# Causality Analysis Advancements and Applications by Subspace-based Techniques

**Xu Huang**

Faculty of Management

Bournemouth University

A thesis submitted in partial fulfilment of the requirements of

Bournemouth University for the degree of

*Doctor of Philosophy*

December 2017

To

My Loving Parents and Grandparents
*who gave me life, raised me with hard works and believe in me*

The Memory of My Late Grandfather
*who was always so proud of me*

AND

Zixuan Yang
*who encourages me, gives me strength, supports and accompanies*
*every step of my big or small dreams …*

# Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university.

Xu Huang  
December 2017

# Acknowledgements

I would like to acknowledge my sincere gratitude to Dr. Hossein Hassani and Dr. Mansi Ghodsi for offering this research opportunity, and your supports, guidance and kindness all along this journey. The experiences and knowledge you generously shared during the last three years are still guiding me every day.

Also, I would like to thank the Faculty of Management and Bournemouth University for providing the full scholarship and amazing facilities for research, and all the staff at Faculty of Management who have been always kind, friendly and helpful.

A special thanks to Dr. George Filis for his expertise, trust and supports that assure the successful completion of this research. Also, I would like to thank the external examiner and the internal examiner, for their valuable time and comments that further improved the quality of this thesis.

# List of Publications

Hassani, H., Ghodsi, M., Huang, X. and Silver, E. (Under Review). Big Data in Banking: A Review and Outlook for the Future.

Hassani, H., Huang, X., Silver, E. and Ghodsi, M. (Under Review). Causality between tourist arrivals and oil prices: evidences from US and EU nations.

Huang, X., Hassani, H. and Ghodsi, M. (2017). Big data and causality. Annals of Data Science, 1-24. Springer.

Huang, X., Hassani, H., Ghodsi, M., Mukherjee, Z. and Gupta, R. (2017). Do trend extraction approaches affect causality detection in climate change studies? Physica A: Statistical Mechanics and its Applications, 469, 604-624. Elsevier.

Ghodsi, Z., Huang, X. and Hassani, H. (2017). Causality analysis detects the regulatory role of maternal effect genes in the early Drosophila embryo. Genomics Data, 11, 20-38. Elsevier.

Hassani, H., Huang, X., Gupta, R. and Ghodsi, M. (2016). Does sunspot numbers cause global temperatures? A reconsideration using non-parametric causality tests. Physica A: Statistical Mechanics and its Applications, 460, 54-65. Elsevier.

Hassani, H., Huang, X., Silver, E. and Ghodsi, M. (2016). A review of Data Mining in crime analysis. Statistical Analysis and Data Mining: The ASA Data Science Journal. John Wiley & Sons. doi:10.1002/sam.11312.

Huang, X., Ghodsi, M. and Hassani, H. (2016). A novel similarity measure based on eigenvalue distribution. Transactions of A. Razmadze Mathematical Institute, 32. Elsevier.

Huang, X. and Ghodsi, M. (2016). A novel mutual association measure based on eigenvalue-based criterion. International Journal of Energy and Statistics, 4(04), 1650017. World Scientific.

Ghodsi, M. and Huang, X. (2015). Causality between energy poverty and economic growth in Africa: evidences from time and frequency domain causality technique. International Journal of Energy and Statistics, 4(03). World Scientific.

Huang, X. (2015). A comparison of association measures: Evidence from stock markets and oil prices. International Journal of Energy and Statistics, 3(03). World Scientific.

# Abstract

Causality analysis remains a fundamental research question and the ultimate objective for many scientific studies. Alongside the increasing speed of data science and technological advancements, as well as the overwhelming existence of complex systems in social science and economics studies, causality analysis has become more complex than ever. The drawbacks of the existing empirical methods (parametric and limited nonparametric approaches) are gradually revealed through implementations. There are increasing number of proofs that the existing methods are limited and fail to catch up the rapid progress of the causality analysis study. Therefore, it is both crucial and time-sensitive to establish the advancements of causality analysis methods by embracing the advanced time series analysis techniques.

Subspace-based techniques adopted in this thesis include Singular Value Decomposition (SVD), Singular Spectrum Analysis (SSA) and Convergent Cross Mapping (CCM). These subspace-based techniques have been proved powerful nonparametric time series analysis techniques with promising performances on various fields, for instance, time series denoising, filtering, forecasting, signal extraction, image processing, etc. This thesis aims to expand the multivariate extension of subspace-based techniques on causality analysis and brings novel contributions to not only the theoretical advancements of causality analysis methods but also broadening the horizon of the corresponding applications in complex systems like climate change, economics and genetic science.

This research project focuses on, but is not limited to the causality detection test. In particular, the thesis initially proposed four novel multivariate analysis methods based on the study of subspace-based techniques: the similarity measure based on eigenvalue distribution; the mutual association measure based on eigenvalue-based criterion; the causality detection method based on multivariate SSA forecasting accuracy; the hybrid causality detection approach by combining SSA and CCM. Moreover, this thesis also introduces CCM in details and expands its implementations in climate change, oil-tourism study, and gene regulatory role detection. The advantages of these methods are that they are nonparametric approaches, assumption free, only two key variables needed, no limitations to nonlinearity or complex dynamics, signal and noise together as a whole as the research object. Both simulations and a number of successful implementations are conducted for the critical evaluation of the proposed advancements with promising robust performances. Specifically, the novel similarity measure overcomes the difficulties of empirical similarity measures through identifying the comparable criterion, and it is proved robust among various types of series. The novel mutual association measure has no restriction on nonlinearity, it performs well with various generated linear and nonlinear association patterns, as well as real data from oil-stock market and oil-tourism studies. SSA causality test, CCM causality and the SSA-CCM hybrid causality tests are comprehensively evaluated by comparing with empirical Granger approaches respectively and with two key variables considered, the results of applications significantly reflect their advantages on nonlinear dynamics and causality detection in complex systems.

In general, this thesis contributes on offering novel solutions to the crucial question of causality analysis. However, causality analysis contains a broad range of integrated disciplines, and it has the characteristics of cross discipline, strong practicality and intimate connection with other academic fields. It is such a broad subject that no study can independently comprise all. Therefore, this research attempts to provide evidence of successful applications in a possibly wide range of subject-s rather than one subject only so to initially evident on the applicability of these novel methods. The applications have covered studies of climate change, oil-stock market, oil-tourism relationship, gene regulatory role detection to date and more future works are in progress. These novel approaches are self-contained to address the corresponding advancements, therefore, they are not comparable between each other, but all contribute differently to the development of causality analysis in a broad sense. These newly proposed approaches offer the interested parties a different angle to resolve the causality analysis questions in a reduced form, data-oriented perspective. It is also expected to open up the research opportunities of nonparametric multivariate analysis through the advanced, inclusive subspace-based techniques that show strong adaptability and capability in the study of complex systems.

# Table of Contents

# List of Figures

# List of Tables

# Nomenclature

| | |
|---|---|
| *ADF* | Augmented Dickey-Fuller |
| *ASCII* | American Standard Code for Information Interchange |
| *CCM* | Convergent Cross Mapping |
| *CDF* | Cumulative Distribution Function |
| *DF-GLS* | Dickey-Fuller Test with Generalised Least Squares Detrended Residuals |
| *DGT* | Detrended Global Temperature |
| *ECG* | Electrocardiogram |
| *EEG* | Electroencephalography |
| *EMD* | Empirical Mode Decomposition |
| *GC* | Granger Causality |
| *GISS* | Goddard Institute for Space Studies |
| *GRN* | Gene Regulatory Network |
| *GT* | Global Temperature |
| *HEN* | Henderson Filter |
| *HP* | Hodrick Prescott Filter |
| *KPSS* | Kwiatkowski-Phillips-Schmidt-Shin |
| $K-S$ *Test* | Kolmogoriv-Smirnov Test |
| *LOESS* | Local Regression |
| *LRF* | Linear Recurrent Formula |
| *MA* | Moving Average |
| *MBA* | Model Based Approach |
| *MI* | Mutual Information |
| *MIC* | Maximal Information Coefficient |
| *MSSA* | Multivariate Singular Spectrum Analysis |
| *NP* | Ng and Perron Test |
| *PCA* | Principal Component Analysis |
| *PP* | Phillips and Perron Test |
| *SIC* | Schwarz Information Criterion |
| *SIDC* | Solar Influences Data Analysis Centre |
| *SS* | Sunspot Number |
| *SSA* | Singular Spectrum Analysis |
| *SVD* | Singular Value Decomposition |
| *TC* | Time Class |
| *VAR* | Vector Autoregression Model |
| *WAV* | Wavelet |

# Chapter 1

# Introduction

This chapter is aimed at clarifying the motivations behind this research and the main aims of the study. The first section introduces the motivation and aim from broad aspects, followed by the research framework including both fundamental philosophy and detailed blueprint of this research. Eventually, the research contributions are summarized in the third section.

## 1.1 Motivation and Aim

Alongside the consistent exploration of causes since the beginning of human history, which is driven by the instinctive desire of knowledge, causality analysis has been one of the fundamental areas of study regardless of the research area (Wold, 1954). In line with the rapid developments of society and technology over the past decades, causality analysis has been extensively exploited on a tremendous amount of subjects that cover almost all aspect of research (Clark et al., 2015; Granger, 1969; Hassani et al., 2010c; Hsiao, 1979; Sugihara et al., 2012). Scholars have been persistently pursuing the truth of changes and relationships, which indicates one of the ultimate purposes of researches – seeking the better understanding of the complexity and answering the WHY question.

On this basis, this study focuses on the time series analysis perspective due to the wide scope of causality analysis. Time series analysis has been widely used by researchers to investigate the dependence relationship, mostly linear, among factors in a complex system. Using time series methods enables researchers to evaluate what happened in the past, and to make current year comparisons among factors, as well as new insights into the future. This will provide the flexibility to address important questions, such as whether the changes of one factor have relationship with the changes of another factor in the current sequence or after specific lag of time.

Another question that frequently arises in various academic disciplines is whether one time series can help in analysing/predicting another. One way to address this question was proposed by Granger (1969) and various related researches have been conducted based on this. However, the definition given by Granger on causality is based on classical time series analysis approaches which are extremely limited by several restrictive assumptions. Firstly, not all relationships can be explained by linear based models. The corresponding revolutions of nonlinear applicability have no significant breakthrough as it is only expended to limited number of nonlinear relationships. Secondly, it is build on the assumption of a particular type of relationship based on a few selected factors, it is even fixed before the actual analysis starts. It is not convincing and satisfying enough to explain these existing relationships in a far more complex system than the assumed model. Moreover, it assumes separability between the variables which can be eliminated from the overall system, and this assumption is often not satisfied for complex systems (Sugihara et al., 2012). In this case, these models are inappropriate for a research that studies the causes in a complex system that generally exists nowadays, let alone achieving accurate analysis on the possible complex nonlinearity where these linear or restricted nonlinear models fail to achieve.

For the past decades, along with the enhancements of knowledge on time series, there have been various developments on time series analysis techniques, among which, the sub-space based techniques have shown remarkable performance and has made significant contributions in terms of incorporating complex circumstance like non-linearity. The representative sub-space based techniques include Singular Value Decomposition (SVD), Singular Spectrum Analysis (SSA) and Convergent Cross Mapping (CCM), which are also the key techniques that this study considers. They have been proved powerful and robust techniques on time series decomposing, filtering, denoising, forecasting as well as multivariate analysis (Alter et al., 2000; Deyle et al., 2013; Hassani et al., 2013b; Rajwade et al., 2013; Silva et al., 2017; Sugihara et al., 2012; Vautard et al., 1992a; Ye et al., 2015).

This study aims at theoretical advancements of causality analysis methods by embracing the advanced subspace-based techniques, consequently, to propose sufficient novel approaches for causality analysis that not only overcome the shortages of empirical methods, but also has no restrictions on linearity with sufficient performance on nonlinear dynamics. The development of a method which is both practical and theoretically robust is of paramount importance in time series analysis. Thus, this study offers critical evaluations through comprehensive simulations as well as sufficient amount of applications that covers diverse disciplines.

## 1.2 Research Framework

This section addresses the framework of this research following the classical research process of philosophy, literature review, methodology and application. Furthermore, this section aims to provide the overall view of this research and clarify the significant internal connections and relationships among all components of this study.

### 1.2.1 Fundamental Philosophy

The fundamental philosophy of this research is the philosophy of causes by Aristotle (384-322 B.C.) in the *Physics* and in the *Metaphysics*, detailed introduction and summary can be found in the work of Falcon (2015), which is also followed as the main reference in this subsection to provide brief information on philosophy of cause. As stated by Falcon (2015), answering the "why" question as the first step of comprehensive understanding of a thing now is considered as a request of explanation. Aristotle claimed to answer the questions that how many kinds of causes exist and what is the definition of cause, so to explain changes in the world. Two types of causes defined by Aristotle are adopted by this research for the development of novel causality analysis methods: **the formal cause** and **the efficient cause**. Specifically, they can be distinguished as follows based on what Aristotle claimed in *Physics*, Book ii:

**The Formal Cause** is "the account of what-it-is-to-be", or "what makes a thing one thing rather than many things".

**The Efficient Cause** is "the primary source of the change or rest", or simply as "initiator of the movement".

Aristotle claimed to offer systematic method of better "learning and understanding things", also in the meantime Aristotle accepts "their lacking of complete understanding of the range of possible causes and their systematic interrelations". As a topic as complex as "causes" existing in the current intricate phenomena of society, also considering the ranges of literature in causality analysis studies, in order to present clear and reliable research process and achievements, these two types of causes are referred as the principles of investigating "causes" and understanding causal relationships based on subspace-based technique in the rest of this study with evidences of both simulation and real world data implementations. As specifically detailed in Fig. 1.1, studies of similarity and association aspects are classified accordingly as the formal cause. In accordance with the concept of philosophy in efficient cause, it contains the empirical causality analysis methods that are based on linear models and also serves as the fundamental philosophy for the newly developed causality detection methods in this research.

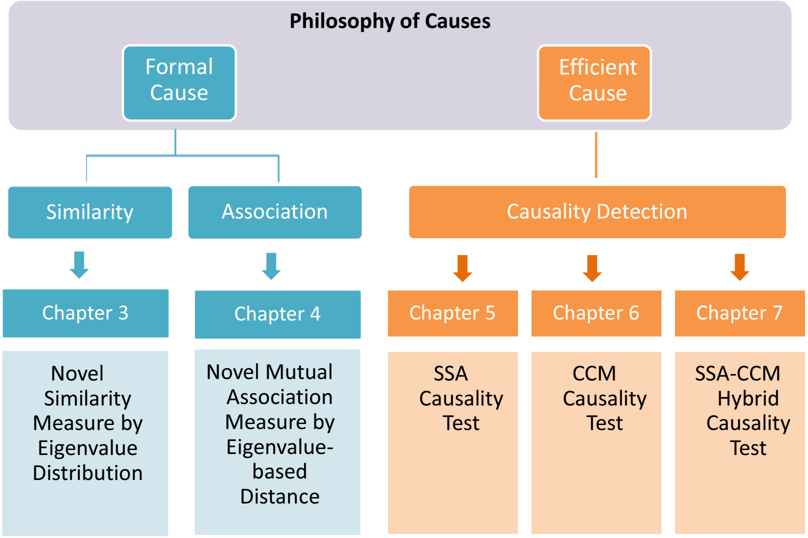


Fig. 1.1 Framework of Research Philosophy.

### 1.2.2 Framework of Study

This research contains both the theoretical advancements of causality analysis methods and corresponding broad implementations from the time series analysis aspect. As the initial inspiration of this research, the advanced subspace-based techniques that are adopted in this research are firstly introduced in **Chapter 2**. Specifically, each subspace-based technique is presented in detail including literature of development history, theoretical formulation, selective review of applications.

The main contributions of this research are introduced respectively in **Chapter 3-7**. To date, four causality analysis methods are initially proposed, which are the novel similarity measure by eigenvalues distribution, the novel mutual association measure by eigenvalue-based distance, the SSA causality test by forecasting accuracy and finally the SSA-CCM hybrid causality test. Two newly proposed methods in **Chapter 3** and **Chapter 4** are built on the fundamentals of similarity and association studies from formal cause philosophy respectively combining with the SVD technique. **Chapter 5** introduces an innovative modification of currently well accepted causality analysis approaches that are based on linear model. It adopts the SSA technique in a multivariate system and address to answer the question that whether one variable can be helpful to analyse or forecast the others. In terms of **Chapter 6**, the CCM causality test, which was firstly introduced by Sugihara et al. (2012), is adopted and addressed here due to its subspace-based feature and innovative information theory concept of reverse engineering to distinguish causality. Moreover, it acts as the fundamental introduction and comparison for the novel hybrid method in **Chapter 7**, where the SSA and CCM techniques are combined to contribute on the causality analysis advancement for complex systems.

Each chapter is self-contained to highlight the corresponding advancement and its satisfying performance in applications. In order to evaluate the performances of proposed methods, following critical evaluations and robust performances of proposed methods by simulation, implementations of different subjects are also summarised by representative data. Moreover, both comparison with corresponding empirical approaches and cross-chapter comparison are achieved to emphasise the significance of each advancement respectively. Finally, the conclusion and discussions of future research are presented in **Chapter 8**.

## 1.3 Contributions of Research

This research focuses on the advanced and relatively new subspace-based techniques while keeping pace with the developments of causality analysis techniques for time series. Considering the amount of subspace-based techniques adopted, the comprehensive theoretical improvements made regarding different format of causes, and the wide range of applications this thesis attempted to cover, here in this section, the contributions of this research are briefly introduced in detail as follows:

1. In the interest of combining the advanced subspace-based techniques so to offer both theoretical and practical advancements on causality analysis. To my knowledge, this study is the first attempt that extends subspace-based techniques into complex multivariate systems from the causality analysis perspective. It brings significant contributes on present literature of subspace-based techniques as well as causality analysis.

2. In terms of causality analysis, this research is not limited to causality detection only. Instead, it explores the causality analysis through different format of causes. It is the first time to my knowledge that a study extends subspace-based techniques to such a comprehensive system of causality analysis, involving similarity measure, association measure, and causality detection. It presents provisional insights into the better understanding of causality in a complex multivariate system, which is ubiquitous nowadays alongside the advancements of technology and developments of human society. Moreover, this initial study not only reflects one step further development to work with multivariate system, but also a significant advancement on causality analysis, not to mention the difficulties due to the complexities of the data, dynamical system and intricate cross relationships, etc.

3. This research further discovers the significant potentials of these subspace-based techniques on causality analysis due to their advantages of being assumption free, being broad applicable and being sensitive to nonlinear dynamics. The advantages of these subspace-based techniques are preserved while the shortages of empirical causality analysis methods are eliminated. More

specifically, **Chapter 3** introduces the novel similarity measure by eigenvalue distribution, this overcomes the difficulties of empirical similarity measure through identifying the comparable criterion that has no limitations of nonlinearity. The novel association measure by eigenvalue-based distance in **Chapter 4** has no restriction on detecting nonlinear or complex associations, also it maintains impressive ability on basic linear association detection. **Chapter 5** proposes a nonparametric SSA causality test that has no assumptions of any linear or specific nonlinear models prior or during the test, such that it is capable of distinguishing any possible relationships, even complex unknown patterns that previous tests cannot achieve. Moreover, in **Chapter 6**, another nonparametric causality test CCM, which also has no prior model assumptions, incorporates information theory and reverse engineering framework to provide a refreshing insight of causality analysis in dynamical systems. The final theoretical development in **Chapter 7** combines SSA and CCM to formulate a hybrid causality test that not only maintains all the advantages above, but also reveals more information of the data itself for assisting the discovery of causal relationship.

4. This research focuses on the aspect of causality analysis for time series only regardless of subjects and how complex the systems and data can be. For empirical researches, it is crucial to recognise significant relevant indicators and factors among all possibilities in a complex system like economics and social science. Hence, it is necessary to draw lessons from previous references as supportive evidences but it also aggravates its limitations in the same time. Nevertheless, the ability to discover new relevant indicators is equivalently significant especially in a complex dynamical system with limited reference or some emerging research areas that are newly discovered with immeasurable prospects. Therefore, this research contributes on offering the data driven, re-

duced form methods for conducting causality analysis with no limitations of subjects in a robust and efficient manner. Moreover, it aims to allow the data speaks for itself and simplify the causality analysis process starting from the essential two factors required.

5. The developments of all novel methods in this research satisfy the requirements of being suitable for broader implementations. Any possible causal links that this research succeeded to firstly identify represent important achievements and practical significance on contributing to causality analysis study up to date and possibly assisting and guiding policy making forward accordingly. Additionally, in terms of the future research, more valuable contributions can be expected through successful implementations on a wider range of subjects, especially for applications of proposed methods on areas that causality analysis is an emerging need but the system itself is relatively new and too complex to be model by traditional methods. This research can offer considerable opportunity of 'impact' as the success of this research will contribute to the wider community of economics and social science, more importantly, it will attract more interests from various areas.

# Chapter 2

# Subspace-based Techniques

This chapter introduces the subspace-based techniques that are adopted in this research. Three key techniques considered include SVD, SSA and CCM. These methods either factorize the high dimensional data into true dimensionality determinate and subspaces or share the identical component of origins in delay embedding theorem by Takens (1981), which is called the Takens' Theorem. Takens (1981) presented that an attractor to describe the dynamics of a system can be reconstructed by a time series through the delay embedding. For instance, consider a time series $Y_t$ of length $n$ at discrete time $t$, so the reconstructed multidimensional state with a time delay of $\tau$ is $S_t = (Y_t, Y_{t-\tau}, Y_{t-2\tau}, ..., Y_{t-(E-1)\tau})$, where $E$ is the embedding dimension accordingly. Apart form the shared elements of subspace concept and delay embedding origin, there are also varying degrees of emphases among these techniques that make one method different from another. Thus, this chapter focuses on the theoretical introduction of these techniques so to provide a critical overall view of these subspace-based techniques themselves together with selective literature review of applications respectively. Moreover, this chapter provides the fundamental knowledge for the advancements forward and novel methods that will be explicated in detail in the following chapters.

## 2.1 Singular Value Decomposition (SVD)

### 2.1.1 Definitions and Algorithms

SVD is a well known common technique for matrix decomposition in linear algebra that was firstly introduced by Beltrami and Jordan in the 1870's. It was presented as the generalised principal axis transformation for Hermitian matrices[1] and firstly proved for rectangular and complex matrices by Eckart and Young (1936). From then on, SVD has been extensively studied and applied due to its interesting and attractive algebraic properties along with the significant geometrical and theoretical insights it brings regarding linear transformations (Kalman, 1996). Note that more details of SVD can be found in (Kalman, 1996; Klema and Laub, 1980; Van Loan, 1976). A brief early history of SVD was summarized by Stewart (1993), which is not reproduced here in this research since it is not the key focus. However, it is of note that SVD is a key step of SSA, which will be comprehensively introduced in the next section. The calculations related to SVD in this research are obtained by R and the following algorithms are mainly based on the work of Stewart (1993).

For a random given matrix $\mathbf{X}$ that has the dimension of $K \times L$, where $K \geqslant L$, the SVD of the matrix can then be written as:

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T, \tag{2.1}$$

[1] Hermitain matrix is a complex square matrix that is identical to its own conjugate transpose that $\mathbf{X} = \overline{\mathbf{X}^T}$ or $x_{ij} = \overline{x_{ji}}$, where $\mathbf{X}$ indicates a random Hermitian matrix while $x_{ij}$ is the element in the $i$-th row and $j$-th column of matrix $\mathbf{X}$.

or

$$\mathbf{X} = \begin{bmatrix} u_1 & u_2 & \cdots & u_K \end{bmatrix} \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \sigma_L \end{bmatrix} \begin{bmatrix} v_1^t \\ v_2^t \\ \vdots \\ v_L^t \end{bmatrix}, \tag{2.2}$$

where the matrix $\mathbf{X}$ is factorized into the product that is made of an $K \times L$ matrix $\mathbf{U}$, an diagonal matrix $\Sigma$ with dimension of $L \times L$, and another $L \times L$ matrix $\mathbf{V}^T$. More specifically, the *singular values* are the non-negative elements on the diagonal of matrix $\Sigma$, which are $\sigma_1, \sigma_2, ..., \sigma_L$ and arranged in descending order of magnitude; for $\mathbf{U}$ and $\mathbf{V}^T$, the columns of $\mathbf{U}$ are the *left singular vectors*, while the rows of $\mathbf{V}^T$ indicates the *right singular vectors*. Note that according to linear algebra, the *singular values* of a $K \times L$ matrix $\mathbf{X}$ are the square roots of the *eigenvalues* of the $L \times L$ matrix $\mathbf{XX}^T$, which will be again addressed in the following section of SSA that containing SVD as a key step. Moreover, rarely in the case of $\mathbf{X}$ is a real symmetric square matrix positive definite (where $K = L$), the *singular values* is equivalent to the *eigenvalues*, as well as the left and right *eigenvectors* / *singular vectors*.

### 2.1.2 Selective Review of Illustrative Applications

As a well known matrix decomposition technique, SVD is extensively adopted to compute the representations for various subspaces and maps that arise in linear algebra (Klema and Laub, 1980). As it is impossible to cover the full implications and review the whole history of developments of SVD, nevertheless, a few landmarks are summarized below to give a brief flavor of its significance.

One of the most significant application of SVD is the least squares solution, for which, the full potential of SVD is realized in the analysis of non-square, possibly rank-deficient matrices (Klema and Laub, 1980) (more details and algorithms can be found in (Arun et al., 1987; Golub and Reinsch, 1970; Lawson and Hanson, 1974; Van Loan, 1976), thus it is not reproduced here in this research).

SVD is also a key element of Principal Component Analysis (PCA) and SSA (more details see (Hassani, 2007; Shlens, 2014; Wall et al., 2003) due to the fact that it can reveal comprehensive information of a matrix to give better understanding of the matrix itself). According to Shlens (2014), SVD and PCA are intimately related while SVD is a more general method of understanding change of basis. The direct relation between PCA and SVD is in the case where principal components are computed from the covariance matrix while SVD performs on the centered data matrix (Shlens, 2014). SVD serves as an important step of SSA, which will be explicated in the following section regarding SSA.

Moreover, SVD uncovers the numerical determination of the true rank of a matrix, which has significant role in data dimensionality reduction (Wall et al., 2003). The singular values that revealed by SVD can help to identify the most important information that a matrix contains, thus these numerical determinations can then be selected and processed so to reduce the size of the original matrix while still capture most of the important features. Due to the fact that SVD can provide the reliable computation for relatively accurate approximation along with much more efficient calculation, it has been widely employed in the studies of dimensionality reduction like image denoising (Demirel et al., 2010; Rajwade et al., 2013; Zhang et al., 2015), clustering (Drineas et al., 2004; Savas and Eldén, 2007), complex data preprocessing (Jha and Yadava, 2011; Konstantinides et al., 1997; Wallace et al., 1992), digital image watermarking (Bao and Ma, 2005; Chandra, 2002; Lai and Tsai, 2010), etc.

SVD also has advantages in its sensitivity and capability regarding weak signals retrieval (Wall et al., 2003), thus it has been extensively adopted in various signal processing studies (see (Kanjilal et al., 1997; Le Bihan and Mars, 2004; Van Der Veen et al., 1993)).

## 2.2 Univariate/Multivariate Singular Spectrum Analysis (SSA/MSSA)

### 2.2.1 Literature Review

SSA is a relatively new technique known for both time series analysis and forecasting. It has been widely applied in a range of different fields and a multitude of fairly precise results has proven it to be a powerful and applicable technique. A concise short description of SSA technique is stated in Hassani (2007) as a nonparametric technique of time series analysis incorporating the elements of classical time series analysis, multivariate statistics, multivariate geometry, dynamical systems, and signal processing. In brief, SSA firstly decomposes a time series into the sum of a small number of independent and interpretable components such as a slowly varying trend, periodic or quasi-periodic components and noise, which is followed by a reconstruction of the original series (Hassani et al., 2009). Note that more coherent and detailed theoretical explanation of SSA technique can be found in (Golyandina et al., 2001; Sanei and Hassani, 2015).

The relative history of SSA can be tracked back to the papers of Broomhead and King (1986), in which the authors reviewed the embedding theorem of Takens (1981) and introduced the singular system analysis and its implementation of nonlinear phenomena in time series analysis. Fraedrich (1986) independently applied the technique and expanded the research of nonlinear dynamics with weather and climate attractors. Since then, the initial idea of SSA has been independently developed in Russia, UK and US by several groups of researchers respectively. In that period, more literature that exploited the methodological aspects and applications of SSA are further conducted (some representative literature see (Danilov and Zhigljavsky, 1997; Elsner and Tsonis, 1996; Vautard and Ghil, 1989; Vautard et al., 1992b)). Furthermore, Golyandina et al. (2001) presented a comprehensive description of the theoretical and practical foundations of SSA along with several examples of applications. It is of note that the fundamental knowledge of this research is significantly built upon the studies of Golyandina et al. (2001) and their relevant post literature to date: methodology and comparison (Hassani, 2007), forecasting with missing value (Golyandina and Osipov, 2007), minimum variance estimator (Hassani, 2010), multivariate extension and causality detection (Hassani and Mahmoudvand, 2013; Hassani et al., 2010c), separability and window length (Hassani et al., 2011), through description (Golyandina and Zhigljavsky, 2013; Sanei and Hassani, 2015), determining the number of eigenvalues (Alharbi and Hassani, 2016).

The SSA technique has also been extensively applied to a broad range of subjects that include but not limited to: industrial production forecasting in EU (Hassani et al., 2009) and UK (Hassani et al., 2013b), analysing Iranian national accounts data (Hassani and Zhigljavsky, 2009), economics and financial market analysis (Hassani et al., 2010a, 2013a; Hassani and Thomakos, 2010; Menezes et al., 2012), daily exchange rate prediction (Hassani et al., 2010b), electrocardiogram (ECG) signal extraction (Ghodsi et al., 2010), inflation forecasting (Hassani et al., 2013c), electricity price (Miranian et al., 2013), gene expression (Ghodsi et al., 2015c; Hassani and Ghodsi, 2014), gold price prediction (Hassani et al., 2015b), US trade forecasting (Silva and Hassani, 2015), US and EU tourist arrivals (Hassani et al., 2017b, 2015c).

In general, SSA has been extensively exploited and proved a powerful technique for time series analysis. The mainstream capabilities of SSA focus on the areas of signal extraction from different resolution, time series smoothing and filtering, specific component extraction from original series, forecasting, change point detection. Nevertheless, SSA is still a developing technique with immeasurable potentials. This research aims to further fulfil SSA regarding the causality analysis aspect so to contribute to the literature of both theoretical developments and practical applications. In the following sections, the theoretical descriptions are summarized with details for both univariate and multivariate SSA.

### 2.2.2 Univariate SSA

The SSA technique is performed in two stages, which are known as decomposition and reconstruction. Embedding and SVD are included in the first stage of decomposition, while the second stage of reconstruction contains grouping and diagonal averaging(see the SSA framework in Fig. 2.1). Note that the brief introduction of univariate SSA below mainly follows Hassani (2007) (more information see (Elsner and Tsonis, 1996; Golyandina et al., 2001; Golyandina and Zhigljavsky, 2013)).

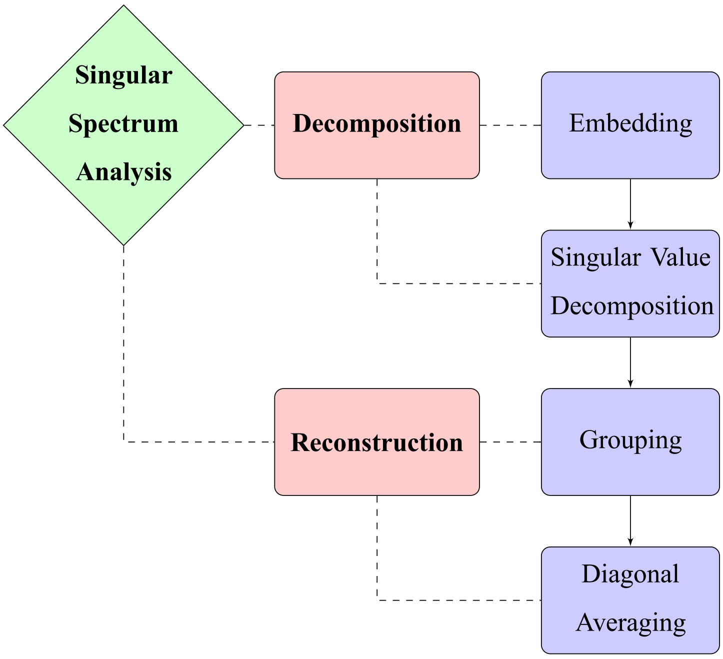


Fig. 2.1 Framework of Singular Spectrum Analysis Technique.

#### Stage One – Decomposition

The main purpose of SSA is to decompose the original series into a sum of series, so that each component in this sum can be identified as either a trend, periodic or quasi-periodic component, or noise (Hassani, 2007). The decomposition stage includes two steps: embedding and SVD[2].

***1st Step: Embedding***

Consider the real-valued nonzero time series $Y_N = (y_1, ..., y_N)$ of sufficient length $N$. Let $K = N - L + 1$, the embedding step transfers this one-dimensional time series $Y_N$ into the multi-dimensional matrix $\mathbf{X} = [X_1, ..., X_K]$ with the lagged vectors $X_i = (y_i, ..., y_{i+L-1})$ that is based on the Takens' theorem. The single setting of the embedding is an integer $L$, which meet the condition of $2 \leq L \leq N/2$ and is named the window length[3]. Thereafter, a trajectory matrix below is obtained while the window length is sufficiently large:

$$\mathbf{X} = [X_1, ..., X_K] = (x_{ij})_{i,j=1}^{L,K} = \begin{pmatrix} y_1 & y_2 & y_3 & \cdots & y_K \\ y_2 & y_3 & y_4 & \cdots & y_{K+1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ y_L & y_{L+1} & y_{L+2} & \cdots & y_N \end{pmatrix}.$$

Note that the trajectory matrix $\mathbf{X}$ is a Hankel matrix, where all the elements along the diagonal $i + j = constant$ are equivalent. It is worthy to be addressed again that the embedding step adopts the delay embedding theorem by Takens (1981), which is the fundamental knowledge of time series embedding that also have been incorporated for many advance time series analysis techniques.

***2ed Step: SVD***

The second step of decomposition stage is performing the SVD of the matrix $\mathbf{X}$ that is achieved in the last embedding step. In brief, the trajectory matrix is factorized into a sum of rank-one bi-orthogonal elementary matrices. The numerical

[2] Note that the detailed introduction of SVD can be found in section 2.1 of Chapter 2.

[3] According to Hassani (2007), the window length $L$ should generally be proportional to the periodicity of the series, also it has to be large enough to be able to retrieve sufficient enough information of the data while satisfying the condition $2 \leq L \leq N/2$ due to the natural feature of a trajectory matrix

determinants are extracted through SVD, and comprehensively represent the whole information of the matrix can possibly contain. More details can be found in section 2.1 of Chapter 2.

Denote $\lambda_1,...,\lambda_L$ as the eigenvalues of $\mathbf{S} = \mathbf{X}\mathbf{X}^T$ in decreasing order of magnitude ($\lambda_1 \geq ...\lambda_L \geq 0$) and $U_1,...,U_L$ the corresponding orthogonal system of the corresponding eigenvectors $V_i$. Thus, if $V_i = \mathbf{X}^T U_i/\sqrt{\lambda_i}$, then the SVD of the trajectory matrix can be written as

$$\mathbf{X} = \mathbf{X}_1 + \cdots + \mathbf{X}_d, \tag{2.3}$$

$$\mathbf{X}_i = \sqrt{\lambda_i} U_i V_i^T (i = 1,...,d), \tag{2.4}$$

where $d = max$(i, such that $\lambda_i > 0) = rank\mathbf{X}$. Note that $\sqrt{\lambda_i}$ are the corresponding *singular value*, $(\sqrt{\lambda_i}, U_i, V_i)$ is the *i*-th *eigentriple* and the set of $\{\sqrt{\lambda_i}\}$ is named the *spectrum*. Note that this terminology also indicates the reason of the name SSA, which is decomposing the original time series so to retrieve and analyse the *spectrum* of *singular values* for the aims of denoising, preprocessing, signal extracting, forecasting, etc.

