biological networks that their identification has immediate applications in cancer prediction and drug development [54]. Indeed, GRN represents the regularity interactions of the genes, proteins, and other molecules that yield activation or suppression of other genes. Modeling or reconstruction of GRNs based on experimental data is called 'network inference' or 'reverse-engineering'.

Different approaches were applied to infer GRNs such as differential equations [55], Boolean networks [56], Bayesian networks [57], and Information theory [58]. The CS method presented in Chapter 3 can be a candidate to infer the GRNs as it is capable of identifying directed couplings in severe practical conditions such as

unidirectional and bidirectional coupling, nonlinear coupling, and time series with small sample sizes.

In this section, we apply the CS method to infer the GRN of E2F1 transcription factor from microarray data. The microarray data analysis by the CS method shows the successful performance of this method for inference of GRNs.

This section is organized as follows. In Section 4.1.2, we introduce the biological background of this biological application. The results of analysis of the biological data is presented in Section 4.1.5, and finally, conclusions are drawn.

4.1.2 Biological Background

4.1.2.1 DNA and Gene

Proteins are the main components of the cell with vital functions in the cell and body, e.g., enzymes. Briefly, we can say that proteins are the essential components to control all the cell processes and reactions. Proteins are produced by the instructions encoded into DNA [59].

DNA is the information carrier molecule in a cell, which contains the genetic instructions. DNA molecules store the needed information for construction of proteins. A gene is a segment of DNA from which the information can be read as recipes to construct a particular protein. The process by which cells produce proteins from genes is called 'gene expression'. As it is shown in Fig. 4.1, gene expression has two major steps:

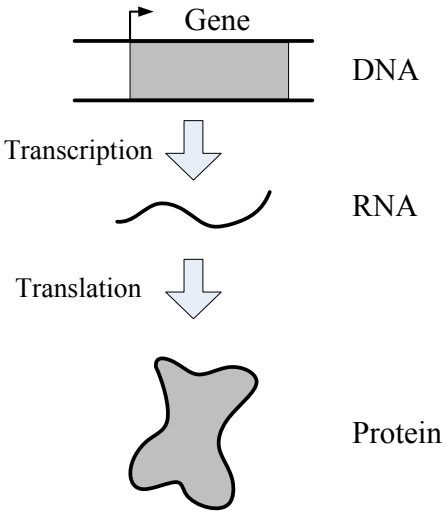


Figure 4.1: Gene expression.

1. **Transcription**: a copy of the encoded information in the gene is created, which is called RNA

2. **Translation**: the decoding phase of RNA where the protein is made from encoded data in the RNA.

4.1.2.2 Gene Regulatory Network

Different cell types in different tissues have similar DNA. Therefore, the cellular differentiation originates from differences in gene expression. In other words, not all the genes are active in all the cells and the combination of switched 'on' and 'off' genes determines the type of the cell. An important question in biology is how the genes are switched on and off, i.e., how genes are regulated.

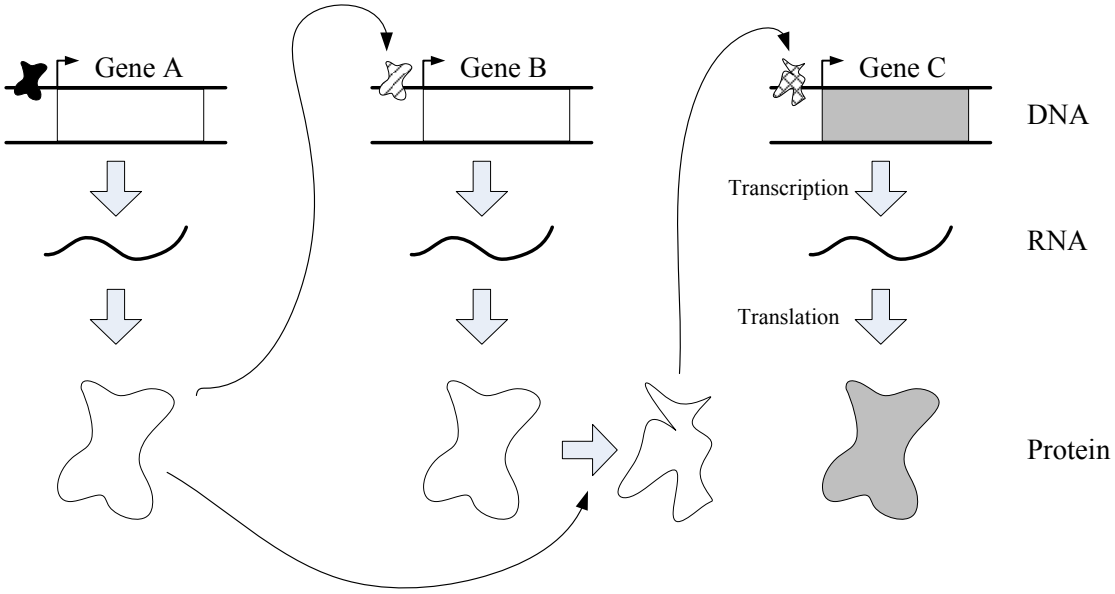
Let us see a simple example of a GRN shown in Fig. 4.2(a). When a gene is being expressed, the produced protein can regulate the expression of other genes directly (A $\rightarrow$ B and B $\rightarrow$ C) or indirectly (A $\rightarrow$ C). A GRN, which is consisting of these interactions, is often modeled as a graph composed of nodes (genes or proteins) and edges (gene-gene, protein-DNA, and protein-protein interactions). Fig.4.2(b) depicts the gene-gene graph representation of the GRN corresponding to Fig. 4.2(a).

Microarray is a technology to answer the question, what genes are expressed in a particular cell type, at a particular time, under particular conditions. Indeed, DNA microarray measures the expression levels of large numbers of genes simultaneously [60]. The data can be observed in a series of time-points to make the dynamics of the system visible.

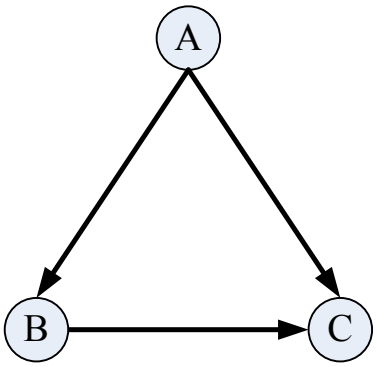
The regulatory interaction between two genes can be considered as a driver-response coupling, i.e., the regulatory and regulated genes are the driver and response, respectively. Provided that the expression level of each gene is observed as a time series, we can identify the regulatory interactions by analyzing the time series. Hence, a GRN can be inferred by identifying the couplings in a set of coupled genes.

4.1.2.3 Biological Data

The advances in high-throughput technologies and providing the quantitative data promote the reconstruction of GRNs through computational methods. Here, we briefly introduce different kinds of data showing cellular interactions, and challenges these data pose for GRN inference.



(a) An example of a GRN in three different levels (DNA, RNA, and protein).



(b) The gen-gene graph representation of the GRN shown above.

Figure 4.2: A gene regulatory network (GRN) and its graph representation.

• Different types of data show various kinds of interactions in the cell:

1. **Transcriptome data** are measured in RNA level by DNA microarrays (also known as gene chip or DNA chip). Indeed, DNA microarray measures the expression levels of large numbers of genes simultaneously [60]. DNA microarray data are more accessible and cheaper, hence, they are traditionally used for reverse-engineering.

2. **Proteom data** show protein-protein interactions in the cell. Remarkably, the total number of proteins is much higher than the number of protein-encoding genes. Hence, proteomic studies are complex and difficult [61].

3. **Interactome data** show the whole set of molecular interactions in cells. For example, protein-DNA interaction measured by ChIP-on-chip technology.

• The data can be observed in two scenarios:

1. **Static** (steady state): for each perturbation (i.e., experimental conditions), the observation is accomplished at the steady state of the biological system. The dynamics of the system is missed herein.

2. **Dynamic** (time series): after perturbation, we do the measurement in a series of time-points and the dynamics of the system is visible at the cost of time, effort, and more expenses.

• There are two challenging problems regarding microarray data that impede all the GRN inference methods:

1. **Dimensionality problem**: we are dealing with thousands of genes, but the number of samples for each gene is small.

2. **Noise**: the signal to noise ratio in biological data is usually extremely low.

4.1.3 Gene Regulatory Network Inference Methods

A number of computational approaches have been proposed for inference of GRNs. Generally, there are two types of models used by inference methods:

1. **Undirected Models**: Some methods are merely looking for associations and dependency between genes. Hence, they ignore the cause-effect relationships, i.e., the obtained graphs by these methods are undirected. The simplest example is the correlation network [62] that the weight of edges in the graph represents the correlation coefficient. Correlation networks consider the linear dependency between genes. Besides the correlation method, the mutual information, which is an information theoretical measures, has been used widely which makes no assumption about the dependencies between the genes [63, 64].

2. **Directed Models**: These models consider causation on top of dependency resulting in a directed graph of GRN. Different approaches have been proposed such as applying differential equations [55], Boolean networks [65], Bayesian networks [57, 66], and information theory [58].

As the main goal of this thesis is directed network inference of GRNs, here we briefly introduce the proposed methods for inference of this kind of networks. Here, it is assumed that the transcriptome data is used.

4.1.3.1 Differential and Difference Equations

Consider we have N genes and x_i denotes the level of transcription of the ith gene. Hence, a GRN can be modeled as a system of differential equations as follows [55]:

$$\frac{dx_i}{dt} = f_i(x_1, \ldots, x_N). \tag{4.1}$$

Different forms of f_i can be considered, however, the linear form is mostly used for simplicity and less required amount of data [67], i.e.,

$$\frac{dx_i}{dt} = \sum_j w_{ij} x_j + p_i. \tag{4.2}$$

As the derivative can amplify the measurement errors, one can use the discrete model and apply the difference equations as follows [68]:

$$x_i(n) = \sum_j w_{ij} x_j(n-1) + p_i. \tag{4.3}$$

Although the linear model is simplified, it cannot cope with nonlinear nature of GRNs and it is not successful in large-scale GRNs.

4.1.3.2 Boolean Network

A set of Boolean (binary) variables that their state is determined by other variables in the network form a Boolean network. Boolean network was initially proposed as a random model of GRNs [65]. '0' and '1' states represent the active (up-regulated) and inactive (down-regulated) genes, respectively. The Boolean network consists of N nodes, representing genes, with k inputs to each node representing regulatory interactions, .i.e.,

$$x_i(n) = f_i^B\left(x_i^1(n-1), \ldots, x_i^k(n-1)\right) \tag{4.4}$$

where x_i^k denotes the binary state of the kth regulator of x_i and f_i^B is a Boolean function made up of Boolean operations such as AND, OR, and NOT.

The first step to infer a Boolean network is converting the transcript data to binary data. However, in the presence of the noise, choosing the appropriate threshold for discretization is not trivial. Furthermore, the inference methods of Boolean networks require large amount of data. As the regulators are represented by k inputs of a node, each of 2^k states of the regulatory nodes should be observed in experimental measurements to be able to find the Boolean function f_i^B. However, this complete data is not available in most cases. Moreover, although the Boolean network is a dynamic model for the GRN, the underlying processes of gene expression cannot be described perfectly in a two-state system.

4.1.3.3 Bayesian Network

Bayesian network (BN) presents a probabilistic model for GRN that considers the transcription level of the gene as a random variable X [57, 66]. This model is represented by a directed acyclic graph (DAG) that each node is associated with a gene and the direction of each edge is from regulatory toward regulated gene. The regulatory genes of X_i are known as its parents, denoted by $\mathrm{Pa}(X_i)$. Let us consider N random variable $X_1, ..., X_N$ representing N genes. By applying the chain rule, the joint probability distribution $P(X1, \ldots, X_N)$ has the form

$$P(X1, \ldots, X_N) = \prod_{i=1}^{N} P(X_i|\mathrm{Pa}(X_i)). \tag{4.5}$$

For instance, the graph of the GRN depicted in Fig. 4.2 is shown again in Fig. 4.3 and the corresponding probability distribution $P(X1, \ldots, X_N)$ is as follows:

$$P(X1, \ldots, X_N) = P(X_A)P(X_B|X_A)P(X_C|X_A, X_B). \tag{4.6}$$

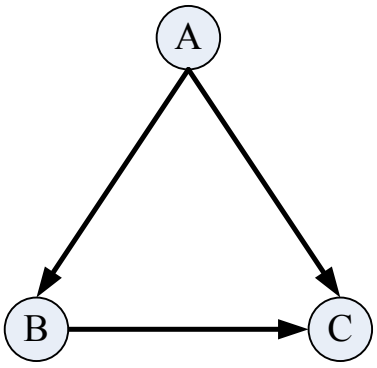


Figure 4.3: Gene regulatory network graph

To model a GRN by a BN, two sets of parameters should be found: (1) network topology, which is determined by the interconnections between the genes; (2) the conditional probabilities $P(X_i|\mathrm{Pa}(X_i))$.

The probabilistic nature of the BN enables us to deal with the inherent noise of the data as well as the incomplete data and hidden variables (e.g., lack of proteomic data) [69]. Furthermore, BN can combine different types of data and it can consider the prior knowledge in the model [70].

Despite the benefits of the BN, the acyclic feature of the DAG prevents the existence of a loop in the inferred network, which commonly exists in GRNs [71]. To resolve this restriction, the Dynamic BN (DBN) or Temporal BN can be applied by using the time series data [72, 73]. The DBN is defined as follows:

$$P(X1(t),\ldots,X_N(t)) = \prod_{i=1}^{N} P(X_i(t)|\mathrm{Pa}(X_i(t-1))) \tag{4.7}$$

and its representation for a given graph is depicted in Fig. 4.4. Although the

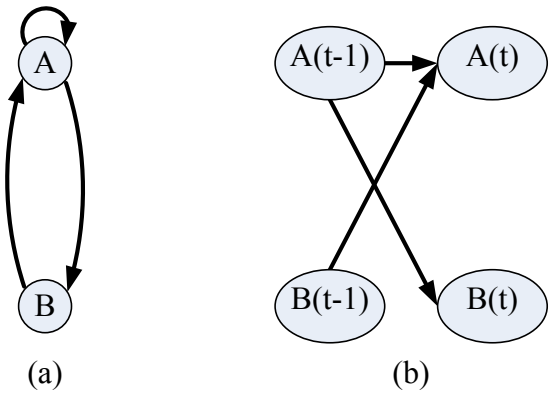


Figure 4.4: (a) Graph with loops; (b) Dynamic Bayesian Network (DBN).

inference of loops is possible by applying the DBN, it considerably increases the data demand of the inference method that is not easily accessible by biological data.

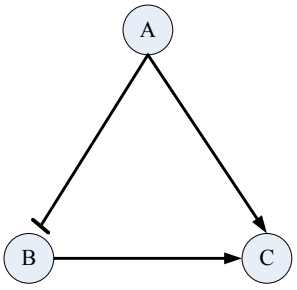
4.1.3.4 Information Theory

Some efforts have been made to infer the GRN through TE [58]. The characterizations of TE that make it a promising tool for network inference of GRN are as follows:

- Discovering one and bi-directional interactions (applicable in directed and cyclic graphs).
- Dealing with time-series to capture the dynamic behavior of the system.
- Handling the continuous data (instead of quantized data that suffers from quantization noise).
- There is no assumption about the underlying distribution of the samples.

Despite the advantages of the TE method, the direction of detected interactions by the TE method depends on the scaling of the data and rescaling may reverses the direction (see Section 2.5). Scaling is commonly used for biological data; hence, this drawback of the TE method can make the results of GRN inference derived by the TE method erroneous.

A summary of these methods is shown in Fig. 4.5



Difference Equations	Boolean Networks	Bayesian Networks	Information Theory
$A_{n+1} = A_n$	$A_{n+1} = A_n$	P(A=0) = 0.4	TE(A→B) > 0
$B_{n+1} = B_n - 0.2\ A_n$	B_{n+1} = **NOT** A_n	P(B=0 | A=0) = 0.3	TE(B→A) = 0
$C_{n+1} = C_n + 0.4\ A_n + 0.3\ B_n$	$C_{n+1} = A_n$ **OR** B_n	P(B=0 | A=1) = 0.9	TE(A→C) > 0
		P(C=0 | A=0,B=0) = 0.8	TE(C→A) = 0
		P(C=0 | A=0,B=1) = 0.3	TE(B→C) > 0
		P(C=0 | A=1,B=0) = 0.4	TE(C→B) = 0
		P(C=0 | A=1,B=1) = 0.1	

Figure 4.5: Different inference methods used for modeling a GRN are presented. In the graph, → denotes 'activation' and ⊣ represents 'deactivation'.

4.1.4 Biological Application of the CS Method

Consider a gene regulatory interaction $D \rightarrow R$ as two coupled genes that the regulatory gene D is the driver gene and the regulated gene R is the response gene. The transcription level of genes D and R are measured in N samples as a time series, denoted by $\{d_n\}$ and $\{r_n\}$, respectively.