The matrices $\mathbf{X}_i$ have rank 1 and the SVD of the trajectory matrix is optimal in the sense that among all the matrices of rank $q < d$, the matrix $\sum_{i=1}^{q} \mathbf{X}_i$ provides the best approximation to the matrix $\mathbf{X}$, so that $\parallel \mathbf{X} - \mathbf{X}^{(q)} \parallel$ is a minimum.

### Stage Two – Reconstruction

As the series has been decomposed in the first stage, a new selectively reconstructed series is obtained accordingly in the second stage. Note that according to the different requirements of studies aiming at diverse objectives, this reconstruction can be noise free, containing specific level of noise, or even just a component or combination of components of particular feature(s) of the original series.

***1st Step: Grouping***

In the grouping step, the elementary matrices is divided into a range of groups and those divided groups are summarised to be matrices within each group. The grouping process can also be interpreted as the grouping of the eigentriples due to the fact that the eigentriple is the only determinant for each matrix component.

Let $I = i_1,...,i_p$ be a group of indices $i_1,...,i_p$. Then the matrix $\mathbf{X}_I$ corresponds to the group $I$ is defined as

$$\mathbf{X}_I = \mathbf{X}_{i_1} + \cdots + \mathbf{X}_{i_p}. \tag{2.5}$$

Thus, by splitting the set of indices $1,...,d$ into disjoint subsets $I_1$,...,$I_g$, grouping is then named of the process of choosing the sets $I_1,...,I_g$, which corresponds to:

$$\mathbf{X} = \mathbf{X}_{I_1} + \cdots + \mathbf{X}_{I_g}, \tag{2.6}$$

in which the contribution of the component $\mathbf{X}_I$ in a given group is measured by the share of the corresponding eigenvalues: $\sum_{i\in I}\lambda_i / \sum_{i=1}^{d}\lambda_i$.

***2ed Step: Diagonal Averaging***

For the last step, it is aimed to transform the grouped product from the step above back to a time series. In order to do so, the regrouped matrix has to be transformed to a Hankel matrix first due to the common matrix algebra fact that a Hankel matrix can be subsequently converted to a time series.

To introduce the diagonal averaging process in brief, consider $z_{ab}$ stands for the $a$-th row, $b$-th column element of a matrix $\mathbf{Z}$, then the $c$-th term of the resulting time series is obtained by averaging $z_{ab}$ over all $a,b$ such that $a+b=c+1$. Thus, by performing the diagonal averaging of all matrix components in the expansion of **X** above, another expansion is obtained below:

$$\mathbf{X} = \widetilde{\mathbf{X}}_{I_1} + \widetilde{\mathbf{X}}_{I_j} + \cdots + \widetilde{\mathbf{X}}_{I_g}, \tag{2.7}$$

where $\widetilde{\mathbf{X}}_{I_j}$ is the diagonalized version of the matrix $\mathbf{X}_{I_j}$. From another point of view, this is actually equivalent to the decomposition of the initial series $Y_N = (y_1, ..., y_N)$ into a sum of $g$ series

$$y_n = \sum_{j=1}^{g} \widetilde{y}_n^{(j)}, \tag{2.8}$$

where $\widetilde{Y}_N^{(j)} = (\widetilde{y}_1^{(j)}, ..., \widetilde{y}_N^{(j)})$ corresponds to the matrix $\widetilde{\mathbf{X}}_{I_j}$. Note that the only setting for the second stage of SSA is the number of eigenvalues $r$[4], so that in what follows in general, two groups of indices $I_1 = \{1, ..., r\}$ and $I_2 = \{r+1, ..., L\}$ are used and associated the group $I = I_1$ with the signal component and the group $I_2$ with noise.

### 2.2.3 MSSA

In brief, MSSA is the multivariate extension of univariate SSA. There is still two stages containing four steps in total for the whole MSSA process while the key difference is that the form of how the trajectory matrix for each time series get organized to be a block trajectory matrix. Therefore, two types of MSSA is presented based on the form of combining the matrices either horizontally or vertically. The following brief theoretical introduction of MSSA mainly follows Sanei and Hassani (2015) (more information can be found in (Hassani et al., 2013b; Hassani and Mahmoudvand, 2013; Patterson et al., 2011)).

[4] Note that for each specific window length $L$, there are $L-1$ choices of numbers of eigenvalues $r$. The general approach of seeking the optimal setting is based on the evaluations of reconstruction or forecasting performances where all possible $L$ and $r$ combinations are evaluated.

### Vertical Form

#### Stage One – Decomposition

Consider $M$ time series with different series length $N_i$: $Y_{N_i}^{(i)} = (y_1^{(i)}, ..., y_{N_i}^{(i)})(i = 1, ..., M)$. In this case, the standard univariate form can be acquired by setting *M = 1*.

***1st Step: Embedding***

The one-dimensional time series $Y_{N_i}^{(i)}$ is firstly transformed into a multidimensional matrix $[X_1^{(i)}, ..., X_{K_i}^{(i)}]$ with vectors $X_j^{(i)}$ that equals to $(y_j^{(i)}, ..., y_{j+L_i-1}^{(i)})^T \in \mathbf{R}^{L_i}$, where $L_i(2 \leq L_i \leq N_i/2)$ is the window length for each series with length $N_i$ and $K_i = N_i - L_i + 1$, respectively. The trajectory matrix that is produced after this step is

$$\mathbf{X}^{(i)} = [X_1^{(i)}, ..., X_{K_i}^{(i)}] = (x_{ab})_{a,b=1}^{L_i,K_i}. \tag{2.9}$$

The above procedure is implied for each series separately, which provides $M$ different $L_i \times K_i$ trajectory matrices $\mathbf{X}^{(i)}(i = 1, ..., M)$. Note that in order to form a new block Hankel matrix in a vertical form, it is required to have $K_1 = \cdots = K_M = K$. Accordingly, this version enables various window length $L_i$ and different series length $N_i$, but equivalent $K_i$ for all series. Therefore, the result of this step is the following block Hankel trajectory matrix:

$$\mathbf{X}_V = \begin{bmatrix} \mathbf{X}^{(1)} \\ \vdots \\ \mathbf{X}^{(M)} \end{bmatrix}.$$

Note that $\mathbf{X}_V$ indicates that the output of the first step is a block Hankel trajectory matrix formed vertically.

***2st Step: SVD***

The SVD of matrix $\mathbf{X}_V$ is then performed in the second step. Denote $\lambda_{V_1}, ..., \lambda_{V_{L_{sum}}}$

as the eigenvalues of $\mathbf{X}_V\mathbf{X}_V^T$, arranged in decreasing order $(\lambda_{V_1} \geq \cdots \geq \lambda_{V_{L_{sum}}} \geq 0)$, where $L_{sum} = \sum_{i=1}^{M} L_i$. The structure of the matrix $\mathbf{X}_V\mathbf{X}_V^T$ is as follows:

$$\mathbf{X}_V\mathbf{X}_V^T = \begin{bmatrix} \mathbf{X}^{(1)}\mathbf{X}^{(1)T} & \mathbf{X}^{(1)}\mathbf{X}^{(2)T} & \cdots & \mathbf{X}^{(1)}\mathbf{X}^{(M)T} \\ \mathbf{X}^{(2)}\mathbf{X}^{(1)T} & \mathbf{X}^{(2)}\mathbf{X}^{(2)T} & \cdots & \mathbf{X}^{(2)}\mathbf{X}^{(M)T} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{X}^{(M)}\mathbf{X}^{(1)T} & \mathbf{X}^{(M)}\mathbf{X}^{(2)T} & \cdots & \mathbf{X}^{(M)}\mathbf{X}^{(M)T} \end{bmatrix}.$$

The structure of the matrix $\mathbf{X}_V\mathbf{X}_V^T$ is similar to the variance-covariance matrix in the classical multivariate statistical analysis literature. The matrix $\mathbf{X}^{(i)}\mathbf{X}^{(i)T}$ for the series $Y_{N_i}^{(i)}$, appears along the main diagonal and the products of two Hankel matrices $\mathbf{X}^{(i)}\mathbf{X}^{(j)T}(i \neq j)$, which are related to the series $Y_{N_i}^{(i)}$ and $Y_{N_j}^{(j)}$, appears in the off-diagonal.

Moreover, let the $U_{V_1},...,U_{V_{L_{sum}}}$ be the orthogonal system of the corresponding eigenvectors $W_{V_1},...,W_{V_{L_{sum}}}$. The SVD of $\mathbf{X}_V$ can be written as

$$\mathbf{X}_V = \mathbf{X}_{V_1} + \cdots + \mathbf{X}_{V_{L_{sum}}}, \tag{2.10}$$

where $\mathbf{X}_{V_i} = \sqrt{\lambda_{V_i}}U_{V_i}W_{V_i}^T$ and $W_{V_i} = \mathbf{X}_V^T U_{V_i}/\sqrt{\lambda_{V_i}}$ $(\mathbf{X}_{V_i} = 0$ if $\lambda_{V_i} = 0)$.

**Stage Two – Reconstruction**

***1st Step: Grouping***

Relatively equivalent to the univariate SSA case, the grouping step splits the matrices $\mathbf{X}_{V_1},\cdots,\mathbf{X}_{V_{L_{sum}}}$ into several disjoint groups and summing the matrices within each group. For instance, the spilt of the set of indices $\{1,\cdots,L_{sum}\}$ into disjoint subsets $I_1,\cdots,I_g$ corresponds to the representation

$$\mathbf{X}_v = \mathbf{X}_{I_1} + \cdots + \mathbf{X}_{I_g}. \tag{2.11}$$

Thus, for an ideal simple case where only signal and noise two components exist, two groups of indices are separated by numbers of eigenvalues index $r$, which are $I_1 = \{1, \cdots, r\}$ and $I_2 = \{r+1, \cdots, L_{sum}\}$ and associate the group $I = I_1$ with signal component and the group $I_2$ with noise.

***2st Step: Diagonal Averaging***

Here in this step, the reconstructed matrix $\widehat{\mathbf{X}}_{V_i}$ is transformed to the form of a Hankel matrix, which can be subsequently converted to a time series. Let $\widetilde{\mathbf{X}}^{(i)}$ be the approximation of $\mathbf{X}^{(i)}$ obtained from the diagonal averaging step. If $\widetilde{x}_{ab}^{(i)}$ stands for the $a$-th row, $b$-th column element of a matrix $\widetilde{\mathbf{X}}^{(i)}$, then the $j$-th term of the reconstructed series $\widetilde{Y}_{N_i}^{(i)} = (\widetilde{y}_1^{(i)}, \cdots, \widetilde{y}_j^{(i)}, \cdots, \widetilde{y}_{N_i}^{(i)})$ is achieved by arithmetic averaging $\widetilde{x}_{ab}^{(i)}$ over all $(a,b)$ such that $a+b-1=j$.

**Horizontal Form**

The key difference of the horizontal MSSA is in the first step, where the structure of the block Hankel matrix is horizontally organized. The rest of the horizontal form algorithms are very similar to those provided above for the vertical form, therefore, only the differences are addressed below without reproducing the identical processes.

Equivalently, donate $M$ different $L_i \times K_i$ trajectory matrices $\mathbf{X}^{(i)}(i = 1, ..., M)$ as the results after the embedding step for $M$ time series with different series length $N_i$: $Y_{N_i}^{(i)} = (y_1^{(i)}, ..., y_{N_i}^{(i)})(i = 1, ..., M)$. To construct a block Hankel matrix in the horizontal form, it is necessary to have $L_1 = L_2 = ... = L_M = L$. Accordingly, there are different values of $K_i$ and series length $N_i$, but similar $L_i$. The result of this step is

$$\mathbf{X}_H = [\mathbf{X}^{(1)} : \mathbf{X}^{(2)} : \cdots : \mathbf{X}^{(M)}]. \tag{2.12}$$

Hence, the structure of the matrix $\mathbf{X}_H\mathbf{X}_H^T$ is as follows:

$$\mathbf{X}_H\mathbf{X}_H^T = \mathbf{X}^{(1)}\mathbf{X}^{(1)^T} + \cdots + \mathbf{X}^{(M)}\mathbf{X}^{(M)^T}. \tag{2.13}$$

Regarding the structure of the matrix $\mathbf{X}_H\mathbf{X}_H^T$, there are not any cross-product between Hankel matrices $X^{(i)}$ and $X^{(j)}$. Thus, in this format, the sum of $\mathbf{X}^{(i)}\mathbf{X}^{(i)^T}$ provides the new block Hankel matrix. Nevertheless, performing the SVD of $\mathbf{X}_H$ yields $L$ eigenvalues, whilst $L_{sum} = \sum_{i=1}^{M} L_i$ eigenvalues are obtained in vertical MSSA.

### 2.2.4 Forecasting with SSA/MSSA

SSA forecasting techniques can be applied on time series that approximately satisfy linear recurrent formula (LRF) (Golyandina et al., 2001). The class of time series governed by LRF is rather wide; it includes harmonics, polynomials and exponential time series and is closed under term-by-term addition and multiplication (Hassani, 2007). There are two types of SSA/MSSA forecasting including the recurrent and vector approaches (Hassani and Mahmoudvand, 2013), while the MSSA also have both horizontal and vertical forms. Thus, it leads to two different univariate SSA forecasting algorithms and accordingly four different MSSA forecasting algorithms as listed below:

$$\begin{cases} SSA \begin{cases} RecurrectSSAForecasting \\ VectorSSAForecasting \end{cases} \\ MSSA \begin{cases} VMSSA \begin{cases} RecurrentVMSSAForecasting \\ VectorVMSSAForecasting \end{cases} \\ HMSSA \begin{cases} RecurrentHMSSAForecasting \\ VectorHMSSAForecasting \end{cases} \end{cases} \end{cases}$$

Consider a time series $Y_T = Y_T^{(1)} + Y_T^{(2)}$ that is ideally formed by signal $Y_T^{(1)}$ and noise $Y_T^{(2)}$, the forecasting aims to predict the signal in the presence of a noise. There are two main assumptions: the signal series admits a recurrent continuation with the assistant of the LRF of a relatively small dimension $d$; and the existence of a specific window length $L$ so that the signal and noise can be properly separated. The SSA recurrent forecasting algorithm is briefly introduced here by following (Hassani, 2007). Donate the time series $Y_T$ satisfies an LRF of order $d$ if there are numbers $\alpha_1, ..., \alpha_d$ so that

$$y_{i+d} = \sum_{k=1}^{d} \alpha_k y_{i+d-k}, \tag{2.14}$$

in which $1 \leq i \leq T - d$. The coefficient $\alpha_1, ..., \alpha_d$ can be achieved based on eigenvectors obtained from the SVD step so to further proceed to the forecasting.

Note that since the forecasting capability of SSA/MSSA is not the main focus of this research except being briefly adopted for part of the implementation of the novel method in Chapter 5, the full version of algorithms and proofs that have been clearly presented in literature are therefore not reproduced here in the main text. However, the algorithms of forecasting with SSA/MSSA are still briefly summarized in the Appendix A as in Golyandina et al. (2001), Hassani and Mahmoudvand (2013) and Sanei and Hassani (2015) for reference.

## 2.3 Convergent Cross Mapping (CCM)

### 2.3.1 Literature Review

CCM was firstly introduced by Sugihara et al. (2012) that aimed at identifying the causality among time series in complex systems. It was based on nonlinear state space reconstruction (more details see (Casdagli et al., 1991)) and was seeking to provide a better understanding of dynamical systems that are not covered by other well established methods such as Granger causality (GC), thus, it contributes as an alternative approach rather than a competition to extend causality analysis especially on complex dynamic systems that is non-separable weakly connected (Sugihara et al., 2012).

As the primary advancement of the causality analysis, GC by Granger (1969) distinguishes causal effect based on that the predictability of the effect variable is weakened due to the exclusion of the expected cause variable of the defined model. It also incorporated the lagged coordinator so to reflect the causal effect from the past on present. However, it is highly model restricted and assumes that the information of the cause variable is independent and unique with no more related information existing in the model/system, which according to Sugihara et al. (2012), it ignores the fact that the cause factor will be redundantly present in the effect variable itself according to the dynamical system theory in (Deyle and Sugihara, 2011; Dixon et al., 1999; Takens, 1981).

As a relatively new technique, CCM has proven to be an advanced non-parametric technique for distinguishing causality in a dynamical system by implementations in a number of subjects. It was applied on the ecological data by Sugihara et al. (2012) where CCM was also initially introduced, in which, the results confirms that the sea surface temperature influences sardine and anchovy population size along with more supporting evidences by other fishery-independent data.

Regarding the climate change studies, Fan et al. (2014) adopted CCM and identified a unidirectional causality from vegetation green-up to spring dust storms in Inner Mongolia, Northern China. Ye et al. (2015) applied CCM to CO2 and temperature from the Vostok ice core, in which, the bidirectional causality was identified so to confirm the positive feedback between temperature and greenhouse gases. Additionally, the long term time series of chlorophyll-a and sea surface temperature are also exploited by Ye et al. (2015) and a unidirectional causality from the sea surface temperature on chlorophyll-a is detected. Moreover, the causality between galactic cosmic rays and global temperature was exploited by Tsonis et al. (2015) using CCM, no measurable causality was found between cosmic rays and the overall global warming trend while significant causal effect of cosmic ray on short term year to year variation in global temperature was identified for better understanding of the factors of climate changes.

Clark et al. (2015) combined the existing CCM technique and incorporated the dewdrop regression to establish the multispatial CCM test that aimed to detect causality from short time series. This significantly extended the capability of causality analysis of nonlinear dynamical systems with very small number of observations or complex systems where experiments are difficult to perform. In total, ten different real world experiments are conducted to evident the performance of this novel method, which cover a diverse range of subjects including soil nitrate and invading plant species, nitrogen addition and plant species composition, competition plots on a soil gradient, plant biomass in old fields, etc. Similarly, the long term causal link between increasing dryness and the grassland dynamics is detected by using CCM in (Brookshire and Weaver, 2015).

CCM has also been applied to biomedical study, McBride et al. (2015) firstly applied CCM for capturing characteristic changes in Electroencephalography (EEG) activity due to cognitive deficits and demonstrate the capability of detecting early stage Alheimer's disease. Moreover, CCM is applied by Margolis et al. (2016) as an alternative approach to detect causality and clarify the directionality of interactions between species in human micro biome studies. A few more implementations have covered the causality analysis studies in social media (Luo et al., 2014), marketing (Dost, 2015), PM2.5 pollution and meteorological factors (Chen et al., 2017), etc.

### 2.3.2 Theoretical Formulation

According to Sugihara and May (1990) and Sugihara et al. (1990), the time series from the same dynamical system are causally linked, which indicates that each variable can identify the state of the others. Moreover, the information of the cause factor will be contained by the effect factor so to be reconstructed by the effect factor while the effect factor cannot be recovered by the cause factor (Sugihara et al., 2012). Note that in this section, CCM is briefly introduced by following primarily Sugihara et al. (2012).

Assume there are two variables $X_t$ and $Y_t$ such that $X_t$ has a causal effect on $Y_t$, where $t = 1,2,...,N$ and $N$ is the total number of observations of two variables. CCM will test the causality by evaluating whether the historical record of $Y_t$ can be used to get reliable reconstructions of $X_t$. Given a library set of $n$ points that are not necessarily equal to the total number of observations $N$, $t = 1,2,...,n$, the lagged coordinates (lag=$\tau$)[5] are adopted to generate an $E$-dimensional embedding state space based on Takens (1981) and Sugihara and May (1990), in which the points are the library vector $X_t$ and prediction vector $Y_t$

$$X_t : \{x_t, x_{t-\tau}, x_{t-2\tau}, \cdots, x_{t-(E-1)\tau}\}, \tag{2.15}$$

$$Y_t : \{y_t, y_{t-\tau}, y_{t-2\tau}, \cdots, y_{t-(E-1)\tau}\}, \tag{2.16}$$

The $E+1$ neighbors of $Y_t$ from the library set $X_t$ will be selected, which actually form the smallest simplex that contains $Y_t$ as an interior point[6]. Accordingly, the forecast is then conducted by the nearest-neighbour forecasting algorithm of simplex projection as listed below by following Sugihara and May (1990) and Sugihara (1994).

[5] More details that explicitly discuss the time lags for CCM can be found in (Ye et al., 2015).

[6] Note that the optimal $E$ will be evaluated and selected based on the forward performances of these nearby points in an embedding state space.

Assume an observed time series $X_t \in R^{m+1}$ and donate the time series value $T_p$ time steps forward be $X_{t+T_p}(1) = Y_t$, so the forecast at $T_p$ is

$$\hat{Y}_t = \sum_{j=1}^{m} C_t(j) X_t(j). \tag{2.17}$$

The SVD solution for $C$ is obtained by historical points from the fitting set or library set $i$ by $B = AC$, where

$$B_i = \omega(\|X_i - X_t\|) Y_i, \tag{2.18}$$

$$A_{ij} = \omega(\|X_i - X_t\|) X_i(j), \tag{2.19}$$

$$\omega(d) = e^{\theta d / \bar{d}}, \tag{2.20}$$

where $\theta \geq 0$, $d$ is the distance between the predictee and the neighbour vector, the scale factor $\bar{d}$ is the average distance between neighbours.

Therefore, by adopting the essential concept of empirical dynamic modeling and generalized Takens' Theorem (Takens, 1981), two manifolds are conducted based on the lagged coordinates of the two variables under evaluation, which are the attractor manifold $M_Y$ constructed by $Y_t$ and respectively, the manifold $M_X$ by $X_t$. The causation will then be identified accordingly if the nearby points on $M_Y$ can be employed for reconstructing observed $X_t$. Note that the correlation coefficient $\rho$ is used for the estimates of cross map skill due to its widely acceptance and understanding, additionally, leave-one-out cross-validation is considered a more conservative method and adopted for all evaluations in CCM.

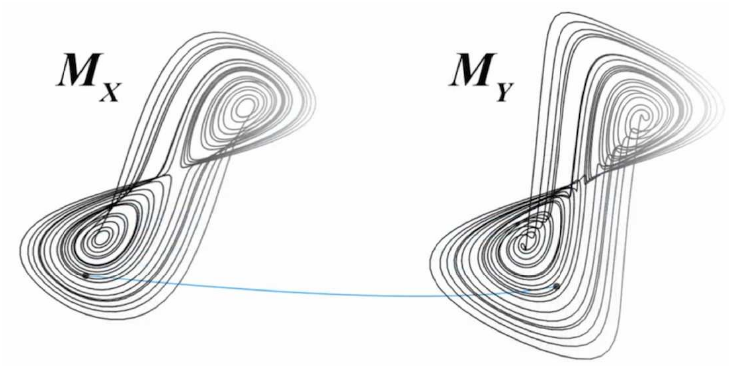


Fig. 2.2 Manifolds of Convergent Cross Mapping Test.

# Chapter 3

# Novel Similarity Measure by Eigenvalue Distribution

Following the “Formal Cause" that was clearly stated in Chapter 1, section 1.2.1, the Formal Cause is “the account of what-it-is-to-be", or “what makes a thing one thing rather than many things" (Falcon, 2015). This chapter specifically exploits the similarity measure aspect of the causality analysis by employing the subspace-based technique. In brief, a novel similarity measure is proposed here that is fundamentally built upon the criterion of eigenvalue distribution, and the relevant subspace based technique adopted here is primarily SVD. Specifically, this chapter is organized as follows: Section 1 contains a brief introduction and reviews of the landmark literature of similarity measure study; Section 2 presents the theoretical formulation of the proposed novel similarity measure; The review of some empirical tests employed as part of the similarity measure are listed in Section 3; Section 4 evaluates the proposed novel method by simulations, as well as the real case scenarios in Section 5; Finally, the discussion and conclusion are summarised in Section 6.

## 3.1 Introduction

The studies of similarity have been overwhelmingly explored and applied in various disciplines on many different formats, for example, numerical values (Hung and Yang, 2004; Mitchell, 2003), images (Roche et al., 1998; Yang et al., 2005), genes (Balasubramaniyan et al., 2005; Daub et al., 2004; Lord et al., 2003), chemical subjects (Barnard and Downs, 1992; Carbó et al., 1980; Nikolova and Jaworska, 2003), words (Huang, 2008; Sahami and Heilman, 2006) and so on. According to Serra and Arcos (2014), the similarity measure is the most essential core element of time series classification and clustering. Therefore, the development of better similarity measure can significantly assist the improvement of data analysis efficiency. As stated by Cha (2007), the similarity measure is closely related to the distance measure, as the distance is defined as a quantitative degree of how far apart two objects are. Consequently, studies of distance and similarity are significantly connected and crucial in terms of solving many pattern recognition related problems, such as clustering technique (Davies and Bouldin, 1979; Jarvis and Patrick, 1973), Taxonomy (Lin et al., 1998; Resnik, 1995), image registration (Penney et al., 1998; Roche et al., 1998), etc.

As one of the crucial difficulties in similarity measure is that the different types of features are not comparable, to overcome this, the corresponding distribution of extracted eigenvalues is considered as the “formal" criterion for developing a novel similarity measure. It is inspired by the subspace-based technique that incorporates the dynamical approach and embedding theorem to transform a one dimensional time series to a multidimensional Hankel matrix. Hankel matrix have many features as a square matrix, where gives a sequence of the one dimensional time series, also defines the dynamical state-space. The explorations of the significance of the empirical distribution of the eigenvalues of the Hankel matrix can be found in (Ghodsi et al., 2015a; Hassani et al., 2014, 2015a).

To my knowledge, this research is the initial attempt of adopting eigenvalue distribution into formulating a similarity measure in the multivariate system. The successful implementation of this novel similarity measure can overcome the limitations of nonlinear dynamic, complex fluctuations and the possibility of distinguishing similarity for particular or selected features. Note that time series under evaluation are embedded into multidimensional matrices and combined either vertically or horizontally to be transformed into a Hankel matrix, where the eigenvalues can be extracted by SVD technique accordingly. Moreover, in order to evaluate the reliability of eigenvalue distribution as the similarity measure, three empirical statistical tests together with the real case scenario are overwhelmingly evaluated. Possible circumstances during the formulation process of the new measure are comprehensively validated with brief introductions and comparisons in following sections.

## 3.2 Theoretical Formulation

### 3.2.1 Eigenvalue Distribution

To overcome the difficulty of existing diverse and incomparable features, the novel similarity measure extracts the corresponding eigenvalue distributions as the formal criterion by considering the elements of time series as a whole without removing any nonlinear or complex features. It is of note that the structures of constructing Hankel matrix containing multiple variables differ, including both horizontal and vertical forms (the corresponding details of constructing Hankel matrix can be found in Chapter 2, section 2.2.3, while the algorithm of extracting eigenvalues can be found in Chapter 2, section 2.1). Thus, a simple example of two time series case is summarized below for reference.

Consider two time series $X$ and $Y$ with different series length $N_X$ and $N_Y$:

$$X = (x_1, x_2, \dots, x_{N_X}),$$

$$Y = (y_1, y_2, \dots, y_{N_Y}).$$

Firstly, transfer the one-dimensional time series $X$ and $Y$ in to multidimensional matrix respectively: $[X_1, ..., X_{K_X}]$ with vectors $X_j$ that equals to $(x_j, ..., x_{j+L_X-1})^T \in \mathbf{R}^{L_X}$; $[Y_1, ..., Y_{K_Y}]$ with vectors $Y_i$ that equals to $(y_i, ..., y_{i+L_Y-1})^T \in \mathbf{R}^{L_Y}$. For which, $L_X (2 \leq L_X \leq N_X/2)$ and $L_Y (2 \leq L_Y \leq N_Y/2)$ are the window lengths for each series and generally $K = N - L + 1$.

Then two trajectory matrices are constructed:

$$\mathbf{X} = [X_1, ..., X_{K_X}] = (x_{mn})_{m,n=1}^{L_X, K_X},$$

$$\mathbf{Y} = [Y_1, ..., Y_{K_Y}] = (y_{pq})_{p,q=1}^{L_Y, K_Y}.$$

In order to construct a block Hankel matrix in the vertical form, thus $K_X = K_Y = K$. Accordingly, this version enables various window length for different series length, but identical $K$ for all series. The result of this step is the following matrix:

$$\mathbf{M}_V = \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix}.$$

Note that $\mathbf{M}_V$ indicates that the output of the first step is a block Hankel trajectory matrix formed in a vertical form.

Then, the SVD of $\mathbf{M}_V$ is performed in the following step. Denote $\lambda_{V_1}, ..., \lambda_{V_{L_X+L_Y}}$ as the eigenvalues of $\mathbf{M}_V \mathbf{M}_V^T$, arranged in decreasing order $(\lambda_{V_1} \geq ... \lambda_{V_{L_X+L_Y}} \geq 0)$ and $U_{V_1}, ..., U_{V_{L_X+L_Y}}$, the corresponding eigenvectors. Note also that the structure of the matrix $\mathbf{M}_V \mathbf{M}_V^T$ is as follows:

$$\mathbf{M}_V\mathbf{M}_V^T = \begin{bmatrix} \mathbf{X}\mathbf{X}^T & \mathbf{X}\mathbf{Y}^T \\ \mathbf{Y}\mathbf{X}^T & \mathbf{Y}\mathbf{Y}^T \end{bmatrix}.$$

The SVD of $\mathbf{M}_V$ can be written as $\mathbf{M}_V = \mathbf{M}_{V_1} + \cdots + \mathbf{M}_{V_{L_X+L_Y}}$, where $\mathbf{M}_{V_g} = \sqrt{\lambda_{V_g}} U_{V_g} V_{V_g}^T$ and $V_{V_g} = \mathbf{M}_V^T U_{V_g} / \sqrt{\lambda_{V_g}}$ $(g = 1, \ldots, L_X + L_Y$ and $\mathbf{M}_{V_g} = 0$ if $\lambda_{V_g} = 0)$.

On the other hand, the horizontal form block Hankel matrix requires $L_X = L_Y = L$ and enables various $K$ for different series length so to construct the matrix $\mathbf{M}_H = [\mathbf{X} : \mathbf{Y}]$. Thus, the structure of the matrix $\mathbf{M}_H\mathbf{M}_H^T$ is as follows:

$$\mathbf{M}_H\mathbf{M}_H^T = \mathbf{X}\mathbf{X}^T + \mathbf{Y}\mathbf{Y}^T.$$

The SVD of $\mathbf{M}_H$ can then yield $L$ eigenvalues: $\mathbf{M}_H = \mathbf{M}_{H_1} + \cdots + \mathbf{M}_{H_L}$, where $\mathbf{M}_{H_g} = \sqrt{\lambda_{H_g}} U_{H_g} V_{H_g}^T$ and $V_{H_g} = \mathbf{M}_H^T U_{H_g} / \sqrt{\lambda_{H_g}}$ $(g = 1, \ldots, L$ and $\mathbf{M}_{H_g} = 0$ if $\lambda_{H_g} = 0)$.

Note that the horizontal form decomposition is proved to produce more reliable and consistent information of eigenvalue distributions, which is evident by the careful consideration and comparison of eigenvalue distributions by both vertically and horizontally formed techniques (the detailed comparisons are available in Appendix B). Hence, all tests in the following sections are based on eigenvalues conducted by decomposition stage of the horizontal form.

### 3.2.2 Novel Similarity Measure

By setting the eigenvalue distribution as the similarity measure criterion, the hypotheses of the novel similarity measure are stated as below:

**Null hypothesis** ($H_0$): there is no significant difference between the eigenvalue distributions of matrices by two tested series.

**Alternative hypothesis** ($H_a$): there is a significant difference between the eigenvalue distributions of matrices by two tested series.

The null hypothesis is rejected when the $p$-value is less than the 5% significance level, and therefore it is concluded that the set of eigenvalues are not similar and consequently two test series are different. While if the $p$-value is very close to or equal to 1, the two tested series are similar as they share very similar or even identical eigenvalue distributions.

As the proposing method of measuring similarity based on eigenvalue distribution is considering a possible implementation of detecting “Formal Cause", different benchmarks of comparison will lead to different results. Consider two random variables $X$ and $Y$, “how similar is $X$ to $Y$" and “how similar is $Y$ to $X$" are two different questions depending on which element is set as the benchmark. However, even it is not expected to receive exactly identical results between comparing $X$ to $Y$ and $Y$ to $X$, the expected final outcomes that define “similar" or “different" should not vary.

Furthermore, even it is the same question to be tested, there are also two types of circumstances determined by with or without the premise of multivariate system. For instance, if the principle is to answer the question of how similar is $Y$ to $X$, the eigenvalue distribution by corresponding matrix $\mathbf{XX}_H$ will be considered as the “benchmark" for further evaluation to compare with eigenvalue distribution extracted by: $\mathbf{YY}_H$ if it is without the premise of multivariate system; $\mathbf{XY}_H$ with the premise of multivariate system. Hence, if the eigenvalue distribution is statistically similar with the “benchmark" eigenvalue distribution, $Y$ will then concluded as similar with $X$. Note that the detailed test results of simulations with and without the premise scenarios will be separately presented in the following sections.

A flowchart is provided in Fig. 3.1 that briefly summarizes the formulation and evaluation process of this proposing similarity measure. Note that in terms of simulation, corresponding process is repeated 1000 times respectively, and the population of tested series are generated by involving random white noises that being maintained at about 10% of the range of tested series. As stated in the flowchart, it is worth to be addressed that the empirical tests that are adopted here for providing statistical measurement on the eigenvalue distributions are Chi-squared Test, Log-likelihood Goodness of Fit Test and Kolmogorov-Smirnov Test, which will be specifically reviewed in the next section.

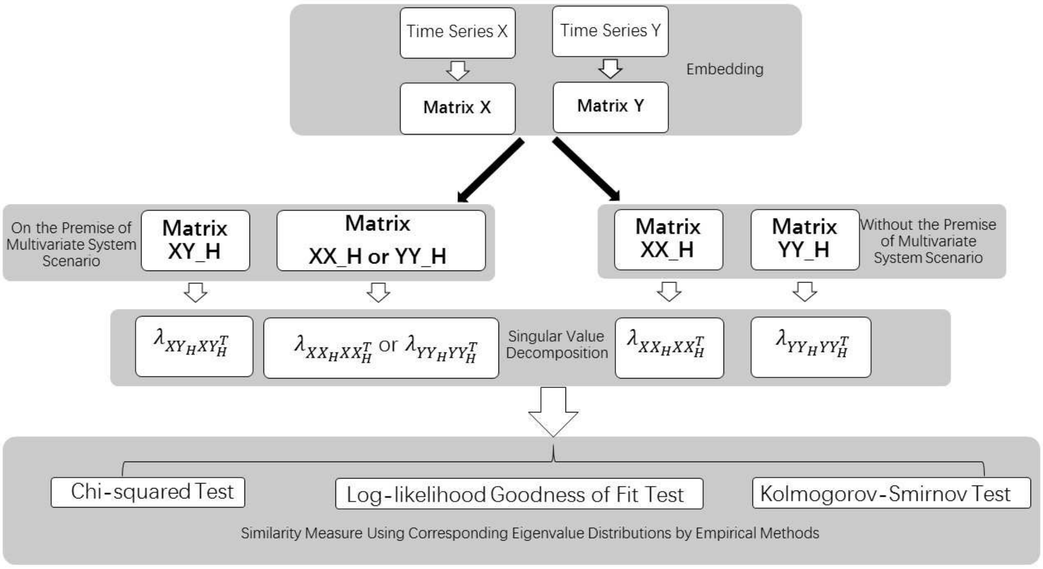

Time Series X
Time Series Y
Embedding
Matrix X
Matrix Y
On the Premise of Multivariate System Scenario
Matrix XY_H
Matrix XX_H or YY_H
Matrix XX_H
Matrix YY_H
Without the Premise of Multivariate System Scenario
$\lambda_{XY_H XY_H^T}$
$\lambda_{XX_H XX_H^T}$ or $\lambda_{YY_H YY_H^T}$
Singular Value Decomposition
$\lambda_{XX_H XX_H^T}$
$\lambda_{YY_H YY_H^T}$
Chi-squared Test
Log-likelihood Goodness of Fit Test
Kolmogorov-Smirnov Test
Similarity Measure Using Corresponding Eigenvalue Distributions by Empirical Methods

Fig. 3.1 The Flowchart of The Novel Similarity Measure by Eigenvalue Distribution.