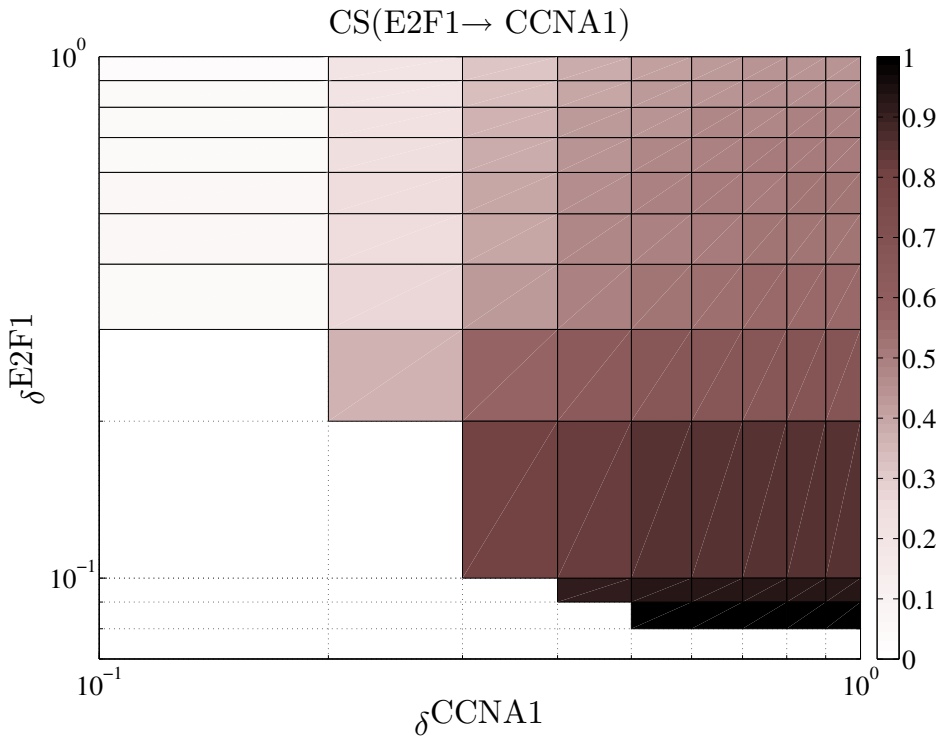
The goal is to identify the direction of the existing coupling between genes D and R from observed samples by using the CS method. Fig. 4.6 depicts the CS derived by analyzing the microarray data to detect the regulatory interactions between E2F1 and CCNA1 genes (for more details about analyzing the microarray data see Sec. 4.1.5). Biological evidences prove the existence of E2F1→CCNA1 regulatory interaction [74, 75]. Fig. 4.6(a) represents CS(E2F1→CCNA1) that we observe a change of color in each column of the CS, meaning that E2F1→CCNA1 exists. On the other hand, Fig. 4.6(b) depicts CS(CCNA1→E2F1) where all the columns of the CS lack the color change; hence, the regulatory interaction CCNA1→E2F1 does not exist.

The standard deviation of the values of $P(\epsilon_o^r|\delta_o^r, \delta^d)$ in each column of the CS, i.e., σ_{cs}, and the corresponding values of $\text{UCI}_{90\%}$ of Figs. 4.6(a) and 4.6(b) are plotted in Figs. 4.7(a) and 4.7(b), respectively. As σ_{cs}(E2F1→CCNA1) lies outside the $\text{UCI}_{90\%}$(E2F1→CCNA1), we confirm the significance of σ_{cs}(E2F1→CCNA1), i.e., E2F1 regulates CCNA1. Conversely, for all values of δ^{E2F1}, σ_{cs}(CCNA1→E2F1) is below the $\text{UCI}_{90\%}$(CCNA1→E2F1); hence, CCNA1 is not a regulatory gene of E2F1.

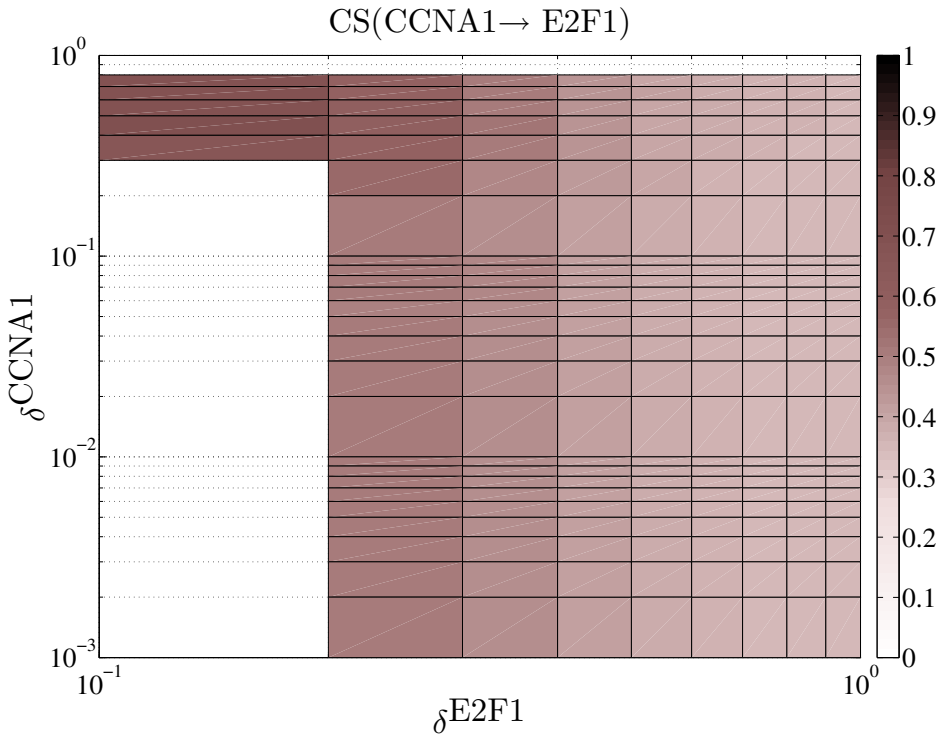
To measure the significance of $\sigma_{\text{cs}}(D \rightarrow R)$ relative to $\text{UCI}_{\alpha\%}(D^p \rightarrow R)$, we add those values of σ_{cs} that are greater than $\text{UCI}_{\alpha\%}$ together as follows

$$\text{SIG}_{\text{cs}}(D \rightarrow R) = \sum_{\delta^r} [\sigma_{\text{cs}}(\delta^r) - \text{UCI}_{\alpha\%}(\delta^r)] \times \text{I}\left(\sigma_{\text{cs}}(\delta^r) - \text{UCI}_{\alpha\%}(\delta^r)\right) \tag{4.8}$$

where $\text{I}(\cdot)$ is an indicator function that $\text{I}(x > 0) = 1$; $\text{I}(x \leq 0) = 0$. We can see in Fig. 4.7(a) that for some values of δ^{CCNA1}, σ_{cs}(E2F1→CCNA1) is greater than $\text{UCI}_{90\%}$(E2F1→CCNA1); therefore, SIG_{cs}(E2F1→CCNA1)> 0 and we can say that the regulatory interaction E2F1→CCNA1 exists. On the other hand, Fig. 4.7(b) shows that σ_{cs}(CCNA1→E2F1) is smaller than $\text{UCI}_{90\%}$(CCNA1→E2F1) for all δ^{E2F1}; therefore, SIG_{cs}(CCNA1→E2F1)= 0 and we conclude CCNA1↛E2F1.

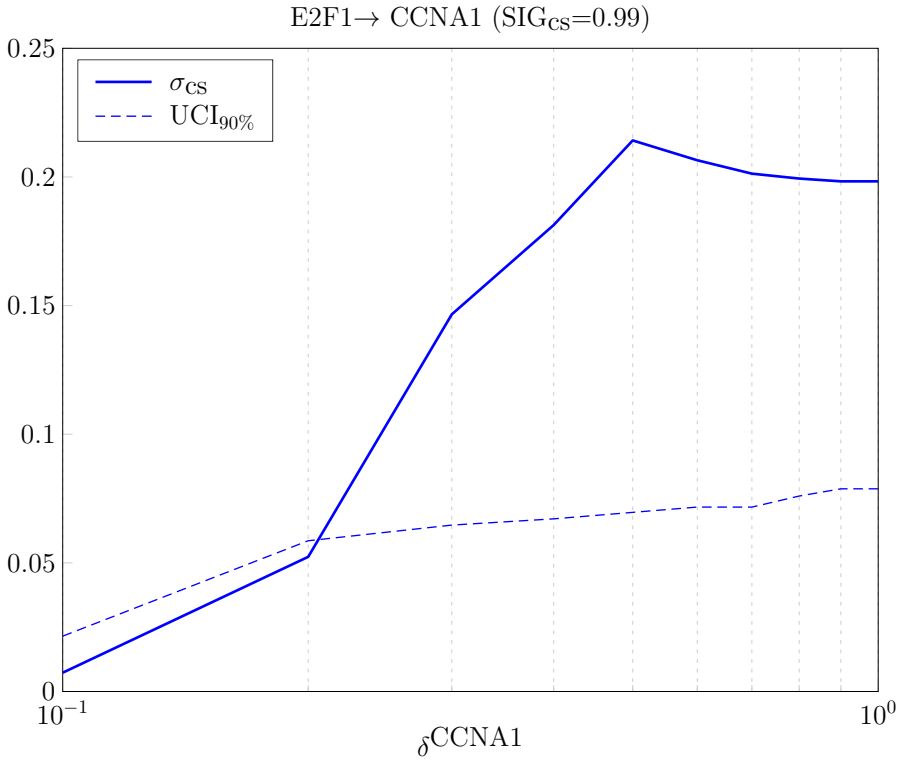


(a)

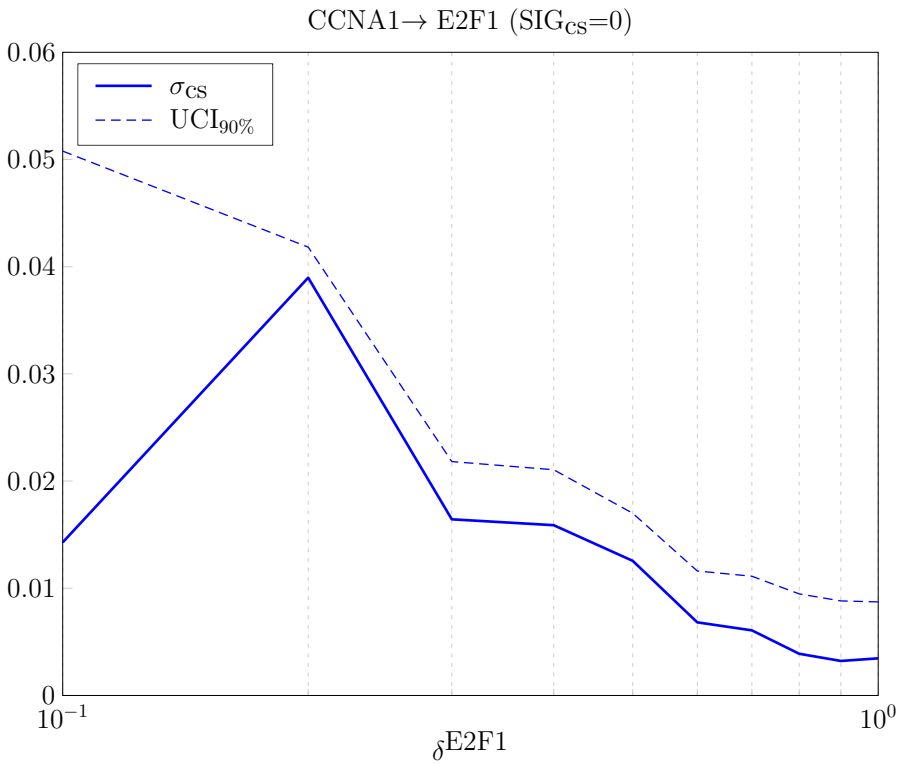


(b)

Figure 4.6: The coupling spectrum (CS) of the regulatory interactions between E2F1 and CCNA2 genes.



(a)



(b)

Figure 4.7: σ_{cs}, UCI$_{90\%}$, and SIG$_{cs}$ of the coupling spectrum shown in Fig. 4.6, corresponding to regulatory interactions between E2F1 and CCNA1.

4.1.5 Biological Results and Discussion

In this section, we apply the CS method for analyzing the microarray data in order to infer the GRNs from transcriptional data. Here, we use the microarray time series data set of the cell cycle in a human cancer cell line (HeLa) studied in [76]. The results of five different experiments are available in this data set. Here, we use the time series of the second and third experiments consisting of 26 and 48 time points, respectively, i.e., we have a total of 74 samples. Because of the differences between experiments, the vector of the lagged values of each experiment are constructed separately and all of them are combined together to be used in the CS method.

The E2F1 gene encodes the E2F1 protein that is a member of the E2F family of transcription factors. This family plays a critical role in control of the cell cycle by regulating the transcription of different target genes. Table 4.1 represents 18 genes recognized as the E2F1 target genes as well as the references providing the biological evidences for these regulatory interactions.

Here, the goal is to apply the CS method for inference of the GRN consisting of the transcriptional regulatory interactions of E2F1 with the target genes (TG) listed in Table 4.1. We calculated $\mathrm{SIG_{cs}}$(E2F1→TG) and $\mathrm{SIG_{cs}}$(TG→E2F1) with n_{min} = 10 and the lag values L_d = L_r = L. The results of the microarray data analysis show that the CS method detects the most number of the interactions with L = 3. Hence, the results obtained by L = 3 are reported in the following. Moreover, we performed the permutation 500 times to obtain $\mathrm{UCI_{90\%}}$. Then, the

Table 4.1: List of E2F1 target genes.

Target gene	Reference	Target gene	Reference
CCNA1	[74], [75]	MYB	[74], [75], [77]
CCNA2	[74], [75]	MYC	[78]
CCNB1	[75]	PCNA	[74]
CCNE1	[74], [75], [77]	POLA2	[74], [75]
CDC2	[74], [75], [77]	RANBP1	[79]
CDC6	[75], [77]	RRM2	[75]
CDKN1A	[75]	TFDP-1	[74], [75]
CDKN2C	[80]	TK2	[74]
DHFR	[74], [75]	TS	[74], [75]

values of SIG_{cs} are sorted and the interactions that their SIG_{cs} is greater than 0.1 are selected. The values of SIG_{cs}(E2F1→TG) and SIG_{cs}(TG→E2F1) for different target genes of E2F1 transcription factor are listed in Table 4.2.

Figure 4.8 depicts the schematic of the E2F1 regulatory network inferred by the CS method. The thickness of the arrows in Fig. 4.8 represent the value of SIG_{cs} in three different intervals: $0.1 \leq SIG_{cs} < 0.3$, $0.3 \leq SIG_{cs} < 0.5$, and $0.5 \leq SIG_{cs}$. Except the target genes CCNE1, CDKN2C, and POLA2, all other E2F1→TG regulatory interactions are successfully detected. On the other hand, the CS method identifies the reverse TG→E2F1 interactions for the target genes CDKN1A, CDC6, CCNE1, CCNA2, RRM2, and CDKN2C.

The transcription factor E2F1 activates itself [74, 77]. This fact can lead us to find a biological evidence for some detected interactions TG→E2F1. The studies of [81] and [82] show that the proteins expressed by CDKN1A and CCNA2 genes bind to E2F1 transcription factor and inhibit its activity. Hence, this inhibition can influence the transcription level of E2F1 gene, and consequently, results the CDKN1A→E2F1 and CCNA2→E2F1 interactions in transcription level.

The regulation of E2F1 transcription factor by Rb protein can provide a biological evidence for CCNE1→E2F1 interaction. Indeed, Rb can bind to E2F1 transcription factor and inhibits its transcriptional activity. However, CCNE1-CDK2 is one of the cyclin-Cdk combinations that phosphorylates Rb during the G1 phase in cell division cycle and prevent it from binding and inactivating E2F1 [83]. Hence,

Table 4.2: List of SIG_{cs}(E2F1→TG) and SIG_{cs}(TG→E2F1) for different target genes (TG) of E2F1 transcription factor. The interactions that their SIG_{cs} is greater than 0.1 (bold numbers) are considered as detected interactions.

TG	E2F1→TG	TG→E2F1	TG	E2F1→TG	TG→E2F1
CCNA1	**0.99**	0	TK2	**0.21**	0
PCNA	**0.98**	0.06	MYC	**0.16**	0
CDKN1A	**0.91**	**0.38**	DP-1	**0.15**	0
CDC6	**0.77**	**0.29**	MYB	**0.14**	0
CCNB1	**0.68**	0	RRM2	**0.12**	**0.14**
TS	**0.53**	0.08	RANBP1	**0.12**	0
CDC2	**0.47**	0	CCNE1	0.03	**0.48**
CCNA2	**0.36**	**0.13**	CDKN2C	0.02	**0.13**
DHFR	**0.34**	0	POLA2	0.02	0

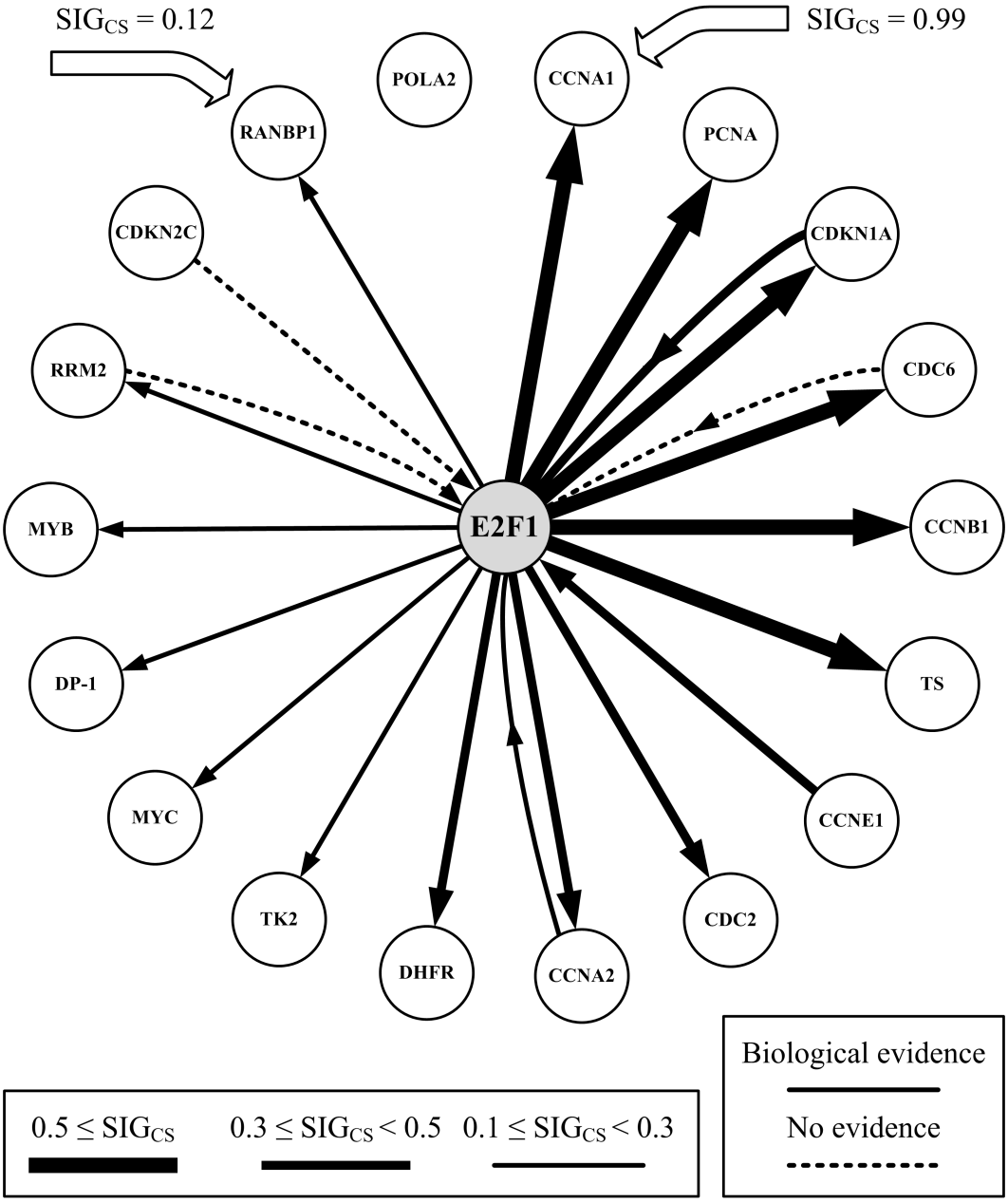


Figure 4.8: GRN of E2F1 transcription factor inferred by the CS method.

CCNE1 can influence the activity of E2F1 protein, and consequently, its transcriptional level. We did not find any biological evidence in the literature for other reverse regulatory interactions TG→E2F1.

According to the GRN inferred by the CS method shown in Fig. 4.8, the sensitivity (true positive rate) and specificity (true negative rate) of the inference method are 86% and 80%, respectively. These values show the high level of the certainty of the CS method for GRN inference.

4.1.6 Conclusion of Biological Application

In this work, we applied the coupling spectrum (CS) method for inference of transcriptional gene regulatory networks. The analysis of the microarray data showed the successful performance of the CS method for inference of GRNs from biological data that their sample size is very small. Here, we compared the inferred GRN of E2F1 transcription factor by the biological supported regulatory network of E2F1. The results revealed that the CS method is able to infer GRNs with high level of certainty.

4.2 Financial Application

4.2.1 Introduction

The Granger causality (G-causality) test [1, 16] is a statistical hypothesis test for identifying causal relationships between time series. This method estimates a linear regression model with lagged values of the time series $\{d_t\}$ (the driver time series) used to predict the future values of $\{r_t\}$ (the response time series) in the presence of lagged values of $\{r_t\}$. If the error of prediction is reduced by inclusion of $\{d_t\}$, $\{d_t\}$ is the Granger-cause of $\{r_t\}$. The assumption of linearity in G-causality test can be violated in real applications and it cannot detect nonlinear causal relationships [39]. Many investigations in the literature provide evidence of linear and nonlinear causality between financial time series [26, 84]. Hence, different nonlinear extensions of G-causality (NLG-causality) were proposed to detect nonlinear causality in financial data [26–28].

In many financial data sets, the direction of causality changes over time. To deal with temporal causality, causality inference methods can be combined with moving window techniques to identify possible causality changes over time. In this work, we extend the CS method by using a moving window technique and compare its performance on a simulated temporal nonlinear causal system to a moving window adaptation of the HJ test proposed in [26]. In Section 3.4, we introduced the relationship between the CS method and the HJ test. Here, We compare their performance on a real data set -the stock prices of Apple Inc. and Microsoft Corporation. The simulated and empirical results show that the CS method is more robust than NLG-causality method.

The financial application of the CS method is presented in the following order: The simulation results and the real data example are presented in Sections 4.2.2 and 4.2.3, respectively. Then, we conclude with a discussion in Section 4.2.4.

4.2.2 Simulation Results

In this work, the goal is to discover causal relationships between two time series where the direction of causality is changing over time. To find the temporal changing causality, we use the overlapped moving window technique to detect the direction of causality in a small period of time.

Here, we evaluate the performance of the HJ and CS methods on simulated data

to detect temporal changing causality. Again consider two time series $\{x_t\}$ and $\{y_t\}$ having a causal relationship by the Hénon map [29]

$$x_n = a - x_{n-1}^2 + bx_{n-2} + c_{y\to x}(x_{n-1}^2 - y_{n-1}^2) \tag{4.9a}$$

$$y_n = a - y_{n-1}^2 + by_{n-2} + c_{x\to y}(y_{n-1}^2 - x_{n-1}^2) \tag{4.9b}$$

where $a = 1.4$, $b = 0.3$ and the initial values of x_0 and y_0 are uniformly distributed in [0, 0.5]. The strength of causalities between $X \to Y$ and $Y \to X$ are controlled by $c_{x\to y}$ and $c_{y\to x}$, respectively. To have a temporal causality in the model, $c_{x\to y}$ and $c_{y\to x}$ change with time as shown in Fig. 4.9(a). As Fig. 4.9(a) represents, there are three combinations of causality: part (i) is a unidirectional causal relation and parts (ii) and (iii) are bidirectional causalities with two different forms of overlapping of the coupling strengths. Here, we use the overlapping window with window length N_w. In each step, the window moves $N_f < N_w$ time points further. In the simulation, we used $N_w = 300$ and $N_f = 60$. The lag-lengths L_d and L_r are set to 2. A significant level of 5% is used for the HJ test and we estimate the $\mathrm{UCI}_{90\%}$ for the CS method.

Figure 4.9 shows the comparison of the CS and HJ methods for 50 trials with different initial values of x_0 and y_0 in equation (4.9). The mean of the σ_{cs} and TVAL values over the 50 trials are plotted. Figure 4.9(b) shows that the outcome of the CS method is consistent with the real causal relationships. In other words, the CS method

1. correctly detects the direction of causality in all three parts (the detected causality $X \to Y$ in part (i) is very weak);

2. distinguishes the strong and weak causality in the bidirectional scenarios;

3. finds for each direction of causality the correct ratios of causality strengths in different parts that are proportional to real ratios.

Figures 4.9(c) and 4.9(d) show the performance of the HJ method for $\epsilon = 0.2$ and $\epsilon = 1$, respectively. For $\epsilon = 0.2$, HJ does not detect any $X \to Y$ causality in part (i). In parts (ii) and (iii), HJ performs as well as CS. However, Fig. 4.9(d) illustrates that increasing ϵ adversely affects the HJ method. In this case

1. the causality $Y \to X$ in part (i) is not detected;

2. weak and strong causalities in bidirectional scenarios are not distinguishable;

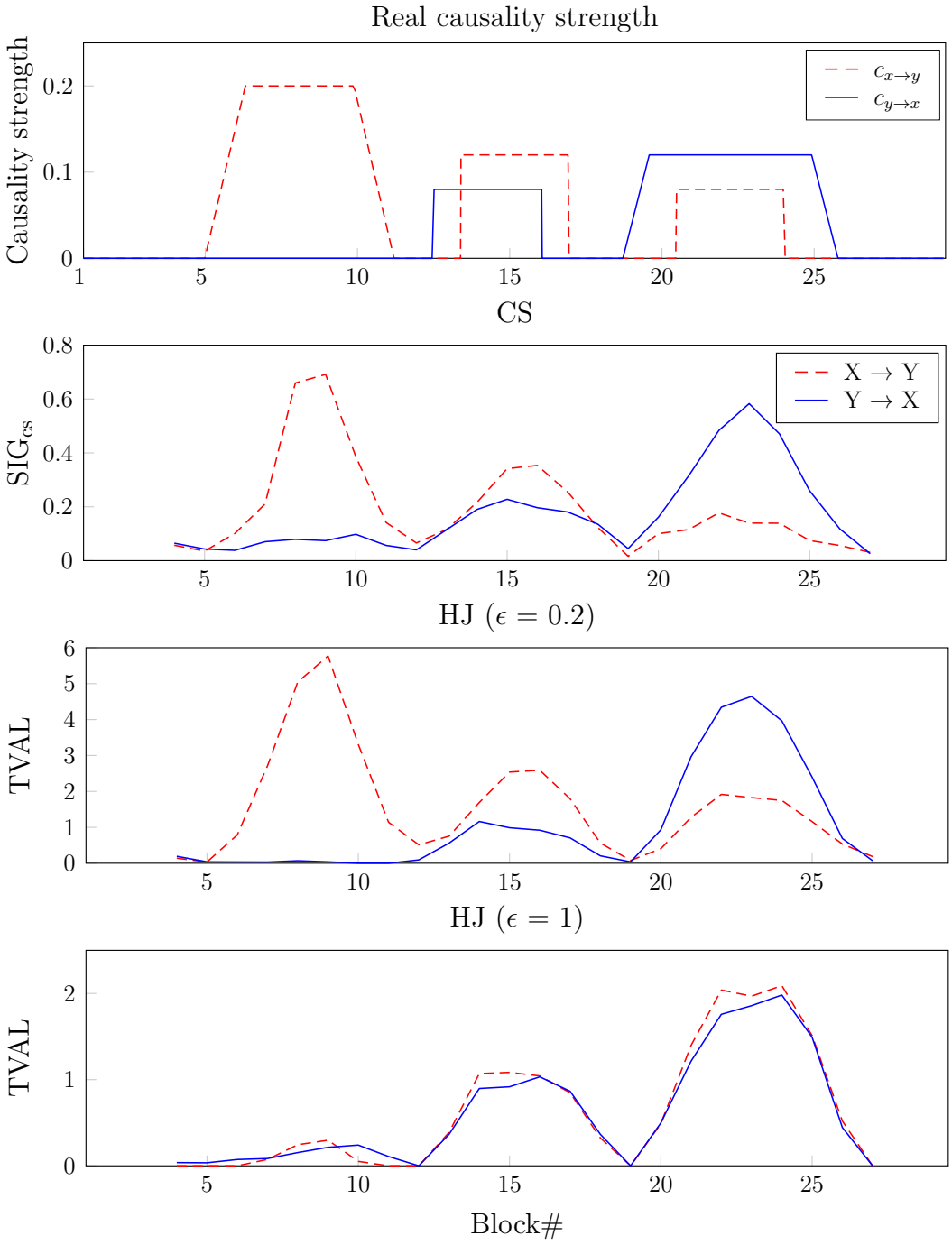


Figure 4.9: Finding temporal causality for the simulated data. (a) The real strength of the temporal causality; (b) CS method; (c) and (d) HJ method for different values of ϵ.

3. for each direction of causality, the ratios of causality strengths in different parts are not proportional to the real ratios.

Indeed, as HJ considers $P(\epsilon_o^r|\delta^r, \delta^d)$ just for a specific value of ϵ, the validity of the HJ results depends severely on ϵ and in all practical applications ϵ will be unknown.

4.2.3 Empirical Results

In this section, we investigate the temporal causality between the stock prices of Apple Inc. (AAPL) and Microsoft Corporation (MSFT). A total of 3199 daily stock prices during the time between January 2000 and August 2012 are used. To render each time series weakly stationary, we carry out a piecewise linear detrending. We select a window length of five month and N_f is the duration of one month. The lag-lengths L_d and L_r are set to 5, i.e., we investigate the causal effect of the stock prices of past five business days on the price of the next business day. In addition, a test significant level of 5% and $\epsilon = 0.7$ are used in the HJ method (this value of ϵ results in larger TVALs). $\text{UCI}_{90\%}$ is estimated for the CS method.

The temporal causalities AAPL→MSFT and MSFT→ AAPL derived by the CS and HJ methods are plotted in Figs. 4.10 and 4.11, respectively. The months in these figures represent the middle month of each block. As an evidence for detected causality, the timeline of AAPL and MSFT major products are depicted by arrows in subplots (a) and (b), respectively. Figures 4.10 and 4.11 reveal the following results:

- The direction of causality between these two companies changes over time. Therefore, to investigate causality between financial time series over a long period of time, we have to use a moving window to deal with this time-varying causality.

- Most of the products of each company affect the other one's stock price immediately or a couple of months after each product release. However, the number of the causal relationships detected by the HJ method is less than that of the CS method.

- There are detected causalities that could be due to other factors other than products releases, e.g., detected causalities in the second half-year of 2008 in MSFT→ AAPL.

- In general, for both methods, it can be concluded that the causal effect of AAPL on MSFT's stock price is greater over time than vice versa.

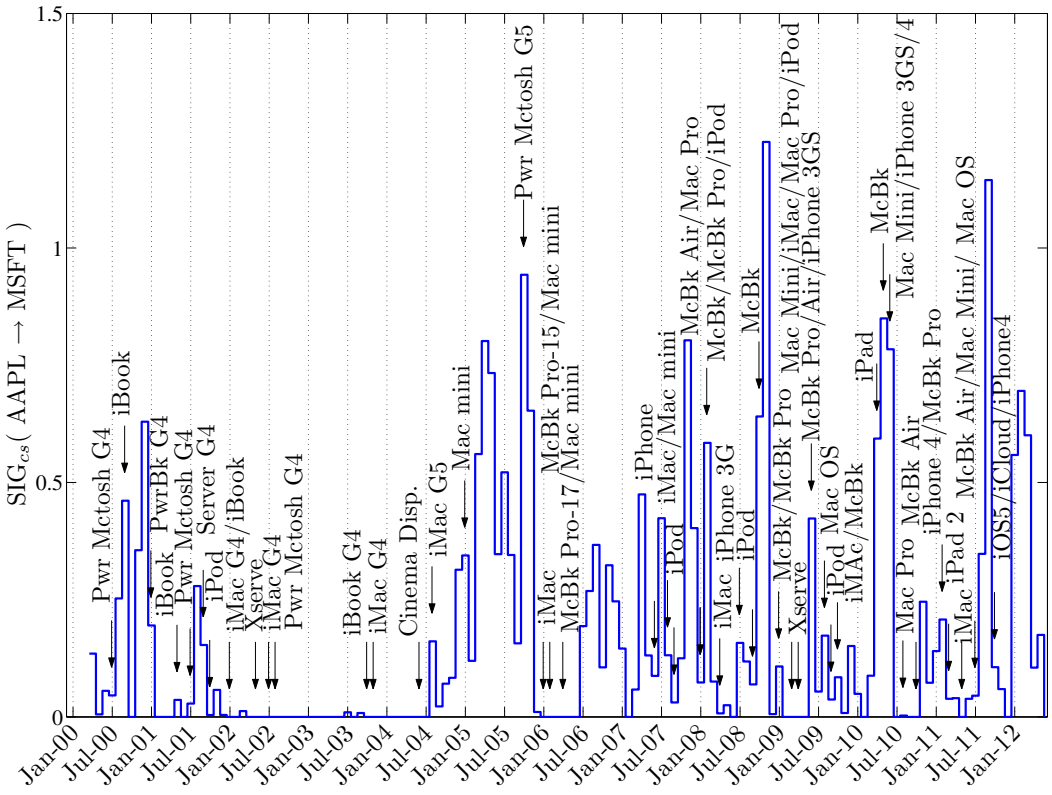
4.2.4 Conclusion of Financial Application

The dynamic causal relationships between many financial time series have a nonlinear and time-varying nature. In this paper, we extended a recently proposed approach called the coupling spectrum (CS) to detect temporal nonlinear causalities between financial time series. We compared two nonlinear causality inference methods, the HJ and CS methods, and used the overlapping moving window technique to deal with temporal causalities. Examination of these two methods on a simulated nonlinear causal relationship showed that due to the generality of the CS parameters over the HJ parameters, the performance of the CS method is more robust than the HJ method. In other words, HJ can be severely affected by its parameter value selection. In the final section we applied the CS and HJ methods to the stock prices of two companies, Apple Inc. and Microsoft Corporation, over a decade to detect the temporal causal effects of their stock prices on each other. We found that the direction of causality changes over time, especially around the advent of new products. Hence, in conclusion, in analyzing causality between financial time series over long periods of time, we have to use moving window techniques to deal with the time-varying causality.