Moreover, in order to ensure the consistency and comparability, the default window length is set as about 1/10 of the time series length. This will be fair number to include almost all significant eigenvalues without containing too much unimportant ones. With a relatively larger window length, the information will be split either flatly or partly flatly by more eigenvalues, and the differences will be split to be less significant to be identified; in contrast, a smaller window length will result in the fewer amount of eigenvalues with more significant differences for all or some of the eigenvalues. Without considering the consistency to be comparable, the most proper window length will be selected heavily depends on the feature of the series being analyzed with the principle of relatively maximizing the significant information with possibly small number of eigenvalues.

## 3.3 Empirical Methods of Comparing Distribution Similarity

In order to evaluate whether the extracted eigenvalues are similar or not to conclude the similarity between two tested series, three empirical statistical tests (Chi-squared Test, Log-likelihood Goodness of Fit Test and Kolmogorov-Smirnov Test) are adopted (note that various distance and similarity measures are comprehensively reviewed and categorized by Cha (2007) for more information). In general, coordinates and the cumulative distribution function (CDF) are the most generally accepted concepts to represent the examined subject. Since the proposed similarity measure is expected to have no assumption or limitation on measuring tested series with only the empirical distributions, some tests that are commonly used to evaluate the consistency with the empirical distributions can not be properly suitable here (i.e. Shapiro-Wilk Test (Shapiro and Wilk, 1965), Hellinger Distance (Bhattacharyya, 1946), Kullback Leibler Divergence (Kullback and Leibler, 1951), Anderson-Darling Test (Anderson and Darling, 1952)). Therefore, only brief introductions of several important and dominant measurements that are referred for formulating the novel similarity measure due to the special feature of eigenvalue distribution are listed respectively as follows.

### 3.3.1 Chi-squared Test

As an improved distance measure comparing to Euclidean distance, the Chi-squared statistic can be simply considered as the summation of squared Euclidean distances of two vectors (by considering them in a $n$ dimensional space domain, where $n$ is the number of observations for both vectors) over the corresponding “coordinates" of the domain vector. The Chi-squared distribution (also known as Helmertian distribution) (Helmert, 1876) is one of the most significantly applied probability distributions, and it is most commonly accepted for measuring the distance or similarity level between two probability distributions. Pearson (1900) adopted the Chi-squared distribution in the goodness of fit domain and conducted the Chi-squared test, which statistically evaluates the observed data about its goodness of fit level and consistency with an expected distribution. Here in this chapter, it is adopted for comparing the eigenvalue distributions as evidence of similarity. The Chi-squared statistic formula is:

$$\chi^2(C,E) = \sum_{i=1}^{Z} \frac{(C_i - E_i)^2}{E_i}, \tag{3.1}$$

where $Z$ is the number of levels of categories; $C$ is the observed frequency and $E$ is the expected count.

Therefore, in terms of Chi-squared test between two tested variables, assume $Z_A$ and $Z_B$ are the number of levels of categorized variables $A$ and $B$, so the degree of freedom can be calculated by $df = (Z_A - 1) \times (Z_B - 1)$. The expected counts/frequencies is computed by

$$E_{Z_{A,B}} = (C_{Z_A} \times C_{Z_B})/n, \tag{3.2}$$

where $C_Z$ refers to observed counts at specific level of category and $n$ indicates the total observation number. Consequently, the corresponding Chi-squared statistics is:

$$\chi^2(A,B) = \sum \frac{(C_{Z_{A,B}} - E_{Z_{A,B}})^2}{E_{Z_{A,B}}}. \tag{3.3}$$

### 3.3.2 Log-likelihood Goodness of Fit Test

The Log-likelihood Goodness of Fit Test is actually based on the commonly used Chi-squared test statistics in Pearson (1900). According to Sokal Robert and James

(1981), the Log-likelihood statistic formula is:

$$G = 2\sum_{i} f_i \cdot ln(\frac{f_i}{q_i}), \tag{3.4}$$

where the $f_i$ refers to the observed frequency, whilst $q_i$ indicates the expected frequency. More specifically, the test is adopted for evaluating whether the eigenvalue distribution of the examined series fit well to the eigenvalue distribution of the benchmark series.

### 3.3.3 Kolmogorov-Smirnov Test

The Kolmogoriv-Smirnov Test (K-S Test) was firstly proposed by Kolmogorov (1933). As a non-parametric statistical test, it quantifies the distance based on the CDF with no assumption about the distribution of data. It can be adopted to examine the similarity level of one distribution to empirical distribution, more importantly, K-S test is also applicable for evaluating the similarity of distributions of two random samples. The K-S test statistic is defined as below, which mainly follows (Hassani and Silva, 2015):

$$D_n = sup_x |F_n(x) - F(x)|, \tag{3.5}$$

where $F$ refers to the theoretical cumulative distribution function, $F_n$ represents the cumulative distribution up to $n$ observations, $sup_x$ indicates the supremum of the set of distances, and $D_n$ refers to the supremum distance reached up to $n$ observations. In terms of the two-sample case of K-S Test, the corresponding test statistic formula is:

$$D_{n,n'i} = sup_x |F_{1,n}(x) - F_{2,n'}(x)|, \tag{3.6}$$

note that $F_{1,n}$ and $F_{2,n'}$ are the corresponding distribution function for two tested samples respectively.

Specifically for the proposed similarity measure method based on eigenvalue distribution, two-sample K-S Test is adopted to determine whether the “benchmark" populations created by the dominate series has consistent eigenvalue distribution as the other series.

## 3.4 Simulation Performance

In order to evaluate the performance of the proposed similarity measure by eigenvalue distribution, various types of simulated series are tested by being separated into two groups of circumstances: the similar group and the different group, additionally the different choices of “benchmark" are also considered in each group. The initials of various types of generated series are listed below for the sake of simplifying the expressions:

| | | |
|---|---|---|
| 1. | $WN$ | White Noise. |
| 2. | $UD[0,1]$ | Uniform Distribution Series [0, 1]. |
| 3. | $UD[-1,1]$ | Uniform Distribution Series [-1, 1]. |
| 4. | $EP[1]$ | Exponential Distribution Series rate 1. |
| 5. | $SINE[-1,1]$ | Sine Wave Series [-1, 1]. |

The robustness of accepting eigenvalue distribution as similarity measure criterion are preliminarily examined in this section, specifically, the test results are summarized in Tables 3.1 and 3.2 by each empirical statistical method adopted as follows.

### 3.4.1 On the Premise of Multivariate System Scenario

Regarding the scenario of with the premise of multivariate system, the similarity of eigenvalue distributions extracted from the matrices $\mathbf{XY}_H$ and $\mathbf{XX}_H$ (or $\mathbf{YY}_H$ determined by which series is considered as the benchmark series) are evaluated, respectively. Note that $\mathbf{XY}_H$ is created from two time series $X_N$ and $Y_N$ simultaneously, and $\mathbf{XX}_H$ (or $\mathbf{YY}_H$) is formed by $X_N$ (or $Y_N$) with itself respectively. The corresponding test results of eigenvalue distributions as novel similarity measure by adopting three different empirical methods are summarized in Table 3.1. Note that the bold number indicates the best performance option in corresponding comparable level.

The Chi-squared test results show positive outcomes as expected for the “similar" group on both numbers of observations scenarios, whilst in terms of the “different" group, the tests can perform better for longer series. However, there are still significantly unexpected results ($p$-value is close to 1) for the UD[0,1] & EP[1] and UD[-1,1] & SINE[-1,1] combinations, especially the results vary greatly for the UD[0,1] & SINE[-1,1] and EP[1] & SINE[-1,1] cases. As mentioned earlier, the population for comparison is created by the “benchmark" series, therefore differences are expected when switching the “benchmark" series, however, opposite results for the same pair of series are not robust as expected, and it is even worse than the cases of indicating “similar" for the groups that are expected to be “different".

In terms of the log-likelihood goodness of fit test results, expected results for the “similar" group are confirmed in accordance with the simulation results. $P$-values are equal to 1, which indicate that it is almost 100% sure to accept the null hypothesis, therefor very similar or identical eigenvalue distributions prove the expected conclusion of “similar". Regarding the expected to be “different" group, both long and short series length, 1000 and 100 observations, show generally consistent significant results, except the UD[0,1] & EP[1] combination. Since UD[0,1] and EP[1] indeed show similar eigenvalue distributions (more information can be found in Appendix B) and the differences are between the tails, the log-likelihood goodness of fit test is not sensitive for detecting differences of distributions with flat tails. However, the advantage of this test can be noticed in the shorter length of observation scenario; the results are almost stable and consistent with the expected results of highly significant statistics.

K-S test show positive results as expected for the “similar" group on both numbers of observations scenarios. In terms of $N = 100$ case for “different" group of combinations, only the UD[0,1] & UD[-1,1] combination can be detected with 10% of significance level, however, the differences between switching dominant series to create “benchmark" populations are not significant. Comparing to the results of previous tests, the inconsistency is worse than less sensitivity of accurate detection, it has to be noticed that the two sample K-S test shows great performance on consistency and stability, even in the quite unstable and greatly varied scenarios that other tests can not even provide uniformed results. In addition, for the “different" group with $N = 1000$ case, almost all results are as expected to be significant (majority is under 5%, only a few are under 10%). Note that the EP[1] & SINE[-1,1] combination is the only one that K-S test could not detect significantly, and this is mostly because that K-S test is not that much sensitive to the differences at tail, also the natural character of eigenvalue distribution for both types of series vary at the tail part with increasing differences when the window length of structuring matrix increases.

Table 3.1 Similarity Measure Evaluation by Three Different Tests on Simulated Groups of Series on the Premise of Multivariate System Scenario.

| | | | Chi-squared Test | | | | Log-likelihood GOF Test | | | | K-S Test | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | N=100<br>L=10 | | N=1000<br>L=100 | | N=100<br>L=10 | | N=1000<br>L=100 | | N=100<br>L=10 | | N=1000<br>L=100 | |
| | | | Y→X | X→Y | Y→X | X→Y | Y→X | X→Y | Y→X | X →Y | Y→X | X→Y | Y→X | X→Y |
| | X | Y | *p*-value | | *p*-value | | *p*-value | | *p*-value | | *p*-value | | *p*-value | |
| **Similar** | UD[0,1] | UD[0,1] | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |
| | UD[-1,1] | UD[-1,1] | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |
| | EP[1] | EP[1] | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |
| | SINE[-1,1] | SINE[-1,1] | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** | **1.00** |
| **Different** | UD[0,1] | UD[-1,1] | 0.14 | **0.00** | 0.00 | 0.00 | **0.04** | 0.04 | 0.00 | 0.00 | 0.07 | 0.05 | **0.00** | **0.00** |
| | UD[0,1] | EP[1] | 0.76 | 0.98 | 0.88 | 1.00 | 0.81 | 1.00 | 0.99 | 1.00 | **0.77** | **0.65** | **0.01** | **0.01** |
| | UD[0,1] | SINE[-1,1] | 0.98 | 0.35 | 1.00 | 0.00 | **0.00** | **0.00** | **0.00** | **0.00** | 0.76 | 0.89 | 0.02 | 0.01 |
| | UD[-1,1] | EP[1] | **0.01** | 0.88 | 0.00 | 0.01 | 0.05 | **0.35** | 0.00 | 0.00 | 0.49 | 0.65 | **0.00** | **0.00** |
| | UD[-1,1] | SINE[-1,1] | 1.00 | 1.00 | 1.00 | 0.99 | **0.00** | **0.00** | **0.00** | **0.00** | 0.11 | 0.41 | 0.10 | 0.10 |
| | EP[1] | SINE[-1,1] | 1.00 | 0.20 | 1.00 | 0.00 | **0.00** | **0.00** | **0.00** | **0.00** | 0.45 | 0.60 | 0.56 | 0.53 |

### 3.4.2 Without the Premise of Multivariate System Scenario

In terms of the scenario without the premise of multivariate system, the similarity measure is performed on the eigenvalue distributions extracted from the matrices $\mathbf{XX_H}$ and $\mathbf{YY_H}$ respectively. To be consistent with the previous evaluation process, the following tests consider both similar and different groups of series and evaluate the performance of similarity measure by 1000 time simulations. Note that this time there is no premise of a multivariate system, therefore, the evaluation by simulated series will have no assumption on benchmark series. Hence, for each pair of series, there is only one test statistic conducted. The default number of observation is 1000 and default window length is 100. All statistical tests results are listed in Table 3.2. Note that the bold number indicates the best performance option in corresponding comparable level.

Table 3.2 Similarity Measure Evaluation by Three Different Tests on Simulated Groups of Series Without the Premise of Multivariate System Scenario.

| | | | Chi-squared Test | | Log-likelihood GOF Test | | K-S Test | |
|---|---|---|---|---|---|---|---|---|
| | | | N=100 | N=1000 | N=100 | N=1000 | N=100 | N=1000 |
| | | | L=10 | L=100 | L=10 | L=100 | L=10 | L=100 |
| | X | Y | *p*-value | *p*-value | *p*-value | *p*-value | *p*-value | *p*-value |
| **Similar** | UD[0,1] | UD[0,1] | 0.99 | 1.00 | 0.99 | 1.00 | 0.99 | 0.99 |
| | UD[-1,1] | UD[-1,1] | 0.99 | 1.00 | 0.99 | 1.00 | 0.99 | 0.98 |
| | EP[1] | EP[1] | 0.99 | 1.00 | 0.99 | 1.00 | 0.98 | 0.93 |
| | SINE[-1,1] | SINE[-1,1] | 0.99 | 1.00 | 0.99 | 1.00 | 0.99 | 0.99 |
| **Different** | UD[0,1] | UD[-1,1] | 0.01 | 0.00 | **0.01** | 0.00 | 0.03 | **0.00** |
| | UD[0,1] | EP[1] | 0.72 | 0.73 | 0.72 | 0.64 | **0.61** | **0.00** |
| | UD[0,1] | SINE[-1,1] | **0.52** | 0.45 | 0.54 | 0.49 | 0.76 | **0.01** |
| | UD[-1,1] | EP[1] | 0.22 | 0.00 | **0.18** | 0.00 | 0.23 | **0.00** |
| | UD[-1,1] | SINE[-1,1] | **0.96** | 0.46 | 0.96 | 0.50 | 0.98 | **0.03** |
| | EP[1] | SINE[-1,1] | **0.58** | **0.49** | 0.59 | 0.50 | 0.99 | 0.57 |

It is worth to be noted that due to the algorithm of applying Chi-square test and Log-likelihood goodness of fit test for two sample test, it is necessary to define one of the tested series as dominant series and re-scale the assumption of distribution in the first place for the further tests. Consequently, for the scenario of without the premise of multivariate system, the simulations of 1000 times are equally shared by both series in one pair of tested series. Therefore, both series have same quantity of chances to be the dominant series to re-scale the assumption distribution. K-S test do not have assumptions on any distribution, hence simulations for two sample test of K-S test here do not have significant difference comparing to the corresponding process of previous scenario on the premise of multivariate system.

According to Table 3.2, all statistical tests provide consistent results on both short and long series for the similar group, which overwhelmingly show *p*-value nearly equal or identical to 1. Consequently, it indicates significantly similar eigenvalue distributions and then the similarity between tested series. However, in terms of the different group, both Chi-squared test and Log-likelihood goodness of fit test could not detect most of the differences properly except the UD[0,1] & UD[-1,1] and UD[-1,1] & EP[1] combinations. It is mostly because of the variation and instability caused by switching dominant series for re-scale distribution assumption. Even for the longer series case, most of the results get smaller *p*-values (which indicates different eigenvalue distributions), they are still not significant enough as expected for the generated different group. K-S test is proved to outperform the other two tests for the long series case, also it can accurately detect the similarity or differences for both simulated groups. Even for the short series case, the results of K-S test are fairly close to the results of the other tests. Unlike the previous test results of log-likelihood goodness of fit test, it does not show good performance on short series this time.

In general, by considering the scenario without the premise of a multivariate system, the K-S test is confirmed as the most proper statistics to be adopted for the new similarity measure based on eigenvalue distribution. Moreover, for both scenarios, it is promising to obtain consistent results as simulative expectations, which convincingly prove the satisfying robust performances of this novel similarity measure on several different types of simulated series.

## 3.5 Evaluation in Real Case Scenario

Following the previous evaluations by simulations, it can be summarized that the eigenvalue distribution can be considered as a proper criterion of measure similarity by adopting suitable statistical test; K-S test outperforms others in the large data size domain with consistent results as simultaneously expected.

Considering the real case scenario, data can be ideally assumed to be formed by signal and noise. Therefore, it is not applicable to simulate noises to form and produce the population of dominate series as the benchmark to measure similarity. Consequently, bootstrap re-sampling technique (Efron, 1979) is adopted to conduct the population of dominate series with specific confidence level and evaluate how similar the other tested variable is to the benchmark population under the specific confidence level circumstance. Note that the newly proposed method can certainly be performed without any re-sampling process if there are already clear information of its population. The corresponding population will only be generated by re-sampling for obtaining the information of its population. Due to the nature of similarity that is discussed previously, the similarity level of $X$ to $Y$ and $Y$ to $X$ are two different questions regarding the differences of the benchmark. Therefore, the re-sampling process will consider two different cases by choosing different original series to create the population.

A flowchart is provided in Fig. 3.2 that briefly summarizes the formulation process under the simultaneous real case scenario by bootstrap re-sampling. For instance, when the principle is to obtain the population of benchmark series $X$, thus, the population of $\lambda_{XX_HXX_H^T}$ (determined by with or without the premise of multivariate system) is conducted, which is formed by eigenvalues distributions within specific confidence interval of K-S statistics. Therefore, if the confidence level is fixed as 95%, the population of eigenvalue distributions will then be conducted that indicate significantly 95% similarity level with benchmark series. To this end, the other series can be evaluated by comparing its corresponding K-S statistics with the range of K-S statistics by the population. Therefore, different similarity levels can be identified respectively with necessary adjustment of confidence level in the bootstrap re-sampling stage.

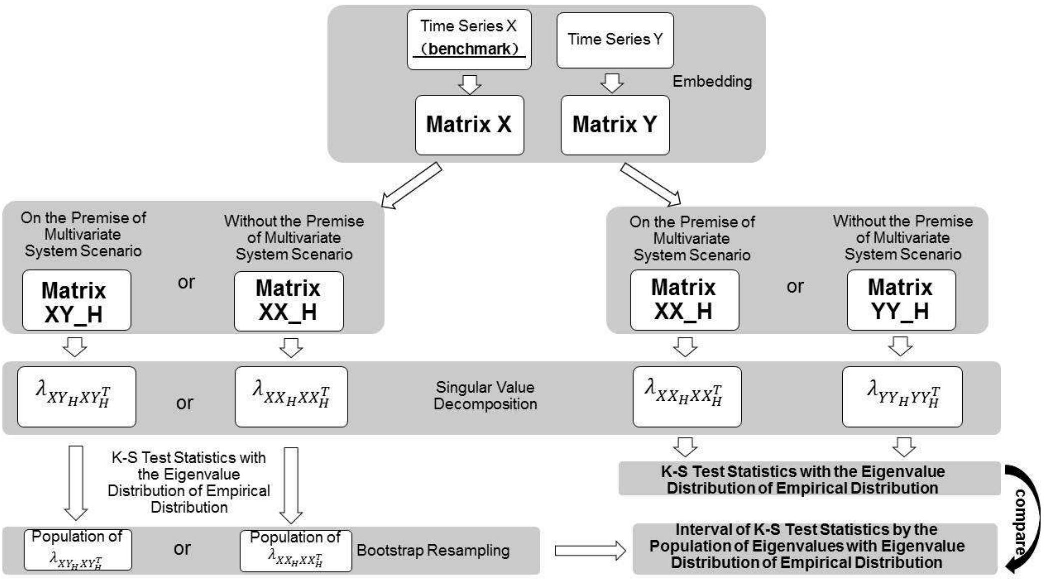

Time Series X
(benchmark)
Time Series Y
Embedding
Matrix X
Matrix Y
On the Premise of Multivariate System Scenario
Without the Premise of Multivariate System Scenario
Matrix XY_H
or
Matrix XX_H
On the Premise of Multivariate System Scenario
Without the Premise of Multivariate System Scenario
Matrix XX_H
or
Matrix YY_H
$\lambda_{XY_H XY_H^T}$
or
$\lambda_{XX_H XX_H^T}$
Singular Value Decomposition
$\lambda_{XX_H XX_H^T}$
$\lambda_{YY_H YY_H^T}$
K-S Test Statistics with the Eigenvalue Distribution of Empirical Distribution
Population of $\lambda_{XY_H XY_H^T}$
or
Population of $\lambda_{XX_H XX_H^T}$
Bootstrap Resampling
K-S Test Statistics with the Eigenvalue Distribution of Empirical Distribution
Interval of K-S Test Statistics by the Population of Eigenvalues with Eigenvalue Distribution of Empirical Distribution
compare

Fig. 3.2 The Flowchart of The Simultaneous Real Case Scenario by Bootstrap Resampling.

The results by representative simultaneous groups of series are provided in Table 3.3. In terms of the similar group, the similar group shows consistent results for both short and long series, in which, 95% significant level indicates tested series share at least 95% of similarity based on the eigenvalue distributions from the corresponding matrices. According to the previous evaluations of K-S test on short and long series for different group, here it is only considered to evaluate the performance on long series in accordance to its previous promising results in simulations (symbol \ for short series in Table 3.3). The 5% significant level refers to that the test statistics does not fit even when the confidence level of bootstrap re-sampling is set as 5%. This significantly indicates that tested series can be considered different as they are not similar even for 5% significant level.

Table 3.3 Simultaneous Real Case Similarity Measure Results by Bootstrap Re-sampling.

| | | | N=100 L=10 | | | | N=1000 L=100 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Y to X | | X to Y | | Y to X | | X to Y | |
| | X | Y | Y/N | Sig Level | Y/N | Sig Level | Y/N | Sig Level | Y/N | Sig Level |
| | UD[0,1] | UD[0,1] | ✓ | 95% | ✓ | 95% | ✓ | 95% | ✓ | 95% |
| **Similar** | UD[-1,1] | UD[-1,1] | ✓ | 95% | ✓ | 95% | ✓ | 95% | ✓ | 95% |
| | EP[1] | EP[1] | ✓ | 95% | ✓ | 95% | ✓ | 95% | ✓ | 95% |
| | UD[0,1] | UD[-1,1] | \ | \ | \ | \ | ✓ | 5% | ✓ | 5% |
| **Different** | UD[0,1] | EP[1] | \ | \ | \ | \ | ✓ | 5% | ✓ | 5% |
| | UD[-1,1] | EP[1] | \ | \ | \ | \ | ✓ | 5% | ✓ | 5% |

Note: ✓ indicates the result is correctly proved by the measure.

## 3.6 Discussion

Although as a novel similarity measure based on eigenvalue distribution with proven robustness and consistent performances, it is also certain that it is still the beginning of developing this new measure. The types of series in simulations are relatively limited, and there are still numerous choices of more complex series or combinations of series haven not been explored. The bootstrap re-sampling by K-S statistics for some real data (especially large size of data that is much longer than the default 1000 observations in simulation) may take a longer time of calculation, which makes it crucial to find a more straight forward process to identify the population information as the benchmark. Also, the performance in short series is not as good as its effort on long series. However, there are also numerous possibilities to improve this novel measure further as the second stage of this research: evaluating more representative data patterns; involving more types of noises with different levels of variations and more options of window lengths; in terms of time series with different frequencies, it can also provide possible solution by adopting SSA technique with specific modification accordingly; one significant implementation area of similarity measure is classification, therefore, the future research can also focus on the evaluations of its performances on classification tasks of time series, images, gene expressions, etc.

In general, by overcoming the difficulties of empirical similarity measures through identifying the comparable criterion, this chapter proposes a novel similarity measure based on eigenvalue distribution by incorporating the SVD technique, which is the initial attempt of adopting this technique in terms of the similarity measure development. The evaluation results are promising and robust as this research have considered many possible circumstances in the formulation process. The robustness of adopting eigenvalue distribution as proper criterion of measuring similarity have been examined; additionally, it is found that K-S test outperforms others in the large data size domain with consistent results as simultaneously expected. Furthermore, the simultaneous real case scenario is evaluated by adopting the bootstrap re-sampling technique to prevent the possible impacts during the process of creating benchmark population. Consistent results are achieved in the simultaneous real case scenario indicating the robust performance of distinguishing various “similar" or “different" groups of series.

This novel similarity measure can work properly on long series, and it does not require any assumption of distributions during the measuring process. The computation is reasonably efficient and can be easily employed by modifying currently available R packages. By considering eigenvalue distribution as the criterion of similarity measure, the amount of computation is significantly reduced for large data set. More importantly, this novel similarity measure can work with time series with different lengths and still identify the significant features for evaluations. In brief, this novel similarity measure contributes to providing a measurement that has no limitations of series length, series with nonlinear features or complex fluctuations, series sharing both signal and noises as similarities, etc. It is absolutely worth looking forward to its developments and implementations on various disciplines in the close future.

# Chapter 4

# Novel Mutual Association Measure by Eigenvalue-based Distance

This chapter explores the association measure aspect of the causality analysis by incorporating the subspace-based technique. A novel association measure based on eigenvalue-based distance is developed and comprehensively evaluated. Specifically, this chapter is formed as follows. A brief introduction of association study is provided in Section 1. Section 2 reviews several well established association measures that are used in linear or nonlinear association detection respectively. The development and formulation process of the novel mutual association measure is presented in Section 3. Section 4 concludes the evaluations of both empirical and novel association measures by simulations. The real data applications are conducted and evaluated in Section 5, and finally the conclusion is summarized in Section 6.

## 4.1 Introduction

Association can be briefly explained as the representation of any relationships, or measurement of independency between tested subjects. Studies of association from a statistical aspect can be tracked back to over one century ago. As one of the domain subjects in the study of multivariate system, the study of association, or identically named correlation analysis has been developed and applied across subjects on various disciplines, for example, economics (Filis et al., 2011), social science (Hajian and Movahed, 2010), chemistry (Chapman, 2012), biology (George et al., 2014), etc. To date, there are several established association measures with advantages in either linear or nonlinear association detection, for instances, Pearson (Pearson, 1895), Spearman (Spearman, 1904), Kendall (Abdi, 2007; Kendall, 1938), Hoeffding's D (Hoeffding, 1948), Distance Correlation (Székely et al., 2007), Mutual Information (Dionisio et al., 2004) and Maximal Information Coefficient (Reshef et al., 2011). However, there are still numerous possibilities for further improvements as none of these measures can master significant performances for the detection of all possible relationships in a broad sense.

This chapter develops a novel association measure that is more sensitive on detecting nonlinear or complex associations without losing the ability on basic linear association detection. This development is inspired by the subspace-based techniques: SVD and SSA, which have been applied and proved with promising performances on time series analysis, forecasting, denoising and multivariate analysis across various disciplines (more details can be found in Chapter 2, section 2.1 and 2.2). This research is the first attempt of incorporating the subspace-based technique with association study from a multivariate system aspect. More specifically, the concept of eigenvalue-based distance (Rodrıguez-Aragón and Zhigljavsky, 2010) is adopted as the criterion of measurement to assist on the development of the novel mutual association measure.

Note that in order to evaluate the reliability of this novel association measure, a few well established association measures are summarized and overwhelmingly considered as comparison. The performances of both empirical and novelly proposed association measures are evaluated by comprehensive simulations involving representative linear and nonlinear relationships. Furthermore, the real data implementations are conducted to evident on the robust performance of the novel mutual association measure in actual application scenario.

## 4.2 Benchmark Empirical Methods

This section briefly summarizes a few empirical association measures that are generally accepted and well established in literature by classifying them into linear and nonlinear domains as follows.

### 4.2.1 Linear Correlation

**Pearson Correlation Coefficient**

The Pearson correlation coefficient (Pearson, 1895) has been generally accepted as the most significant and well known measurement index to examine the correlation relationship between tested variables. The calculation process is easy and it has been applied for the majority of practical implementations in terms of association study. The Pearson correlation coefficient, $\rho$, between two random variables $X$ and $Y$ each containing $n$ observations is defined as:

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}, \tag{4.1}$$

where $E$ is the expected value operator, $\mu_X$, $\sigma_X$ and $\mu_X$, $\sigma_Y$ are expected value and standard deviation of random variables $X$ and $Y$, respectively.

Consequently, under null hypothesis circumstance, the Pearson correlation coefficient $\rho$ computed by the formula above follows the $t$-distribution with degree of freedom of $n-2$. The $t$ statistic is calculated by:

$$t = \frac{\rho(X,Y)\sqrt{n-2}}{\sqrt{1-\rho^2(X,Y)}}. \tag{4.2}$$

Pearson correlation coefficient $\rho$ satisfies $-1 \leq \rho \leq 1$ with the special cases of perfect linear dependence that equals to -1 or 1. It measures the direction and the dependence level between two tested variables in a linear domain. However, if two variables cannot be identified correlation by Pearson correlation, one cannot deny the possibility that they are associated in a nonlinear domain.

### Spearman Rank Correlation

Spearman rank correlation (Spearman, 1904) is another well accepted measure of correlation relationship between two variables. It is a nonparametric test and used ranked values to evaluate the association level based on the underlying assumption of a monotonic relationship. The monotonic relationship assumption is the major difference comparing to Pearson correlation, which is built on satisfying the much restrictive linear relationship. Therefore, assume that two variables $X_i$ and $Y_i$ are the original variables expecting to be evaluated, where $i$ is the paired score and $i \in [1,n]$ as $n$ is the number of observations for each variable. In addition, $x_i$ and $y_i$ are their corresponding ranked values. As a reminder, the Spearman rank correlation $s$ is calculated by the formula listed below:

$$s = 1 - \frac{6\sum(x_i - y_i)^2}{n(n^2-1)}, \tag{4.3}$$

where $n$ is the number observations.

Therefore, under the null hypothesis circumstance, the Spearman correlation coefficient can be estimated by the $t$-distribution with degree of freedom of $n-2$.

The $t$ statistics is then calculated by:

$$t = \frac{s\sqrt{n-2}}{\sqrt{1-s^2}}. \tag{4.4}$$

The Spearman correlation coefficient has the values that satisfy $-1 \leq s \leq 1$, where values -1 and 1 refer to perfect monotonic relationship, whilst $s = 0$ indicates monotonically independent random variables. It has been a significant improvement that the Spearman rank correlation coefficient extends the restriction of linearity to monotonic relationship. However, it is less sensitive to outliers and the results of "independent" tendency still cannot reject the possibility of nonlinear association.

### Kendall $\tau$ Rank Correlation Coefficient

Kendall correlation is firstly proposed by Kendall (1938) as an updated version of rank correlation measure. It considered the possible differences of ranking orders corresponding to random observers and developed the index $\tau$ to represent the new rank correlation coefficient as below by following (Kendall, 1938):

$$\tau = \frac{actual\ score}{maximum\ possible\ score}, \tag{4.5}$$

where, in terms of $n$ observations, the *actual score* is the number of different pairs between these two ordered sets, called the symmetric difference distance (Abdi, 2007).

Therefore, the *maximum possible score* can be calculated by

$$maximum\ possible\ score = (n-1) + (n-2) + \cdots + 1 = \frac{n(n-1)}{2}. \tag{4.6}$$

As an alternative rank correlation measure comparing to Spearman rank correlation, the Kendall rank correlation can also detect possible monotonic relationships. As the standard deviation of $\tau$ can be computed by $\sigma_\tau = \frac{1}{3}\sqrt{\frac{2(2n+5)}{n(n-1)}}$, therefore, the null hypothesis test process of obtaining significant test statistics is introduced below by following (Abdi, 2007):

$$Z_\tau = \frac{\tau}{\sigma_\tau}, \tag{4.7}$$

where $Z_\tau$ is a normal distributed statistics, also satisfies mean of 0 and standard deviation of 1. Kendall rank correlation directly illustrates the identical and different pairs and generally it indicates similar interpretations as the Spearman rank correlation. However, it is less sensitive to error and shows better performance with smaller sample size. Moreover, it is still not applicable to nonlinear association detection.

### 4.2.2 Nonlinear Association

**Hoeffding's D Test**

Hoeffding's D test is another nonparametric test of independence between two random variables proposed and named after Hoeffding (Hoeffding, 1948). The major difference between Hoeffding's D test and classical linear association measure methods like Pearson and Spearman is that it can detect some level of nonlinearity beyond the monotonic association relationship. It is based on ranked value similar as Spearman, however, the difference is that it measures the joint ranked values of two examined variables together.

As a reminder, assume two random variables $X$ and $Y$ with $n$ observations each, in which $x_i$ and $y_i$ have the ranks representing as $RX_i$ and $RY_i$ respectively ($i \in [1,n]$). Additionally, $Q_i$ refers to the number of points with both $x$ and $y$ values less than their corresponding $i$th point. Therefore, $Q_i = \sum_{j=1}^{n} \phi(x_j, x_i)\phi(y_j, y_i)$ and the Hoeffding's D statistic can be calculated as the formula listed below:

$$D = \frac{A - 2(n-2)B + (n-2)(n-3)C}{n(n-1)(n-2)(n-3)(n-4)}, \tag{4.8}$$

in which by setting as follows:

$$\begin{aligned} A &= \sum_{i=1}^{n}(RX_i - 1)(RX_i - 2)(RY_i - 1)(RY_i - 2) \\ B &= \sum_{i=1}^{n}(RX_i - 2)(RY_i - 2)Q_i. \\ C &= \sum_{i=1}^{n} Q_i(Q_i - 1) \end{aligned} \tag{4.9}$$

### Distance Correlation

Distance correlation is proposed by Székely et al. (2007) as a new measure of dependence between random vectors, which also claimed to be designed for detecting nonlinearity. It adopted the empirical concept of Euclidian distance together with sample moments. It is stated by Székely et al. (2009) that it is easy to calculate and can be apply to sample sizes $n \geq 2$ without restrictions on matrix inversion or estimation of parameters.

Assume two random variables $X$ and $Y$ with $n$ observations each, for which the pairwise Euclidean distances $a_{ij}$ and $b_{ij}$ (where $i, j = 1, \dots, n$) can be calculated by:

$$\begin{aligned} a_{ij} &= |x_i - x_j| \\ b_{ij} &= |y_i - y_j|. \end{aligned} \tag{4.10}$$

Therefore, transformed distance matrices $A_{ij}$ and $B_{ij}$ can be defined by:

$$\begin{aligned} A_{ij} &= a_{ij} - \frac{1}{n}\sum_{i=1}^{n} a_{ij} - \frac{1}{n}\sum_{j=1}^{n} a_{ij} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} \\ B_{ij} &= b_{ij} - \frac{1}{n}\sum_{i=1}^{n} b_{ij} - \frac{1}{n}\sum_{j=1}^{n} b_{ij} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}. \end{aligned} \tag{4.11}$$

Then the distance covariance can be calculated by following

$$V_{xy}^2 = \frac{1}{n^2} \sum_{i,j=1}^{n} A_{ij} B_{ij}, \tag{4.12}$$

and the distance correlation can be computed by

$$R^2 = \frac{V_{xy}^2}{V_x V_y}, \tag{4.13}$$

which satisfies $0 \leq R \leq 1$ and is employed to measure the correlation between $X$ and $Y$. Another significant difference of distance correlation is that $R = 0$ indicates independence unlike the other tests, and the index cannot be negative.

### Mutual Information

According to Hassani et al. (2010a), the mutual information (MI) are applied to measure the information that two tested variables $S$ and $W$ share with each other, or the same concept that to measure how much knowing one of these variables reduces the uncertainty about the other. The MI can be expressed as the following formula in accordance with (Hassani et al., 2010a):

$$I(S;W) = H(S) - H(S|W) = H(W) - H(W|S) = H(S) + H(W) - H(S,W). \tag{4.14}$$

where $H(S)$ and $H(W)$ are the marginal entropies, $H(S|W)$ and $H(W|S)$ are the conditional entropies, and $H(S,\ W)$ is the joint entropy of $S$ and $W$. The MI defined above takes a value between 0 and infinity, $0 \leq I(S,\ W) \leq +\infty$, which makes the comparisons difficult between different samples (Hassani et al., 2010a). In this context, Dionisio et al. (2004) among others, defined and used a standard measure for the MI:

$$\lambda = (1 - exp[-2I(S,\ W)])^{\frac{1}{2}}. \tag{4.15}$$

Note that $\lambda$ captures the overall dependence, both linear and nonlinear, between $S$ and $W$.

Additionally, Hassani et al. (2010a) defined the MI of two continuous random variables $S$ and $W$ as below:

$$I(S;W) = \int_S \int_W P(s,w) log\Big(\frac{P(s,w)}{P(s)P(w)}\Big) d_w d_s, \tag{4.16}$$

where $P(s,w)$ is the joint probability distribution function of $S$ and $W$, and $P(s)$ and $P(w)$ are the marginal probability distribution functions of $S$ and $W$, respectively. In brief, the MI is more general concept of measuring the mutual association level comparing to other correlation coefficients, and it certainly has no restrictions of nonlinearity. However, as stated by Hassani et al. (2010a), the comparisons are extremely difficult due to the fact that it does not have a uniformed range of values as the criterion for the index to be comparable across different groups of variables.

### Maximal Information Coefficient

According to Reshef et al. (2011), Maximal Information Coefficient (MIC) is a recently proposed measure of association based on the MI which measures that if a relationship between two random variables exists, a grid can be drawn on the scatter plot of the two variables for partitioning the data points and encapsulating this relationship. The details of definition and calculation of MIC are listed as follows, which here mainly follows (Reshef et al., 2011).

Specifically, for a give finite set $C$ of ordered pairs, the $x$ and $y$ values of $C$ are partitioned into $x$ and $y$ bins respectively, which is defined as an $x-by-y$ grid. As $C|_G$ refers to the distribution induced by the points in $C$ on the cells of $G$, the MI of $C|_G$ can be expressed as $I(C|_G)$. Therefore, the MIC of a set $C$ of two variables $X$

and $Y$ with $n$ observations each can be computed by:

$$MIC(X,Y) = max_{|X||Y|<B} \frac{maxI(C|_G)}{log(min(|X||Y|))}, \tag{4.17}$$

where $|X|$ and $|Y|$ are the number of bins for each variable respectively, and default setting of $B = n^{0.6}$ provides the upper bound of the size of the grids. The MIC measure of association will result in a coefficient in the range of $[0,1]$, which plays better criterion of association measure than MI. Furthermore, it has no restrictions of applicability on linear or nonlinear association relationships. However, as a relatively new method, it was challenged regarding the weaker performances on small size of samples and its inconsistent power for functional relationships at identical noise level (Gorfine et al., 2012; Simon and Tibshirani, 2014).

## 4.3 Theoretical Formulation

In this section, a new mutual association measure, which is built on the eigenvalue-based distance, is introduced with detailed formulation process. Note that the relative subspace-based techniques that have been adopted here are SVD and SSA, which have been introduced with details in Chapter 2, section 2.1 and 2.2, therefore, it is not reproduced here in this section.

### 4.3.1 Eigenvalue-based Distance

Eigenvalue-based approach is combined with image processing by considering digital image as matrix of grey level or color values (Rodrıguez-Aragón and Zhigljavsky, 2010). In which, the authors proposed the relatively new method for image denoising by combining MSSA technique and modified Frobenius distance formula based on eigenvalues. One of their significant research outcomes that is adopt here is the concept of eigenvalue-based distances between images. The eigenvalue-based distance for image processing proved with promising performances in image denoising and can be widely applied for face recognition and verification as another competitive approach (Rodrıguez-Aragón and Zhigljavsky, 2010). The theoretical formula of eigenvalue-based distance is listed below accordingly based on the work of Rodrıguez-Aragón and Zhigljavsky (2010).

Briefly, the eigenvalue-based distance introduced by Rodrıguez-Aragón and Zhigljavsky (2010) is built on the trajectory matrices of the images and their SVD expansions. Assume there are two trajectory matrices $\mathbf{X}^{(1)}$ and $\mathbf{X}^{(2)}$ of size $g \times q$, which are associated with two corresponding images $\mathbf{I}^{(1)}$ and $\mathbf{I}^{(2)}$ of the same size $h \times w$. In order to compare with uniform standard, these two matrices are firstly normalized by the formulation process below:

$$\begin{aligned} \mathbf{Y}_1 &= \mathbf{X}^{(1)}/\sqrt{tr(\mathbf{X}^{(1)}(\mathbf{X}^{(1)})^T)} \\ \mathbf{Y}_2 &= \mathbf{X}^{(2)}/\sqrt{tr(\mathbf{X}^{(2)}(\mathbf{X}^{(2)})^T)}. \end{aligned} \tag{4.18}$$

Then the corresponding eigenvalues of matrices $\mathbf{Y}_1\mathbf{Y}_1^T$ and $\mathbf{Y}_2\mathbf{Y}_2^T$ (where both of them are nonnegative definite) can be obtained by SVD, that are represented by $\lambda_1 \geq \cdots \geq \lambda_g$ and $\mu_1 \geq \cdots \geq \mu_g$ respectively. Note that $tr(\mathbf{Y}_1\mathbf{Y}_1^T) = tr(\mathbf{Y}_2\mathbf{Y}_2^T) = 1$, consequently, for all $i$, $\sum_{i=1}^{g} \lambda_i = \sum_{i=1}^{g} \mu_i = 1$ and corresponding eigenvalues satisfy $\lambda_i \geq 0, \mu_i \geq 0$.

A joint trajectory matrix based on $\mathbf{Y} = \binom{\mathbf{Y}_1}{\mathbf{Y}_2}$ is created to analyze two images simultaneously by:

$$\mathbf{Y}\mathbf{Y}^T = \begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} (\mathbf{Y}_1\mathbf{Y}_2) = \begin{pmatrix} \mathbf{Y}_1\mathbf{Y}_1^T & \mathbf{Y}_1\mathbf{Y}_2^T \\ \mathbf{Y}_2\mathbf{Y}_1^T & \mathbf{Y}_2\mathbf{Y}_2^T \end{pmatrix}. \tag{4.19}$$

Consequently, the eigenvalues of the joint trajectory matrix above can be donated as $\upsilon_1 \geq \cdots \geq \upsilon_{2g} \geq 0$, where $\upsilon$ satisfy $\sum_{i=1}^{2g} \upsilon_i = 2$ in accordance of $tr(\mathbf{Y}\mathbf{Y}^T) = tr(\mathbf{Y}_1\mathbf{Y}_1^T) + tr(\mathbf{Y}_2\mathbf{Y}_2^T) = 2$.

As Thompson and Therianos (1972) proved that for any matrix of joint form like $\mathbf{Y}\mathbf{Y}^T$, there exists the following relationship between corresponding eigenvalues for any positive integer $k$:

$$\sum_{j=1}^{k}\lambda_j+\sum_{j=1}^{k}\mu_j\geq\sum_{j=1}^{k}\upsilon_j. \tag{4.20}$$

By defining the cumulative distribution function on the integers $\{1,\ldots,g\}$ or $\{1,\ldots,2g\}$ respectively, so to have

$$\begin{aligned}F_1(t)&=\sum_{j=1}^{[t]}\lambda_j\\F_2(t)&=\sum_{j=1}^{[t]}\mu_j,\\F(t)&=\frac{1}{2}\sum_{j=1}^{[t]}\upsilon_j\end{aligned} \tag{4.21}$$

which indicate the inequality $F_1(t)+F_2(t)-2F(t)\geq 0$ for all $t\geq 0$. Finally the distance based on eigenvalue can be formulated as:

$$G(t)=F_1(t)+F_2(t)-2F(t). \tag{4.22}$$

More specifically, the natural definition of the eigenvalue-based distance between two images $\mathbf{I}^{(1)}$ and $\mathbf{I}^{(2)}$ can be expressed by:

$$d_1(\mathbf{I}^{(1)},\mathbf{I}^{(2)})=\int_0^k G(t)dt=\sum_{j=1}^{k}(\lambda_j+\mu_j-\upsilon_j). \tag{4.23}$$

### 4.3.2 Novel Mutual Association Measure

The new mutual association measure is then obtained based on the fundamental concept of eigenvalue-based distance. Assume there are two random series $X$ and $Y$, which have the same number of observations $n$. Firstly, these two random series

$X$ and $Y$ are transformed into two dimensional trajectory matrices $\mathbf{M_X}$ and $\mathbf{M_Y}$ by multiplying their transpose series respectively, which can be expressed as:

$$\begin{aligned}\mathbf{M_X} &= XX^T \\ \mathbf{M_Y} &= YY^T.\end{aligned} \tag{4.24}$$

Considering the concept is built on the relationships between eigenvalues, those two trajectory matrices are normalized before further formulation, which followed the normalization algorithm below:

$$\begin{aligned}\mathbf{NM_X} &= \mathbf{M_X}/\sqrt{tr(\mathbf{M_X M_X^T})} \\ \mathbf{NM_Y} &= \mathbf{M_Y}/\sqrt{tr(\mathbf{M_Y M_Y^T})}.\end{aligned} \tag{4.25}$$

Therefore, the joint matrix $\mathbf{NM}$ is created by combining $\mathbf{NM_X}$ and $\mathbf{NM_Y}$:

$$\mathbf{NM} = \begin{pmatrix}\mathbf{NM_X} \\ \mathbf{NM_Y}\end{pmatrix}. \tag{4.26}$$

Note that in terms of forming this joint matrix, horizontally and vertically formed structures do not have difference for the next step of transforming to trajectory matrices as they will show symmetric feature and provide identical eigenvalues.

The joint matrix $\mathbf{NM}$ then get transformed into trajectory matrix by multiplying its transpose matrix:

$$\mathbf{TM} = \mathbf{NM} \cdot \mathbf{NM^T}, \tag{4.27}$$

where $\xi_1 \geq \xi_2 \geq \ldots \geq \xi_n \geq 0$ donate the corresponding eigenvalues of $\mathbf{TM}$.

In fact by combining the two random series $X$ and $Y$, the final joint trajectory matrix will provide two significant eigenvalues $\xi_1$ and $\xi_2$, which are the first two in order. The identical (or closely associated) features will be able to presented by the first eigenvalue $\xi_1$ without any information left. In other words, the second eigenvalue $\xi_2$ indicates the information of "distance" between these two series, which also represents the not associated information between these two series. More specifically, if the two random series are identical, the eigenvalues of the joint matrix **TM** will show $\xi_1 = 2$ and $\xi_2 = \cdots = \xi_n = 0$. Additionally, on the contrary of the perfectly identical scenario, meaning if these two series are not associated at all, the corresponding $\xi_1$ and $\xi_2$ of **TM** will be both extremely close or equal to 1.

In summary, denote $\varphi(X,Y)$ as the mutual association index between two random variables $X$ and $Y$. The definition formula of $\varphi(X,Y)$ is written accordingly as

$$\varphi(X,Y) = 1 - \xi_2, \tag{4.28}$$

in which $\varphi(X,Y)$ satisfies $0 \leq \varphi(X,Y) \leq 1$. Specifically, $\varphi(X,Y) = 1$ indicates $X$ and $Y$ are most significantly associated (identical); $\varphi(X,Y) = 0$ refers that there is almost no association between $X$ and $Y$.

## 4.4 Simulation Performance

The performances of both empirical and newly proposed association measures are summarized below by simulations, in which different representative linear and nonlinear relationships or patterns are simulated for investigation and comparison. For each specific relationship (linear or nonlinear), a group of series with 200 observations for each specific population correlation values are generated and then repeated this process 1000 times. All statistics results are summarized and listed for comparison in Table 4.1, where 2.5% and 97.5% quartile, mean and standard deviation of test statistics from corresponding simulations are provided. Note that all simulations are obtained by R program with corresponding packages, in which representative nonlinear patterns are adopted by referring to the codes in (Boigelot, 2011).

According to the results in Table 4.1 by simulations of representative groups of series, the results obtained are in line with those previous literature of Reshef et al. (2011) and Clark (2013). In more details, Pearson and Spearman coefficients work properly on simulated linear group as usual with promising results and small standard deviation, while it is noticed that the standard deviation of both Pearson and Spearman correlation coefficients slightly increase when the population correlation coefficients converge to 0; Kendall coefficient provides coefficients with higher variations comparing to corresponding populations and the standard deviations are higher than both Spearman and Pearson with the same slight increasing trend as population coefficients converging to 0; in terms of the simulated nonlinear groups, all three linear measures cannot pick up any relationships, which provide coefficients equal or very close to 0.

In terms of the empirical nonlinear association measures, Hoeffding' s D test, MI and MIC cannot provide proper association indices for simulated linear groups comparing to other measures, whilst Distance Correlation is the only one can possibly measure the association by providing relatively closer indices if one takes no account of the direction of correlation. These results also confirm the findings by Clark (2013) that the Hoeffding's D and MIC appeared to get more differences away from the defined level of population whilst the Distance Correlation got much less. It is also noticed that for Hoeffding's D test, MI and MIC, conversely, their standard deviations show tendency of slight decreasing when the population correlation coefficients converge to 0, which may indicate that these association measures do detect some level of linear relationship, while the consistency and accuracy level are not stable as the other measures. Regarding the results of simulated nonlinear groups: Hoeffding' s D test shows very limited capacity of detecting any possible relationships; Distance Correlation is relatively more sensitive on nonlinear patterns; MI and MIC detect different levels of association for different linear patterns, in which, MI shows relatively significant estimates for quadratic and cross patterns, additionally, due to its algorithm of discrediting data for calculation, it give same value for the cluster pattern, whilst MIC gives significant estimates for wave and provides relatively significant indices for quadratic, cross and circle patterns.

Considering the performance of the novel mutual association measure proposed here in this chapter, it is worth to be noted that for the simulated linear group, the novel mutual association measure achieves solid and consistent indices with corresponding standard deviations extremely close to 0, which indicates that the results are almost precisely identical to the absolute values of corresponding generated population correlation coefficients. As a mutual association measure built on the eigenvalue-based distance, the brief concept of this measure is identifying the information shifted to the second eigenvalue of a matrix formed by a multivariate system. Therefore, the mutual association measure does not consider the direction of effect. Actually, in general, the direction of effect can easily be noticed by a simple time series diagram. Comparing to the other widely accepted association measures, the novel mutual association measure provides consistent and satisfying results for 1000 times of simulations with extremely low variations.

Regarding the performance of novel mutual association on nonlinear patterns, it is noticed that only trapezoid pattern can be detected with relatively significant statistics. It cannot provide more significant evidences for other nonlinear patterns, whilst considering the other empirical linear association measures, the novel mutual association measure is relatively more sensitive than Hoeffding's D Test, with fairly less significant results for quadratic and cross. It is also of note that in terms of the trapezoid pattern, it has not been significantly detected by any other listed measures except the novel mutual association method.

In general, according to the evaluations of all listed association measures, there is no measure that can well perform for detecting both linear and nonlinear relationships with also relatively accurate estimates. The highlight point for novel mutual association measure is that it does not get effected by noise and shows consistent and precise estimates for all linear simulations with very close to 0 variation, and it can detect the trapezoid pattern with significant estimates that all previously listed measure could not achieve.

Table 4.1 Evaluations of Association Measures by Simulated Groups of Series.

| | | **Mutual Association** | | | | **Pearson** | | | | **Spearman** | | | | **Kendall** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Population | 2.5% | mean | 97.5% | St.D | 2.5% | mean | 97.5% | St.D | 2.5% | mean | 97.5% | St.D | 2.5% | mean | 97.5% | St.D |
| **Linear** | 0.80 | 0.80 | 0.80 | 0.80 | 0 | 0.78 | 0.79 | 0.82 | 0.01 | 0.76 | 0.78 | 0.81 | 0.01 | 0.51 | 0.59 | 0.67 | 0.04 |
| | 0.60 | 0.60 | 0.60 | 0.60 | 0.00 | 0.56 | 0.60 | 0.64 | 0.02 | 0.54 | 0.58 | 0.62 | 0.02 | 0.30 | 0.41 | 0.51 | 0.05 |
| | 0.40 | 0.40 | 0.40 | 0.40 | 0.00 | 0.34 | 0.39 | 0.45 | 0.03 | 0.33 | 0.38 | 0.44 | 0.03 | 0.14 | 0.26 | 0.37 | 0.06 |
| | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | -0.06 | 0 | 0.06 | 0.03 | -0.06 | 0.00 | 0.06 | 0.03 | -0.14 | 0.00 | 0.12 | 0.07 |
| | -0.40 | 0.40 | 0.40 | 0.40 | 0.00 | -0.45 | -0.40 | -0.34 | 0.03 | -0.44 | -0.38 | -0.33 | 0.03 | -0.38 | -0.26 | -0.14 | 0.06 |
| | -0.60 | 0.60 | 0.60 | 0.60 | 0.00 | -0.64 | -0.60 | -0.56 | 0.02 | -0.62 | -0.58 | -0.54 | 0.02 | -0.51 | -0.41 | -0.30 | 0.06 |
| | -0.80 | 0.80 | 0.80 | 0.80 | 0.00 | -0.82 | -0.79 | -0.77 | 0.01 | -0.81 | -0.79 | -0.76 | 0.01 | -0.67 | -0.59 | -0.51 | 0.04 |
| **Non-linear** | wave | 0.01 | 0.07 | 0.16 | 0.04 | -0.06 | 0.00 | 0.06 | 0.03 | -0.06 | 0.00 | 0.07 | 0.03 | -0.16 | 0.00 | 0.17 | 0.08 |
| | trapezoid | 0.52 | 0.63 | 0.71 | 0.05 | -0.06 | 0.00 | 0.05 | 0.03 | -0.06 | 0.00 | 0.05 | 0.03 | -0.11 | 0.00 | 0.10 | 0.06 |
| | diamond | 0.01 | 0.05 | 0.14 | 0.04 | -0.04 | 0.00 | 0.04 | 0.02 | -0.05 | 0.00 | 0.05 | 0.02 | -0.11 | 0.00 | 0.12 | 0.06 |
| | quadratic | 0.01 | 0.12 | 0.33 | 0.09 | -0.07 | 0.00 | 0.08 | 0.04 | -0.08 | 0.00 | 0.08 | 0.04 | -0.20 | 0.00 | 0.19 | 0.11 |
| | cross | 0.01 | 0.12 | 0.33 | 0.09 | -0.09 | 0.00 | 0.08 | 0.05 | -0.08 | 0.00 | 0.07 | 0.04 | -0.21 | 0.00 | 0.23 | 0.12 |
| | circle | 0.01 | 0.08 | 0.23 | 0.06 | -0.04 | 0.00 | 0.04 | 0.02 | -0.03 | 0.00 | 0.03 | 0.01 | -0.14 | 0.00 | 0.13 | 0.07 |
| | cluster | 0.01 | 0.04 | 0.10 | 0.03 | -0.02 | 0.00 | 0.01 | 0.01 | -0.05 | 0.00 | 0.04 | 0.02 | -0.09 | 0.00 | 0.09 | 0.05 |

| | | **Hoeffding's D Test** | | | | **Distance Correlation** | | | | **Mutual Information** | | | | **MIC** | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Population | 2.5% | mean | 97.5% | St.D | 2.5% | mean | 97.5% | St.D | 2.5% | mean | 97.5% | St.D | 2.5% | mean | 97.5% | St.D |
| **Linear** | 0.80 | 0.23 | 0.26 | 0.28 | 0.02 | 0.73 | 0.75 | 0.78 | 0.01 | 0.30 | 0.39 | 0.48 | 0.05 | 0.47 | 0.51 | 0.56 | 0.03 |
| | 0.60 | 0.09 | 0.11 | 0.14 | 0.01 | 0.51 | 0.55 | 0.59 | 0.02 | 0.14 | 0.21 | 0.28 | 0.04 | 0.28 | 0.32 | 0.37 | 0.02 |
| | 0.40 | 0.03 | 0.04 | 0.06 | 0.01 | 0.31 | 0.36 | 0.42 | 0.03 | 0.06 | 0.11 | 0.17 | 0.03 | 0.18 | 0.21 | 0.24 | 0.02 |
| | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.04 | 0.06 | 0.08 | 0.01 | 0.01 | 0.04 | 0.09 | 0.02 | 0.12 | 0.13 | 0.15 | 0.01 |
| | -0.40 | 0.03 | 0.04 | 0.06 | 0.00 | 0.31 | 0.36 | 0.41 | 0.03 | 0.06 | 0.11 | 0.17 | 0.03 | 0.18 | 0.21 | 0.24 | 0.02 |
| | -0.60 | 0.09 | 0.11 | 0.14 | 0.01 | 0.51 | 0.55 | 0.60 | 0.02 | 0.13 | 0.21 | 0.29 | 0.04 | 0.28 | 0.32 | 0.36 | 0.02 |
| | -0.80 | 0.23 | 0.25 | 0.28 | 0.01 | 0.73 | 0.76 | 0.78 | 0.01 | 0.29 | 0.39 | 0.49 | 0.05 | 0.47 | 0.51 | 0.56 | 0.02 |
| **Non-linear** | wave | 0.01 | 0.01 | 0.02 | 0.00 | 0.33 | 0.40 | 0.48 | 0.04 | 0.22 | 0.37 | 0.51 | 0.08 | 0.96 | 0.99 | 1.00 | 0.01 |
| | trapezoid | 0.00 | 0.00 | 0.00 | 0.00 | 0.15 | 0.20 | 0.27 | 0.03 | 0.03 | 0.08 | 0.14 | 0.03 | 0.17 | 0.19 | 0.22 | 0.01 |
| | diamond | 0.00 | 0.00 | 0.00 | 0.00 | 0.15 | 0.20 | 0.28 | 0.03 | 0.06 | 0.12 | 0.19 | 0.04 | 0.13 | 0.15 | 0.17 | 0.01 |
| | quadratic | 0.09 | 0.10 | 0.11 | 0.00 | 0.46 | 0.51 | 0.56 | 0.03 | 0.71 | 0.78 | 0.86 | 0.04 | 0.64 | 0.69 | 0.74 | 0.03 |
| | cross | 0.04 | 0.05 | 0.05 | 0.00 | 0.30 | 0.36 | 0.45 | 0.04 | 0.59 | 0.76 | 0.91 | 0.09 | 0.55 | 0.57 | 0.58 | 0.01 |
| | circle | 0.03 | 0.04 | 0.05 | 0.00 | 0.12 | 0.17 | 0.26 | 0.04 | 0.02 | 0.04 | 0.07 | 0.02 | 0.55 | 0.56 | 0.57 | 0.01 |
| | cluster | 0.00 | 0.00 | 0.00 | 0.00 | 0.07 | 0.10 | 0.14 | 0.02 | 1.38 | 1.38 | 1.38 | 0 | 0.12 | 0.13 | 0.14 | 0.01 |

## 4.5 Evaluation by Applications

Considering the complexity of real data and the restricted nonlinear relationships simulations can offer for evaluation, here in this section two cases of real data are considered for further investigation and comparison. Note that all preconditions of each measure are satisfied respectively. It is also worth to be highlighted that no assumptions or models are made on data that are undertaking tests as the aim of this chapter is proposing a novel association measure and evaluating the performances by comparing to empirical linear and nonlinear association measures from the statistical data analysis point of view.

### 4.5.1 Oil Prices and Stock Markets

Economists have shown great interests to the investigations or analyses on relationship between stock market and oil prices for the recent decades due to the significant role of crude oil on impacting economy worldwide. Economists have extended the research on different representative variables or its transformations, like stock market returns and changes in oil prices (Kilian and Park, 2009), oil supply and demand shocks (Jung and Park, 2011), etc. Moreover, the research of relationship between stock market and oil prices has extended from focusing on the U.S. market itself to a world wide range of countries or regions.

The data employed here are the monthly stock indices and Brent crude oil prices (BRT) in dollars (per barrel) respectively[1]. The Brent crude oil prices data is available at FRED (2015) which is noted with the resource of U.S. Energy Information Administration. Both the oil-importing and oil-exporting countries are considered with corresponding stock market indices respectively (see Table 4.2). The selection of countries are based on the choices of relative literatures and the principle of be-

[1] The West Texas Intermediate (WTI) oil price index is also considered for all tests and the results are very close with minor differences under 0.01 level.

ing comparable within the group and the availability from the database. It contains 10 oil-importing countries including USA (DJIA), Japan (NIKKEI 225), Germany (DAX), France (CAC 40), UK (FTSE 100), Italy (FTSE MIB), China (SHCOM-P), Korea (KOSPI), India (BSE Sensex) and Netherlands (AEX General); and 9 oil-exporting countries are included, which are Saudi Arabia (Tadawul All Share), Kuwait (KWSEIDX), Mexico (MXICP 35), Noway (OSEAX), Russia (MICEX), Indonesia (JKSE), Brazil (Bovespa), Venezuela (IBVC) and Canada (S&P/TSX 60). All stock market indices are monthly data [2] from CEIC (2015) and note that the time ranges for each group of countries are also decided based on the principle of being comparable within the group and the availability from the database.

According to the performances of association measures in Table 4.2, three linear association measures (Pearson, Spearman and Kendall) generally provide diverse significant coefficients across countries under evaluation. The results of Pearson and Spearman are not always similar, while Kendall generally provides much less significant results. More specifically, oil-exporting countries show positive relationship in general, which is also applicable for oil-importing countries except Japan and Italy.

In terms of empirical nonlinear association measures, which do not consider direction of effect in general and focus on the level of association only, it is noticed that the results of Hoeffding' s D Test in general show the lowest levels of association comparing to other measures. The rest of the nonlinear measures are sensitive enough to identify the possible association regardless of countries with more significant results even comparing to general linear association measures. Additionally, the results of DisCorr and MIC are very similar that further indicate the existence of association between evaluated variables.

[2] Note that adopting monthly data instead of daily sequence (which is primarily used by firm-level analysis) aims to cover relatively longer time range across a number of countries for a general evaluation while maintaining sufficient frequency to reflect the fluctuations.

Table 4.2 Comparison of Association Measures on Analyses of Stock Market and Oil Prices Data.

| | **Stock Market and Oil Prices Data** | | | **Association Measures** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Country | Stock Index | Time Range | Pearson | Spearman | Kendall | HoefD | DisCorr | MI | MIC | **Mutual Association** |
| Oil-importing | USA | DJIA | 1987.06 - 2015.07 | 0.74*** | 0.79*** | 0.57*** | 0.29*** | 0.75 | 1.52 | 0.73 | **0.91** |
| | Japan | NIKKEI 225 | 1987.06 - 2015.07 | -0.49*** | -0.61*** | -0.39*** | 0.14*** | 0.53 | 1.38 | 0.54 | **0.63** |
| | Germany | DAX | 1987.06 - 2015.07 | 0.73*** | 0.74*** | 0.53*** | 0.25*** | 0.76 | 1.53 | 0.71 | **0.91** |
| | France | CAC 40 | 1988.01 - 2015.07 | 0.43*** | 0.54*** | 0.35*** | 0.12*** | 0.55 | 1.45 | 0.57 | **0.84** |
| | UK | FTSE 100 | 1987.06 - 2015.07 | 0.62*** | 0.61*** | 0.44*** | 0.17*** | 0.64 | 1.44 | 0.62 | **0.87** |
| | Italy | FTSE MIB | 2003.09 - 2015.07 | -0.42*** | -0.47*** | -0.31*** | 0.10*** | 0.51 | 1.78 | 0.45 | **0.85** |
| | China | SHCOMP | 1990.12 - 2015.07 | 0.65*** | 0.78*** | 0.53*** | 0.24*** | 0.75 | 1.59 | 0.68 | **0.89** |
| | Korea | KOSPI | 1987.06 - 2015.07 | 0.91*** | 0.73*** | 0.53*** | 0.26*** | 0.92 | 1.48 | 0.84 | **0.96** |
| | India | BSE Sensex | 1990.01 - 2015.07 | 0.86*** | 0.81*** | 0.60*** | 0.39*** | 0.93 | 1.57 | 0.94 | **0.94** |
| | Netherlands | AEX General | 2002.11 - 2015.07 | 0.11 | 0.12 | 0.07 | 0.02*** | 0.29 | 1.93 | 0.40 | **0.92** |
| Oil-exporting | Saudi Arabia | Tadawul All Share | 2007.01 - 2015.07 | 0.27*** | 0.22** | 0.16** | 0.03*** | 0.29 | 1.46 | 0.32 | **0.96** |
| | Kuwait | KWSEIDX | 2002.05 - 2015.07 | 0.39*** | 0.32*** | 0.23*** | 0.08*** | 0.47 | 1.44 | 0.64 | **0.92** |
| | Mexico | MXICP 35 | 1987.06 - 2015.07 | 0.93*** | 0.86*** | 0.67*** | 0.47*** | 0.95 | 1.72 | 0.94 | **0.96** |
| | Noway | OSEAX | 2002.05 - 2015.07 | 0.71*** | 0.69*** | 0.53*** | 0.29*** | 0.77 | 1.39 | 0.73 | **0.96** |
| | Russia | MICEX | 2002.05 - 2015.07 | 0.78*** | 0.70*** | 0.54*** | 0.27*** | 0.79 | 1.31 | 0.74 | **0.96** |
| | Indonesia | JKSE | 1987.06 - 2015.07 | 0.88*** | 0.80*** | 0.60*** | 0.36*** | 0.92 | 1.59 | 0.93 | **0.93** |
| | Brazil | Bovespa | 1987.06 - 2015.07 | 0.92*** | 0.87*** | 0.67*** | 0.48*** | 0.94 | 1.72 | 0.99 | **0.96** |
| | Venezuela | IBVC | 2007.01 - 2015.07 | 0.21** | 0.23** | 0.16** | 0.08*** | 0.31 | 1.67 | 0.87 | **0.33** |
| | Canada | S&P/TSX 60 | 2002.05 - 2015.07 | 0.76*** | 0.70*** | 0.54*** | 0.29*** | 0.78 | 1.31 | 0.76 | **0.96** |

*Note*: ***, ** and * refer to the significant level of 1%, 5% and 10% respectively (not applicable for DisCorr, MI, MIC and Mutual Association).

It is worth highlighting that the novel mutual association measure identifies strong association across countries which other measures cannot or only partly achieve, especially for the cases of Netherlands, Italy and France. For example, Netherlands shows the highest MI statistics and minimum coefficients for HoefD, DisCorr and MIC respectively. Additionally, it is the only country to have no significant association detected by all three linear association measures, whilst the mutual association measure detects 0.92 significance level of association, which cannot be successfully identified (or only be partly achieved) by any other nonlinear association.

In general, the mutual association measure is proved to be a reliable method that offers alternative approach of conducting association analysis between stock market and oil prices in a complex economics system environment. The attempt of gathering these association measures and conducting the evaluations can help in providing a broad view of understanding the possible association between stock market and oil prices from the statistically data oriented aspect. It can also be a significant help when considering the process of establishing a relatively more suitable model for relationship investigation and data prediction.

### 4.5.2 Oil Prices and Tourist Arrivals

The emerging concerns of oil price and its impacts on diverse aspects of economy have been studies by numerous researchers recent decades with well established scientific literatures (Hamilton, 1996). Among which, the relationship between oil price and tourism has drawn significant attentions. A critical review of the studies of tourism and oil can be found in the work of Becken (2011).

The data used here are at monthly frequency covering the period from January 1996 to December 2015 of both US and nine European countries, including Austria, Italy, Germany, Greece, Netherland, Portugal, Spain, Sweden, and the UK. In terms of the data, US tourist arrivals were obtained from the US Department of Commerce, National Travel & Tourism Office, while data of European countries were obtained from Eurostat. Data of oil prices containing both WTI crude oil spot price and Europe BRT spot price (note that both in the unit of dollars per barrel) were obtained from the EIA (2016). Fig. 4.1 shows the time series plot of the monthly oil prices, whilst, Fig. 4.2 presents the time series plots of the monthly tourist arrivals. It can be observed that the tourist arrivals for the ten countries considered clearly shows significant feature of cycle with possible existing trend.

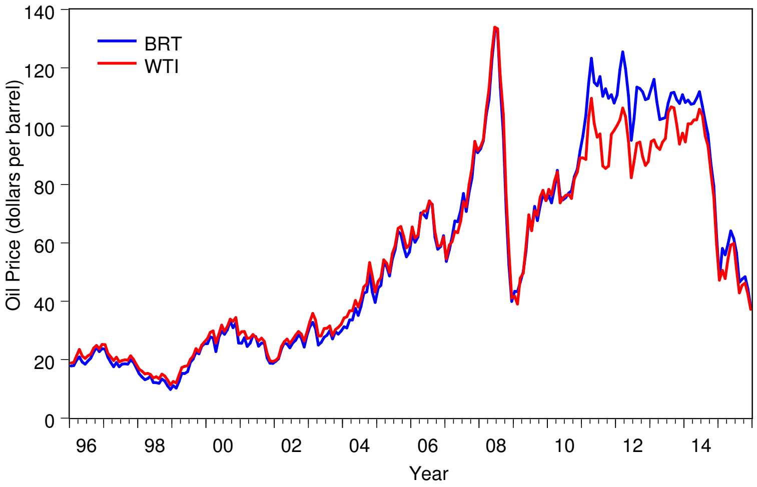


Fig. 4.1 Monthly Oil Price Data (BRT and WTI) from 1996 to 2015.

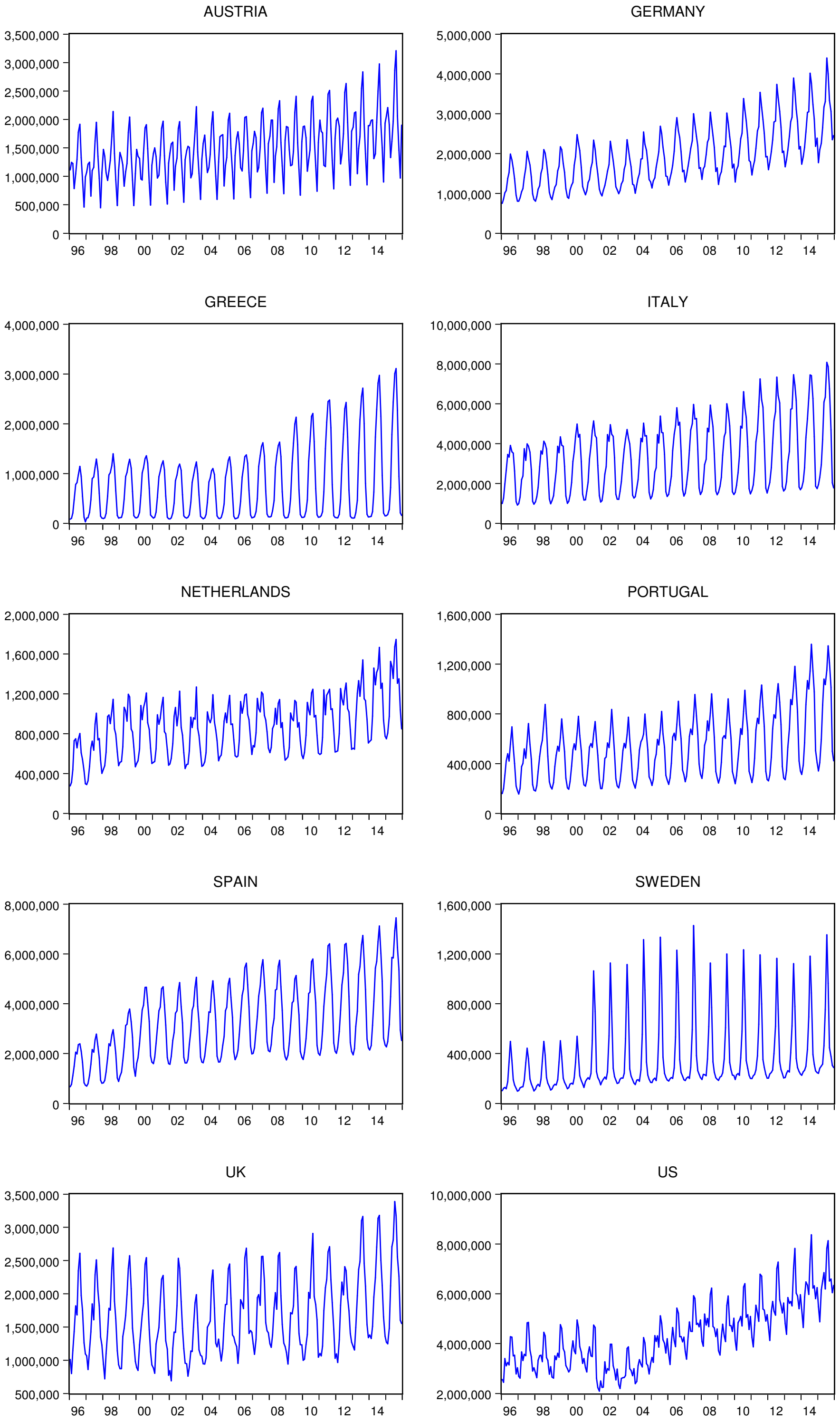


Fig. 4.2 Monthly Tourists Arrivals from 1996 to 2015 by Countries.