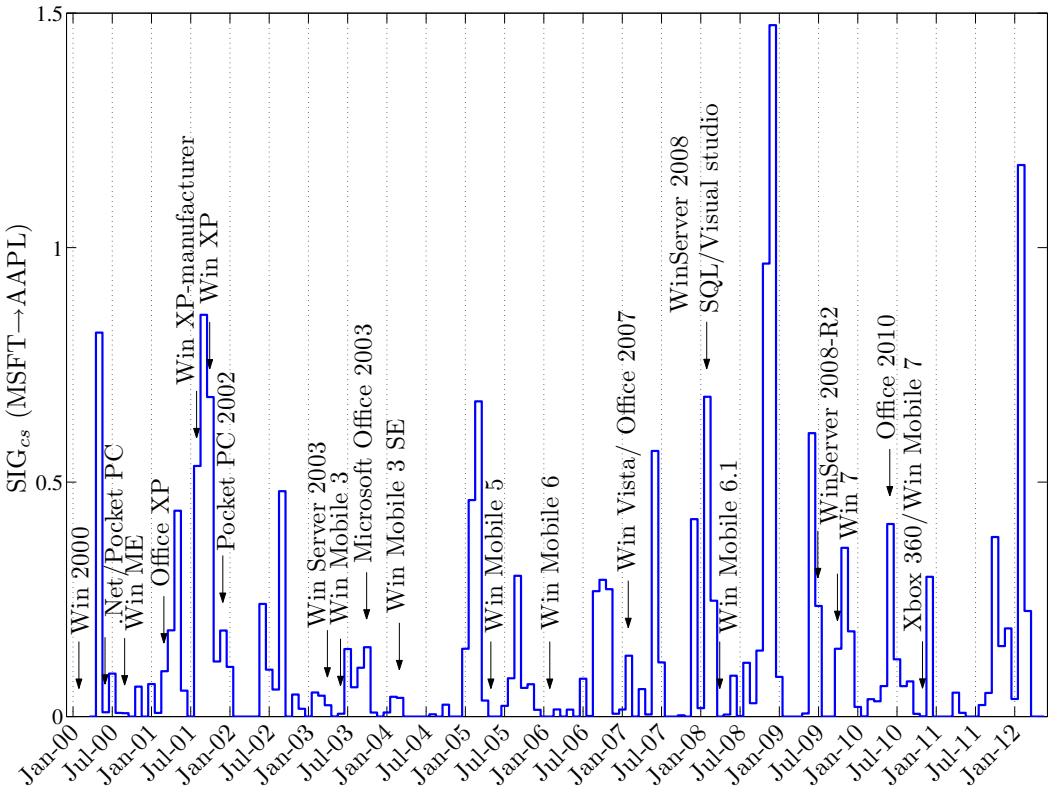
4.3 Conclusion

In this chapter, we applied the CS method to biological and financial applications. In the case of biological application,the results of applying the CS method for inference of the regulatory interactions between E2F1 and its target genes showed that the CS method can detect this regulatory network with a high level of certainty.

To study the performance of the CS method for finding the time-varying causal relationships between financial time series, we combined the proposed CS method by moving window techniques. Then, we performed the windowed-CS method to detect the temporal causalities between the stock prices of Apple Inc. and Microsoft Corporation for more than a decade. Moreover, we compared the CS results with those of the nonlinear Granger causality test (HJ test). The results revealed that the time-varying causality should be considered in long-term financial data analysis. Moreover, the detected causal relationships found by the CS method were more consistent with the advent of new products of theses two companies than those of the HJ test.

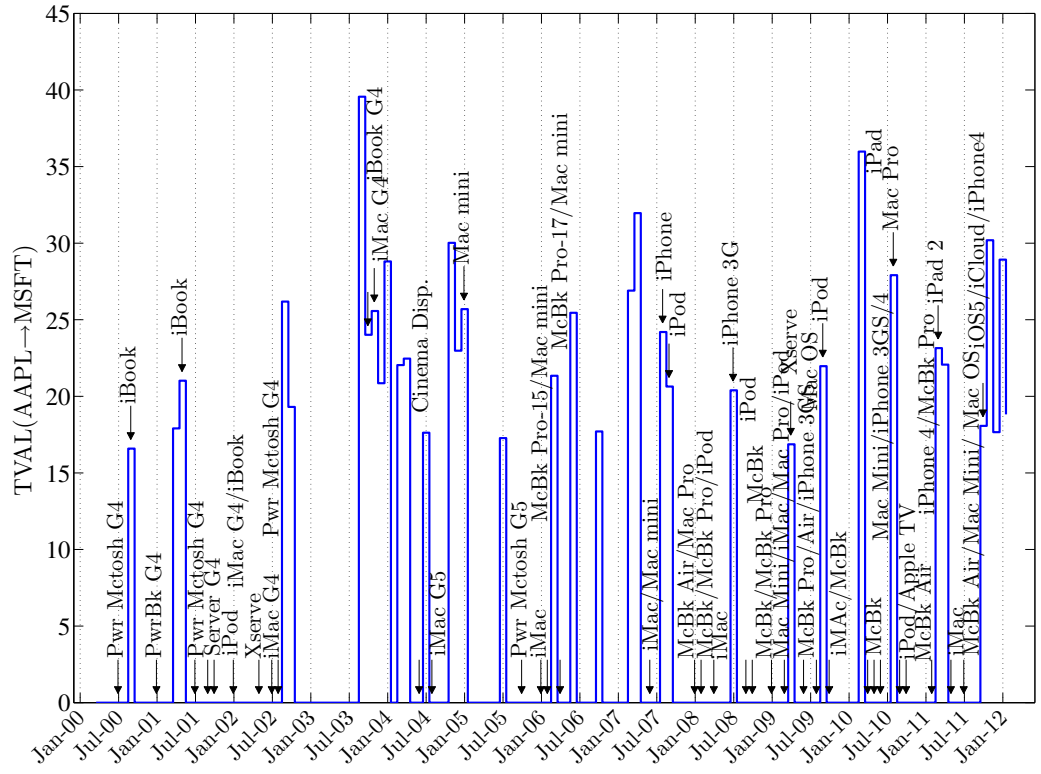


(a) The temporal causality AAPL→MSFT.

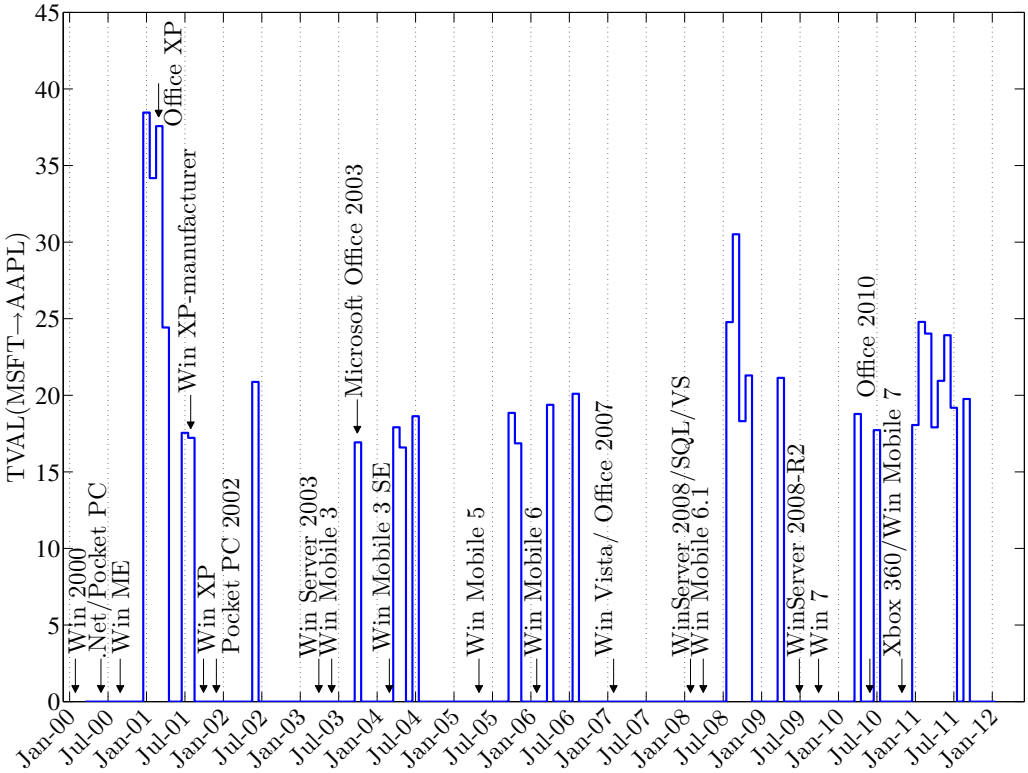


(b) The temporal causality MSFT→AAPL.

Figure 4.10: The temporal causal effect of the stock prices of past five business days on the price of the next business day between the stock prices of Apple Inc. (AAPL) and Microsoft Corporation (MSFT) detected by the CS method.



(a) The temporal causality AAPL→MSFT.



(b) The temporal causality MSFT→AAPL.

Figure 4.11: The temporal causal effect of the stock prices of past five business days on the price of the next business day between the stock prices of AAPL and MSFT detected by the HJ test.

Chapter 5

Distributional Causality Inference

5.1 Introduction

As mentioned in Sec. 1.8, we can generally categorize the causal relationship between two processes X and Y as deterministic or distributional causality. Consider two time series x_t and y_t observed from processes X and Y, respectively, where X causes Y ($X \to Y$). In the absence of the noise, a deterministic causality means that the future of y_t is a deterministic function of the lagged values of x_t (and maybe as well as y_t). In the case of distributional causality, however, the lagged values of x_t affect the underlying probability distribution of y_t. In other words, the future of y_t cannot be obtained from the lagged values of x_t (and maybe as well as y_t), but its statistical properties can.

Because of the importance of the temporal dependencies in the moments of financial data, studying the distributional causality has many applications in econometrics. For instance, in modeling of financial time series exhibiting time-varying volatility clustering, Autoregressive Conditional Heteroscedasticity (ARCH) model and its various extensions are now commonly used. For a survey of ARCH model see [85] and [86]. The progressive development of multivariate and nonlinear extensions of ARCH model implies that financial data can influence the underlying probability distribution of each other in an extremely complicated manner. Therefore, the inference of distributional causal relationships between financial time series can provide the insight needed to thoroughly analyze the financial data.

Many of currently proposed causality inference methods identify the causality based on the correlation between the time series, e.g., linear Granger causality test

([16]). However, in the case of distributional causality, the autocorrelations and cross correlations between the time series can be zero, e.g., in ARCH model-based data. Therefore, these kind of inference methods will fail in the analysis of uncorrelated data. Even though some methods, such as the nonlinear extension of Granger causality presented in [26], can handle uncorrelated time series, they only detect the existence of causality $X \to Y$, but not the type of it, i.e., deterministic or distributional causality. In other words, these methods do not identify the type of influenced moment or other statistical properties of the underlying distribution of the effect time series. As an example, assume that the financial indicator X causes the stock price of the company Y distributionally, e.g., X influences the volatility of the stock price of Y. In this case, the currently proposed causality inference methods detect the causal relationship $X \to Y$; but they do not determine whether the mean of Y or its volatility is affected by X. To answer these questions, which are fruitful for many applications such as portfolio selection and risk management, we need to develop new causality inference methods for detection of distributional causality.

In this chapter, we propose a new non-parametric method for inference of distributional causality between two time series[1]. This method is capable of identifying the type of the moments or distribution parameters influenced by the distributional causal relationship, e.g., mean, volatility, or higher order statistics. In the first step, we propose a method to estimate the temporal moments or distribution parameters of the underlying distribution of the time series. Then, we propose our new method for detection of distributional causality from the estimated temporal moments or distribution parameters.

In the next step, we numerically compare our method with a non-distributional causality inference method which is the modified nonlinear Granger causality test presented in [27], denoted by NLG-Diks test. The simulation results show that in applications with a large sample size of data, e.g., financial data, and contingent on the existence of the distributional causality, our proposed method can be superior to the NLG-Diks method. Accordingly, we present a guideline for using the proposed method and the NLG-Diks test in different situations.

We also use our method to study daily Standard and Poor's 500 index (S&P500)

[1]The results of this work were published as a conference paper entitled '*A new method for detecting non-linear causality in time series*' in proceeding of the Complex Data Modeling and Computationally Intensive Statistical Methods for Estimation and Prediction (SCo2013), Milan, Italy, Sep. 2013. Moreover, the journal paper of this work is under preparation for submitting to the *Journal of Econometrics*.

stock return and percentage change in its volume. Consistent with the results of the NLG-Diks test, we find that return causes the volume change, but we also realize that the return affects the mean of the volume change and not its volatility.

The chapter is structured as follows. Section 5.2 introduces the proposed method for inference of the distributional causality. The performance of the proposed method and the comparison with the NLG-Diks test studied by simulations are presented in Section 5.3. Next, the performance of the proposed method for analyzing the empirical financial data is studied in Section 5.4. Finally, we conclude with a discussion.

5.2 Distributional Causality

Consider two time series x_t and y_t with N samples observed from processes X and Y, respectively. We denote the underlying probability density function of y_t by $f_y(\Theta_{y,t})$ where the distribution parameter $\Theta_{y,t}$ depends instantaneously on the lagged value of x_t. By considering the maximum time lag of 1 we can say that $\Theta_{y,t} = h_{\Theta_y}(x_{t-1})$ where we assume $h_{\Theta_y}(\cdot)$ is a continuous deterministic function. Here, the process X causes the process Y. Note, however, that instead of direct influence on y_t, X affects the parameters of the underlying distribution of Y, here denoted by $X \to \Theta_y$. We refer to this case by saying X causes Y distributionally (in contrast with X being a deterministic cause of Y). Obviously, such causal relationships can be linear or nonlinear.

5.2.1 Estimation of $\Theta_{y,t}$

To detect distributional causality between X and Y, it is first necessary to estimate the time-varying parameter $\Theta_{y,t}$ from x_t and y_t. The continuity of $h_{\Theta_y}(\cdot)$ indicates that for two different time points t and t' and for any $\epsilon > 0$, there exists $\delta_x > 0$ such that $|x_{t-1} - x_{t'-1}| < \delta_x \Rightarrow |\Theta_{y,t} - \Theta_{y,t'}| < \epsilon$. From continuity of $h_{\Theta_y}(\cdot)$, it is understood that when x_{t-1} is close to $x_{t'-1}$, $\Theta_{y,t}$ would be close to $\Theta_{y,t'}$. Therefore, the underlying distribution of y_t and $y_{t'}$ are approximately equivalent, i.e., $f_y(\Theta_{y,t})$ is close to $f_y(\Theta_{y,t'})$. Consequently, the subset of the samples $\{y_{t'}\}$ can be treated as the observed samples of the distribution $f_y(\Theta_{y,t})$. Hence, we present the following estimation method for $\Theta_{y,t}$:

1. Find the time index t' of the neighbor points of x_{t-1} such that $|x_{t-1} - x_{t'-1}| < \delta_x$.

2. Construct the subset $\{y_{t'}\}$.

3. Estimate $\Theta_{y,t}$ from $\{y_{t'}\}$.

The estimation of $\Theta_{y,t}$ is denoted by $\widehat{\Theta}_{y,t}(Y_t|X_{t-1} = x_{t-1})$. In this method, δ_x can be a fixed number or determined by the standard deviation of the time series x_t, σ_x. For instance, one can use $\delta_x = \sigma_x/m$ where $m \geq 1$.

In addition to x_t, the lagged value of y_t can also affect the distribution of the future of y_t, i.e., $\Theta_{y,t} = h_{\Theta_y}(x_{t-1}, y_{t-1})$. Therefore, when studying the causal relation $X \to \Theta_y$, the causal effect of Y on $\Theta_{y,t}$ (denoted by $Y \to \Theta_y$) should be excluded. For this reason, let us focus on detection of $Y \to \Theta_y$ by estimating $\widehat{\Theta}_{y,t}(Y_t|Y_{t-1} = y_{t-1})$. In this case, we first find the time indexes t'' such that $|y_{t-1} - y_{t''-1}| < \delta_y$. Then, we estimate $\widehat{\Theta}_{y,t}(Y_t|Y_{t-1} = y_{t-1})$ from the subset $\{y_{t''}\}$. Similarly, for $X \to \Theta_y$, the time indexes t' are found such that $|x_{t-1} - x_{t'-1}| < \delta_x$. Then, $\widehat{\Theta}_{y,t}(Y_t|X_{t-1} = x_{t-1})$ is estimated from the subset $\{y_{t'}\}$. To exclude the causal effect of y_t on its future values, we simply exclude the common members of $\{y_{t'}\}$ and $\{y_{t''}\}$. In other words, we define $\{y_{t_\Delta}\} = \{y_{t'}\} - \{y_{t''}\}$. Using $\{y_{t_\Delta}\}$, $\widehat{\Theta}_{y,t}(Y_t|X_{t-1} = x_{t-1}, Y_{t-1} \neq y_{t-1})$ is estimated rather than $\widehat{\Theta}_{y,t}(Y_t|X_{t-1} = x_{t-1})$. Note that another way to obtain $\{y_{t_\Delta}\}$ is to find time indexes t_Δ such that $|x_{t-1} - x_{t_\Delta-1}| < \delta_x$ and $|y_{t-1} - y_{t_\Delta-1}| \geq \delta_y$. Hence, the steps of the parameter estimation can be modified as follows:

1. Find the time indexes t_Δ such that $|x_{t-1} - x_{t_\Delta-1}| < \delta_x$ and $|y_{t-1} - y_{t_\Delta-1}| \geq \delta_y$.

2. Construct the subset $\{y_{t_\Delta}\}$.

3. Estimate $\widehat{\Theta}_{y,t}(Y_t|X_{t-1} = x_{t-1}, Y_{t-1} \neq y_{t-1})$ from $\{y_{t_\Delta}\}$.

Again, we suggest choosing $\delta_y = \sigma_y/m$ where σ_y is the standard deviation of the time series y_t and $m \geq 1$.