Table 4.3 summarizes the results of both empirical and novel association measures adopted. It is observed that the empirical methods show much less significant results whilst the novel measure achieves significant evidences regardless of the countries and type of oil price index.

Table 4.3 Comparison of Association Measures on Analyses of Oil Prices and Tourist Arrivals by Countries.

| **Country** | **Association Measures by Oil Prices** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **Mutual Association** | | **Pearson** | | **Spearman** | | **Kendall** | |
| | BRT | WTI | BRT | WTI | BRT | WTI | BRT | WTI |
| **Austria** | 0.869 | 0.882 | 0.384*** | 0.371*** | 0.394*** | 0.388*** | 0.271*** | 0.269*** |
| **Germany** | 0.905 | 0.916 | 0.596*** | 0.586*** | 0.640*** | 0.624*** | 0.447*** | 0.436*** |
| **Greece** | 0.720 | 0.728 | 0.270*** | 0.267*** | 0.223*** | 0.218*** | 0.149*** | 0.147*** |
| **Italy** | 0.839 | 0.852 | 0.344*** | 0.343*** | 0.357*** | 0.349*** | 0.249*** | 0.245*** |
| **Netherland** | 0.871 | 0.886 | 0.391*** | 0.387*** | 0.398*** | 0.388*** | 0.273*** | 0.269*** |
| **Portugal** | 0.853 | 0.866 | 0.382*** | 0.381*** | 0.383*** | 0.377*** | 0.264*** | 0.262*** |
| **Spain** | 0.874 | 0.887 | 0.474*** | 0.478*** | 0.511*** | 0.502*** | 0.358*** | 0.354*** |
| **Sweden** | 0.750 | 0.763 | 0.268*** | 0.278*** | 0.474*** | 0.464*** | 0.332*** | 0.324*** |
| **UK** | 0.851 | 0.868 | 0.286*** | 0.287*** | 0.275*** | 0.268*** | 0.188*** | 0.183*** |
| **US** | 0.919 | 0.929 | 0.699*** | 0.681*** | 0.728*** | 0.716*** | 0.511*** | 0.501*** |
| | **HoefD** | | **DisCorr** | | **MI** | | **MIC** | |
| | BRT | WTI | BRT | WTI | BRT | WTI | BRT | WTI |
| **Austria** | 0.053*** | 0.052*** | 0.382 | 0.371 | 0.156 | 0.146 | 0.329 | 0.281 |
| **Germany** | 0.141*** | 0.132*** | 0.601 | 0.589 | 0.328 | 0.314 | 0.459 | 0.462 |
| **Greece** | 0.017*** | 0.016*** | 0.285 | 0.277 | 0.145 | 0.136 | 0.341 | 0.324 |
| **Italy** | 0.043*** | 0.040*** | 0.332 | 0.327 | 0.219 | 0.238 | 0.388 | 0.384 |
| **Netherland** | 0.048*** | 0.046*** | 0.375 | 0.367 | 0.126 | 0.139 | 0.324 | 0.330 |
| **Portugal** | 0.049*** | 0.047*** | 0.379 | 0.371 | 0.142 | 0.147 | 0.328 | 0.331 |
| **Spain** | 0.086*** | 0.083*** | 0.447 | 0.445 | 0.221 | 0.224 | 0.429 | 0.428 |
| **Sweden** | 0.087*** | 0.082*** | 0.310 | 0.311 | 0.257 | 0.259 | 0.463 | 0.463 |
| **UK** | 0.023*** | 0.021*** | 0.272 | 0.265 | 0.072 | 0.065 | 0.278 | 0.286 |
| **US** | 0.219*** | 0.203*** | 0.737 | 0.721 | 0.426 | 0.429 | 0.604 | 0.608 |

**Note:** *,**,*** indicate the significance of 1%,5%,10% respectively.

Specifically, the results are very similar between BRT and WTI, also the coefficients of Pearson and Spearman show close levels of association. Kendall correlation, Distance Correlation, MI and MIC in general reflect similar levels of significance, whilst Hoeffding's D Test has generally the lowest level of sensitivity across countries and types of oil price index. The novel mutual association measure, on the other hand, can detect the possible association and provide consistent and significant results for all countries considered. It greatly outperforms the empirical methods and indicates the significant mutual association between tourist arrivals and oil price with evidences of nine European countries and US. More specifically, US shows the highest level of mutual association, followed by Germany, Spain, Netherland, Austria, UK, Portugal, Italy, Sweden, then Greece in descending order. The initially adopted novel mutual association measure successfully proves the advantage on nonlinear association detection in complex system like oil-tourism studies. It is sensitive enough to confirm the crucial relationship between the tourist arrivals and oil prices by relatively less amount of data so to contribute to the existing literature regarding the association study of the complex economical systems.

## 4.6 Discussion

Considering the crucial importance of association study in better understanding multivariate systems across various disciplines, this chapter proposed a novel mutual association measure by combing the subspace-based techniques. The performance of the novel association measure was evaluated with comparisons by simulations as well as the cases of real data. The performances achieved are significantly promising and it has to be highlighted that it has valuable potentials on nonlinear association studies in complex systems across numerous subjects.

This research is the first attempt of incorporating eigenvalue-based distance concept with a multivariate system into the development of an alternative or better performed association measure. The mutual association measure currently may not master on identifying all simulated nonlinear patterns, whilst it gives highest significant reaction for trapezoid nonlinear pattern without losing the ability on linear association detection that other measures cannot achieve. Considering the applications on real data cases, it is evidenced that the novelly developed association measure is a reliable, sensitive, assumption free approach that can outperform or at least being alternative method comparing to the empirical measures in the study of a complex economical system. Moreover, considering the limited nonlinear patterns that the simulation covers and the complexity of real data that researchers frequently encounter, it is possible that the novel mutual association measure has not been able to fully present its advantages on complex data, large size of data, data with complex noise, etc., which it inherits from the advanced subspace-based techniques.

In general, researchers never stop on perusing the better solutions. There also should not be a restriction of one specific association measure due to the fact that the advantages of different measures vary significantly. The variety of measurements with their own strength also offers more options for the association analysis of random groups of series in a complex system like economics and social science. This research succeeds satisfying evidences from both simulations and real cases that this novel method can identify complex associations which empirical methods may fail to detect. The advantage of this method is that it does not restrict on the domain of either linearity or nonlinearity, but consider the associated information of the series as a whole. Additionally, the calculations are very efficient and convenient even for large size data computation. However, this chapter is still considered a temporary summary about the beginning of this development. There will be many possibilities for further improvements as the next stage of this particular study, for example, expanding the nonlinear associations or combinations of linearity and nonlinearity for simulated evaluation, developing the direction of association into the current index outcome, decomposing the data into representative components for comprehensive association measure by specific elements, etc.

# Chapter 5

# SSA Causality Test based on Forecasting Performance

Following the research philosophy in Chapter 1, section 1.2.1, the "Efficient Cause" by Aristotle is briefly claimed as the "primary source of the change" or simply as "initiator of the movement". Like the nature of "Efficient Cause", a question that frequently arise in time series analysis is whether one variable can help in analyzing and predicting another variable. In this chapter an innovative modification is proposed on the currently well established causality analysis approach which is primarily based on linear models. Specifically, the SSA forecasting performance is adopted to form the frame of SSA causality test so to address and answer the question that whether one variable can be helpful to analyse or forecast the others.

This chapter is organized as follows: Section 1 presents the theoretical formulation of SSA causality test; a brief review of benchmark empirical methods is summarized in Section 2; Section 3 evaluates and compares the performances by real data application; Finally, the discussion is concluded in Section 4.

## 5.1 Theoretical Formulation

Granger (1969) formalized the causality concept and claimed causality if the elimination of one variable from a system is harmful for explaining the other variable. Similarity, in terms of this research that forms causality analysis based on SSA forecasting performance, the criterion is defined as the improvement of out-of-sample forecasting by multivariate system comparing to univariate scenario. Specifically, the causality analysis is obtained by comparing the forecast values obtained by the univariate procedure–SSA and multivariate process–MSSA (see Fig. 5.1). Consequently, if the forecasting errors using MSSA are significantly smaller than those of univariate SSA, it is concluded that there is a causal relationship detected between these series. As a nonparametric technique, the SSA causality test is able to capture possible nonlinearities using a data-driven approach without specifying any known functional nonlinear model to the relationship, which in turn, could be incorrectly specified in the first place. Detailed introduction is presented below which mainly follows Hassani et al. (2010c). Note that the theoretical introduction of SSA, MSSA and corresponding forecasting algorithms can be found in Chapter 2, section 2.2 and Appendix A.

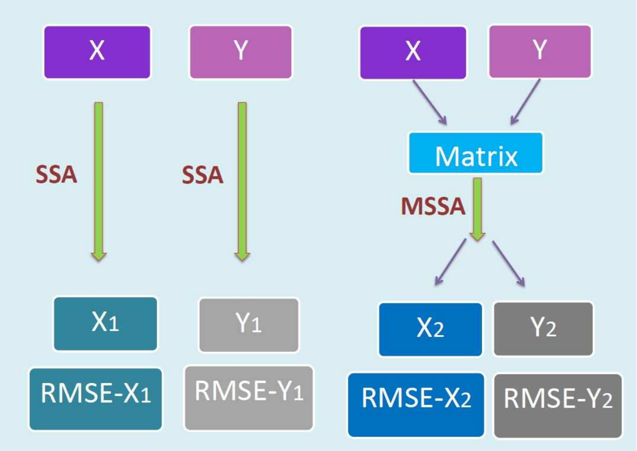


Fig. 5.1 Flowchart of Cause Detection based on SSA Forecasting Accuracy.

Let us consider the procedure for constructing vectors of forecasting error for out-of-sample tests in a two variable case $X_N$ and $Y_N$ by both univariate and multivariate SSA techniques respectively. Firstly, the series $X_N = (x_1, ..., x_N)$ is divided into two separate subseries $X_R$ and $X_F$ that satisfy $X_N = (X_R, X_F)$, where $X_R = (x_1, ..., x_R)$ and $X_F = (x_{R+1}, ..., x_N)$. Same procedure is also conducted for $Y_N$. The subseries $X_R$ and $Y_R$ are used in the reconstruction step to provide the noise-free series $\tilde{X}_R$ and $\tilde{Y}_R$. The noise-free series are then used for forecasting the subseries $X_F$ and $Y_F$ with the help of the forecasting algorithms (see Appendix A) of SSA and MSSA respectively. For variable $X_N$, two different forecasting values of $\hat{X}_F = (\hat{x}_{R+1}, ..., \hat{x}_N)$ by SSA and MSSA are then used for computing the forecasting errors accordingly, which will be the same process in terms of variable $Y_N$. Therefore, in a multivariate system like this, the vectors of forecasts obtained can be used in computing the forecasting accuracy and therefore conducting the causality analysis between the two variables.

The length of out-of-sample does not have specific limitation, generally considering the simulation scenario, the length of time series for reconstruction will take 2/3 of the whole series and the rest 1/3 is considered as out-of-sample for constructing forecasting error. The separate point to define the out-of-sample size for different series can be chosen respectively, whilst it is important that when it goes to comparing the performances of different techniques based on constructed forecasting error of one specific series, the sizes of reconstruction and out-of-sample for all techniques should be identical. In addition, the choices of window length $L$ and the referring options of numbers of eigenvalues $r$ should also be carefully evaluated in practice of SSA causality test respectively. Considering this as the first attempt of application, also in order to conduct the most accurate results, all the possibilities of $L$ and its referring choices of $r$ should be applied for both univariate SSA and MSSA processes, then the optimal ones with best performance of forecasting will be chosen to construct the finally cause detection procedure.

Consequently, define the criterion $F_{X|Y} = \Delta X_F|Y/\Delta X_F$ (where $\Delta X_F$ and $\Delta X_F|Y$ indicate the RMSE of out-of-sample forecasting $X_F$ without and with $Y$) corresponding to the forecast of the series $X_N$ in the presence of the series $Y_N$. Specifically, if $F_{X|Y}$ is small, then having information obtained from the series $Y$ can help to achieve better forecasts of the series $X$. If $F_{X|Y} < 1$, it is concluded that the information provided by the series $Y$ can be regarded as useful or supportive for forecasting the series $X$. Alternatively, if the values of $F_{X|Y} \geq 1$, then either there is no detectable causality between $X$ and $Y$ or the performance of the univariate SSA is better than of the MSSA (this may happen, for example, when the series $Y$ has structural breaks misdirecting the forecasts of $X$).

## 5.2 Benchmark Empirical Methods

### 5.2.1 Time Domain Granger Causality Test

GC (Granger, 1969) is the most widely accepted approach for causality analysis and is extensively used in a number of disciplines[1]. The regression formulation of GC states that vector $X_i$ is the cause of vector $Y_i$ if the past values of $X_i$ are helpful in predicting the future value of $Y_i$, two regressions are considered as follows:

$$\begin{aligned} Y_i &= \sum_{t=1}^{T} \alpha_t Y_{i-t} + \varepsilon_{1i}, && (5.1) \\ Y_i &= \sum_{t=1}^{T} \alpha_t Y_{i-t} + \sum_{t=1}^{T} \beta_t X_{i-t} + \varepsilon_{2i}, && (5.2) \end{aligned}$$

where $i = 1, 2, \cdots, N$ ($N$ is the number of observations), $T$ is the maximal time lag, $\alpha$ and $\beta$ are vectors of coefficients, $\varepsilon$ is the error term. The first regression is the model that predicts $Y_i$ by using the history of $Y_i$ only, while the second regression

[1] Note that a brief review of its development process can be found in Appendix C for your reference.

represents the model of $Y_i$ is predicted by the past information of both $X_i$ and $Y_i$. Therefore, if the second model is a significantly better model than the first one, existence of causality is concluded.

### 5.2.2 Frequency Domain Causality Test

The frequency domain causality test is the extension of time domain GC test that identifies the causality between different variables for each frequency. It is firstly proposed by Geweke (1982), and it permits to investigate causality dynamics at different frequencies rather than relying on a single statistics as is in the case with the conventional time domain analysis (Ciner, 2011). Breitung and Candelon (2006) improved this approach by calculating the GC for each individual frequency component separately instead of computing a single GC measure for the entire relationship, which make it possible to determine whether the predictive power is concentrated at the quickly fluctuating components or at the slowly fluctuating components (Croux and Reusens, 2013). A brief introduction of this test is summarized below mainly following the work of Ciner (2011); Geweke (1982).

It is assumed that two dimensional vector containing $X_i$ and $Y_i$ (where $i = 1, 2, \cdots, N$ and $N$ is the number of observations) with a finite-order Vector Autoregression Model (VAR) representative of order $p$,

$$\Theta(R)\begin{pmatrix} Y_i \\ X_i \end{pmatrix} = \begin{pmatrix} \Theta_{11}(R) & \Theta_{12}(R) \\ \Theta_{21}(R) & \Theta_{22}(R) \end{pmatrix}\begin{pmatrix} Y_i \\ X_i \end{pmatrix} + \mathscr{E}_i, \tag{5.3}$$

where $\Theta(R) = I - \Theta_1 R - \ldots - \Theta_p R_p$ is a $2 \times 2$ lag polynomial and $\Theta_1, \ldots, \Theta_p$ are $2 \times 2$ autoregressive parameter matrices, with $R^k X_i = X_{i-k}$ and $R^k Y_i = Y_{i-k}$. The error vector $\mathscr{E}$ is white noise with zero mean, and $E(\mathscr{E}_i \mathscr{E}_i') = \mathbf{Z}$, where $\mathbf{Z}$ is positive definite matrix. The moving average (MA) representative of the system is

$$\begin{pmatrix} Y_i \\ X_i \end{pmatrix} = \Psi(R)\eta_i = \begin{pmatrix} \Psi_{11}(R) & \Psi_{12}(R) \\ \Psi_{21}(R) & \Psi_{22}(R) \end{pmatrix} \begin{pmatrix} \eta_{1i} \\ \eta_{2i} \end{pmatrix}, \tag{5.4}$$

with $\Psi(R) = \Theta(R)^{-1}\mathbf{G}^{-1}$ and $\mathbf{G}$ is the lower triangular matrix of the Cholesky decomposition $\mathbf{G}'\mathbf{G} = \mathbf{Z}^{-1}$, such that $E(\eta_t\eta_t') = I$ and $\eta_i = \mathbf{G}\mathscr{E}_i$. The causality test developed by Geweke (1982) can be written as:

$$C_{X\Rightarrow Y}(\gamma) = log\left[1 + \frac{|\Psi_{12}(e^{-i\gamma})|^2}{|\Psi_{11}(e^{-i\gamma})|^2}\right]. \tag{5.5}$$

However, according to this framework, no Granger causality from $X_i$ to $Y_i$ at frequency $\gamma$ corresponds to the condition $|\Psi_{12}(e^{-i\gamma})| = 0$, this condition leads to

$$|\Theta_{12}(e^{-i\gamma})| = |\Sigma_{k=1}^{p}\Theta_{k,12}\cos(k\gamma) - i\Sigma_{k=1}^{p}\Theta_{k,12}\sin(k\gamma)| = 0, \tag{5.6}$$

where $\Theta_{k,1,2}$ is the $(1,2)th$ element of $\Theta_k$, such that a sufficient set of conditions for no causality is given by Breitung and Candelon (2006) as follows:

$$\begin{array}{c} \Sigma_{k=1}^{p}\Theta_{k,1,2}\cos(k\gamma) = 0 \\ \Sigma_{k=1}^{p}\Theta_{k,1,2}\sin(k\gamma) = 0 \end{array}. \tag{5.7}$$

Hence, the null hypothesis of no causality at frequency $\gamma$ can be tested by using a standard F-test for the linear restrictions in (5.7), which follows an $F(2, B-2p)$ distribution, for every $\gamma$ between 0 and $\pi$, with $B$ begin the number of observations in the series.

### 5.2.3 Limitations of Granger Causality

In spite of the tremendous amount of implementations of GC on a broad range of subjects, it was mainly developed for linear stochastic systems. The limitation is significant considering the fact that not all relationships can be explained by linear based models. Even the revolutions of nonlinear applicability have been continuously exploited, there are still just limited types of nonlinearity can be so far included for analysis. Nevertheless, it still builds on the assumption of a particular type of relationship that is even fixed before the actual studies start.

Moreover, it is questionable to build a model based on a few selected factors so to expect a satisfying explanation of a phenomenon in a system that is actually far more complex. It again made the assumption of including only these key influential factors based on particular judgements. The way to form this model may lead to outcomes that vary significantly, which can be considered an important limitation that affects the level of reliability.

Another limitation is that by eliminating this variable out of the model, which will be used to determine causality based on whether it is harmful for the analysis or forecasting of the others, GC assumes that this particular variable is separable from this system. Therefore, any possibly existing underlying causality will affect the outcome of the analysis. Not to mention that the issue of separability is often not satisfied especially in deterministic dynamical system (Sugihara et al., 2012).

## 5.3 Evaluation by Real Data Application

### 5.3.1 Global Temperature and Sunspot Number

Global warming has been one of the most crucial subjects of research that has both short and long term environmental and economic implications. As a result, there is growing interest among scientists worldwide to identify the factors that affect the rate of change in global temperature, as it not only allows us to predict global warming, but also takes measures to control it. The connection between solar activity and global warming has been well established in the scientific literature (for example, see (Eichler et al., 2009; Ineson et al., 2011; Lean and Rind, 1998; Lockwood, 2012; Scafetta, 2009, 2014; Scafetta and West, 2005)).

An indication of solar activity is given by the sunspot number (SS), which appear as dark spots on the surface of the Sun. Temperatures in the dark centers of sunspots drop to about 3700 K (compared to 5700 K for the surrounding photosphere). They are magnetic regions on the Sun, with the strength of a magnetic field which is thousands of times stronger than the Earth's magnetic field. Sunspots typically last for several days, although very large ones may live for several weeks[2]. The causality between SS and global temperature (GT) has been explored in many scientific work using different causality detection techniques. A recent paper of Gupta et al. (2015) analyzed whether sunspot numbers cause global temperatures based on monthly data covering the period 1880:1-2013:9. The authors find that standard time domain GC test fails whilst the frequency domain causality test detects emergence of causality running from SS to GT only recently.

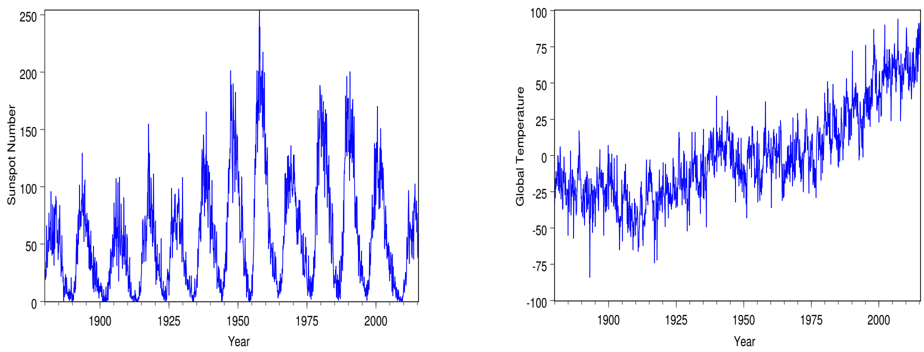
Since the data of SS and GT contain many complex dynamic fluctuations, also there is a high possibility of the existence of non-stationary features, this poses difficulty in deriving convincing results on causality using parametric techniques. However, it makes a great application to evaluate the performance of the novel SSA causality test[3].

[2] Further details can be found at: http://solarscience.msfc.nasa.gov/feature1.shtml.

[3] A cautionary note that the Earth's climate is regulated by anthropogenic emissions like CO2, volcanoes and other greenhouse gases, which need to be factored in as well to properly identify the contribution of solar activity (Scafetta, 2014). Ignoring these issues could also lead to spurious, in other words, more significant influence from SS on GT. However, data on CO2 emissions were only available at annual frequency, in the case here in this research, the objective was extending the work of Gupta et al. (2015) over the up to date sample period to evaluate the performance of the reduced form, data driven, newly proposed technique. In addition, while this is only analyzing causality

### 5.3.2 Data and Unit Root Test

The GT and SS data are at monthly frequency covering the period from January 1880 to May 2015, with the start and up to date end points being updated based on the paper of Gupta et al. (2015). The data for GT were obtained from the Goddard Institute for Space Studies (GISS) (GISS, 2015) and the SS data were obtained from the Solar Influences Data Analysis Centre (SIDC) (SIDC, 2015). Fig. 5.2 presents the time series plots of the original variables, in which it can be observed that the existence of the possible trend in GT which is further addressed with details in Chapter 7[4].



(a) Sunspot Number (SS)

(b) Global Temperature (GT)

Fig. 5.2 Monthly SS and GT from 1880M01 to 2015M05.

By following the work of Gupta et al. (2015) with updated data, different unit root tests were conducted to verify the stationarity of the series in Table 5.1. Additionally, structural breaks were detected in the full sample at 1936M03 and 1986M12 by the test proposed by Bai and Perron (2003), whereby the break test was applied to the GT equation of the VAR comprising GT and SS. The test will

and not correlation between SS and GT, the evidence of causality between SS and GT should not be associated with positive correlation between these two variables. The sign of this relationship is beyond the scope of this application.

[4] Note that the trend and its effects on causality analysis is comprehensively studied in Chapter 7.

not be reproduced here since it is not the key focus of this research (please refer to the paper of Gupta et al. (2015) for more details). Note that all the tests included in this and the following sections/chapters will apply for the full sample as well as all sub-samples for comparison.

Based on the results from the Kwiatkowski-Phillips-Schmidt-Shin (KPSS), augmented Dickey-Fuller (ADF), Dickey-Fuller test with Generalised Least Squares detrended residuals (DF-GLS), Phillips and Perron (PP), and Ng and Perron (NP) unit root tests, the null of a unit root is overwhelmingly rejected (except for KPSS test the null of being stationary, it cannot be overwhelmingly rejected), for the total sample of SS. However, for total sample of GT, while all the tests support that the variable is trend-stationary, the ADF and DF-GLS test tends to suggest non-stationarity of the series when the unit root test-equation has only a constant (or neither a constant and trend in case of the ADF test). The PP and the NP tests, though, indicate stationarity even under the assumption of constant only (and neither a constant and trend in case of the PP test). Given the nature of GT, it is evident that the unit root equation should in fact include a trend. In general, for sub-sample A and sub-sample B, overwhelming evidences of stationary have been obtained (especially based on the results of NP test, which have stronger power compared to the other tests (Gupta et al., 2015)). For sample C, while GT is found to be stationary in general at 1% level, the evidence of stationarity, is slightly weaker for SS, barring the PP and NP tests, at 5% level of significance. In summary, for the full sample and all sub-samples, it can be concluded that both series are stationary, whilst GT in general is trend-stationary, especially for sub-sample C.

Table 5.1 Unit Root Test Results of Original Data of SS and GT.

| Sample Size | Series | Methods | None | | Intercept | | Intercept and Trend | |
|---|---|---|---|---|---|---|---|---|
| | | | Level | Decision | Level | Decision | Level | Decision |
| Total Sample<br>(1625 Obs)<br>1880:1-2015:5 | GT | KPSS | ——— | ——— | 4.234*** (31) | I(1) | 0.686***(30) | I(1) |
| | | ADF | -1.301 (17) | I(1) | -1.296 (17) | I(1) | -3.443** (24) | I(0) |
| | | PP | -5.966*** (12) | I(0) | -5.964*** (12) | I(0) | -18.499*** (23) | I(0) |
| | | DF-GLS | ——— | ——— | -1.315 (6) | I(1) | -6.639***(3) | I(0) |
| | | NP | ——— | ——— | -31.142***(12) | I(0) | -516.032*** (23) | I(0) |
| | SS | KPSS | ——— | ——— | 0.464**(15) | I(1) | 0.101 (15) | I(0) |
| | | ADF | -2.499**(3) | I(0) | -4.055***(3) | I(0) | -4.109***(3) | I(0) |
| | | PP | -2.526** (14) | I(0) | -3.189** (12) | I(0) | -4.027*** (0) | I(0) |
| | | DF-GLS | ——— | ——— | -3.383***(3) | I(0) | -6.879***(3) | I(0) |
| | | NP | ——— | ——— | -47.323***(14) | I(0) | -71.585***(14) | I(0) |
| Sub-sample A<br>(674 Obs)<br>1880:1-1936:2 | GT | KPSS | ——— | ——— | 0.455*(19) | I(1) | 0.430***(19) | I(0) |
| | | ADF | -2.710***(3) | I(0) | -7.207***(2) | I(0) | -7.228***(2) | I(0) |
| | | PP | -4.313***(2) | I(0) | -13.397***(14) | I(0) | -13.424***(14) | I(0) |
| | | DF-GLS | ——— | ——— | -6.325***(2) | I(0) | -7.076***(2) | I(0) |
| | | NP | ——— | ——— | -234.149***(14) | I(0) | -275.304***(14) | I(0) |
| | SS | KPSS | ——— | ——— | 0.053(21) | I(0) | 0.051(21) | I(0) |
| | | ADF | -1.819*(3) | I(1) | -3.451***(3) | I(0) | -3.447**(3) | I(0) |
| | | PP | -3.226***(18) | I(0) | -6.075***(8) | I(0) | -6.075***(8) | I(0) |
| | | DF-GLS | ——— | ——— | -3.043***(3) | I(0) | -3.322**(3) | I(0) |
| | | NP | ——— | ——— | -52.499***(8) | I(0) | -57.985***(8) | I(0) |
| Sub-sample B<br>(609 Obs)<br>1936:3-1986:11 | GT | KPSS | ——— | ——— | 0.794***(17) | I(1) | 0.321***(16) | I(1) |
| | | ADF | -7.121***(1) | I(0) | -7.211***(1) | I(0) | -7.515***(1) | I(0) |
| | | PP | -12.979***(13) | I(0) | -13.102***(13) | I(0) | -13.678***(13) | I(0) |
| | | DF-GLS | ——— | ——— | -3.287***(2) | I(0) | -6.454***(1) | I(0) |
| | | NP | ——— | ——— | -92.270***(13) | I(0) | -229.775***(13) | I(0) |
| | SS | KPSS | ——— | ——— | 0.061(18) | I(0) | 0.052(18) | I(0) |
| | | ADF | -1.690*(2) | I(1) | -2.720*(2) | I(1) | -2.741(2) | I(1) |
| | | PP | -1.932*(11) | I(1) | -3.600***(2) | I(0) | -3.614**(2) | I(0) |
| | | DF-GLS | ——— | ——— | -2.718***(2) | I(0) | -2.754*(2) | I(1) |
| | | NP | ——— | ——— | -24.056***(2) | I(0) | -24.089***(2) | I(0) |
| Sub-sample C<br>(342 Obs)<br>1986:12-2015:5 | GT | KPSS | ——— | ——— | 1.835*** (14) | I(1) | 0.083 (13) | I(0) |
| | | ADF | -0.475 (3) | I(1) | -3.410**(3) | I(0) | -7.064***(1) | I(0) |
| | | PP | -1.063 (22) | I(1) | -6.518***(8) | I(0) | -10.546***(9) | I(0) |
| | | DF-GLS | ——— | ——— | -0.899 (3) | I(1) | -5.899***(1) | I(0) |
| | | NP | ——— | ——— | -15.362***(8) | I(0) | -121.714***(9) | I(0) |
| | SS | KPSS | ——— | ——— | 0.464**(15) | I(1) | 0.101 (15) | I(0) |
| | | ADF | -0.960 (3) | I(1) | -1.870 (3) | I(1) | -2.229 (3) | I(1) |
| | | PP | -1.526(14) | I(1) | -2.898**(2) | I(0) | -4.027***(0) | I(0) |
| | | DF-GLS | ——— | ——— | -1.174(3) | I(1) | -1.427(3) | I(1) |
| | | NP | ——— | ——— | -8.847**(2) | I(0) | -11.959 (1) | I(1) |

[a] The *, ** and *** indicate significance at the 10%, 5% and 1% respectively.

[b] The critical values are as follows:(1)None: -2.566, -1.941 and -1.616 for ADF and PP at 1%, 5% and 10% level of significance, respectively; (2)Intercept: -3.434, -2.863 and -2.567 (-2.566, 1.941, 1.617) [-13.8, -8.1 and -5.7] {0.739, 0.463, 0.347} for ADF and PP (DF-GLS) [NP] {KPSS} at 1%, 5% and 10% level of significance, respectively;(3)Intercept and Trend: -3.963, -3.412 and -3.128 (3.48, 2.89, 2.57) [-23.80, -17.3 and -14.2] {0.216, 0.146, 0.119} for ADF and PP (DF-GLS) [NP] {KPSS} at 1%, 5% and 10% level of significance respectively.

[c] Numbers in parentheses for ADF, PP and DF-GLS tests indicates lag-lengths selected based on the Schwarz Information Criterion (SIC). For the NP test and the KPSS test, based on the Bartlett kernel spectral estimation method, the corresponding numbers are the Newey-West bandwidth.

### 5.3.3 Time Domain Causality Test Results

Given the significant and empirical role of time domain GC test, here the corresponding tests are conducted for total sample as well as three sub-samples. Note that all tests satisfy the preconditions of time domain GC test with results by the optimal lag respectively. According to the results in Table 5.2, the null hypothesis that SS does not Granger cause GT cannot be rejected for both full and sub-samples, which confirm the statements by Gupta et al. (2015). The full sample causality cannot be relied upon due to structural breaks, as GC test assumes constancy of parameters during the sub-sample, which is of course not the case with structural breaks. In summary, the empirical time domain GC test fails to identify any causal links between SS and GT despite the significant connections evident by literature[5].

Table 5.2 Time Domain GC Test Results of Original SS and GT.

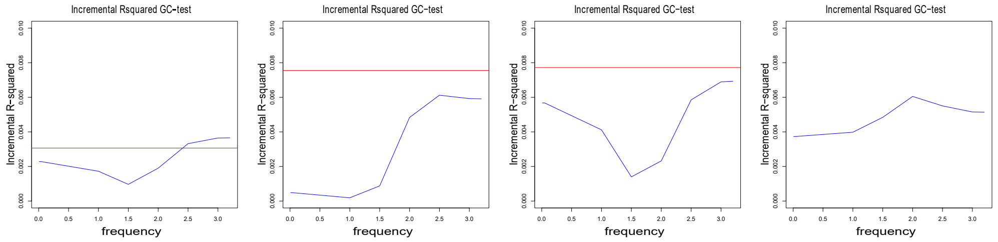
| Sample and Number of Observations | Total sample (1625 Obs) | | Subsample A (674 Obs) | | Subsample B (609 Obs) | | Subsample C (342 Obs) | |
|---|---|---|---|---|---|---|---|---|
| Referring periods | 1880:1-2015:5 | | 1880:1-1936:2 | | 1936:3-1986:11 | | 1986:12-2015:5 | |
| Causality direction | SS→ GT | | SS→ GT | | SS→ GT | | SS→ GT | |
| | $F$ | $p$-value | $F$ | $p$-value | $F$ | $p$-value | $F$ | $p$-value |
| Original series | 1.011 | 0.364 | 0.947 | 0.388 | 1.137 | 0.321 | 1.587 | 0.206 |

### 5.3.4 Frequency Domain Causality Test Results

The frequency domain causality results for the original series of SS and GT are listed below in Fig. 5.3. Note that the optimal lag-structures are maintained for all tests and the test statistics (blue) along with the corresponding 5% critical values (red) for each particular frequencies are adopted to evaluate the possible causal links from SS to GT. Therefore, when the test statistics (blue) is above or very close to the 5% critical value (red), the causality is detected for that particular (range of)

[5] Note that given the weak evidence of stationarity for SS for sub-sample C, the GC test is repeated with first differences of SS and GT. The null of non-causality still continued to hold with $p$-value of 0.728. In order to be robust to the non-normal errors of Holmes and Hutton (1990), the nonparametric rank GC test is also considered with the null of non-causality cannot be rejected.

frequency. The horizontal axis gives the parameter $\omega$ to calculate the corresponding frequency $f$ by $f = 2\pi/\omega$.



(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 5.3 Frequency Domain Causality Result for Original SS and GT.