It is noteworthy that if the parameter $\Theta_{y,t}$ is a function of q lagged values of x_t, i.e., $\Theta_{y,t} = h_{\Theta_y}(\mathbf{X}_{t-1}^{(q)})$ where $\mathbf{X}_{t-1}^{(q)} = [x_{t-1}, \cdots, x_{t-q}]$, then to estimate $\Theta_{y,t}$, we can find the time indexes t_Δ such that $\|\mathbf{X}_{t-1}^{(q)} - \mathbf{X}_{t_\Delta-1}^{(q)}\| < \delta_x$ and $|y_{t-1} - y_{t_\Delta-1}| \geq \delta_y$. Here, $\|\mathbf{X}_{t-1}^{(q)} - \mathbf{X}_{t_\Delta-1}^{(q)}\|$ denotes the distance between the vectors $\mathbf{X}_{t-1}^{(q)}$ and $\mathbf{X}_{t_\Delta-1}^{(q)}$. In this case, δ_x should be set appropriately based on the type of the norm used for measuring the distance between $\mathbf{X}_{t-1}^{(q)}$ and $\mathbf{X}_{t_\Delta-1}^{(q)}$.

5.2.2 Causality inference method

5.2.2.1 The proposed method

In this section, we propose a method to detect the causal effect of X on the underlying distribution of Y. Since this method identifies the distributional causality, we call it DC method. In the DC method, we detect the distributional causality by studying the distribution of the estimated parameter $\widehat{\Theta}_{y,t}(Y_t|X_{t-1} = x_{t-1}, Y_{t-1} \neq y_{t-1})$, denoted by $f(\widehat{\Theta}_{y,t})$. In order to distinguish the existence or lack of causality, we first permute the time series x_t, denoted by x_t^p, to destroy any possible existing causality between X and Y. Let us denote the parameter $\Theta_{y,t}$ estimated from x_t^p and its distribution by $\widehat{\Theta}^p_{y,t}(Y_t|X^p_{t-1} = x^p_{t-1}, Y_{t-1} \neq y_{t-1})$ and $f(\widehat{\Theta}^p_{y,t})$, respectively. Now, by comparing $f(\widehat{\Theta}_{y,t})$ and $f(\widehat{\Theta}^p_{y,t})$ one can decide the existence or lack of distributional causality between X and Y.

First, assume that the process X does not influence the parameter $\Theta_{y,t}$, denoted by $X \nrightarrow \Theta_y$. In this case, there is no relationship between x_t and y_t, and hence, the subset $\{y_{t_\Delta}\}$ derived by the time indexes of the neighbors of x_{t-1} is just a random selection of the samples of time series y_t. Hence, permutation of x_t cannot have a significant effect on this random selection. Consequently, it is likely that the estimated parameters $\widehat{\Theta}_{y,t}(Y_t|X_{t-1} = x_{t-1}, Y_{t-1} \neq y_{t-1})$ and $\widehat{\Theta}^p_{y,t}(Y_t|X^p_{t-1} = x^p_{t-1}, Y_{t-1} \neq y_{t-1})$ have the same distribution, i.e., $f(\widehat{\Theta}_{y,t})$ is close to $f(\widehat{\Theta}^p_{y,t})$. Therefore, if we measure the distance of these two distributions, it should be a small value.

On the other hand, provided that $X \rightarrow \Theta_y$, $\{y_{t_\Delta}\}$ is not a random selection of the samples of y_t. Hence, permutation of x_t can severely affect the estimated parameter of $f_y(\Theta_{y,t})$, and consequently, it is unlikely that $f(\widehat{\Theta}_{y,t})$ and $f(\widehat{\Theta}^p_{y,t})$ be close. As a result, there should be a large distance between these two distributions.

By permuting the time series x_t several times, say N_p times, and estimating $\widehat{\Theta}^p_{y,t}(Y_t|X^p_{t-1} = x^p_{t-1}, Y_{t-1} \neq y_{t-1})$ for each permutation, we will have N_p distributions $f_i(\widehat{\Theta}^p_{y,t})$ for $1 \leq i \leq N_p$ that all of them are close to each other. Hence, the average distribution of $f_i(\widehat{\Theta}^p_{y,t})$s, denoted by $\bar{f}(\widehat{\Theta}^p_{y,t})$, can be used as a distribution representing the non-causal scenario. Accordingly, we can detect the existence of causality in the original time series by measuring the distance between $f(\widehat{\Theta}_{y,t})$ and $\bar{f}(\widehat{\Theta}^p_{y,t})$. If this distance is close to zero, we conclude $X \nrightarrow \Theta_y$; otherwise, $X \rightarrow \Theta_y$.

5.2.2.2 Measuring the distance between two distributions

A common measure of the distance between two distributions $p(x)$ and $q(x)$ is the Kullback-Leibler divergence (KL), introduced in [87], defined by

$$\mathrm{KL}(p(x)\|\ q(x)) = \int_{-\infty}^{\infty} \ln\left(\frac{p(x)}{q(x)}\right) p(x)dx. \tag{5.1}$$

$\mathrm{KL}(p(x)\|\ q(x))$ is always non-negative and there is no upper bound on its value. Moreover, this measure is asymmetric, i.e., $\mathrm{KL}(p(x)\|\ q(x)) \neq \mathrm{KL}(q(x)\|\ p(x))$. To have a symmetric divergence measure, we can use $\mathrm{KL}(p(x)\|\ q(x)) + \mathrm{KL}(q(x)\|\ p(x))$. However, this measure still does not have an upper bound. To have a bounded symmetric divergence measure, the Jensen-Shannon divergence (JS) is proposed by [88], defined by

$$\mathrm{JS}(p(x)\|\ q(x)) = \tfrac{1}{2}\left[\mathrm{KL}(p(x)\|\ m(x)) + \mathrm{KL}(q(x)\|\ m(x))\right] \tag{5.2}$$

where $m(x) = \frac{p(x)+q(x)}{2}$ and we have $0 \leq \mathrm{JS}(p(x)\|\ q(x)) \leq \ln(2)$. In this work, JS is used as a measure of the distance between two distributions.

5.2.2.3 Significance threshold of JS

As mentioned in Section 5.2.2.1, to detect the causal effect $X \to \Theta_y$, we estimate $\Theta_{y,t}$ by the original and permuted time series and obtain the corresponding distributions $f(\widehat{\Theta}_{y,t})$ and $\bar{f}(\widehat{\Theta}^p_{y,t})$, respectively. To obtain the distance of these two distributions, $\mathrm{JS}(f(\widehat{\Theta}_{y,t})\|\bar{f}(\widehat{\Theta}^p_{y,t}))$ is calculated, denoted by JS_Θ. If JS_Θ has a significant value, we conclude that $X \to \Theta_y$ exists; otherwise, $X \nrightarrow \Theta_y$. Thus, we need to set a threshold against which the significance of JS_Θ is decided. If JS_Θ is greater than this threshold, we decide that the distance is significant and therefore causality exists; otherwise, we decide the causality is non-existent.

Recall that the permutation of x_t is performed N_p times and the corresponding distribution $f_i(\widehat{\Theta}^p_{y,t})$ is obtained for each permutation. The distance of all distributions $f_i(\widehat{\Theta}^p_{y,t})$ with the average distribution $\bar{f}(\widehat{\Theta}^p_{y,t})$ can be used to find a threshold for the significance level of JS_Θ. For this reason, we first calculate $\mathrm{JS}(f_i(\widehat{\Theta}^p_{y,t})\|\bar{f}(\widehat{\Theta}^p_{y,t}))$ for $1 \leq i \leq N_p$, denoted by $\mathrm{JS}^i_{\Theta^p}$. Then, the $\alpha\%$ one-sided confidence interval ($\mathrm{CI}_{\alpha\%}$) of $\mathrm{JS}^i_{\Theta^p}$ values can be considered as a threshold, denoted by $\mathrm{JS}_{\mathrm{THR}}$. The percentile bootstrap method ([52]) can be used to find $\mathrm{CI}_{\alpha\%}$, i.e., the values of $\mathrm{JS}^i_{\Theta^p}$ are sorted and $\lceil \frac{\alpha}{100} N_p \rceil$-th sorted value is considered as $\mathrm{JS}_{\mathrm{THR}}$ (see Sec. 3.2 for more information about the percentile bootstrap method).

Now, to make a decision on the existence of causality, if $\mathrm{JS}_\Theta > \mathrm{JS}_{\mathrm{THR}}$, the distance between $f(\widehat{\Theta}_{y,t})$ and $\bar{f}(\widehat{\Theta}^p_{y,t})$ is significant, hence, $X \to \Theta_y$. Otherwise, $f(\widehat{\Theta}_{y,t})$ and $\bar{f}(\widehat{\Theta}^p_{y,t})$ are close to each other, and consequently, we conclude $X \nrightarrow \Theta_y$.

5.3 Simulations

In this section, numerical results are used to illustrate different aspects of the DC method and to study its performance. Comparison with the NLG-Diks test is also provided from which some recommendations are made about when to use each method. We take note that our method is more general than NLG-Dicks in that it provides more information such as whether the time series x_t affects the mean or volatility of the time series y_t, while the NLG-Dicks test can only find the existence or lack of the causal effect of X on Y.

Consider a modified bivariate ARCH(q) process defined by

$$x_t = \sigma_{x,t} \times z_{x,t} + by_{t-1} \tag{5.3a}$$

$$y_t = \sigma_{y,t} \times z_{y,t} \tag{5.3b}$$

where

$$\sigma^2_{x,t} = \sigma^2_{y,t} = a_0 + a_1 x^2_{t-1} + \cdots + a_q x^2_{t-q}. \tag{5.4}$$

$\sigma^2_{x,t}$ and $\sigma^2_{y,t}$ are the instantaneous variance of x_t and y_t, respectively. Moreover, $z_{x,t}$ and $z_{y,t}$ are independent and identically distributed random variables with zero mean and unit variance, i.e., iid(0,1). Typically, the standard normal distribution is used for $z_{x,t}$ and $z_{y,t}$. In this bivariate process, the time series x_t influences $\sigma^2_{y,t}$ (i.e., $X \to \sigma^2_y$) and y_t affects the the mean of x_t, denoted by $\mu_{x,t}$ (i.e., $Y \to \mu_x$).

First, Let us demonstrate the performance of the parameter estimator for ARCH(1) with $a_0 = 1$, $a_1 = 0.4$, and $b = 0.2$. The simulation is performed with $N = 2000$ samples, $\delta_x = \sigma_x/4$, and $\delta_y = \sigma_y/4$. Figure 5.1 depicts the actual and estimated variances corresponding to detection of both directions of causality. For more visibility, the first 200 out of 2000 samples are shown. As it can be seen in Fig. 5.1(a), in the case of the existence of causality, the parameter estimator, almost always, accurately estimates the time-varying variance. The estimation is erroneous for very large values of $\sigma^2_{y,t}$ due to large value of x_{t-1}. Indeed, large values of x_{t-1} are observed infrequently. Hence, we can find only a small number of the neighbors of x_{t-1}, and consequently, the subset $\{y_{t_\Delta}\}$ has only a few members. Therefore, the larger the

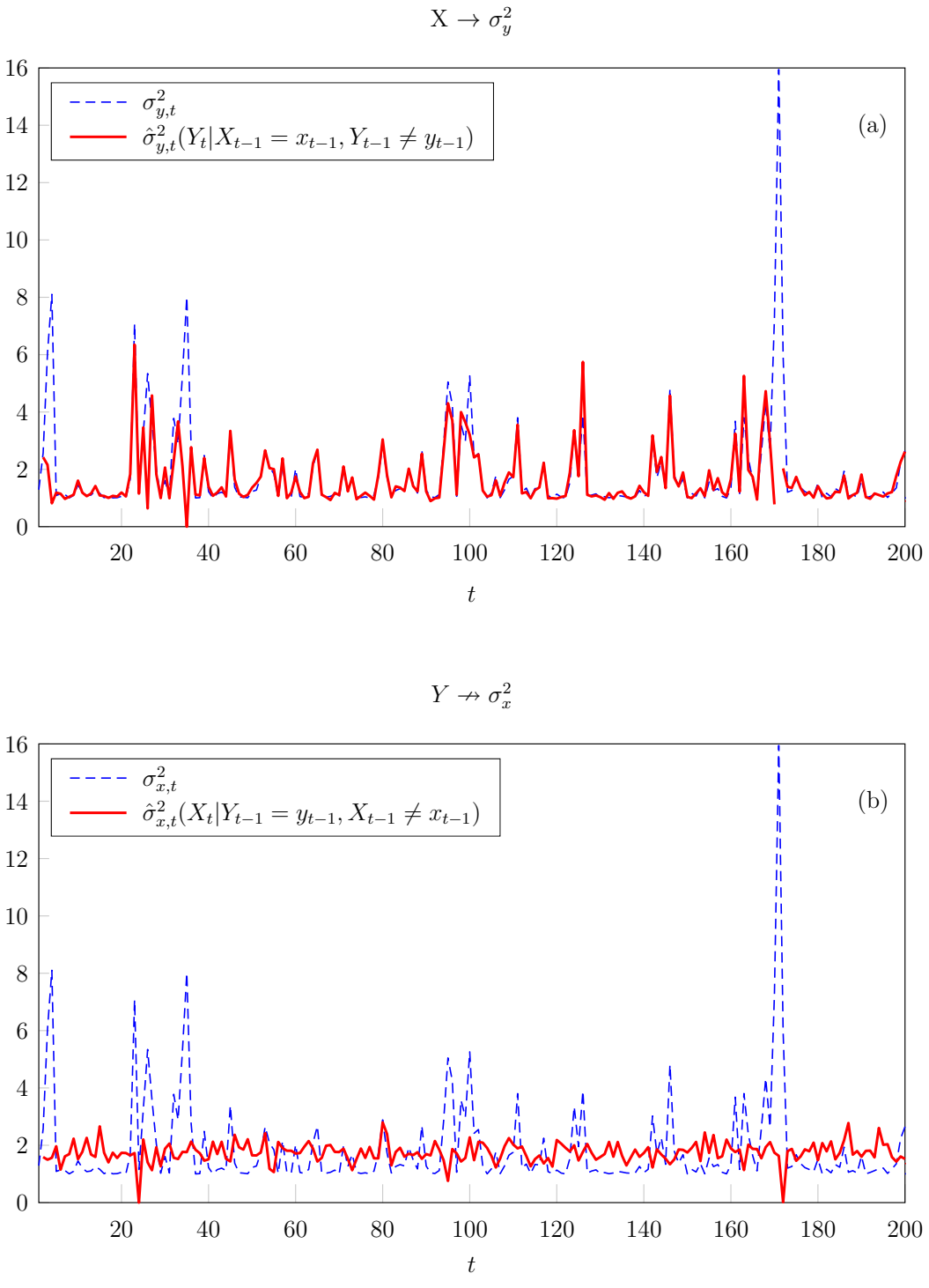


Figure 5.1: Actual and estimated values of the time-varying variance of the ARCH model for (a) $X \to \sigma_y^2$ and (b) $Y \nrightarrow \sigma_x^2$.

value of x_{t-1}, the smaller the size of $\{y_{t_\Delta}\}$, and consequently, the more erroneous is our estimation. In some cases, $\{y_{t_\Delta}\}$ has just one member or even it becomes empty. Hence, the estimated variance is zero or undefined, e.g., the estimated variances of the time points 35 and 172 in Fig. 5.1(a). In these cases, we can use $|y_t|^2$ as an estimation of $\sigma^2_{y,t}$.

Figure 5.1(b) illustrates $\sigma^2_{x,t}$ and $\hat{\sigma}^2_{x,t}(X_t|Y_{t-1} = y_{t-1}, X_{t-1} \neq x_{t-1})$ for detection of $Y \to \sigma^2_x$. As in the ARCH model (5.3), σ^2_x is not affected by y_{t-1}, no relationship is visible between the actual and estimated variances in Fig. 5.1(b). The same situation is visible for estimation of $\mu_{x,t}$ and $\mu_{y,t}$, i.e., for the existing causality $Y \to \mu_x$, $\mu_{x,t}$ and $\hat{\mu}_{x,t}(X_t|Y_{t-1} = y_{t-1}, X_{t-1} \neq x_{t-1})$ are matched together (except for the large values of $\mu_{x,t}$). However, for the reverse direction that there is no causal effect on the mean of Y, $\mu_{y,t}$ and $\hat{\mu}_{y,t}(Y_t|X_{t-1} = x_{t-1}, Y_{t-1} \neq y_{t-1})$ are totally different.

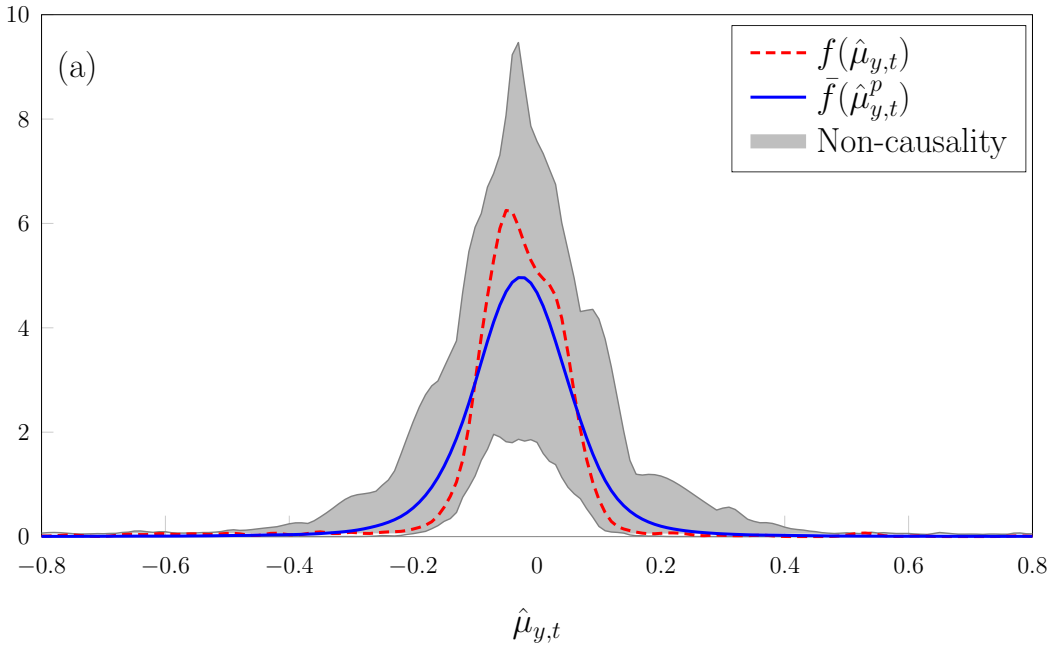
Now, let us investigate the properties of $f(\widehat{\Theta})$ where Θ is the mean or variance. This distribution is estimated by the Gaussian kernel density estimator with optimal bandwidth of the Gaussian kernel ([89]). $N_p = 200$ permutations are performed and $\text{CI}_{95\%}$ is considered as JS_{THR}.

Figure 5.2(a) demonstrates $f(\hat{\mu}_{y,t})$ and $\bar{f}(\hat{\mu}^p_{y,t})$ for detection of $X \to \mu_y$. In the bivariate process (5.3), $\mu_{y,t}$ is not affected by x_t; hence, permutation does not change the distribution of $\hat{\mu}_{y,t}$ severely. Therefore, $f(\hat{\mu}_{y,t})$ and $\bar{f}(\hat{\mu}^p_{y,t})$ are very close to each other, and consequently, their distance is a small value, i.e., $\text{JS}_{\mu_y} = 0.05$. In this case, $\text{JS}_{\mu_y} < \text{JS}_{\text{THR}}$ showing $X \nrightarrow \mu_y$. On the other hand, for the reverse direction of causality $Y \to \mu_x$, Fig. 5.2(b) depicts that $f(\hat{\mu}_{x,t})$ and $\bar{f}(\hat{\mu}^p_{x,t})$ are totally different. For this case, $\text{JS}_{\mu_x} = 0.3$ and $\text{JS}_{\text{THR}} = 0.07$, i.e., JS_{μ_x} is greater than JS_{THR} indicating the large distance between $f(\hat{\mu}_{x,t})$ and $\bar{f}(\hat{\mu}^p_{x,t})$. Therefore, we conclude $Y \to \mu_x$.

Figures 5.3(a) and 5.3(b) represent $f(\widehat{\sigma^2})$ and $\bar{f}(\widehat{\sigma^2}^p)$ for detecting $X \to \sigma^2_y$ and $Y \to \sigma^2_x$, respectively. Since in process (5.3) we have $X \to \sigma^2_y$, the distributions $f(\widehat{\sigma^2_{y,t}})$ and $\bar{f}(\widehat{\sigma^2_{y,t}}^p)$ differ considerably in Fig. 5.3(a). Here, $\text{JS}_{\sigma^2_y} = 0.5$ is larger than $\text{JS}_{\text{THR}} = 0.08$ showing the existence of causality $X \to \sigma^2_y$. However, in Fig. 5.3(b), $\text{JS}_{\sigma^2_x} = 0.02$ is smaller than $\text{JS}_{\text{THR}} = 0.07$ indicating the closeness of the distributions and lack of causality, i.e., $Y \nrightarrow \sigma^2_x$. The results shown in Figs. 5.2 and 5.3 illustrate how the distance between the distributions $f(\widehat{\Theta})$ and $\bar{f}(\widehat{\Theta}^p)$ represents the existence or lack of distributional causality.

Different permutations of x_t result in slightly different distributions $f_i(\widehat{\Theta}^p)$. The gray areas in Figs. 5.2 and 5.3 represent the area in which all these distributions

$X \to \mu_y$,$\mathrm{JS}_{\mu_y} = 0.05$, $\mathrm{JS}_{\mathrm{THR}} = 0.08$, $\mathrm{S}_{\text{min-max}} = 0$



$Y \to \mu_x$, $\mathrm{JS}_{\mu_x} = 0.3$, $\mathrm{JS}_{\mathrm{THR}} = 0.07$, $\mathrm{S}_{\text{min-max}} = 0.23$

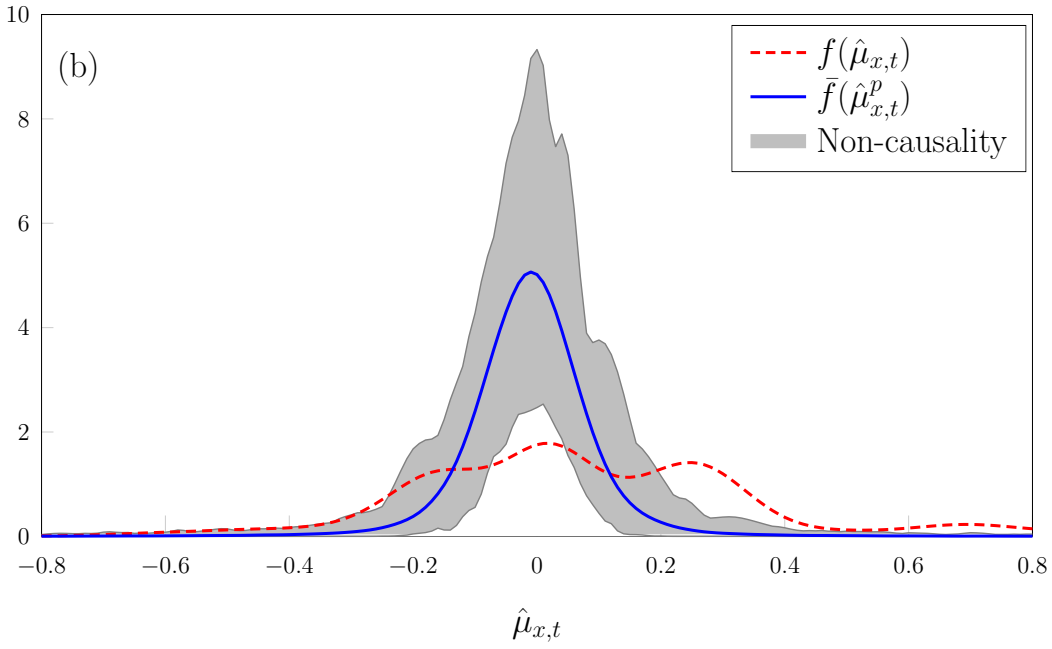


Figure 5.2: Distributions $f(\hat{\mu})$ and $\bar{f}(\hat{\mu}^p)$ are depicted for the distributional causal relationships $Y \to \mu_x$ realized by the ARCH(1) model. (a) and (b) show the causal effect of one time series on the mean of the other one. The lack of causality in (a) results in the closeness of $f(\hat{\mu}_{y,t})$ and $\bar{f}(\hat{\mu}^p_{y,t})$, and consequently, $\mathrm{JS}_{\mu} < \mathrm{JS}_{\mathrm{THR}}$ and $\mathrm{S}_{\text{min-max}} = 0$. In the case of the existence of the causal effect in (b), $f(\hat{\mu}_{x,t})$ and $\bar{f}(\hat{\mu}^p_{x,t})$ are totally different and $\mathrm{JS}_{\mu_x} > \mathrm{JS}_{\mathrm{THR}}$ and $\mathrm{S}_{\text{min-max}} > 0.05$.

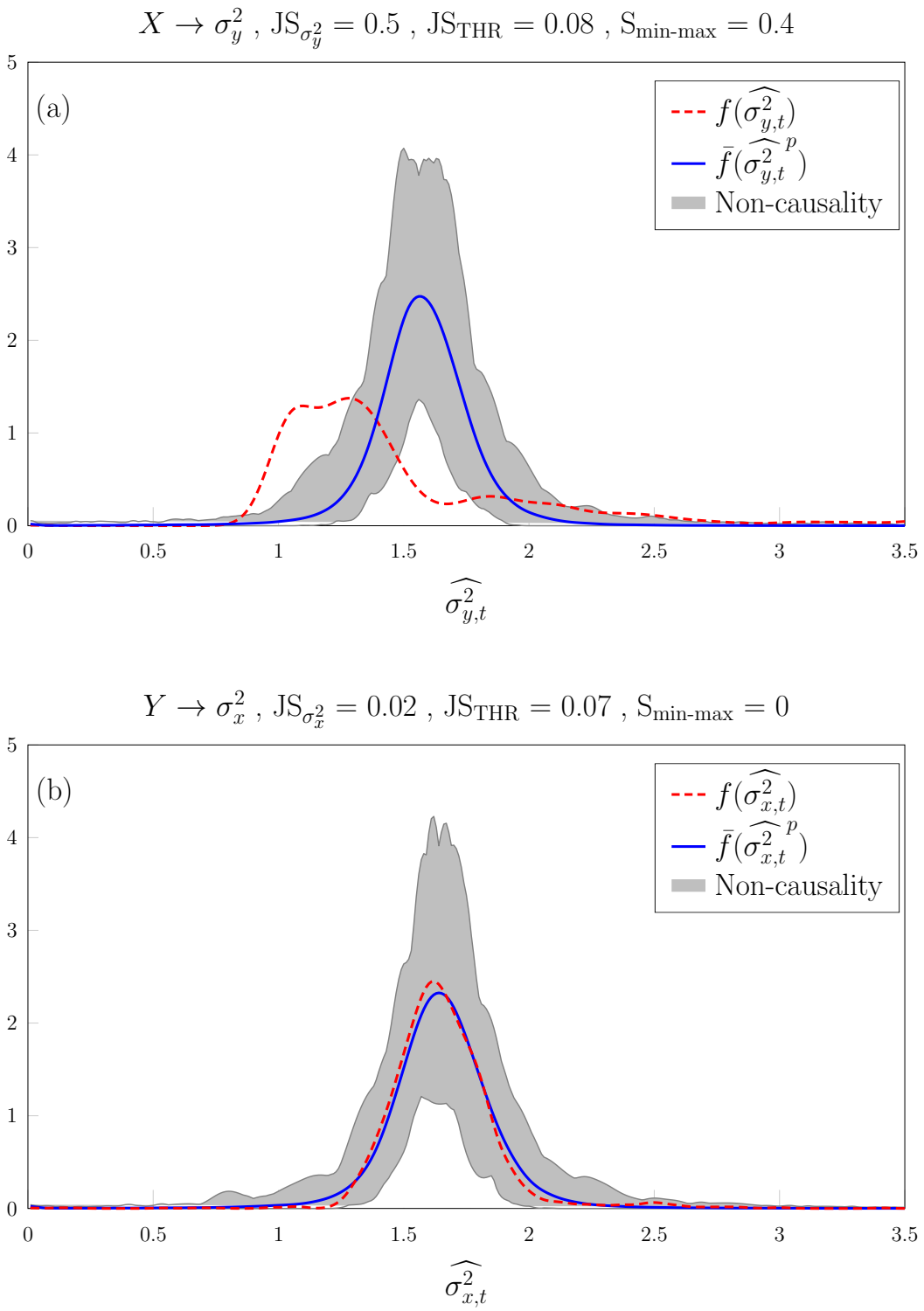


Figure 5.3: Actual and estimated values of the time-varying variance of the ARCH model for (a) $X \to \sigma_y^2$ and (b) $Y \nrightarrow \sigma_x^2$.

lie. We call this area the non-causality area because in a non-causal scenario, we expect the distribution $f(\widehat{\Theta})$ to be in the gray area. In other words, it should be indistinguishable from permuted distributions. When causality exists, however, we expect $f(\widehat{\Theta})$ not to lie entirely inside the gray area. Figure 5.2 and 5.3, clearly demonstrates this idea.

The upper border of the non-causality area represents $f_{\max}(\widehat{\Theta}^p) = \max_{1\le i\le N_p}\{f_i(\widehat{\Theta}^p)\}$. Similarly, the lower border is $f_{\min}(\widehat{\Theta}^p) = \min_{1\le i\le N_p}\{f_i(\widehat{\Theta}^p)\}$. Provided that $X \nrightarrow \Theta_y$, we expect $f(\widehat{\Theta})$ to be always between $f_{\min}(\widehat{\Theta}^p)$ and $f_{\max}(\widehat{\Theta}^p)$. However, for $X \to \Theta_y$, $f(\widehat{\Theta})$ is not entirely inside the non-causality area, i.e., it would be below $f_{\min}(\widehat{\Theta}^p)$ or above $f_{\max}(\widehat{\Theta}^p)$ in some parts. According to the relative position of $f(\widehat{\Theta})$ compared to $f_{\min}(\widehat{\Theta}^p)$ and $f_{\max}(\widehat{\Theta}^p)$, we define a new measure $\mathrm{S}_{\text{min-max}}$ to determine weather $f(\widehat{\Theta})$ is entirely inside the non-causality area. $\mathrm{S}_{\text{min-max}}$ is defined as the area enclosed between $f(\widehat{\Theta})$ and $f_{\max}(\widehat{\Theta}^p)$ where $f(\widehat{\Theta}) > f_{\max}(\widehat{\Theta}^p)$, plus the area enclosed between $f(\widehat{\Theta})$ and $f_{\min}(\widehat{\Theta}^p)$ where $f(\widehat{\Theta}) < f_{\min}(\widehat{\Theta}^p)$. In the case of the non-causal relationship, $\mathrm{S}_{\text{min-max}} = 0$ (see Figs. 5.2(a) and 5.3(b)). However, in presence of causality, $\mathrm{S}_{\text{min-max}} > 0$. For example, in Figs. 5.2(b) and 5.3(a) corresponding to $Y \to \mu_y$ and $X \to \sigma_y^2$, respectively, $\mathrm{S}_{\text{min-max}}$ equals 0.23 and 0.4, respectively. It can be shown that always $\mathrm{S}_{\text{min-max}} \le 2$. To obtain an upper bound on $\mathrm{S}_{\text{min-max}}$, for $X \to \Theta_y$ assume that the non-zero parts of $f(\widehat{\Theta}_{y,t})$ and $\bar{f}(\widehat{\Theta}^p_{y,t})$ are not overlapped and all of $f_i(\widehat{\Theta}^p_{y,t})$ distributions are equal. Therefore, $f_{\min}(\widehat{\Theta}^p)$ and $f_{\max}(\widehat{\Theta}^p)$ are same as $\bar{f}(\widehat{\Theta}^p_{y,t})$. In this case, $\mathrm{S}_{\text{min-max}}$ has its maximum value that is equal to the summation of the area under $f(\widehat{\Theta}_{y,t})$ and $\bar{f}(\widehat{\Theta}^p_{y,t})$. Hence, the maximum value of $\mathrm{S}_{\text{min-max}}$ is 2. However, in most of the cases, the observed value of $\mathrm{S}_{\text{min-max}}$ is less than 1. Based on the simulation results and provided that $\mathrm{JS}_\Theta > \mathrm{JS}_{\mathrm{THR}}$, $\mathrm{S}_{\text{min-max}} > 0.05$ is sufficiently significant to conclude the existence of causality. As we will see in the following simulation results, the extra condition of $\mathrm{S}_{\text{min-max}} > 0.05$ can reduce the false detection rate of the DC method.

The normal distribution is too light-tailed for modeling the financial data. Hence, in the following simulations, we use the student's t-distribution with degree of freedom $\nu > 2$ as $z_{x,t}$ and $z_{y,t}$ in bivariate ARCH model (5.3), denoted by t-ARCH. The samples of the student's t-distribution are normalize by its standard deviation (i.e., $\sqrt{\frac{\nu}{\nu-2}}$) to be unit variance.

Now, let us investigate the true detection rate (TDR) and false detection rate (FDR) of the DC method against the sample size and compare them with those

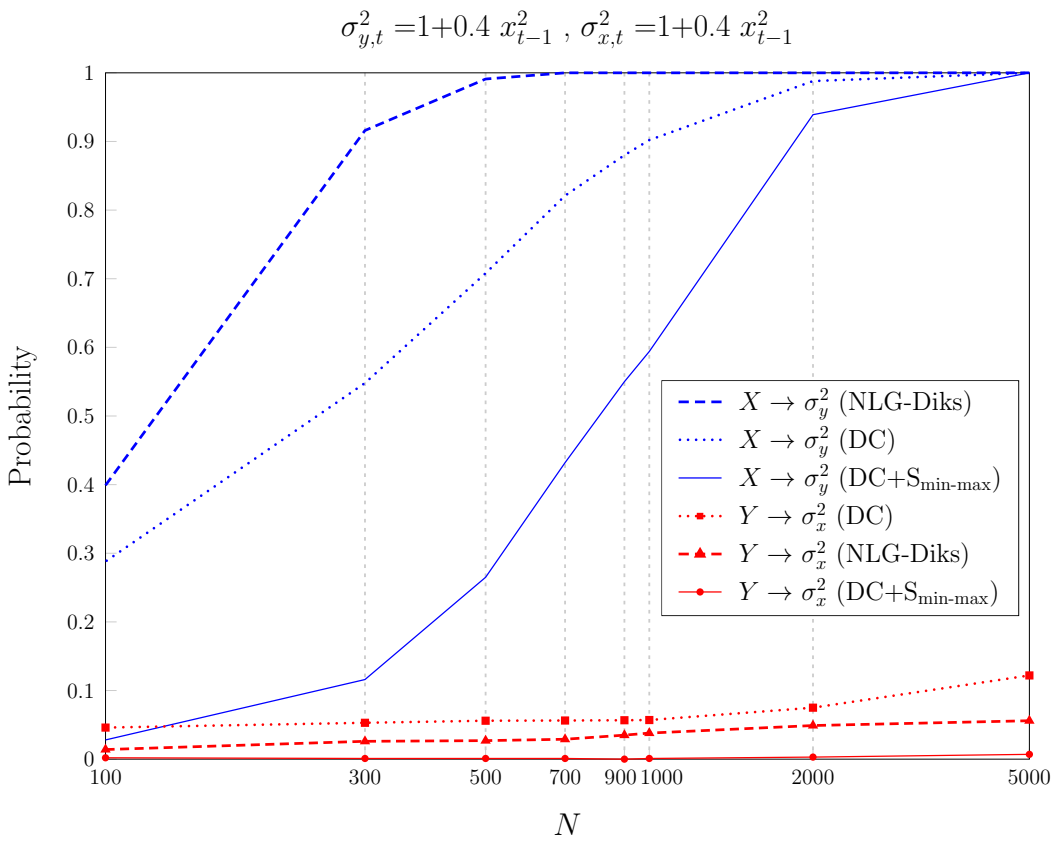


Figure 5.4: Comparison of the true and false detection rates of the DC method (with and without conditioning on $S_{\text{min-max}}$) and those of the NLG-Diks test.

of the NLG-Diks test. Here, we simulate the bivariate t-ARCH process $X \to \sigma_y^2$ for different sample sizes with $a_0 = 1$, $a_1 = 0.4$, $b = 0$, and $\nu = 4$ in equation (5.3). For each sample size, 1000 different data sets are generated and for each of them 1000, 500, and 300 permutations are performed for N =100 –1000, N = 2000, and N = 5000, respectively. $CI_{95\%}$ is used as JS_{THR} and $\delta_x = \sigma_x/10$ and $\delta_y = \sigma_y/10$. The DC method is used to detect the distributional causality for both directions of $X \to \sigma_y^2$ and $Y \to \sigma_x^2$. As there is no causal effect on $\mu_{x,t}$ and $\mu_{y,t}$, we do not show the results of mean any more. Since the current setup realizes $X \to \sigma_y^2$ and $Y \nrightarrow \sigma_x^2$, the detection probabilities corresponding to $X \to \sigma_y^2$ and $Y \to \sigma_x^2$ determine the TDR and FDR, respectively. Similarly, we find the TDR and FDR of the NLG-Diks test with the approximate optimal asymptotic bandwidth presented in [27] that is $\epsilon_N = \min(8N^{-2/7}, 1.5)$.