For the full sample, significant causal link is confirmed for frequency that is greater than 2.45 corresponding to a cycle length between 2 and 2.6 months. Whilst, in terms of the sub-samples, no significant causality can be identified for any frequency and the frequency domain test fails to prove that SS has any significant causal effects on GT in the sub-samples.

### 5.3.5 SSA Causality Test Results

Against this background of lack of evidence of causality in the time and frequency domains, it is now focused on identifying the causality by the SSA-based approach. As previously mentioned in section 5.1, in order to conduct the SSA causality test for the SS and GT data, the out-of-sample size for each sub-sample series is 1/3 of the whole series. In addition, before the last step which determines causality by causality criterion $F_{GT|SS}$ as clarified in section 5.1, all the forecasting results of both SSA and MSSA steps are the optimal choice chosen respectively after considering all the possibilities of window length $L$ and its corresponding choices of number of eigenvalues $r$.

Table 5.3 summarizes the results of SSA causality test. As what is mentioned in section 5.1, if the causality criterion $F_{GT|SS} \geq 1$, then either there is no detectable causality between GT and SS or the performance of the univariate SSA is better than of the MSSA, this may happen, for example, when one of the series has structural breaks misdirecting the forecasts; If $F_{GT|SS} < 1$, then it is concluded that the information provided by SS can be regarded as useful or supportive for forecasting GT.

Table 5.3 SSA Causality Test Results of Original SS and GT.

| Sample and Number of Observations | Total sample (1625 Obs) | Subsample A (674 Obs) | Subsample B (609 Obs) | Subsample C (342 Obs) |
|---|---|---|---|---|
| Referring periods | 1880:1-2015:5 | 1880:1-1936:2 | 1936:3-1986:11 | 1986:12-2015:5 |
| Test statistics | $F_{GT\|SS}$ | $F_{GT\|SS}$ | $F_{GT\|SS}$ | $F_{GT\|SS}$ |
| Original Series | 0.998 | 0.284 | 0.399 | 0.308 |

According to the results in Table 5.3, when the whole sample is considered, the test statistics is very close to 1 and could not provide strong information to determine the causality between GT and SS. This is possibly affected by the structural breaks that were detected in GT, which misleads the forecasts. Comparing with the empirical evidence of Gupta et al. (2015), whereby the authors detected causality only in for the full-sample, the SSA causality test provides strong evidence of causality for all the sub-samples as well, to go on with the weak evidence of causality for the full-sample. In more details, sub-sample A show the strongest effect comparing to other sub-samples followed by sub-sample C with slightly weaker causal effect from SS to GT and the weakest causal effect holds for sub-sample B.

In summary, the SSA causality test show that SS has predictive ability for GT for the all three sub-samples, over and above the full-sample, even if the latter result can be ignored due to structural instability. SSA causality test is able to capture possible nonlinearities that could exist in the data generating processes of the GT and SS, but also, in the relationship between GT and sunspot activity, for instance due to the structural breaks. It outperforms both time domain and frequency domain GC test and more importantly, highlights that SS has always been important in understanding the emerging trend of global warming. In other words, researchers working on global warming need to rely on a nonlinear data-driven, i.e., nonparametric approach while seeking the better understanding of the driven factors of global climate change.

## 5.4 Discussion

SSA causality test is primarily inspired by the framework of the well established causality analysis approach. Instead of building a linear or restricted nonlinear model, the innovative advancement adopts the SSA and MSSA forecasting performance to address whether including another variable is beneficial for analysing or predicting the other variable. It initially set the out-of-sample forecasting as the determinate criterion and all steps only proceed with the optimal performances achieved for the evaluation.

It is a nonparametric technique and has no assumptions of any linear or specific nonlinear models prior or during the test. Therefore, it is capable of distinguishing possible nonlinear relationships that other empirical tests cannot achieve, even for complex unknown nonlinearities. To date, SSA causality test is initially designed as a reduced form, data driven test that seeks better understanding of the causality relationship between a network that is as simple as only two variables included. It does not require a complex model that contains a number of factors. Moreover, there is no restriction of the same length of tested variables. In general, it can overcome many drawbacks that the currently established approaches fall short at, with straightforward way of thinking and implementing.

However, this is only the beginning of the development of this method and there are a number of aspects that will need further intensive researches so to get improved. Firstly, it is undoubtable that researchers who have been masterful on the empirical approaches will feel difficult to embrace the data driven technique. However, with the rapid advancements of technology and information science, there is no doubt that this type of method will be at least alternative way of causality analysis, especially at the circumstance of more frequently researchers have to work with a complex system containing large volume of complex data. Regarding the test itself, there is a concern about minor difference outcomes, which is due to the fact that the corresponding optimal performances are selected for every step of forecasting. The comparisons between optimal forecasting performances do not have a shared ground such as a same window length or identical number of eigenvalues. Finding a critical ground rule of comparison so to measure the level of assistance the other variable can bring along will be beneficial future research direction. However, it is worth to be highlighted that, with this significantly strict criterion to clarify causality, even very minor improvement of forecasting can possibly indicate a significant causality considering the complexity of the system and the extremely simple straightforward way of conducting this causality analysis with only two variables.

# Chapter 6

# CCM Causality Test

## 6.1 Introduction

Following the theoretical formulation and literature review of CCM in Chapter 2 section 2.3, In this chapter a few real data applications are conducted for the performance evaluation of this sub-space based technique on causality analysis. It is worth highlighting that the significant advantages of embracing this novel non-parametric technique are that no prior model assumptions are made; this technique is initially designed for better understanding of causal relationships in complex dynamical system; it can distinguish statistically significant causality by considering only two key variables instead of building a complex model by incorporating many possible influential variables based on regression modelling; CCM has remarkable sensitivity at detecting causal links within complex systems whilst not being limited to linearity or nonlinearity; and the calculation itself is efficient and comparatively straight forward. Note that this technique serves as an alternative approach rather than a competition to contribute on the literature of causality analysis studies. Moreover, another advancement of CCM on causality analysis is its reverse engineering framework. Instead of evaluating the influences of the cause on the effect factor, causality is determined by a good reconstruction of the cause from the effect factor as CCM believes that there must be information left behind in the effect factor from its cause.

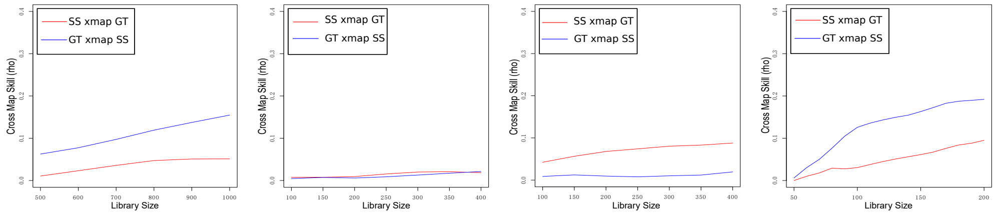
In the following sections, a diverse collection of real data about climate change, energy and tourism, and gene regulatory role are adopted for the evaluation of C-CM, respectively. This chapter intends to reflect the inherent efficiency and power of CCM in relation to empirical tests so as to promote its use in future including but not limited to the areas of listed implementations. Note that all the test results are obtained by the corresponding optimal embedding dimension, which specifically is determined by the nearest neighbor forecasting performance; range of library sizes are set identical within one corresponding data case for the sake of further comparisons; leave-one-out cross validation is applied for the best choice with optimal performance at specific library size. These conditions apply for all CCM causality tests in this thesis and there will be no more repetitive specification unless there are exceptional circumstances.

## 6.2 Evaluation by Real Data Applications

### 6.2.1 Global Temperature and Sunspot Number

Following the applications of GC and SSA causality tests in Chapter 5 section 5.3, a further extension of the causality analysis between SS and GT is obtained here by CCM. In brief, the same data set of GT and SS are adopted, with monthly frequency covering the period from 1880:01 to 2015:05. Equivalently, the sub-samples divided based on the structural breaks are also considered for all following tests. Note that more details of the data as well as the unit root tests results can be found in Chapter 5 section 5.3, thus, they are not reproduced here in this section.

The results of CCM causality tests between SS and GT on the total sample and all sub-samples respectively are listed as follows in Fig. 6.1.



(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 6.1 CCM Result for SS and GT.

The cross map skill index reflects the reconstruction ability of the fact factor to the cause factor for both directions respectively. Therefore, according to the results in Fig. 6.1, CCM indicates significant ability of cross mapping considering the total sample, in which, positive outcome reflects identified causal link from SS to GT. However, regarding each sub-sample, sub-sample A cannot detect obvious causation from SS to GT; sub-sample B shows opposite causation from GT to SS, which is considered misleading results due to the nature of the data; sub-sample C reflects significant causation from SS to GT[1]. Note that due to the long time span of the data and the wide scale of library size considered for more comprehensive analyses and comparisons, the increasing significance level along with larger library size is reasonable as more information is adopted for cross validation test and then for cross mapping.

Comparing to the results of SSA causality test in Chapter 5, the satisfying aspects are that CCM identifies significant causality for total sample and latest sub-sample C. It requires less calculation and is very sensitive to possible causal relationships in a complex system even with long time span and possible structural breaks. Also, there is no need to define window length and number of eigenvalues for a ground

[1] It possibly contains the influence of trend, which will be comprehensively discussed in Chapter 7.

rule of comparison like SSA causality test. However, the misleading results for subsample B reflects that the CCM outcome of causality can be significantly influenced by trend or complex noise due to the fact that a misleading direction of causality is much worse than no (or very weak) causality detected with the correct direction.

### 6.2.2 Oil Price and Tourist Arrivals

Given the energy-intensive nature of the tourism industry, the harmful effects of oil price fluctuations on transportation, production costs, economic uncertainty and disposable income have long been discussed (Becken, 2008). For countries that are heavily reliant on the tourism industry (the third largest industry in the world, after oil and automobiles) and unevenly exposed to sudden fluctuations in oil prices, the relationship between oil prices and tourist arrivals remains an important research topic which has drawn significant attention (Becken and Lennox, 2012).

The accurate detection of causality between oil prices and tourist arrivals can help the tourism planning process and aid in improving the quality and accuracy of tourist arrival forecasts and related managerial decisions (Goh, 2012). Literature indicates negative effects between oil price and tourism evidenced overwhelmingly by factors like inflation, consumer price indices, oil production, tourism income, and industrial production indices (Becken and Lennox, 2012). A considerable amount of recent causality analysis applications relating to tourist arrivals reflect that GC approach continues to remain a key method for assessing causality between tourist arrivals and influential variables (more details see (Antonakakis et al., 2015; Massidda and Mattana, 2013; Pérez-Rodríguez et al., 2015; Tang and Abosedra, 2016; Tang and Tan, 2013, 2015; Tsui and Fung, 2016)). However, it is pertinent to note again few drawbacks underlying the GC approach. Firstly, these tests continue to be initially conducted based on a complex model involving many variables, and the principle theory underlying the model has not been improved much from the simple linear or some assumed specific regressions. Secondly, the conclusion of causality is only obtained by following a certain order of tests which are all restricted by various assumptions, this makes the process extremely unreliable under real world conditions.

This section further evaluates the oil-tourism relationship and efficiently investigates the existence of causal links with an advanced non-parametric space-based method CCM. Accordingly, a reduced form, data-driven investigation is conducted to find significant evidences of oil-tourism causal relationships on a global scale by involving only the two key variables - oil price and tourist arrivals. To the best of my knowledge, the literature shows that CCM is yet to be exploited for evaluating causality between tourism and related variables, and this research marks the introductory and successful adoption of CCM for identifying causality between oil price and tourist arrivals.

### Data and Descriptive Analysis

Following the novel mutual association measure introduced in Chapter 4, section 4.5.2 adopted the oil prices and tourist arrivals data as real data application to evaluate its performance. Here in this section, the same groups of data are considered for the causality analysis by CCM along with the comparisons by the empirical GC approaches. In brief, the data of tourist arrivals and oil prices (both BRT and WTI) (EIA, 2016) are at monthly frequency covering the period from January 1996 to December 2015 of both US and nine European countries (including Austria, Italy, Germany, Greece, Netherland, Portugal, Spain, Sweden, UK). Note that the full details of the data can be found in Chapter 4 section 4.5.2, which are not reproduced here.

The summary of descriptive statistics are listed in Table 6.1, which confirms the similarity between BRT and WTI oil prices data. In terms of tourist arrivals, all countries generally show quite close level of Skewness and Kurtosis except Sweden is relevantly higher.

Table 6.1 Descriptive Statistics of Oil Price and Tourist Arrivals.

**Oil Prices**

| | Obs | Mean | Median | Max | Min | Std. Dev. | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|---|
| BRT | 240 | 56.41 | 49.22 | 132.72 | 9.82 | 35.24 | 0.47 | 1.85 |
| WTI | 240 | 54.78 | 49.06 | 133.88 | 11.35 | 31.19 | 0.40 | 1.89 |

**Tourist Arrivals**

| | Obs | Mean | Median | Max | Min | Std. Dev. | Skewness | Kurtosis |
|---|---|---|---|---|---|---|---|---|
| Austria | 240 | 1481894 | 1434455 | 3205966 | 446240 | 504448 | 0.39 | 3.21 |
| Germany | 240 | 1918394 | 1788583 | 4401682 | 747141 | 724552 | 0.75 | 3.29 |
| Greece | 240 | 765847 | 564523 | 3107955 | 29856 | 710611 | 1.11 | 3.66 |
| Italy | 240 | 3343953 | 3277084 | 8084209 | 907367 | 1709118 | 0.50 | 2.45 |
| Netherland | 240 | 870900 | 864200 | 1745779 | 275000 | 284180 | 0.34 | 2.79 |
| Portugal | 240 | 539796 | 522395 | 1359284 | 155438 | 256280 | 0.70 | 3.03 |
| Spain | 240 | 3229314 | 2934373 | 7443749 | 671109 | 1533209 | 0.51 | 2.42 |
| Sweden | 240 | 357927 | 239902 | 1428207 | 98357 | 289081 | 1.93 | 5.97 |
| UK | 240 | 1668020 | 1541000 | 3390515 | 692120 | 582239 | 0.59 | 2.64 |
| US | 240 | 4325374 | 4222034 | 8364940 | 2094287 | 1292787 | 0.59 | 2.88 |

In order to evaluate the stationarity of data, three different unit root tests including KPSS, ADF and PP are conducted and summarized in Table 6.2. The results are overwhelmingly suggesting trend stationary for all variables, whilst, the PP test indicates stationary for a few countries regarding tourist arrivals data. In general, the variables are concluded non-stationary with one unit root.

Table 6.2 Unit Root Test Results of Oil Price and Tourist Arrivals.

| Variables | Series | Methods | None | | Intercept | | Intercept and Trend | |
|---|---|---|---|---|---|---|---|---|
| | | | Level | Decision | Level | Decision | Level | Decision |
| Oil Prices<br>(240 Obs)<br>1996:1-2015:12 | BRT | KPSS | ——— | ——— | 1.675***(11) | I(1) | 0.139*(11) | I(0) |
| | | ADF | -10.284***(0) | I(1) | -10.264***(0) | I(1) | -10.294***(0) | I(1) |
| | | PP | -10.279***(4) | I(1) | -10.258***(4) | I(1) | -10.283***(4) | I(1) |
| | WTI | KPSS | ——— | ——— | 1.663***(11) | I(1) | 0.166**(11) | I(1) |
| | | ADF | -10.104***(0) | I(1) | -10.083***(0) | I(1) | -10.109***(0) | I(1) |
| | | PP | -10.104***(0) | I(1) | -10.083***(0) | I(1) | -10.109***(0) | I(1) |
| Tourists Arrivals<br>(240 Obs)<br>1996:1-2015:12 | Austria | KPSS | ——— | ——— | 1.458***(15) | I(1) | 0.144*(27) | I(0) |
| | | ADF | -3.938***(14) | I(1) | -16.637***(11) | I(1) | -17.093***(11) | I(0) |
| | | PP | -49.801***(23) | I(1) | -9.945***(31) | I(0) | -10.345***(24) | I(0) |
| | Germany | KPSS | ——— | ——— | 2.305***(9) | I(1) | 0.115 (1) | I(0) |
| | | ADF | -2.524***(13) | I(1) | -3.581***(13) | I(1) | -3.825***(13) | I(1) |
| | | PP | -12.185***(16) | I(1) | -4.832***(5) | I(0) | -5.169***(0) | I(0) |
| | Greece | KPSS | ——— | ——— | 0.755***(3) | I(1) | 0.058(2) | I(0) |
| | | ADF | -4.411***(11) | I(1) | -4.791***(11) | I(1) | -4.985***(11) | I(1) |
| | | PP | -4.056***(5) | I(0) | -5.414***(6) | I(0) | -5.529***(6) | I(0) |
| | Italy | KPSS | ——— | ——— | 1.079***(5) | I(1) | 0.014(2) | I(0) |
| | | ADF | -3.527***(13) | I(1) | -4.403***(13) | I(1) | -4.527***(13) | I(1) |
| | | PP | -2.828***(3) | I(0) | -6.291***(4) | I(0) | -6.604***(4) | I(0) |
| | Netherland | KPSS | ——— | ——— | 1.744***(8) | I(1) | 0.084(4) | I(0) |
| | | ADF | -2.976***(13) | I(1) | -3.496***(13) | I(1) | -3.503***(13) | I(1) |
| | | PP | -14.361***(3) | I(1) | -5.952***(2) | I(0) | -6.548***(1) | I(0) |
| | Portugal | KPSS | ——— | ——— | 1.653***(7) | I(1) | 0.111(1) | I(0) |
| | | ADF | -4.077***(12) | I(1) | -4.658***(12) | I(1) | -4.848***(12) | I(1) |
| | | PP | -2.101**(6) | I(0) | -5.731***(5) | I(0) | -5.672***(6) | I(0) |
| | Spain | KPSS | ——— | ——— | 1.991***(8) | I(1) | 0.071(1) | I(0) |
| | | ADF | -2.353**(12) | I(1) | -2.857*(12) | I(0) | -3.469**(13) | I(0) |
| | | PP | -2.306**(4) | I(0) | -5.646***(4) | I(0) | -6.118***(5) | I(0) |
| | Sweden | KPSS | ——— | ——— | 1.052***(2) | I(1) | 0.161**(9) | I(1) |
| | | ADF | -5.708***(13) | I(1) | -6.117***(13) | I(1) | -6.104***(13) | I(1) |
| | | PP | -3.940***(14) | I(0) | -5.961***(19) | I(0) | -5.794***(24) | I(0) |
| | UK | KPSS | ——— | ——— | 0.818***(5) | I(1) | 0.090(3) | I(0) |
| | | ADF | -4.889***(12) | I(1) | -4.981***(12) | I(1) | -5.196***(12) | I(1) |
| | | PP | -10.446***(4) | I(1) | -5.821***(1) | I(0) | -6.387***(2) | I(0) |
| | US | KPSS | ——— | ——— | 1.825***(11) | I(1) | 0.392***(9) | I(1) |
| | | ADF | -3.591***(12) | I(1) | -3.928***(12) | I(1) | -4.074***(12) | I(1) |
| | | PP | -19.331***(6) | I(1) | -3.796***(8) | I(0) | -7.063***(8) | I(0) |

[a] The *, ** and *** indicate significance at the 10%, 5% and 1% respectively.

[b] The critical values are as follows:(1)None: -2.574, -1.942 and -1.616 for ADF and PP at 1%, 5% and 10% level of significance, respectively; (2)Intercept: -3.457, -2.873 and -2.573 {0.739, 0.463, 0.347} for ADF and PP {KPSS} at 1%, 5% and 10% level of significance, respectively;(3)Intercept and Trend: -3.996, -3.428 and -3.137 {0.216, 0.146, 0.119} for ADF and PP{KPSS} at 1%, 5% and 10% level of significance respectively.

[c] Numbers in parentheses for ADF and PP tests indicates lag-lengths selected based on the SIC. For the KPSS test, based on the Bartlett kernel spectral estimation method, the corresponding numbers are the Newey-West bandwidth.

### Causality Tests Results

Here in this section, the causality tests are conducted for tourist arrivals and both BRT and WTI oil prices respectively for each country. Corresponding results are summarized in Table 6.3 by different causality detection techniques.

Table 6.3 Summary of Causality Tests Results on Oil Price and Tourist Arrivals.

| **Method** | **Country** | **Oil Prices** | | | |
|---|---|---|---|---|---|
| | | **BRT** | | **WTI** | |
| | | → | ← | → | ← |
| | | Yes/No | Yes/No | Yes/No | Yes/No |
| **Time Domain** | **Austria** | No(0.68) | No(0.56) | No(0.81) | No(0.34) |
| | **Germany** | No(0.52) | No(0.27) | No(0.29) | No(0.17) |
| | **Greece** | No(0.54) | No(0.36) | No(0.46) | No(0.44) |
| | **Italy** | No(0.60) | No(0.98) | No(0.67) | No(0.74) |
| | **Netherland** | No(0.30) | No(0.83) | No(0.29) | No(0.65) |
| | **Portugal** | No(0.38) | No(0.41) | No(0.72) | No(0.31) |
| | **Spain** | No(0.62) | No(0.24) | No(0.54) | No(0.12) |
| | **Sweden** | No(0.21) | No(0.55) | No(0.14) | No(0.93) |
| | **UK** | No(0.63) | No(0.95) | No(0.53) | No(0.82) |
| | **US** | No(0.48) | No(0.85) | No(0.53) | No(0.48) |
| **Frequency Domain** | **Austria** | No | No | No | No |
| | **Germany** | No | No | No | No |
| | **Greece** | No | No | No | No |
| | **Italy** | No | No | No | No |
| | **Netherland** | No | No | No | No |
| | **Portugal** | No | No | No | No |
| | **Spain** | No | No | No | No |
| | **Sweden** | No | No | No | No |
| | **UK** | No | No | No | No |
| | **US** | No | No | No | No |
| **CCM** | **Austria** | No | Yes | No | Yes |
| | **Germany** | No | Yes | No | Yes |
| | **Greece** | No | Yes | No | Yes |
| | **Italy** | No | Yes | No | Yes |
| | **Netherland** | No | Yes | No | Yes |
| | **Portugal** | No | Yes | No | Yes |
| | **Spain** | No | Yes | No | Yes |
| | **Sweden** | No | Yes | No | Yes |
| | **UK** | No | Yes | No | Yes |
| | **US** | No | Yes | No | Yes |

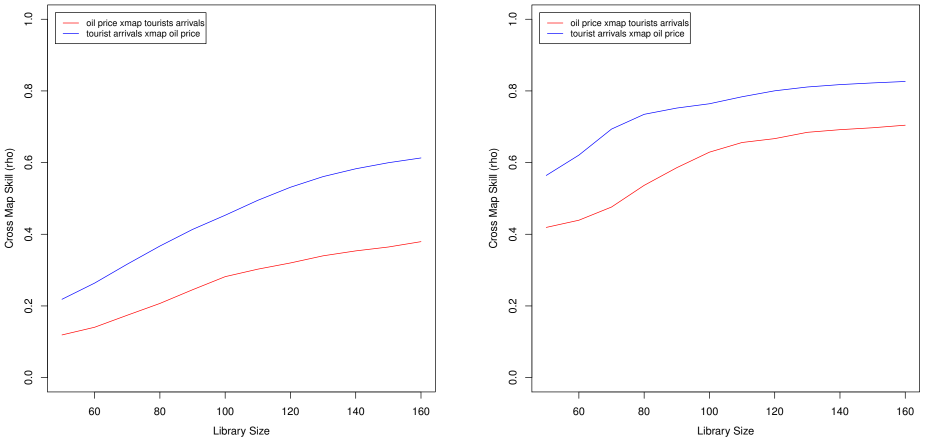
**Notes:** → indicates tourist arrivals causes oil price;
← indicates oil price causes tourist arrivals;
numbers in () indicate $P$-values.

Regarding the time domain GC test, note that all tests conducted satisfy the preconditions of the test with results by the corresponding optimal lag. The results indicate that no causal link can be detected regardless of countries and types of oil price index (the null of non-causality still cannot be rejected at 10% significant level for all countries considered). In brief, time domain GC test fails to detect any causality between tourist arrivals and oil prices for both US and nine European countries.

The frequency domain GC test furthermore evaluates the possible causality at specific frequencies with corresponding optimal lag-structures maintained for all tests. No significant causality can be identified for any frequency[2], and the frequency domain test also fails to prove the causal links between tourist arrivals and oil prices regardless of the countries and types of oil price index.

Finally, the non-parametric subspace-based technique - CCM shows the existence of one-directional causality from oil prices to tourist arrivals for all countries when the empirical methods fail to detect same. The results of CCM tests between tourist arrivals and oil prices (BRT) of US and UK are listed in Fig. 6.3, and the details of the rest of the test results can be found in Appendix D.2. As the initial attempt of incorporating CCM in oil-tourism causality studies, this application establishes a reduced form, data-driven investigation and contributes to existing literature through the successful and introductory application of an advanced method, and via the uncovering of significant causal links from oil prices to tourist arrivals in US and nine European countries, alongside the comparison of empirical and novel methods.

[2] A full detailed frequency domain GC test results can be found in Appendix D.1.



(a) UK

(b) US

**Notes:** The cross map skill index reflects the reconstruction ability of the fact factor to the cause factor for both directions respectively; the blue line above the red line indicates significant cross map skill of tourist arrivals on oil price, which means causality from oil price to tourist arrivals.

Fig. 6.3 CCM Causality Results of Tourists Arrivals and Oil Prices (BRT) for UK and US.

## 6.2.3 Gene Regulatory Role

Here in this section, the implementation of CCM is extended to the gene regulatory role study.

### Introduction

Segmentation in *Drosophila melanogaster* is a particularly well studied process which highlights the role of gene regulatory network (GRN) in the earliest stage of development (Lewis, 1978). There have been considerable attempts to portrait a picture of the interactions presented between regulators in this GRN (see (Baird-Titus et al., 2006; Berleth et al., 1988; Karlebach and Shamir, 2008; Liu and Jack, 1992; Lopes et al., 2012; Niessing et al., 2002)). Recently, the availability of more data on molecular mechanisms of regulatory interactions has made it possible to study these interactions in more quantitative depth. Hence, for the first time to the best of my knowledge, this section seeks to evaluate the dynamical interactions of this system from a statistical causality point of view so to further contribute on the literature of both gene regulatory role and causality analysis. Specifically, as an advanced non-parametric method that is designed for a dynamical system involving complex interactions, CCM is adopted for the analysis along with the empirical GC approaches for comparisons. It is of note that the detected regulatory link can be either inductive (i.e. increasing the protein concentration of one gene raises the protein concentration of the other gene), or inhibitory (i.e. increasing the protein concentration of one gene decreases the protein concentration of the other gene). Any efforts at identifying the nature of the detected interaction would require more extensive research and that objective is beyond the mandate of this implementation (Davidson and Levin, 2005).

As the best-studied transcriptional network in *Drosophila* development, segmentation GRN contains three fundamental types of genes which play a crucial role in *Drosophila* development: maternal effect genes, gap genes and pair rule genes (Bieler et al., 2011). Among them, the maternal effect genes including bicoid (*bcd*) [3] and caudal (*cad*) must be addressed as the most important factors since they respectively determine most aspects of anterior and posterior axis of an adult fruit fly and more importantly, they commence the sequential activation of segmentation GRN (Berleth et al., 1988; Bieler et al., 2011; Copf et al., 2004). It is imperative to note that since providing robust genetic evidence is an important step in reporting genetic causality, among all the interactions between regulators in segmentation GRN, this application has been narrowed down to the interactions between *bcd* and *cad*, *bcd*

[3] In what follows, the italic lower-case *bcd* represents either the gene or mRNA and Bcd refers to protein. This can be applied for all other genes mentioned in this section (for example, *cad* and Cad).

and Kruppel (*kr*) and *cad* and *kr* genes which their interactions have been previously accredited via laboratory experimental evidences. Accordingly, extracting these links using CCM will provide the credit to step further and apply these methods to find the unknown regulatory links between other genes.

### Data

The quantitative *bcd*, *cad* and *kr* gene expression profiles representing the protein concentrations of these genes in wild-type *Drosophila* embryos are achieved using the confocal scanning microscopy of fixed embryos immunostained for segmentation proteins and is available via FlyEx database[4]. The applied antibody allows the visualisation of the proteins under study. Note that such quantification relies on the assumption that the actual protein concentrations detected by the antibodies and the fluorescence intensities are linearly related to the embryo's natural protein concentration (Pisarev et al., 2009; Poustelnikova et al., 2004).

To this aim, a 1024 × 1024 pixel confocal image with 8 bits of fluorescence data was obtained for each embryo which then transformed into an American Standard Code for Information Interchange (ASCII) table. The ASCII table contains the fluorescence intensity levels attributed to each nucleus in the 10% of longitudinal strips (i.e. only the nuclei correspondents to the central 10% strip consists of the 45–55% of the dorsoventral (D–V) axis are selected) along the Anterior-Posterior (A-P) axis and is unprocessed for any noise reduction methods. By adopting from Surkova et al. (2008), Fig. 6.5 shows an example of a confocal image with the 10% longitudinal strip, in which the white horizontal lines depict the 10% strip utilised to collect data from. Fig. 6.6 presents a typical example profile achieved by flourocence antibodies technique containing noisy Bcd, Cad and Kr for a specific embryo at time class 14(1), in which the x-axis shows the position of the nuclei along the

[4] Available at: http://urchin.spbcas.ru/flyex/

Anterior-Posterior (A-P) axis of the embryo and y-axis shows the fluorescence intensity levels[5].

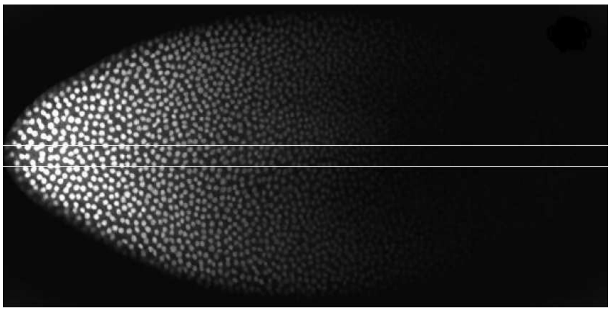


Fig. 6.5 Confocal Image of An Embryo at Time Class 14(1) with 10% Longitudinal Strip.

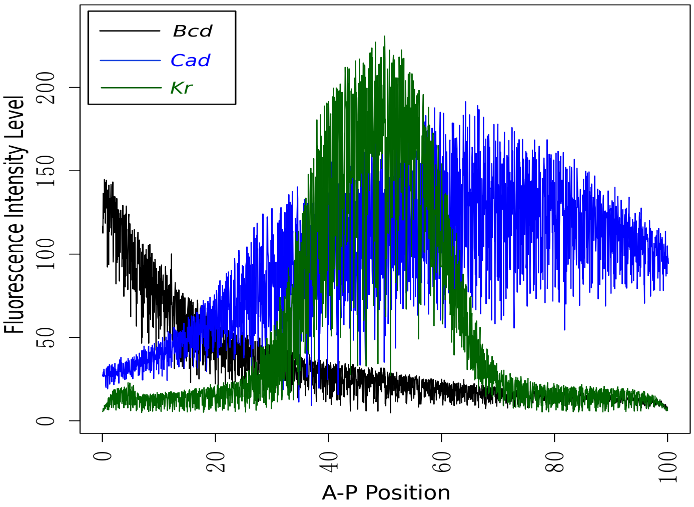


Fig. 6.6 A Typical Example of Noisy Bcd, Cad and Kr for Embryo *ms26* at Time Class 14(1).

Since the segment determination starts from cleavage cycle 10 and lasts until the end of cleavage cycle 14A (when proteins synthesised from maternal transcripts

[5] Note that the data is highly volatile and there is significantly high possibility that it will influence the successful identification of a cause-and-effect relationship. Chapter 7 will further extend the research on this aspect.

begin to appear up to the onset of gastrulation) the data has been categorised to five main cycles of 10 to 14A. Additionally, as the cleavage cycle 14A is considerably longer in time, to facilitate the analysis, temporal classes 1 to 8 have been considered as the subgroups of this cleavage cycle (Pisarev et al., 2009; Poustelnikova et al., 2004). It should also be noted that each class of data contains a different number of embryos. Table 6.4 presents the summary of the number of embryos studied per each time class (TC). Note that the expression profile of each embryo has a different length of data where the length in this table reports the average.

Table 6.4 Different Time Classes and Embryos Studied Per Each Time Class.

| Time Class (TC) | N | Length | SD |
|---|---|---|---|
| 10 | 5 | 127 | 18.83 |
| 11 | 12 | 276 | 25.83 |
| 12 | 15 | 489 | 97.18 |
| 13 | 47 | 1224 | 78.56 |
| 14(1) | 28 | 2318 | 143.87 |
| 14(2) | 15 | 2315 | 86.83 |
| 14(3) | 20 | 2367 | 141.05 |
| 14(4) | 17 | 2309 | 119.16 |
| 14(5) | 14 | 2301 | 126.96 |
| 14(6) | 18 | 2347 | 103.74 |
| 14(7) | 13 | 2007 | 229.61 |
| 14(8) | 12 | 1600 | 311.21 |

Note: N = Number of embryos studied per each time class; Length = The average length of data of expression profiles; SD= Standard deviation of length of data.

### Causality Test Results

A summary of the causality test results by CCM and the empirical GC approaches is listed in Table 6.5 that illustrates the findings of the causality analysis on Bcd and Cad profiles. Note that for all evaluations, the corresponding requirements for each test are satisfied, also the optimal outcomes are selected. More specifically, differentiations are taken accordingly for stationarity prior to the tests. The co-integration test is also conducted for those groups of variables having equivalent one unit root. Note that none of the tested groups showed significant results in indicating co-integration, thus the results are not reproduced here as it is not the main focus of this research. Moreover, the corresponding optimal lag is selected for obtaining the test result for each group of profiles; the $p$-values reported for time domain GC test are the average $p$-values attained for embryos studied each corresponding time class.

Table 6.5 Causality Tests Results for Bcd on Cad Profiles.

| **Time Class (TC)** | **Time Domain GC** | | **Frequency Domain GC** | **CCM** |
|---|---|---|---|---|
| | YES/NO | p-value | YES/NO | YES/NO |
| 10 | NO | 0.68 | NO | YES |
| 11 | NO | 0.71 | NO | YES |
| 12 | NO | 0.89 | NO | YES |
| 13 | NO | 0.89 | NO | YES |
| 14(1) | NO | 0.95 | NO | YES |
| 14(2) | NO | 0.98 | NO | YES |
| 14(3) | NO | 0.98 | NO | YES |
| 14(4) | NO | 0.94 | NO | YES |
| 14(5) | NO | 0.95 | NO | YES |
| 14(6) | NO | 0.96 | NO | YES |
| 14(7) | NO | 0.81 | NO | YES |
| 14(8) | NO | 0.79 | NO | YES |

Note: "Yes" stands for the detected regulatory link and "No" means the regulatory link could not be detected by the adopted test.

According to the results in Table 6.5, the regulatory link between Bcd and Cad can be detected by neither time domain nor frequency domain GC tests. Since the length of the data under study vary between different time classes, specifically, time class 10 to 13 and 14(7-8) have shorter lengths comparing to the time classes

14(1-6), which may be the reason of getting slightly smaller *p*-values for time class 11 to 13 and 14(8) comparing to the rest of the sub classes of time class 14. The frequency domain test shows less sensitivity to the data length possibly because this method focuses on each individual frequency component rather than the entire series. Nevertheless, CCM accomplished to overwhelmingly identify the regulatory relationship between Bcd and Cad in expression profiles despite the various length and highly volatile feature of the data.

In addition, Table 6.6 and Table 6.7 summarise the causality test results of identifying the regulatory link between Bcd and Kr profiles and Cad and Kr profiles, respectively. Similar to the results reported in Table 6.5, CCM efficiently identify the regulatory relationship whilst the empirical GC approaches both fail to detect.

Table 6.6 Causality Tests Results for Bcd on Kr Profiles.

| **Time Class (TC)** | **Time Domain GC** | | **Frequency Domain GC** | **CCM** |
|---|---|---|---|---|
| | YES/NO | p-value | YES/NO | YES/NO |
| 12 | NO | 0.71 | NO | YES |
| 13 | NO | 0.66 | NO | YES |
| 14(1) | NO | 0.89 | NO | YES |
| 14(2) | NO | 0.93 | NO | YES |
| 14(3) | NO | 0.97 | NO | YES |
| 14(4) | NO | 0.94 | NO | YES |
| 14(5) | NO | 0.95 | NO | YES |
| 14(6) | NO | 0.92 | NO | YES |
| 14(7) | NO | 0.81 | NO | YES |

Note: "Yes" stands for the detected regulatory link and "No" means the regulatory link could not be detected by the adopted test.

Table 6.7 Causality Tests Results for Cad on Kr Profiles.

| **Time Class** | **Time Domain GC** | | **Frequency Domain GC** | **CCM** |
|---|---|---|---|---|
| | YES/NO | p-value | YES/NO | YES/NO |
| 12 | NO | 0.39 | NO | YES |
| 13 | NO | 0.78 | NO | YES |
| 14(1) | NO | 0.84 | NO | YES |
| 14(2) | NO | 0.89 | NO | YES |
| 14(3) | NO | 0.94 | NO | YES |
| 14(4) | NO | 0.91 | NO | YES |
| 14(5) | NO | 0.87 | NO | YES |
| 14(6) | NO | 0.82 | NO | YES |
| 14(7) | NO | 0.75 | NO | YES |

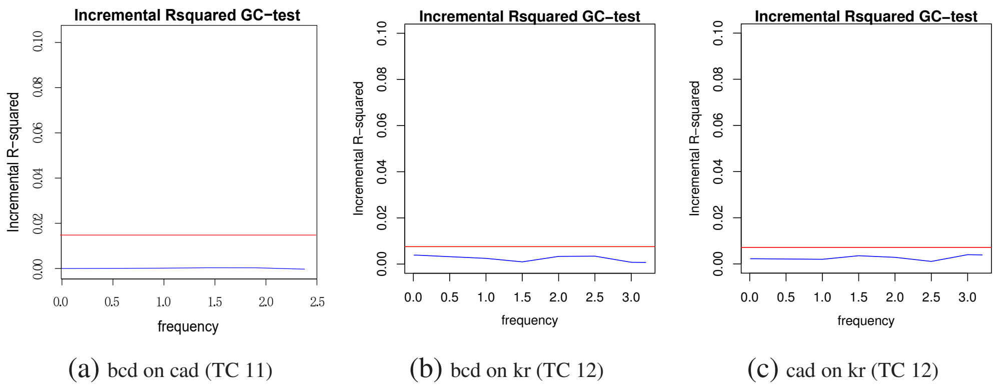
Note: "Yes" stands for the detected regulatory link and "No" means the regulatory link could not be detected by the adopted test.

Fig. 6.7 depicts an example of the results obtained by frequency domain GC test for Bcd–Cad, Bcd–Kr and Cad–Kr profile pairs respectively [6]. In these figures, the blue line represents the statistic test of each specific frequency, and the red line represents the 5% critical value for all the frequencies. The horizontal axis gives the parameter $w$ to calculate the corresponding frequency $f$ by $f = 2\pi/w$. Therefore, when the test statistics is above or very close to the 5% critical value, the causality is detected for that corresponding frequency. As the component of each frequency is considered separately for identifying possible causal link, the impacts of relatively less information are significantly reduced.

As previously mentioned in the beginning of this chapter, CCM test results are obtained with the corresponding optimal embedding dimension $E$ for each pair of gene expression profiles based on the nearest neighbor forecasting performance by simplex projection. Fig. 6.9 presents the examples of the CCM test result for Bcd–Cad, Bcd–Kr and Cad–Kr respectively [7]. In general, the higher ability of factor $X$

[6] The frequency domain GC test results for all considered pairs of genes related to all different time classes can be found in Appendix D.3.

[7] The CCM test results for all considered pairs of genes related to all different time classes can be found in Appendix D.4.

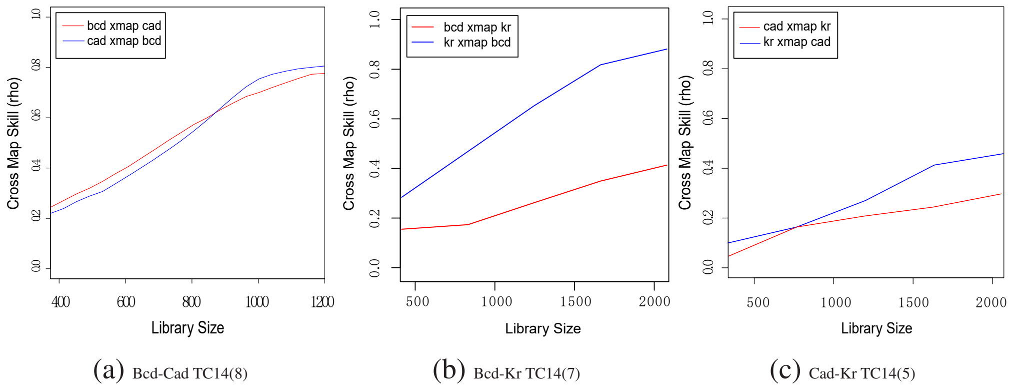


(a) bcd on cad (TC 11) (b) bcd on kr (TC 12) (c) cad on kr (TC 12)

Note: The blue line represents the statistic test of each specific frequency, and the red line represents the 5% critical value for all the frequencies.

Fig. 6.7 Example Frequency Domain GC Test Results for Bcd, Cad and Kr.

on reconstructing the attractor reflects more significant causal effects of the attractor on $X$. As can be seen, the results of CCM reflect close relationships between Bcd and Cad, whilst Bcd shows more significant relationship with Kr comparing to Cad. The crossmap abilities of Bcd and Cad on Kr are fairly similar, however, Kr clearly indicates higher reconstruction ability on Bcd comparing to Cad.



(a) Bcd-Cad TC14(8) (b) Bcd-Kr TC14(7) (c) Cad-Kr TC14(5)

Note: For the left(middle)[right] results, the red line indicates the reconstruction ability of Bcd(Bcd)[Cad] crossmap Cad(Kr)[Kr], while the blue line represents the performance of Cad(Kr)[Kr] on crossmapping Bcd(Bcd)[Cad].

Fig. 6.9 Example CCM Test Results for Bcd, Cad and Kr.

Even though the regulatory role of *bcd* on *cad*, *bcd* on *kr* and *cad* on *kr* genes have been previously reported through several genetics experiments, in practice they have not been validated using any causality detection methods. The CCM method established consistent significant evidences to identify the regulatory links between these expression profiles while the empirical GC approaches both failed to prove the same. As the initial attempt of extending CCM causality analysis technique on gene regulatory role studies, this application has provided satisfying performance of the adopted technique as well as its convincing capability to be further adapted to explore the regulatory relationships among other challenging GRNs.

## 6.3 Discussion

This chapter challenges the relatively new subspace-based technique CCM on causality analysis performance with a diverse range of real data applications. In general, the consistent and significant evidences presented herewith prove that this advanced nonparametric and assumption free technique is a robust, solid and efficient method that can produce reliable evidences of causality with only two key variables, even at the circumstance of extremely complex and nonlinear scenarios as witnessed in the above implementations. It shows satisfying performances on seeking the solution of a complex research question–causality. Moreover, similar to the SSA causality test introduced in Chapter 5, it is a reduced form, straight forward, data driven approach that is initially designed as an ideal method for better understanding and distinguishing causality in complex systems.

Following the literature review of successful CCM applications that are summarized in Chapter 2 section 2.3.1, CCM shows its significant sensitivity, adaptability and capability as a reliable causality analysis technique through the above implementations. It is of note that this research incorporates CCM for the first time with the corresponding areas of causality detection applications in this chapter. As such, this research also aims to motivate further developments and increased applications of CCM for other causality analyses whilst assisting policy and decision making in a broader range of subjects, where the multivariate analysis of nonlinear or complex systems can be of utmost importance.

As a relatively new subspace-based technique for causality detection, CCM surely has the immeasurable potential of further advancements. The possible directions of future researches may include the extension of panel data applicable version of CCM, improving the current cross map skill measure index, etc. Moreover, as briefly mentioned in the above applications, there are still concerns regarding the possible influences by trend, complex noise as well as the lower efficiency when works with very large size of data. Nevertheless, the next chapter of this research will focus on eliminating the possible influences by trend or complex noise through incorporating the data preprocessing techniques.

# Chapter 7

# SSA-CCM Hybrid Causality Test

In relation to the strict evaluations by a diverse range of implementations in Chapter 6, it is of note that the composition/structure of the data can generally lead to widely different outcomes for causality analysis despite the real information contained by the data itself, especially for the empirical approaches. The novel techniques in this research did indeed prove their advantages whilst the possible influences by existing trend or complex noises should still be exploited and further clarified so to complete this research. Here in this chapter, a hybrid approach is proposed that incorporates the data preprocessing technique with CCM so to conduct causality analysis in the manner of better understanding the data as well as the causality relationship in a complex system.

## 7.1 Introduction

The hybrid method that is introduced here is the SSA-CCM hybrid causality test. More specifically, SSA technique is firstly adopted for the data preprocessing, then the processed data work as the input of CCM test so to establish a causality analysis. The most important reason of applying SSA for data preprocessing is its full-featured data processing superiority. As detailed introduced in Chapter 2 section

2.2, SSA can decompose the data into significant components that represent particular features respectively. This can assist on the elimination of specific component of the data that is in relation to the negative influences on accurately identifying causality. By introducing the preprocessing step into the CCM technique, this allows CCM to perform its best level without being affected by the insignificant components of the data due to the fact that CCM can be too sensitive so to get influenced and lead to a less accurate reveal of the overall truth.

It is also of note that the data preprocessing is necessary for some particular areas of research, for instance, among the applications introduced in the previous chapters of this research, the long time span trend in climate change study, the complex noise in gene expression profiles. Thus, in this chapter, the performance of the SSA-CCM hybrid causality test is evaluated by the application of these two examples respectively. In specific, SSA is firstly performed for trend extraction before the reveal of the causality between SS and GT in section 7.2, the comparisons of a few empirical trend extraction methods and the empirical GC approaches are also conducted for the purpose of the comprehensive evaluation. In terms of the gene expression profiles that contains complex noise, SSA is applied for noise filtering before proceeding on the causality analysis along with the comparisons with the empirical GC approaches in section 7.3. Note that each section is self-contained to address the corresponding advancement of the hybrid causality analysis method and its satisfying performances.

## 7.2 Sunspot Number and Global Temperature

As initially introduced in Chapter 5 section 5.3, as well as Chapter 6 section 6.2.1, here the same group of SS and GT data is adopted for illustrating the advancement of SSA-CCM hybrid causality approach. It has been previously mentioned that the existing trend[1] of GT may lead to inaccurate conclusion of causality. In order to reveal the truth and conduct precise understanding of the causality, the existing trend is considered misleading information that requires a prior step of data preprocessing. The following sections compare SSA with 7 more trend extraction methods on their performances of preparing data for causality analysis by not only CCM, but also the empirical GC approaches.

### 7.2.1 Trend Extraction Methods and Extracted Data

Following the detailed introduction of the SS and GT data in Chapter 5 section 5.3, the detrended GT (DGT) is obtained here aiming to remove the possible misleading effects of the trend on the causality detection. The SSA-CCM hybird approach is applied, whilst a few selected and representative trend extraction methods are adopted and compared in this section by mainly following Alexandrov et al. (2012)[2]. In brief, despite the SSA technique that is presented in the original version of the hybrid approach, the trend extraction methods employed in this section cover almost all aspects of trend extraction studies to date, including Model Based Approach (MBA), Empirical Mode Decomposition (EMD), Wavelet (WAV), Local Regression (LOESS), Henderson Filter (HEN) and Hodrick Prescott Filter (HP).

More specifically, the MBA refers to a family of methods that commonly sharing the reliance upon time series models for the trend estimation; the EMD technique decomposes the signal into a collection of intrinsic mode functions with a trend; the SSA technique embeds the data into multidimensional matrix, then applies SVD technique to decompose the data into representative components; the WAV technique conducts the wavelet transformation and transfers the time series into multi-scale decompositions, where the details of scales represents the different

[1] The detailed unit root test results can be found in Chapter 5 section 5.3.3.

[2] Note that a comprehensive review and theoretical introductions of the trend extraction methods can be found in (Alexandrov et al., 2012), thus it is not reproduced here since it is not the primary objective of this research.

features of time series for further reconstruction; the LOESS is based on nearest neighbors weights, where it allows the smoothing with fitted degree of polynomials considering the weights estimated accordingly based on the neighborhood; the HEN technique minimize smoothing with respect to a third degree polynomial within the span of the filter; the HP technique builds up the over long time period framework of trend and cycle with average 0, in which the measure of the smoothness of the trend is the sum of the squares of its second difference, while the cycles are deviations from trend.

Note that the detrended GT by different methods respectively are listed as, for instance, DGT(MBA), DGT(EMD), etc. Similarly, the trend extracted by different methods are noted as, for example, Trend(MBA), Trend(EMD), etc. The DGT series and corresponding extracted trend series are summarized below in Fig. 7.1 and Fig. 7.2 respectively. Note that all detrended series and corresponding extracted trend series will be adopted for CCM and empirical GC tests respectively with considerations of both total sample (1880:01-2015:05) and sub-samples (1880:01-1936:03, 1936:03-1986:12, 1986:12-2015:05).

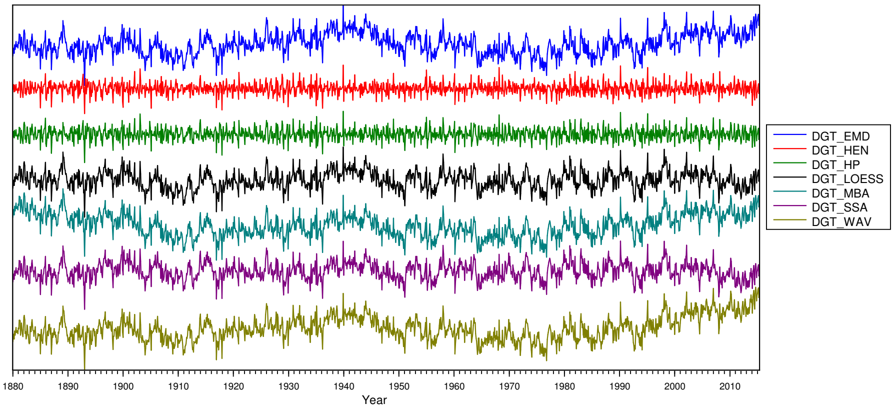


Fig. 7.1 DGT by Different Trend Extraction Methods.

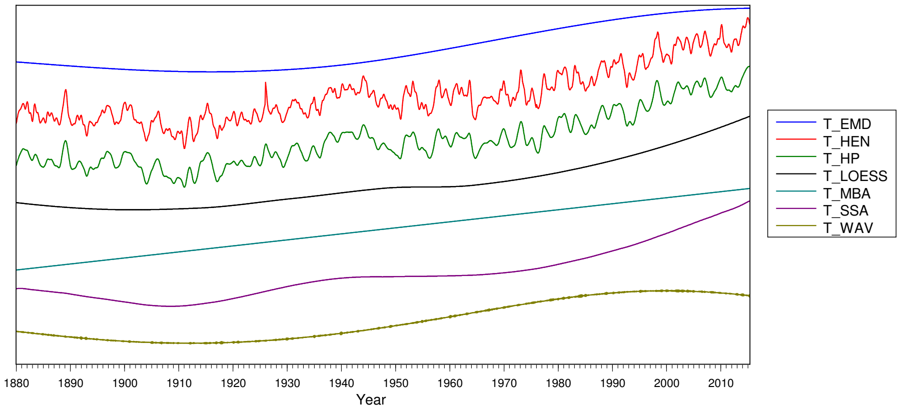


Fig. 7.2 Extracted Trend of GT by Different Trend Extraction Methods.

### 7.2.2 Evaluation of Performances in Causality Analysis

In this section, the causality tests are conducted for the original series, different detrended series, and extracted trend series respectively. The corresponding results are summarized below by different causality detection techniques.

#### Time Domain GC Test Results

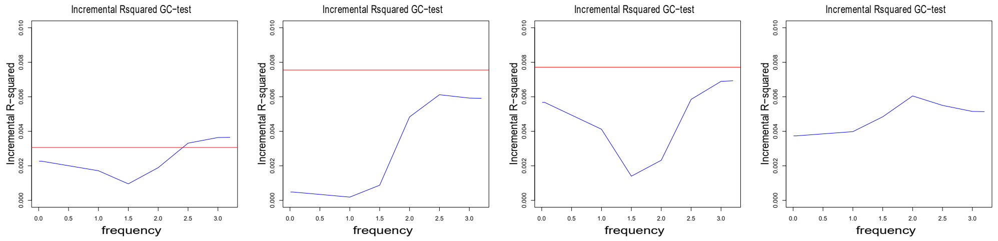
Given the significant and empirical role of time domain GC test, here the tests of SS and different DGT as well as extracted trend series by different techniques are conducted respectively in Table 7.1. Note that all tests conducted satisfy the preconditions of time domain GC test with results by the corresponding optimal lag. The null hypothesis that SS does not Granger cause GT in general cannot be rejected for the sub-sample A except for DGT by HEN. The results also point out that the null of non-causality still continued to hold at 5% significant level for all trend and detrended series of sub-sample B and sub-sample C. The overall causality from SS to GT considering the total sample is proved by DGT(HEN), DGT(LOESS) and DGT(SSA) at 5% significant level. Therefore, comparing to the insignificant results of the original series, the trend extraction is confirmed helpful on time domain GC test, more specifically, DGT by HEN, LOESS and SSA indicate the significant causal link from SS to GT at the total sample level.

Table 7.1 Summary of Time Domain GC Test Results by Trend Extraction Techniques.

| | Total Sample | | Subsample A | | Subsample B | | Subsample C | |
|---|---|---|---|---|---|---|---|---|
| Tested Series | 1880:1-2015:5 | | 1880:1-1936:2 | | 1936:3-1986:11 | | 1986:12-2015:5 | |
| | F | p-value | F | p-value | F | p-value | F | p-value |
| Original | 1.0107 | 0.3642 | 0.947 | 0.3884 | 1.1374 | 0.3213 | 1.5871 | 0.2062 |
| DGT(EMD) | 1.9439 | 0.0842 | 1.6239 | 0.1825 | 1.4771 | 0.1953 | 0.6153 | 0.6055 |
| DGT(HEN) | 1.5184 | **0.0458** | 1.7268 | **0.0133** | 1.2998 | 0.2023 | 0.6559 | 0.5797 |
| DGT(HP) | 1.1439 | 0.1875 | 1.4968 | 0.0522 | 1.1201 | 0.2956 | 0.6757 | 0.5675 |
| DGT(LOESS) | 1.6184 | **0.0298** | 1.6243 | 0.1824 | 2.2294 | 0.0837 | 0.6179 | 0.6039 |
| DGT(MBA) | 1.3569 | 0.2287 | 1.6244 | 0.1824 | 1.4767 | 0.1955 | 0.6164 | 0.6048 |
| DGT(SSA) | 1.6201 | **0.0295** | 1.5080 | 0.1981 | 2.2310 | 0.0835 | 0.6188 | 0.6032 |
| DGT(WAV) | 1.9514 | 0.0831 | 1.6192 | 0.1836 | 1.4557 | 0.2025 | 0.9852 | 0.5374 |
| Trend(EMD) | 0.6276 | 0.9689 | 0.6934 | 0.9399 | 0.9475 | 0.5892 | 0.5359 | 0.9499 |
| Trend(HEN) | 0.8384 | 0.6107 | 1.2114 | 0.2983 | 1.1247 | 0.3212 | 1.0686 | 0.4127 |
| Trend(HP) | 1.5676 | 0.1661 | 0.9720 | 0.4432 | 0.9772 | 0.4309 | 0.5552 | 0.6449 |
| Trend(LOESS) | 0.5487 | 0.9552 | 0.7468 | 0.7978 | 0.6677 | 0.8723 | 0.4067 | 0.9527 |
| Trend(MBA) | 0.1212 | 0.9416 | 0.4413 | 0.7235 | 0.4869 | 0.6147 | 0.5125 | 0.6739 |
| Trend(SSA) | 0.6943 | 0.9624 | 1.2123 | 0.1642 | 1.0918 | 0.3507 | 0.5151 | 0.9986 |
| Trend(WAV) | 1.5173 | **0.0434** | 1.2843 | 0.1547 | 1.4776 | 0.0646 | 1.1989 | 0.3094 |

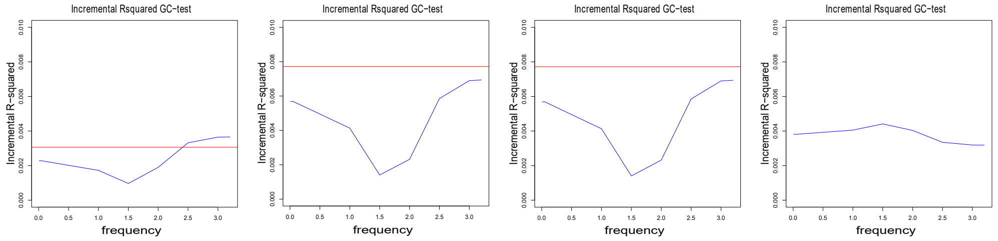
### Frequency Domain GC Test Results

The following figures present the frequency domain GC test results for DGT by each trend extraction methods respectively with specific results of all sub-samples. Identically, for each test, the optimal lag-structure is assured and having greater test statistics (blue) than the corresponding 5% critical values (red) indicates possible causal links from SS to GT within corresponding frequency range.



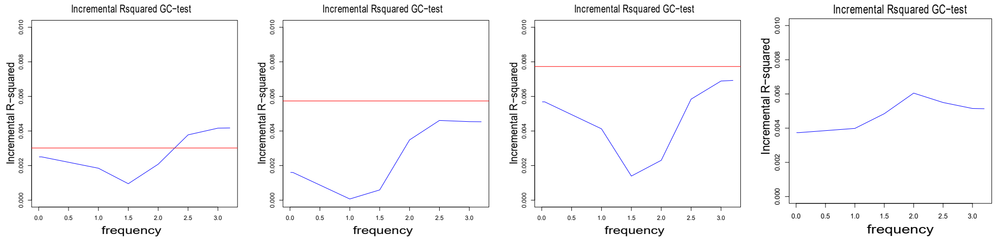
(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.3 Frequency Domain GC Test Results for DGT(MBA).



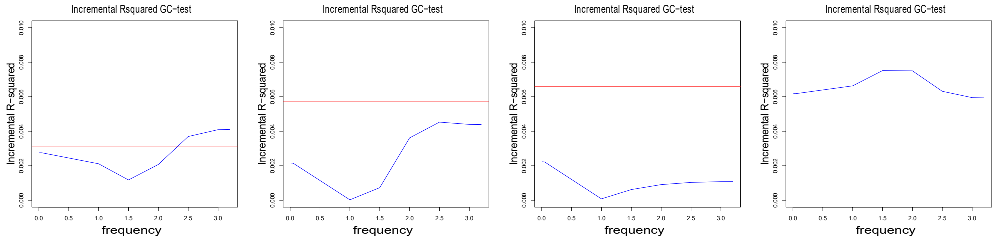
(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.4 Frequency Domain GC Test Results for DGT(EMD).



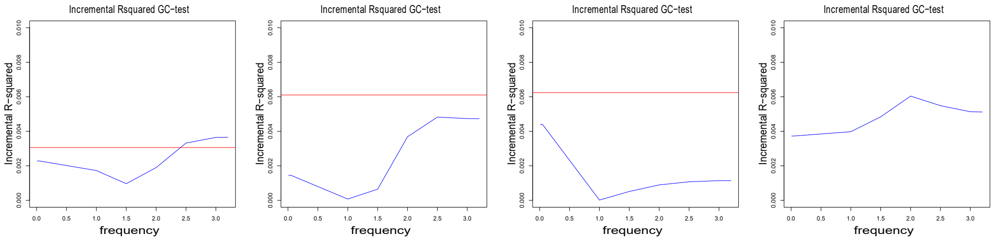
(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.5 Frequency Domain GC Test Results for DGT(SSA).



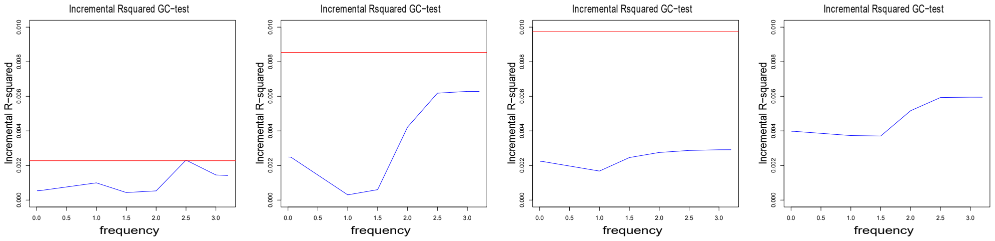
(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.6 Frequency Domain GC Test Results for DGT(WAV).



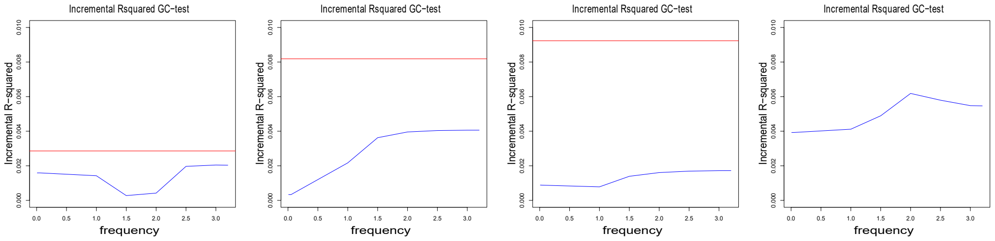
(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.7 Frequency Domain GC Test Results for DGT(LOESS).



(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.8 Frequency Domain GC Test Results for DGT(HEN).

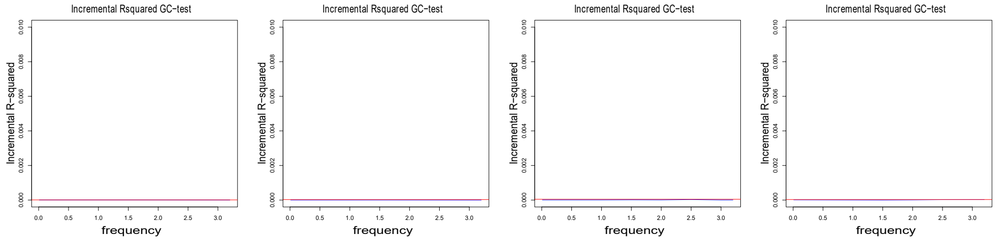


(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.9 Frequency Domain GC Test Results for DGT(HP).

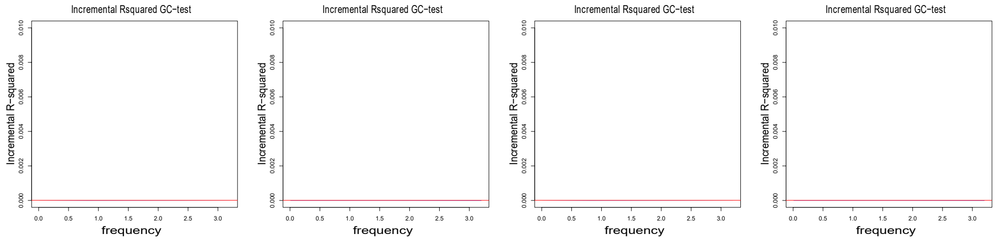
The results of DGT by seven different trend extraction methods above could not reflect significant differences on influencing the frequency domain GC test. In terms of the total sample, weak causality is identified in general except the cases of DGT by HP and HEN filter. The test statistics vary for each trend extraction method considering each sub-sample respectively, however, there are no significant evidences of showing causality for all sub-samples by different trend extraction methods. In general, given the evidences from seven different trend extraction methods adopted, the possible existing trend of GT do not have significant influence on frequency domain GC test and detrending cannot assist or affect significantly on frequency domain GC test for the research of causal link between SS and GT in climate change study.

Furthermore, the following figures present the frequency domain GC test results for trend extracted by each trend extraction methods respectively with specific results of all sub-samples, followed by the summary of all frequency domain GC test results that are listed in Table 7.2. It is noticed that in general no causality can be detected by the extracted trend series regardless of the sub-samples and trend extraction methods, except that the significant causal link at short cycle frequency is detected at sub-sample B of the trend series extracted by SSA.



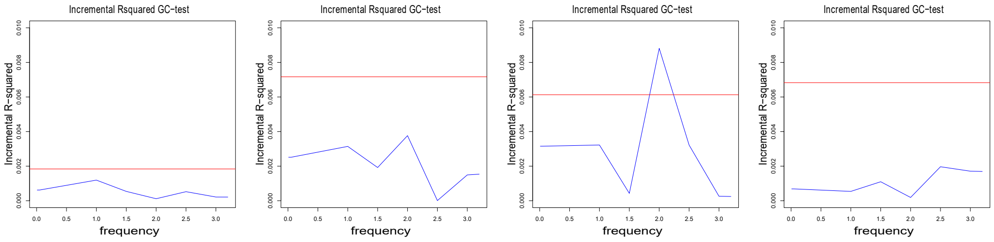
(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.10 Frequency Domain GC Test Results for Trend(MBA).



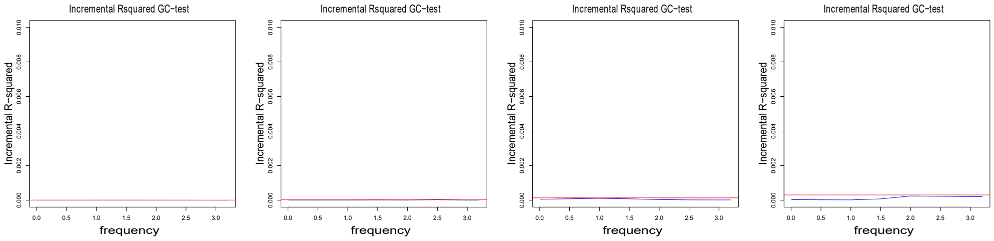
(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.11 Frequency Domain GC Test Results for Trend(EMD).



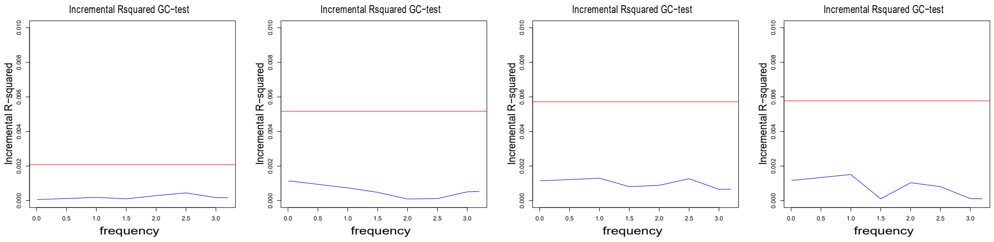
(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.12 Frequency Domain GC Test Results for Trend(SSA).



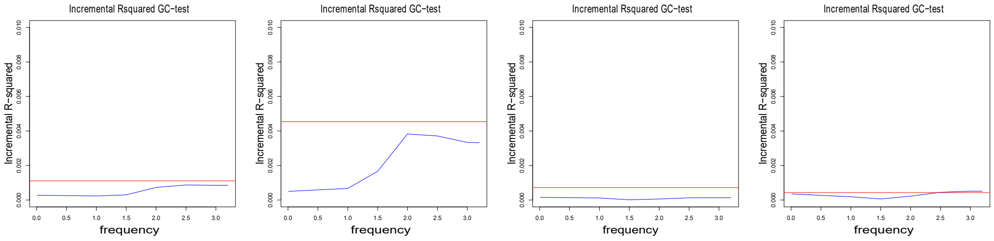
(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.13 Frequency Domain GC Test Results for Trend(WAV).



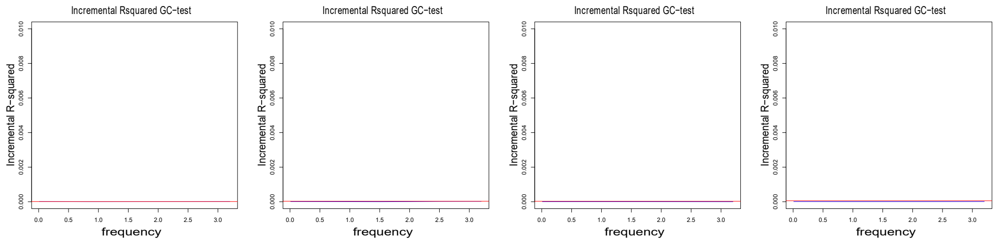
(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.14 Frequency Domain GC Test Results for Trend(LOESS).



(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.15 Frequency Domain GC Test Results for Trend(HEN).



(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.16 Frequency Domain GC Test Results for Trend(HP).

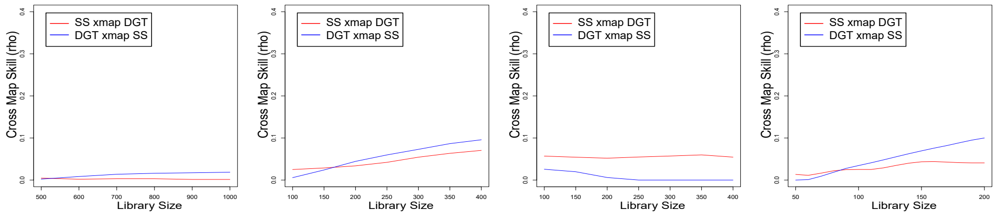
Table 7.2 Summary of Frequency Domain GC Test Results by Trend Extraction Techniques.

| | Total Sample | Sub-sample A | Sub-sample B | Sub-sample C |
|---|---|---|---|---|
| Tested Series | 1880:1-2015:5 | 1880:1-1936:2 | 1936:3-1986:11 | 1986:12-2015:5 |
| Original | YES(short cycle) | NO | NO | NO |
| DGT(MBA) | YES(short cycle) | NO | NO | NO |
| DGT(EMD) | YES(short cycle) | NO | NO | NO |
| DGT(SSA) | YES(short cycle) | NO | NO | NO |
| DGT(WAV) | YES(short cycle) | NO | NO | NO |
| DGT(LOESS) | YES(short cycle) | NO | NO | NO |
| DGT(HEN) | YES(week) | NO | NO | NO |
| DGT(HP) | NO | NO | NO | NO |
| Trend(MBA) | NO | NO | NO | NO |
| Trend(EMD) | NO | NO | NO | NO |
| Trend(SSA) | NO | NO | YES(short cycle) | NO |
| Trend(WAV) | NO | NO | NO | NO |
| Trend(LOESS) | NO | NO | NO | NO |
| Trend(HEN) | NO | NO | NO | NO |
| Trend(HP) | NO | NO | NO | NO |

### CCM Causality Test Results

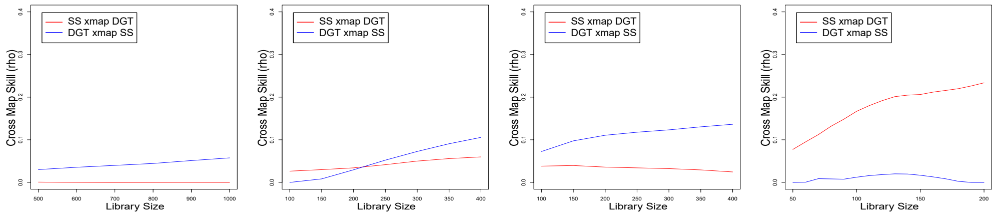
Following the CCM test for original series listed in Fig. 6.1 of Chapter 6 section 6.2, the figures listed below present all the CCM test results for DGT by different trend extraction methods. Identically, all tests are obtained by the optimal embedding dimension respectively; by optimal outcome based on cross validation results;

with identical library size within one corresponding sample size for the sake of comparisons.