As Fig. 5.4 shows, without considering $S_{\text{min-max}}$, the NLG-Diks method is superior to the DC method, i.e., the TDR and FDR of the NLG-Diks method are

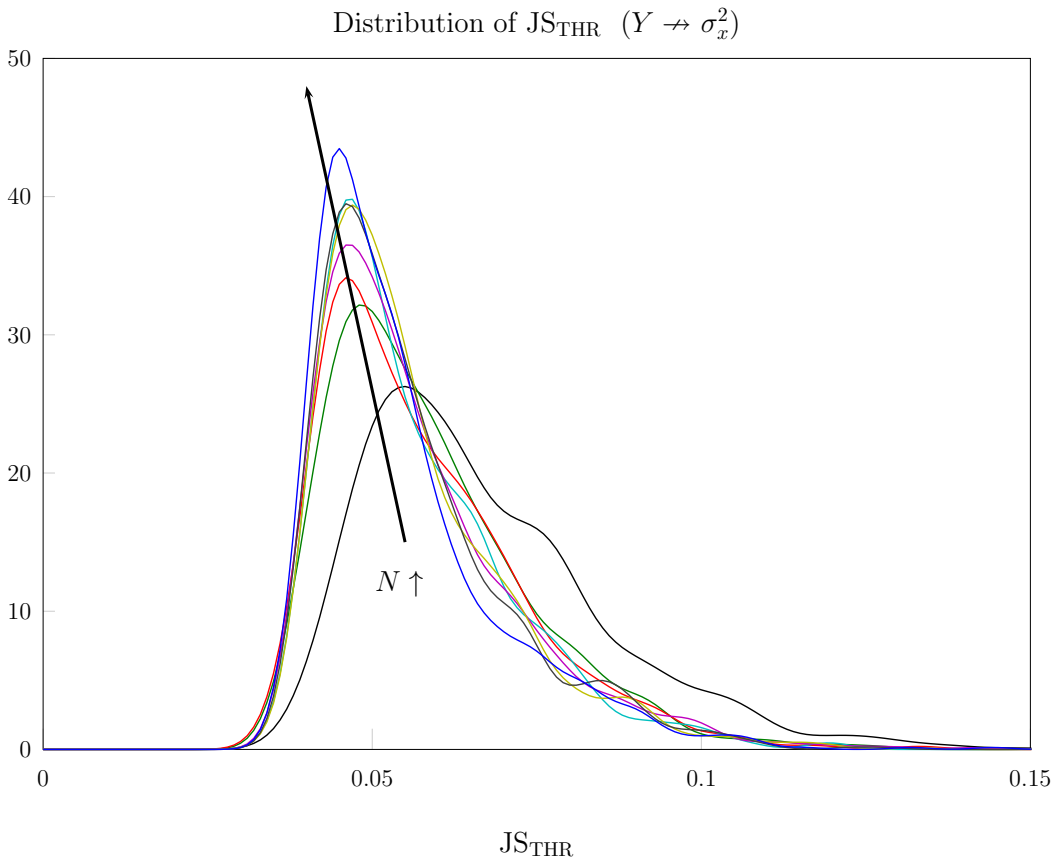


Figure 5.5: Distribution of JS_{THR} against different sample sizes (N = 100–5000) shows that increasing the sample size yields smaller values of JS_{THR}.

higher and lower than those of the DC method, respectively. Moreover, Fig. 5.4 demonstrates that the FDR of both methods increases with the sample size. For NLG-Diks test, this increase is reported in [27].

To explain the reason of the increase in the FDR of the DC method against the sample size, the distribution of JS_{THR} is depicted in Fig. 5.5 for different sample sizes. This figure shows that by increasing N, distribution of JS_{THR} has smaller mean and variance. Therefore, larger sample sizes result in smaller values of JS_{THR}. The reason for this phenomenon is that by increasing N, the deviation of distributions $f_i(\widehat{\Theta}^p)$ for $1 \leq i \leq N_p$ from each other reduces, and consequently, their distance from their average distribution decreases as well. Therefore, the value of $JS^i_{\Theta^p}$ and consequently $CI_{\alpha\%}$ or JS_{THR} reduces. On the other hand, in the case of the lack of causality, the values of JS_Θ does not change significantly with N. Therefore, for the non-causal scenario, the ratio of JS_Θ/JS_{THR} and consequently the FDR increases with N.

One way to deal with increasing the FDR for the large sample sizes is considering the additional condition $S_{\text{min-max}} > 0.05$. That is to say, we will conclude the existence of causality if $JS_{\Theta} > JS_{\text{THR}}$ and $S_{\text{min-max}} > 0.05$. As Fig. 5.4 shows for $N \geq 2000$, this additional condition reduces the FDR to 1% and it becomes 10-15 times smaller than that of the NLG-Diks test. However, for $N < 2000$, the conditioning on $S_{\text{min-max}}$ reduces both TDR and FDR of the DC method. Hence, this extra condition is not recommended for the small sample size. Here, we can present the following guidelines for using the proposed DC method and the NLG-Diks test in different situations.

- If the goal is just detecting the causal effect of X on Y, use the NLG-Diks test for $N < 2000$ and apply the DC method with conditioning on $S_{\text{min-max}}$ for $N \geq 2000$. Provided that no causality is found by the DC method for large sample sizes, apply the NLG-Diks test to detect other kinds of nonlinear causality rather than the distributional causality.

- In the second case, the goal is further than just detecting $X \rightarrow Y$. In other words, the goal is detecting the distributional causality $X \rightarrow \Theta_y$ and determining the type of Θ_y, i.e., mean, volatility, variance, or any other moments or higher-order statistics of $f_y(\Theta_{y,t})$. In this case, the NLG-Diks test is not applicable and it cannot identify the type of Θ_y. Therefore, we have to use the DC method. In this case, for $N < 2000$ exclude the conditioning on $S_{\text{min-max}}$ and for $N \geq 2000$ apply the condition $S_{\text{min-max}} > 0.05$.

5.4 Financial data

In this section, we apply the DC method to detect the causal relationships between the daily volume change and return data for the Standard and Poor's 500 index (S&P500). For the period between January 1950 and December 1990, the studies of [27] indicates that returns are influencing future volume changes strongly. However, their results show that the evidence for volume causing returns is considerably weak.

To keep our results comparable with those of [27], the same period of time is considered here for the daily stock prices of S&P500. We study the effect of one time series on the mean and volatility of the other one. Periods without trading activity due to for example weekends or holidays are excluded in stock prices and the remaining parts of the original time series are reconnected afterwards. Hence,

Table 5.1: The results of the DC method for inference of the distributional causality between the daily volume change and return of the S&P500 index.

	return$\rightarrow \mu_{\text{volume}}$	volume$\rightarrow \mu_{\text{return}}$	return$\rightarrow \sigma_{\text{volume}}$	volume$\rightarrow \sigma_{\text{return}}$
JS_{Θ}	0.15	0.1	0.08	0.12
JS_{THR}	0.08	0.07	0.095	0.13
$\text{S}_{\text{min-max}}$	0.075	0.035	0	0.02

we have 10307 samples for this period of time. The stock returns are calculated by $r_t = 100 \ln\left(\frac{P_t}{P_{t-1}}\right)$ where P_t is the daily closing prices. Moreover, to make the volume stationary, the percentage change of volume derived by differencing is considered here, i.e., $v_t = 100 \ln\left(\frac{V_t}{V_{t-1}}\right)$ where V_t is the daily volume. $N_p = 100$ permutations are performed and $\text{CI}_{95\%}$ is used to obtain JS_{THR}. Moreover, $\delta_x = \sigma_x/4$ and $\delta_y = \sigma_y/4$ are considered in parameter estimation. The results obtained with the DC method are summarized in Table 5.1.

Examination of causality on the volatility of the volume change and return denotes that in both cases we have $\text{JS}_{\sigma} < \text{JS}_{\text{THR}}$. Consequently, we conclude that return$\nrightarrow \sigma_{\text{volume}}$ and volume$\nrightarrow \sigma_{\text{return}}$. On the other hand, for both cases of the causal effect on the mean of the volume change and return, $\text{JS}_{\mu} > \text{JS}_{\text{THR}}$. However, the condition $\text{S}_{\text{min-max}} > 0.05$ is satisfied only for return$\rightarrow \mu_{\text{volume}}$. For the causation of the volume change on the mean of the return, the value of $\text{S}_{\text{min-max}} = 0.035$ provides a weak evidence for the existence of causality; hence, this result indicates the lack of coupling, i.e., volume$\nrightarrow \mu_{\text{return}}$.

Although our results are consistent with the results presented in [27], our results yield more information about the nature of the causality between the volume change and return. Indeed, instead of merely finding the causal effect of return on the volume change, we can say that return does not influence the volatility of the volume change and its effect is on the mean of the volume change. As the causal relationship return→volume is discovered in [27] by excluding the linear causality between return and volume change, we can conclude that the return affects the mean of the volume change nonlinearly.

5.5 Conclusion

In this chapter, we proposed a new method for inference of the distributional causality (DC method) between two time series. This method is able to detect the type of the moments or distribution parameters influenced by the distributional causality, e.g., mean or volatility. For distributional causality inference, we proposed a method for estimation of the time-varying moments or distribution parameters. Then, we detected the existence of distributional causality according to the distribution of the temporal estimations of the moments or distribution parameters.

To study the performance of our method, we analyzed the simulated and empirical data and the results were compared with those of the NLG-Diks method. The simulation results showed that the DC method is superior to the NLG-Diks method for large sample size of data. Then, we used the DC method to find the causality between daily S&P500 stock return and percentage change in its volume. The result was consistent with the result of the NLG-Diks test, i.e., return→ volume change. However, our method also revealed that return affects the mean of volume change and not its volatility.

Chapter 6

Conclusion and Future Work

In this chapter, we first summarize the contributions of this dissertation and conclude our work. Then, new problems are described for future research directions.

6.1 Conclusion and Summary of the Contributions

The main focus of this thesis was detection of causal relationships or couplings between different processes or systems. As these couplings or causal relationships are inherently hidden in the underlying dynamics of the system and are not necessarily accessible, we developed methods to discover these interactions by some observations of the system measured in the form of a time series. That is to say, for detection of the coupling between the driver system D and the response system R ($D \to R$), instead of manipulating D to see its effect on R, we observed the outcome of these systems and inferred the existence of coupling or causality based on these observed data.

In Chapter 3, we proposed a new method called the coupling spectrum (CS) for inference of the predictive causality or directed coupling in a deterministic system. In this method, we introduced a conditional probability which shows the effect of the past samples of the driver system D on the future value of the response system R. It was observed that this method identifies the direction of coupling in different scenarios such as unidirectional and bidirectional couplings, nonlinear dynamics, identical and nonidentical subsystems, multivariate systems, small and large sample sizes, weak and strong couplings, and in the presence of the noise. Unlike the transfer entropy method that the direction of the discovered coupling may change by scaling of the data, the CS method is scaling invariant.

Two applications of the CS method for inference of the existing couplings in

biological systems and financial data were studied in Chapter 4. In the first part of this chapter, the CS method was used for inference of the regulatory interactions between genes. Hence, the microarray data were analyzed by the CS method for inference of gene regulatory network of E2F1 transcription factor. We discovered the regulatory interactions between E2F1 transcription factor and 18 known target genes (TG) of E2F1, studied by biological experiments. We discovered 15 out of 18 known E2F1→TG interactions. We also detected 6, previously unknown, reverse interactions TG→E2F1. Further investigations revealed biological evidences for 3 of these reverse interactions.

The second application of the CS method studied in Chapter 4 is about the causality inference between financial data. We studied the causal relationships between the stock prices of Apple Inc. and Microsoft Corporation over more than a decade. Since we analyzed these data over a long period of time, we combined the CS method with overlapped moving window technique to detect time-varying causality. In this part, we compared the results of the CS method with that of the HJ test (a nonlinear extension of the Granger causality test). The outcome of these studies proved the existence of time-varying causality between financial data over a long period of time. Therefore, the moving window techniques should be applied for discovering the time-varying causality in financial data. Moreover, the results showed that most of the products of each company influence the other one's stock price immediately or a couple of months after each product release.

Finally, we proposed a new method for inference of the distributional causality (DC method) between two time series in Chapter 5. This method is able to detect the type of the moments or distribution parameters influenced by the distributional causality, e.g., mean, variance, or higher order statistics. For inference of the distributional causality, we have to deal with time-varying moments or distribution parameters. Therefore, in the first step, we proposed a method for temporal estimation of these moments or parameters. Then, according to these estimated moments or parameters we detected the existence of distributional causality. To study the performance of our method, we analyzed the simulated and empirical data and the results were compared with those of the NLG-Diks method (a modified version of the HJ test). The simulation results showed that contingent on the existence of the distributional causality, the DC method is superior to the NLG-Diks method for large sample size of data. Then, we used the DC method to find the causal rela-

tionships between daily S&P500 stock return and percentage change in its volume. The result was consistent with the result of the NLG-Diks test, i.e., return→ volume change. However, our method provided more information about the distributional nature of this relationship, i.e., it revealed that return affects the mean of volume change and not its volatility.

6.2 Future Research Directions

In the following, we define some research problems which can be studied to extend the scope of our presented works.

6.2.1 Improving the Multivariate CS Method

As presented in Sec. 3.3, the CS method is extendable to multivariate scenario. However, multivariate cases demand a significantly larger sample size of data. Consequently, in applications with small sample size, e.g., analyzing microarray data for gene regulatory network inference, we cannot directly use the multivariate CS method for detection of indirect regulatory interactions. Therefore, improving the proposed multivariate CS method or finding another extension of the CS method to multivariate scenario is important for applying the CS method to applications with the small sample sizes.

6.2.2 Bootstrapping for Causality Inference

In the cases that we are dealing with the severe conditions, such as small sample size of data or weak couplings, it is possible that we (do not) detect coupling in the case of $(D \to R)$ $D \nrightarrow R$. Therefore, under these difficult conditions, it is necessary to validate the outcome of the causality inference method.

One standard way in statistics usable in this case is the bootstrapping method [52, 90]. This method is used for estimating the distribution of an estimator or test statistic by resampling the data. The simplest method of resampling a time series is sampling the data points randomly with replacement. However, the data derived from coupled systems are usually highly correlated, hence, random selection of the samples does not generate correlated data. Therefore, we have to use block bootstrapping [90], which divides the data into blocks of observations and samples the blocks randomly with replacement. We can use non-overlapped or overlapped

blocks [91]. Therefore, we are interested to implement a block bootstrap method for the validation of the CS method or other kinds of causality inference methods.

6.2.3 Reducing the False Detection Rate of the DC method

As mentioned in Sec. 5.3, the false detection rate (FDR) of the DC method increases with the sample size. In our proposed method, we used an extra condition based on $S_{\text{min-max}}$ to mitigate the adverse effect of the sample size on FDR. However, this extra condition severely reduces the true detection rate (TDR) of the DC method for small sample sizes. In fact, by this extra condition, the DC method becomes inapplicable for very small sample size of data. Hence, finding another method for reducing the effect of the sample size on the FDR (instead of $S_{\text{min-max}}$) which has less destructive effect on TDR is necessary.

6.2.4 Applications of the Proposed Methods in Other Disciplines

Inference of the cause-effect relationships or couplings have applications in various disciplines such as process and control engineering [92,93], chemical engineering [94], and neuroscience [95]. As a future work of this research, we can apply the proposed CS and DC methods to theses applications and compare the performance of these methods with that of other existing methods.

Bibliography

[1] C. Granger, "Testing for causality. A personal viewpoint," *Journal of Economic Dynamics and Control*, vol. 2, no. C, pp. 329–352, 1980.

[2] R. Riegelman, "Contributory cause: unnecessary and insufficient," *Postgrad Med*, vol. 66, no. 2, pp. 177–179, Aug 1979.

[3] P. Suppes, Ed., *A Probabilistic Theory of Causality*. New York: North-Holland, 1970.

[4] I. J. Good, "A causal calculus," *British Journal of the Philosophy of Science*, vol. 11, pp. 305–318, 1961.

[5] R. Otte, "A critique of Suppes' theory of probabilistic causality," *Synthese*, vol. 48, no. 2, pp. 167–189, Aug. 1981.

[6] P. Smyth, "Belief networks, hidden markov models, and markov random fields: A unifying view," *Pattern Recognition Letters*, vol. 18, no. 11-13, pp. 1261–1268, 1997.

[7] M. I. Jordan, Ed., *Learning in Graphical Models (Adaptive Computation and Machine Learning)*. MIT Press, 1998.

[8] F. B. Jensen, *Bayesian Networks and Decision Graphs*. Springer, 2001.

[9] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[10] J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, California: MorganKaufmann, 1988.

[11] J. Pearl and T. S. Verma, "A theory of inferred causation," in *Principles of Knowledge Representation and Reasoning*. Morgan Kaufmann, 1991.

[12] D. Heckerman and R. Shachter, "A decision-based view of causality," Microsoft Research, Tech. Rep. MSR-TR-94-11, Mar. 1994.

[13] D. Heckerman and R. Shachter, "A definition and graphical representation for causality," pp. 262–273.

[14] N. Wiener, "The theory of prediction," in *Modern Mathematics for Engineers*, E. F. Beckenbach, Ed. NY, USA: McGraw-Hill, 1956.

[15] J. Lizier and M. Prokopenko, "Differentiating information transfer and causal effect," *European Physical Journal B*, vol. 73, no. 4, pp. 605–615, 2010.

[16] C. W. J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, vol. 37, no. 3, pp. 424–438, Jul. 1969.

[17] K. Pyragas, "Weak and strong synchronization of chaos," *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics*, vol. 54, no. 5, pp. R4508–R4511, Nov 1996.

[18] C. Schafer, M. G. Rosenblum, H. H. Abel, and J. Kurths, "Synchronization in the human cardiorespiratory system," *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics*, vol. 60, no. 1, pp. 857–870, Jul 1999.

[19] A. G. Balanov, N. B. Janson, D. E. Postnov, and O. V. Sosnovtseva, *Synchronization: From Simple to Complex*. Berlin: Springer, 2009.

[20] C. Schafer, M. G. Rosenblum, J. Kurths, and H. H. Abel, "Heartbeat synchronized with ventilation," *Nature*, vol. 392, no. 6673, pp. 239–240, Mar 1998.

[21] L. Glass, "Synchronization and rhythmic processes in physiology," *Nature*, vol. 410, no. 6825, pp. 277–284, Mar 2001.

[22] J. Jamsek, A. Stefanovska, and P. V. McClintock, "Nonlinear cardio-respiratory interactions revealed by time-phase bispectral analysis," *Phys Med Biol*, vol. 49, no. 18, pp. 4407–4425, Sep 2004.

[23] L. M. Pecora and T. L. Carroll, "Synchronization in chaotic systems," *Phys. Rev. Lett.*, vol. 64, no. 8, pp. 821–824, Feb 1990.

[24] S. J. Schiff, P. So, T. Chang, R. E. Burke, and T. Sauer, "Detecting dynamical interdependence and generalized synchrony through mutual prediction in a neural ensemble," *Phys. Rev. E*, vol. 54, pp. 6708–6724, Dec 1996.

[25] M. Le Van Quyen, J. Martinerie, C. Adam, and F. Varela, "Nonlinear analyses of interictal eeg map the brain interdependences in human focal epilepsy," *Physica D: Nonlinear Phenomena*, vol. 127, no. 3-4, pp. 250–266, 1999.

[26] C. Hiemstra and J. D. Jones, "Testing for Linear and Nonlinear Granger Causality in the Stock Price-Volume Relation," *The Journal of Finance*, vol. 49, no. 5, pp. 1639–1664, 1994.

[27] C. Diks and V. Panchenko, "A new statistic and practical guidelines for nonparametric Granger causality testing," *Journal of Economic Dynamics and Control*, vol. 30, no. 9-10, pp. 1647–1669, 2006.

[28] M. Dhamala, G. Rangarajan, and M. Ding, "Estimating Granger Causality from Fourier and Wavelet Transforms of Time Series Data," *Phys. Rev. Lett.*, vol. 100, p. 018701, Jan 2008.

[29] M. Wiesenfeldt, U. Parlitz, and W. Lauterborn, "Mixed state analysis of multivariate time series," *Int. J. Bifurcation Chaos*, vol. 11, no. 8, pp. 2217–2226, 2001.

[30] L. Bauwens, S. Laurent, and J. V. Rombouts, "Multivariate GARCH models: a survey," Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), CORE Discussion Papers, 2003.

[31] Y. Chen, G. Rangarajan, J. Feng, and M. Ding, "Analyzing Multiple Nonlinear Time Series with Extended Granger Causality," *Physics Letters A*, vol. 324, May 2004.

[32] J. Arnhold, P. Grassberger, K. Lehnertz, and C. Elger, "A robust method for detecting interdependences: application to intracranially recorded EEG," *Physica D: Nonlinear Phenomena*, vol. 134, no. 4, pp. 419 – 430, 1999.

[33] M. C. Romano, M. Thiel, and C. Kurths, J.and Grebogi, "Estimation of the direction of the coupling by conditional probabilities of recurrence," *Physical Review E*, vol. 76, no. 3, 2007.

[34] M. G. Rosenblum, A. Pikovsky, and J. Kurths, *Synchronization – A universal concept in nonlinear sciences*. Cambridge: Cambridge University Press, 2001.

[35] R. Q. Quiroga, T. Kreuz, and P. Grassberger, "Event synchronization: A simple and fast method to measure synchronicity and time delay patterns," *Phys. Rev. E*, vol. 66, p. 041904, Oct 2002.

[36] T. Schreiber, "Measuring information transfer," *Phys. Rev. Lett.*, vol. 85, no. 2, pp. 461–464, 2000.

[37] R. Marschinski and H. Kantz, "Analysing the information flow between financial time series," *The European Physical Journal B - Condensed Matter and Complex Systems*, vol. 30, no. 2, pp. 275–281, Nov. 2002.

[38] F. E. Harrell, *Springer Series in Statistics: Regression Modeling Strategies*, corrected ed. New York, NY: Springer New York, Jan. 2010.

[39] W. A. Brock, "Causality, chaos, explanation and prediction in economics and finance," in *Beyond Belief: Randomness, Prediction, and Explanation in Science*, J. Casti and A. Karlqvist, Eds. Boca Raton, FL: CRC Press, 1991, pp. 230–279.

[40] D. A. Smirnov and R. G. Andrzejak, "Detection of weak directional coupling: Phase-dynamics approach versus state-space approach," *Phys. Rev. E*, vol. 71, no. 3, p. 036207, 2005.

[41] F. W. King, *Hilbert Transforms*. Cambridge: Cambridge University Press, 2009.

[42] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ: Wiley-Interscience, July 2006.

[43] J. L. Massey, "Causality, feedback and directed information," *Proc. Int. Symp. Inf. Theory Applic. (ISITA-90)*, pp. 303–305, 1990.

[44] A. Kaiser and T. Schreiber, "Information transfer in continuous processes," *Physica D: Nonlinear Phenomena*, vol. 166, no. 1-2, pp. 43–62, 2002.

[45] D. W. Scott, *Multivariate Density Estimation: Theory, Practice, and Visualization*, 1st ed. New York: Wiley, Sep. 1992.

[46] D. W. Scott and S. R. Sain, *"Multi-Dimensional Density Estimation"*. Amsterdam: Elsevier, 2004, pp. 229–263.

[47] P. O. Amblard and O. J. J. Michel, "The relation between Granger causality and directed information theory: a review," *CoRR*, vol. abs/1211.3169, 2012.

[48] M. Palus and A. Stefanovska, "Direction of coupling from phases of interacting oscillators: An information-theoretic approach," *Phys. Rev. E*, vol. 67, p. 055201, May 2003.

[49] M. Lungarella, K. Ishiguro, Y. Kuniyoshi, and N. Otsu, "Methods for quantifying the causal structure of bivariate time series," *Int. J. Bifurcation Chaos*, vol. 17, no. 3, pp. 903–921, 2007.

[50] M. Palus and M. Vejmelka, "Directionality of coupling from bivariate time series: How to avoid false causalities and missed connections," *Phys. Rev. E*, vol. 75, p. 056211, May 2007.

[51] H. Pi and C. Peterson, "Finding the embedding dimension and variable dependencies in time series," *Neural Comput.*, vol. 6, pp. 509–520, May 1994.

[52] B. Efron, *The Jackknife, the bootstrap and other resampling plans*. Philadelphia, PA, USA: SIAM, 1982.

[53] M. Itoh, T. Yang, and L. Chua, "Conditions for impulsive synchronization of chaotic and hyperchaotic systems," *Int. J. Bifurcation Chaos Appl. Sci. Eng.*, vol. 11, no. 2, pp. 551–560, 2001.

[54] D. Bernardo, M. Thompson, T. Gardner, S. Chobot, E. Eastwood, A. Wojtovich, S. Elliott, S. Schaus, and J. Collins, "Chemogenomic profiling on a genome-wide scale using reverse-engineered gene networks," *Nature Biotechnology*, vol. 23, no. 3, pp. 377–383, 2005.

[55] T. Chen, H. He, and G. Church, "Modeling gene expression with differential equations." *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 29–40, 1999.

[56] S. Bornholdt, "Boolean network models of cellular regulation: Prospects and limitations," *Journal of the Royal Society Interface*, vol. 5, no. SUPPL. 1, pp. S85–S94, 2008.

[57] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian Networks to Analyze Expression Data.„" *Journal of Computational Biology*, no. 3-4, pp. 601–620.

[58] T. Tung, T. Ryu, K. Lee, and D. Lee, "Inferring gene regulatory networks from microarray time series data using transfer entropy," 2007, pp. 383–388.

[59] H. Lodish, A. Berk, C. Kaiser, M. Krieger, M. Scott, A. Bretscher, H. Ploegh, and P. Matsudaira, *Molecular Cell Biology*, 6th ed. W. H. Freeman, 2007.

[60] F. Emmert-Streib and M. Dehmer, Eds., *Analysis of Microarray Data: A Network-based Approach*. Wiley VCH Publishing, 2008.

[61] A. Pandey and M. Mann, "Proteomics to study genes and genomes," *Nature*, vol. 405, no. 6788, pp. 837–846, 2000.

[62] J. Stuart, E. Segal, D. Koller, and S. Kim, "A gene-coexpression network for global discovery of conserved genetic modules," *Science*, vol. 302, no. 5643, pp. 249–255, 2003.

[63] A. Butte and I. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements." *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 418–429, 2000.

[64] R. Steuer, J. Kurths, C. Daub, J. Weise, and J. Selbig, "The mutual information: Detecting and evaluating dependencies between variables," *Bioinformatics*, vol. 18, no. SUPPL. 2, pp. S231–S240, 2002.

[65] S. A. Kauffman, "Metabolic stability and epigenesis in randomly constructed genetic nets," *Journal of Theoretical Biology*, vol. 22, no. 3, pp. 437–467, Mar. 1969.

[66] A. Hartemink, D. Gifford, T. Jaakkola, and R. Young, "Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks." in *Pacific Symposium on Biocomputing*, 2001, pp. 422–433.

[67] P. D'haeseleer, X. Wen, S. Fuhrman, and R. Somogyi, "Linear modeling of mrna expression levels during cns development and injury." *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 41–52, 1999.

[68] E. van Someren, L. Wessels, and M. Reinders, "Linear modeling of genetic networks from experimental data." *Proc. Int. Conf. Intell. Syst. Mol. Biol. (ISMB)*, vol. 8, pp. 355–366, 2000.

[69] G. Cooper, "A bayesian method for learning belief networks that contain hidden variables," *Journal of Intelligent Information Systems*, vol. 4, no. 1, pp. 71–88, 1995.

[70] A. V. Werhli and D. Husmeier, "Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge," in *Stat. Appl. Genet. Mol. Biol, 6:Article 15.* The Berkeley Electronic Press, May 2007.

[71] T. Lee et al., "Transcriptional regulatory networks in saccharomyces cerevisiae," *Science*, vol. 298, no. 5594, pp. 799–804, 2002.

[72] R. Van Berlo, E. Van Someren, and M. Reinders, "Studying the conditions for learning dynamic bayesian networks to discover genetic regulatory networks," *Simulation*, vol. 79, no. 12, pp. 689–702, 2003.

[73] B. Perrin, L. Ralaivola, A. Mazurie, S. Bottani, J. Mallet, and F. d'Alché Buc, "Gene networks inference using dynamic bayesian networks," *Bioinformatics*, vol. 19, no. SUPPL. 2, pp. ii138–ii148, 2003.

[74] P. Lavia and P. Jansen-Dürr, "E2F target genes and cell-cycle checkpoint control," *BioEssays*, vol. 21, no. 3, pp. 221–230, 1999.

[75] J. DeGregori, "The genetics of the E2F family of transcription factors: shared functions and unique roles," *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, vol. 1602, no. 2, pp. 131 – 150, 2002.

[76] M. Whitfield, G. Sherlock, A. Saldanha, J. Murray, C. Ball, K. Alexander, J. Matese, C. Perou, M. Hurt, P. Brown, and D. Botstein, "Identification of genes periodically expressed in the human cell cycle and their expression in tumors," *Molecular Biology of the Cell*, vol. 13, no. 6, pp. 1977–2000, 2002.

[77] Y. Takahashi, J. Rayman, and B. Dynlacht, "Analysis of promoter binding by the E2F and pRB families in vivo: Distinct E2F proteins mediate activation and repression," *Genes and Development*, vol. 14, no. 7, pp. 804–816, 2000.

[78] S. Hiebert, M. Lipp, and J. Nevins, "E1A-dependent trans-activation of the human MYC promoter is mediated by the E2F factor." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, no. 10, pp. 3594–3598, 1989.

[79] B. Di Fiore, G. Guarguaglini, A. Palena, R. M. Kerkhoven, R. Bernards, and P. Lavia, "Two E2F sites control growth-regulated and cell cycle-regulated transcription of the Htf9-a/RanBP1 gene through functionally distinct mechanisms," *J. Biol. Chem.*, vol. 274, no. 15, pp. 10 339–10 348, Apr 1999.

[80] A. Blais, D. Monté, F. Pouliot, and C. Labrie, "Regulation of the human cyclin-dependent kinase inhibitor p18INK4c by the transcription factors E2F1 and Sp1," *Journal of Biological Chemistry*, vol. 277, no. 35, pp. 31 679–31 693, 2002.

[81] L. Delavaine and N. B. La Thangue, "Control of E2F activity by p21Waf1/Cip1," *Oncogene*, vol. 18, no. 39, pp. 5381–5392, Sep 1999.

[82] M. Xu, K. A. Sheppard, C. Y. Peng, A. S. Yee, and H. Piwnica-Worms, "Cyclin A/CDK2 binds directly to E2F-1 and inhibits the DNA-binding activity of E2F-1/DP-1 by phosphorylation," *Mol. Cell. Biol.*, vol. 14, no. 12, pp. 8420–8431, Dec 1994.

[83] J. W. Harbour, R. X. Luo, A. Dei Santi, A. A. Postigo, and D. C. Dean, "Cdk phosphorylation triggers sequential intramolecular interactions that progressively block Rb functions as cells move through G1," *Cell*, vol. 98, no. 6, pp. 859–869, Sep 1999.

[84] N. Yörük, C. Erdem, and M. Erdem, "Testing for linear and nonlinear Granger Causality in the stock price-volume relation: Turkish banking firms' evidence," *Applied Financial Economics Letters*, vol. 2, no. 3, pp. 165–171, 2006.

[85] T. Bollerslev, R. Chou, and K. Kroner, "ARCH modeling in finance. A review of the theory and empirical evidence," *Journal of Econometrics*, vol. 52, no. 1-2, pp. 5–59, 1992.

[86] A. Silvennoinen and T. Teräsvirta, "Multivariate GARCH models," Stockholm School of Economics, Working Paper Series in Economics and Finance 669, 2008.

[87] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statistics*, vol. 22, pp. 79–86, 1951.

[88] J. Lin, "Divergence measures based on the shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[89] B. W. Silverman, *Density Estimation for Statistics and Data Analysis*. Chapman and Hall/CRC, April 1986.

[90] W. Härdle, J. Horowitz, and J. P. Kreiss, "Bootstrap Methods for Time Series," *International Statistical Review / Revue Internationale de Statistique*, vol. 71, no. 2, pp. 435–459, 2003.

[91] P. Hall, "Resampling a coverage process," *Stochastic Process Applications*, vol. 19, pp. 259–269, 1985.

[92] L. Desborough and R. Miller, "Increasing customer value of industrial control performance monitoring - honeywell's experience," vol. 98, no. 326, 2002, pp. 169–189.

[93] L. Chiang and R. Braatz, "Process monitoring using causal map and multivariate statistics: Fault detection and identification," *Chemometrics and Intelligent Laboratory Systems*, vol. 65, no. 2, pp. 159–178, 2003.

[94] M. Bauer, J. Cox, M. Caveness, J. Downs, and N. Thornhill, "Finding the direction of disturbance propagation in a chemical process using transfer entropy," *IEEE Transactions on Control Systems Technology*, vol. 15, no. 1, pp. 12–21, 2007.

[95] R. Vicente, M. Wibral, M. Lindner, and G. Pipa, "Transfer entropy-a model-free measure of effective connectivity for the neurosciences," *Journal of Computational Neuroscience*, vol. 30, no. 1, pp. 45–67, 2011.