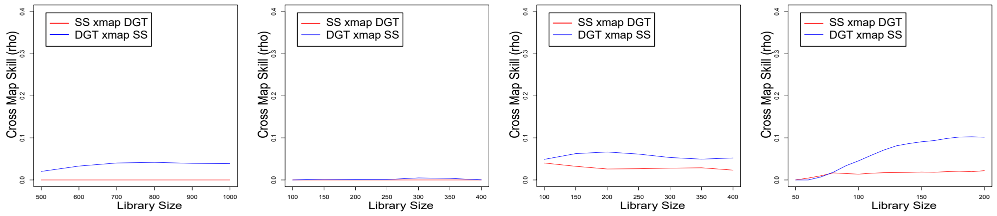
(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.17 CCM Result for DGT(MBA).



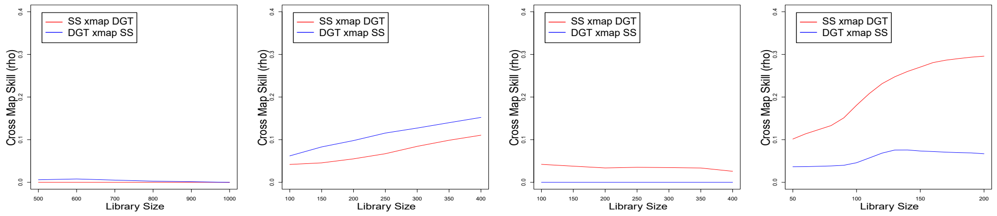
(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.18 CCM Result for DGT(EMD).



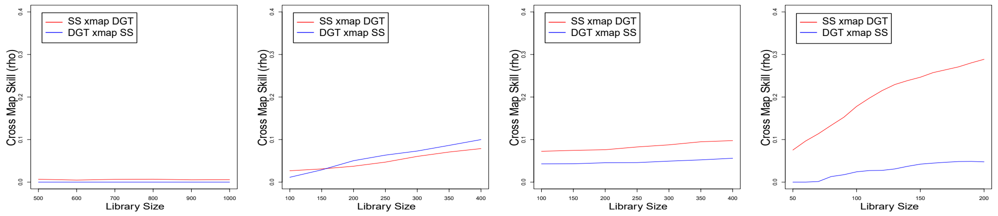
(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.19 CCM Result for DGT(SSA).



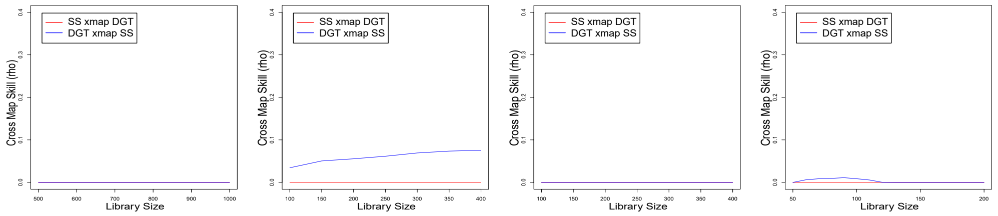
(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.20 CCM Result for DGT(WAV).



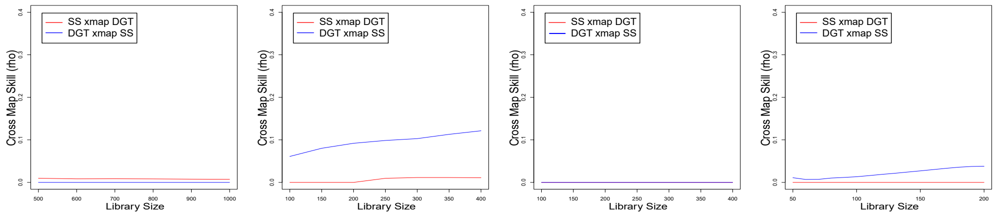
(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.21 CCM Result for DGT(LOESS).



(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

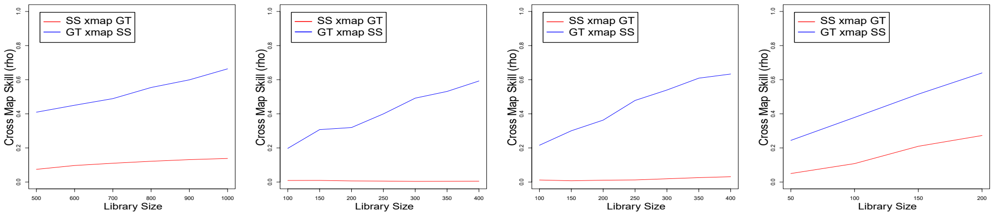
Fig. 7.22 CCM Result for DGT(HEN).



(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

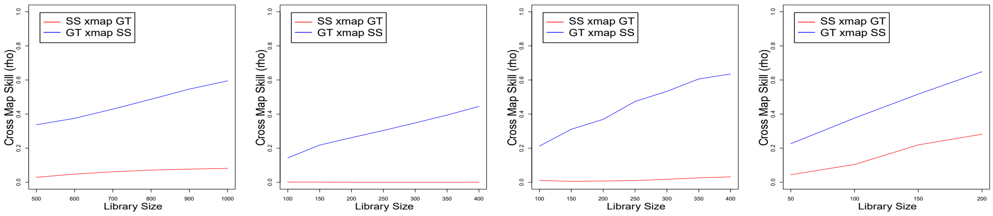
Fig. 7.23 CCM Result for DGT(HP).

Furthermore, the figures listed below show all the CCM test results for trend series extracted by different trend extraction methods respectively with considerations of both total sample and sub-samples.



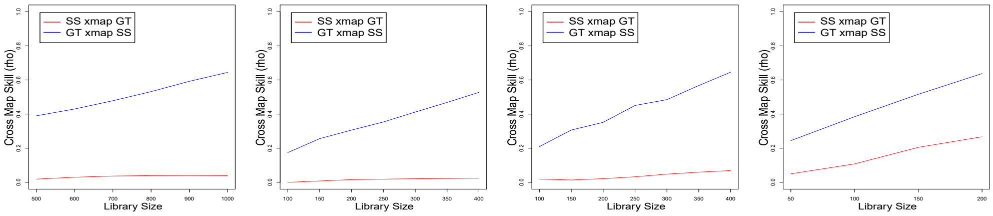
(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.24 CCM Result for Trend(MBA).



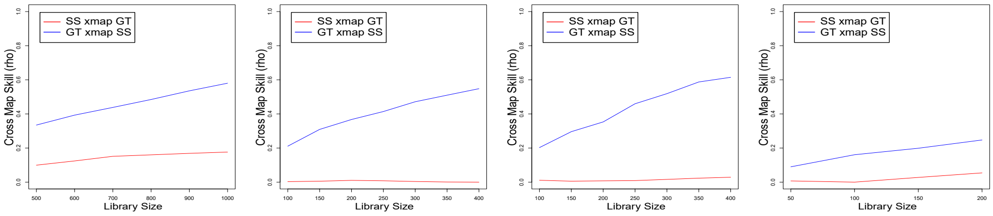
(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.25 CCM Result for Trend(EMD).



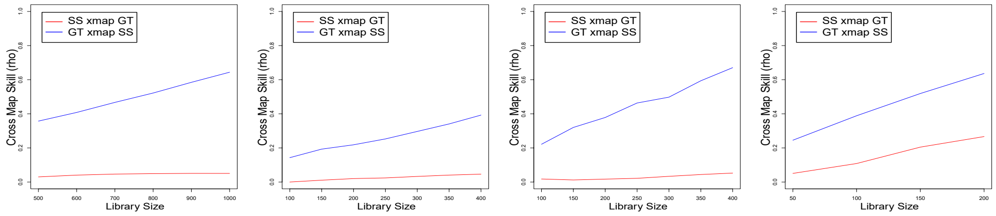
(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.26 CCM Result for Trend(SSA).



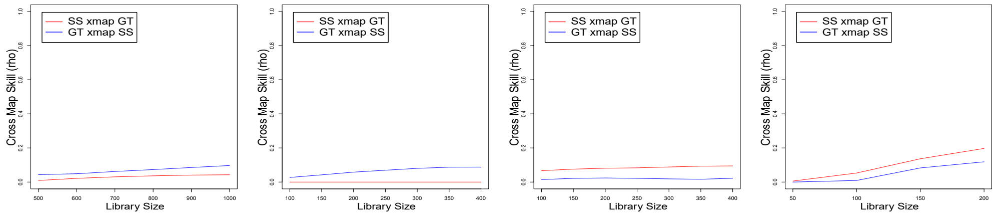
(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.27 CCM Result for Trend(WAV).



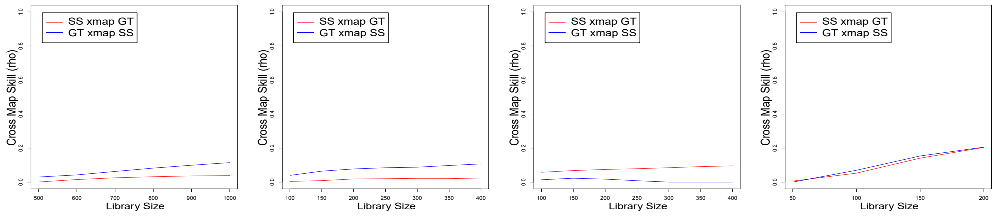
(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.28 CCM Result for Trend(LOESS).



(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.29 CCM Result for Trend(HEN).



(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.30 CCM Result for Trend(HP).

If the CCM results of original GT are recalled for comparison, in which, the total sample and sub-sample C both indicate significant causation from SS to GT, whilst sub-sample A reflects no causation and sub-sample B shows wrong direction. In terms of the DGT and trend series in line with the corresponding CCM results listed above, significant differences are generally obtained among seven different trend extraction methods. In order to provide further analyses on the corresponding effects of each trend extraction method on CCM causality test, the results are summarized below in Table 7.3.

Table 7.3 Summary of CCM Causality Test Results by Trend Extraction Techniques.

| | Total Sample | Sub-sample A | Sub-sample B | Sub-sample C |
|---|---|---|---|---|
| Tested Series | 1880:1-2015:5 | 1880:1-1936:2 | 1936:3-1986:11 | 1986:12-2015:5 |
| Original | YES | NO | Wrong Direction | YES |
| DGT(MBA) | YES(very weak) | YES(weak) | Wrong Direction | YES(weak) |
| DGT(EMD) | YES | YES(weak) | YES | Wrong Direction |
| DGT(SSA) | YES | YES(very weak) | YES | YES |
| DGT(WAV) | YES(very weak) | YES | Wrong Direction | Wrong Direction |
| DGT(LOESS) | NO | YES(weak) | Wrong Direction | Wrong Direction |
| DGT(HEN) | NO | YES | NO | YES(very weak) |
| DGT(HP) | NO | YES | NO | YES(weak) |
| Trend(MBA) | YES(strong) | YES(strong) | YES(strong) | YES(strong) |
| Trend(EMD) | YES(strong) | YES(strong) | YES(strong) | YES(strong) |
| Trend(SSA) | YES(strong) | YES(strong) | YES(strong) | YES(strong) |
| Trend(WAV) | YES(strong) | YES(strong) | YES(strong) | YES |
| Trend(LOESS) | YES(strong) | YES | YES(strong) | YES(strong) |
| Trend(HEN) | YES(week) | YES(week) | Wrong Direction | Wrong Direction |
| Trend(HP) | YES(week) | YES(week) | Wrong Direction | YES(very week) |

Evidence of significant causality in general is found in terms of trend series, which strongly reflect the causal link from SS to the emerging trend of global warming. Only trend by HP and HEN show relatively weaker causality, while the corresponding DGT by HEN and HP also fail on providing positive results of existing causal links. However, regarding the total sample that was proved significant by original series, DGT (as well as trend series) by MBA, EMD, SSA and WAV continue to hold significant results. Thus, DGT by LOESS, HEN and HP may work fine on extracting the existing trend, but it will possibly remove or reduce the causal effects between SS and GT to be captured by CCM.

Regarding the sub-sample A, all CCM results in general show significant evidences of causality, regardless of whether the level of causation is weak or strong. This is an impressive effect detected as no causal link can be identified when the original series are considered. It can be concluded that the existing trend series and DGT together lead to weaken or mislead the results for sub-sample A on causality analyses. This can be generally improved to obtain more significant outcomes by extracting the trend with all representative trend extraction methods listed here (regardless the level of significance for different methods).

In terms of sub-sample B, which indicated a wrong direction causality in case of the original series, only DGT by EMD and SSA show positive results indicating causality. The others fail to detect or, even worse, provide misleading results of wrong direction of causation.

For the most recent sub-sample C that possibly reflects the most meaningful conclusion, DGT by SSA manage to hold the significant result that keep the consistency of the results by original series, more importantly, with reasonable level of significance. The CCM results of DGT by SSA show emerging causality relationship from SS to GT that also prove the significant predictive ability of SS on GT.

As an advanced non-parametric subspace-based technique for causality analysis, CCM outperforms the empirical GC approaches with not only the original series but also the DGT and extracted trend series. Nevertheless, the existing trend of GT has also been proved affecting the outcomes of CCM. According to the comparisons of different trend extraction methods in terms of the study of SS and GT in climate change, SSA outperforms the others on providing better preprocessed series for CCM test with significant causal link detected regardless of the DGT and trend series for all sub-samples as well as the total sample. More importantly, it indicates the emerging causal effects from SS to GT, which contributes on explaining the tendency of global warming recent decades. Therefore, the hybrid SSA-CCM approach is proved a robust, reliable method that stands out among all considered trend extraction techniques.

### 7.2.3 Data Decomposition by SSA and Corresponding Effects on Causality Analysis

All previous comparisons are conducted by the original SS together with the trend and DGT respectively, in line with the SSA technique is found the most appropriate trend extraction method on data preprocessing for causality analyses in climate change study. In this section, both the SS and GT are further decomposed by SSA into representative components: trend, cycle and noise. All causality tests are then obtained by different components respectively to provide further comprehensive understanding of the causal link between SS and GT, which may contribute on target the most significant component that dominates the causality analysis in a complex system like climate changes study.

#### Data Decompositions by SSA

As can be seen in Fig. 7.31 and Fig. 7.32, the original GT and SS are decomposed by SSA into the representative components of trend, cycle and noise respectively. Note that the window length is selected as $N/2$ where $N$ is the number of observations, combinations of eigenvalues are selected due to the features of representative components[3].

[3] All operations are conducted by R with the corresponding package (RSSA), there are also different alternative packages available in R, as well as another software CaterpillarSSA.

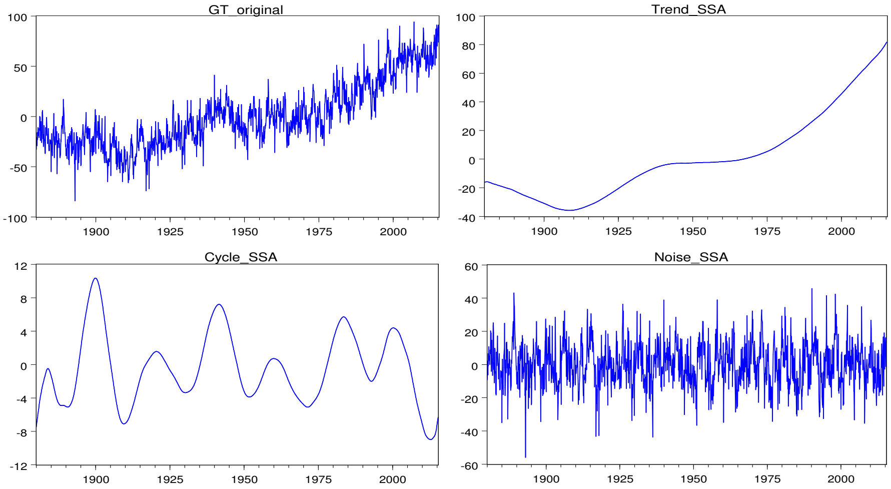


Fig. 7.31 Decompositions of GT by SSA (1880M01 to 2015M05) .

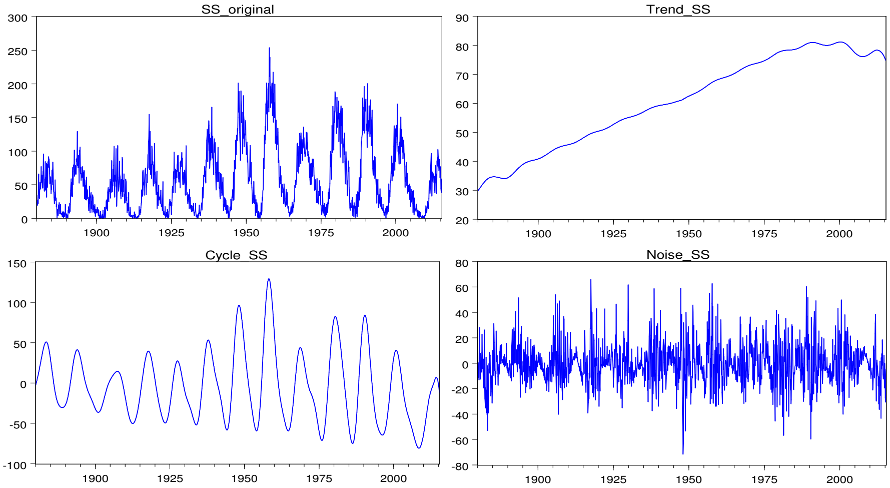
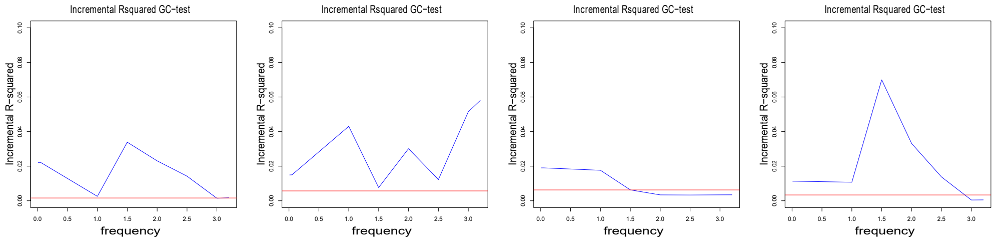


Fig. 7.32 Decompositions of SS by SSA (1880M01 to 2015M05) .

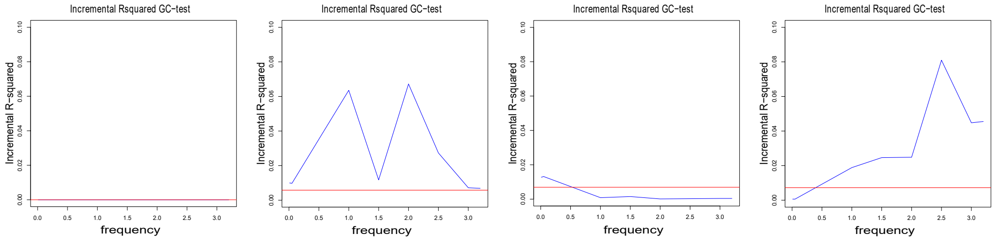
### Causality Test Results of the Data Decompositions by SSA

The causality tests are conducted on each group of components respectively with considerations of not only the total sample but also all sub-samples. The detailed results of frequency domain and CCM tests are listed below, followed by the brief summary of causality test results by components in Table 7.4.



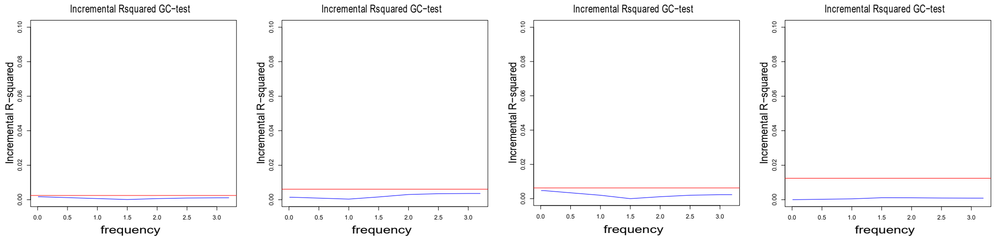
(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.33 Frequency Domain Causality Result for Trends.



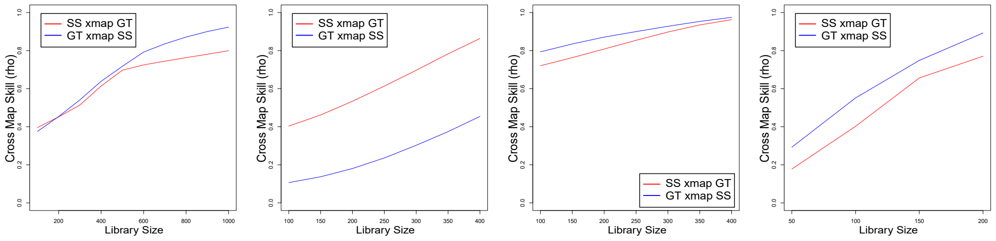
(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.34 Frequency Domain Causality Result for Cycles.



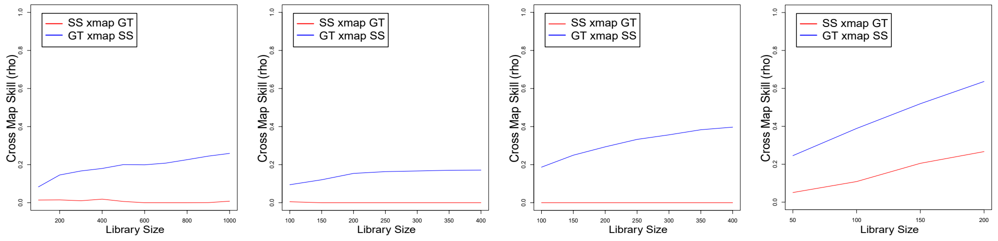
(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.35 Frequency Domain Causality Result for Noises.



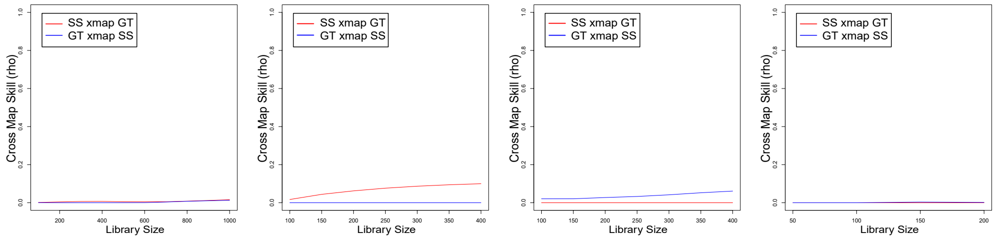
(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.36 CCM Causality Result for Trends.



(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.37 CCM Causality Result for Cycles.



(a) total sample (b) sub-sample A (c) sub-sample B (d) sub-sample C

Fig. 7.38 CCM Causality Result for Noises.

Table 7.4 Summary of Causality Tests Results of Data Decompositions by SSA.

| | Decomposition | Total Sample<br>1880:1-2015:5 | Sub-sample A<br>1880:1-1936:2 | Sub-sample B<br>1936:3-1986:11 | Sub-sample C<br>1986:12-2015:5 |
|---|---|---|---|---|---|
| Time Domain | Trend | YES(5.2009***) | YES(8.5674***) | YES(4.4397***) | YES(3.3258***) |
| | Cycle | YES(11.5135***) | YES(7.6866***) | NO(0.9371) | YES(9.7739***) |
| | Noise | NO(1.7220) | NO(0.9646) | NO(1.5214) | NO(0.4616) |
| Frequency Domain | Trend | YES | YES | YES | YES |
| | Cycle | NO | YES | YES | YES |
| | Noise | NO | NO | NO | NO |
| CCM | Trend | YES(week) | Wrong Direction | YES(week) | YES |
| | Cycle | YES | YES | YES | YES |
| | Noise | NO | Wrong Direction | YES(week) | NO |

For the data decompositions that different components represent representative features, it is of note that even the generally accepted parametric method – time domain GC test – achieves the significant results for both trend and cycle series (except for the sub-sample B of the cycle component). For the frequency domain GC and CCM causality tests, the causal link is generally proved significant for both the trend and cycle series. Only the total sample of cycle fails to be detected by frequency domain GC test, and sub-sample A of trend component shows misleading results of opposite direction of causation by CCM. The causality is overwhelmingly rejected for the noise component by both time domain and frequency domain GC tests, except the sub-sample B by CCM test. In case of the trend component, the overall significant results indicate the long term causal effect from SS to GT, while the CCM results provide comprehensive analyses that the predictive ability from the other direction is also relatively strong. This is due to the feature of the component and it is worth highlighting that the results by CCM (considering the gap between two cross mapping skill indices) reflect the wrong direction - very weak SS to GT - slightly stronger SS to GT across the sub-samples in line with week causation detected over the total sample.

In terms of the cycle component, the results by different causality tests are generally significant. It is found that no causality can be detected by time domain GC test in sub-sample B. The corresponding results by frequency domain GC test is also weak (only in limited range of long cycle). CCM is the most sensitive technique that shows the highly significant causality of sub-sample B comparing to the other tests. It is worthwhile to note that the week causality detected by CCM on the noise component at sub-sample B. This possibly indicates some irregular or nonlinear patterns of causal relationships that is also successfully captured by CCM. More importantly, in case of the CCM results of the cycle component across the time scale, it again proves that the emerging causal effects of SS on GT. This indicates the tendency of global warming due to sunspot activity, especially for the recent decades.

In general, by decomposing the data into representative components with SSA technique, even the applicability of the basic parametric method (time domain GC test) is significantly improved. The causality conclusion is overwhelmingly obtained, while the CCM shows the strongest ability of capturing the possible causal effects regardless of the time scale and components. The data preprocessing is again proved absolutely necessary for the causality analyses of the complex system like climate change, and the clear evidence here in this research indicate that data preprocessing is able to significantly improve the performances of generally all causality tests, whilst the SSA-CCM hybrid approach stands the strongest position among all combinations with proved remarkable capabilities on causality analysis.

### 7.2.4 Discussion

Chapter 5 and 6 illustrate that these non-parametric subspace-based methods outperformed the generally accepted techniques when investigating the causal link between SS and GT. Section 7.2 here further extends this literature by considering the possible effects of the existing trend on causality detection technique performances. As the initial attempt to my best knowledge, 7 different mainstream well-established trend extraction methods are incorporated and evaluated for comprehensive analysis of the effects of trend on causality analysis by both empirical and advanced techniques. Moreover, for the first time, the data is further decomposed by SSA into representative components of trend, cycle and noise for obtaining the most detailed knowledge of the relationship between data components and causality analysis techniques.

This section successfully answers a crucial question that the existing trend has impacts on the causality analyses outcomes, even for the advanced non-parametric techniques that already outperformed the generally accepted methods with significant advantages. It is also confirmed that trend extraction will absolutely contribute on assisting the causality analyses in climate change study, whilst the causality analyses will still lead to conclusions with great contrast by using different trend extraction methods. Among which, the SSA trend extraction is identified as the most reliable method for data preprocessing, while CCM shows outstanding performance among all causality tests adopted. The emerging causal effects from SS to GT, especially for recent decades, are overwhelmingly proved, which reflects the better understanding of the tendency of global warming. Broadly, this research proves to contribute on the current literature of causality analysis with the most detailed causality analysis by trend extraction techniques as well as representative components. Nevertheless, this research verifies and demonstrates the powerful SSA-CCM hybrid approach that shows outstanding capabilities in causality analysis due to its robust performances on not only the dominating components like trend and cycle but also the less important component of noise.

## 7.3 Gene Regulatory Role

Following the application of gene regulatory role in Chapter 6 section 6.2.3, CCM has been proved a satisfying causality analysis tool on identifying the regulatory links between original gene expression profiles whilst the empirical approaches fail to detect. Considering the fact that gene expression profiles are exceedingly noisy (Golyandina et al., 2012), there are still no misleading conclusions of causality demonstrated by CCM. However, it has been noted that some of the levels of significance are relatively small. Therefore, here in this section, the SSA-CCM hybrid approach is again applied. Specifically, SSA works as the data preprocessing technique to filter the noise before the causality analysis step is conducted by CCM. Both empirical GC approaches are also considered for causality detection by the corresponding filtered series as contextual comparisons. This section of research aims to further validate the SSA-CCM hybrid approach and its performance on causality analysis for data containing complex noise. Moreover, broadly, it contributes on the literature of causality analysis through the better understanding of the effect of noise on causality detection by the example of gene regulatory role detection.

### 7.3.1 Data and Noise Filtering

Note that the details of brief literature review and data can be found in section 6.2.3 in Chapter 6, which is therefore not reproduced here. Also, the identical groups of data are adopted here in this section along with the comparisons of the empirical GC approaches. As it has been shown in Fig. 6.6, the profile achieved by flourocence antibodies technique is highly volatile[4] and in such cases, establishing a cause-and-effect relationship is more challenging and demands applying a noise filtering step prior to causality analysis, for which, SSA is applied here to produce noise filtered series, and a few more gene expression signal extraction examples by SSA can be found in (Ghodsi et al., 2015b,c; Hassani and Ghodsi, 2014; Holloway et al., 2006).

Fig. 7.39 illustrates the output of signal extraction from the original noise data by SSA[5]. Note that the x-axis shows the position of the nuclei along the A-P axis of the embryo and y-axis shows the fluorescence intensity level. It has been evident that the SSA method provides a relatively smooth signal line with correlation below 0.10 which credits the satisfactory level of separation between noise and signal using SSA (Ghodsi et al., 2015b).

[4] Although confocal scanning microscopy is a generally employed technique for measuring the gene expression profiles, its use in systems biology studies presents a number of challenges such as the considerable amount of noise entering data after quantifying the fluorescence intensity. Possible errors in instrument functionality, sample preparation and mathematical treatment of data have been considered as the most common sources of noise (Myasnikova et al., 2009).

[5] The optimal noise filtering by SSA on these gene expression profiles follow the method proposed by Alharbi and Hassani (2016) and Alharbi et al. (2016), where the optimal number of eigenvalues are selected based on the skewness, coefficient of variation and correlation coefficient between eigenvalues.

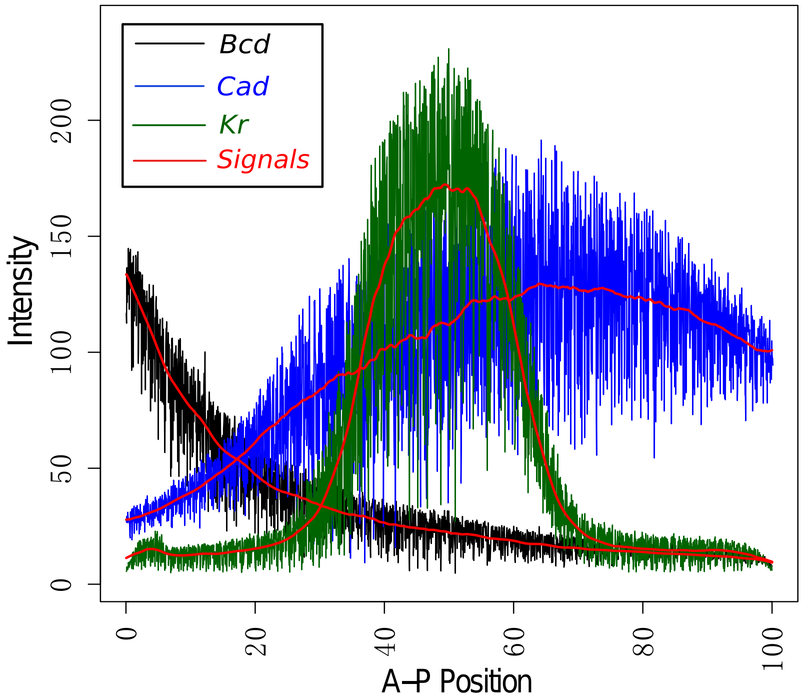


Fig. 7.39 A Typical Example of Bcd, Cad and Kr for Embryo *ms26* at Time Class 14(1) with Extracted Signals in Red.

### 7.3.2 Causality Test Results

This section provides a summary of the CCM causality test before and after filtering the expression profiles using SSA, along with the contextual comparisons of the empirical GC approaches. Table 7.5 illustrates the findings of the causality detection analysis on Bcd and Cad profiles. As previously clarified in Chapter 6, note that for all evaluations, the corresponding requirements for each test are satisfied, also the optimal outcomes are selected. Moreover, the corresponding optimal lag is selected for obtaining the test result for each group of profiles; the *p*-values reported for time domain GC test are the average *p*-values attained for embryos studied each corresponding time class. It is of note that even the test results for original series have been listed in Chapter 6, which are still reproduced here for the sake of clear and convenient comparison.

Table 7.5 Causality Tests Results for Noisy and Filtered Bcd on Cad Profiles.

| Time Class | Time Domain GC | | | | Frequency Domain GC | | CCM | |
|---|---|---|---|---|---|---|---|---|
| | Noisy Series | | Filtered Series | | Noisy Series | Filtered Series | Noisy Series | Filtered Series |
| | YES/NO | p-value | YES/NO | p-value | YES/NO | YES/NO | YES/NO | YES/NO |
| 10 | NO | 0.68 | NO | 0.45 | NO | YES | YES | YES |
| 11 | NO | 0.71 | NO | 0.33 | NO | YES | YES | YES |
| 12 | NO | 0.89 | NO | 0.32 | NO | YES | YES | YES |
| 13 | NO | 0.89 | NO | 0.24 | NO | YES | YES | YES |
| 14(1) | NO | 0.95 | YES | 0.05 | NO | YES | YES | YES |
| 14(2) | NO | 0.98 | YES | 0.04 | NO | YES | YES | YES |
| 14(3) | NO | 0.98 | YES | 0.01 | NO | YES | YES | YES |
| 14(4) | NO | 0.94 | YES | 0.01 | NO | YES | YES | YES |
| 14(5) | NO | 0.95 | YES | 0.00 | NO | YES | YES | YES |
| 14(6) | NO | 0.96 | YES | 0.00 | NO | YES | YES | YES |
| 14(7) | NO | 0.81 | YES | 0.00 | NO | YES | YES | YES |
| 14(8) | NO | 0.79 | YES | 0.04 | NO | YES | YES | YES |

Note: "Yes" stands for the detected regulatory link and "No" means the regulatory link could not be detected by the adopted test.

According to Table 7.5, it is evident that there is a significant difference in results before and after reducing the noise from the profiles. The regulatory link between Bcd and Cad can be detected by neither time domain nor frequency domain tests in the presence of noise. Accordingly, it is clear that the filtering capability displayed by SSA is indeed advantageous for causality detection analysis.

Nevertheless, as can be seen, the feasibility of capturing the regulatory link for CCM method has not been affected by noise and the results achieved by this test confirm the regulatory relationship between Bcd and Cad in expression profiles with and without noise. However, regardless of the time class, the index representing the ability of cross mapping is relatively smaller on average for noisy series than filtered series.

It is of note that the length of the data under study vary between different time classes. Time class 10 to 13 and 14(7-8) have shorter lengths comparing to the time classes 14(1-6), which may be the reason of getting slightly smaller p-values for time class 11 to 13 and 14(8) comparing to the rest of the sub classes of time class 14. Yet, the frequency domain test shows less sensitivity to the data length possibly because this method identifies the possible regulative link for each individual frequency component rather than the entire series.

Furthermore, the p-values for both noisy and filtered data of all the embryos in different time classes are summarised in Fig. 7.40 and Fig. 7.41 as box and whisker diagram respectively. They follow the standard formate of box plot on displaying the distribution of the p-values based on maximum, upper quartile, median, lower quartile, and minimum. Specifically, circle refers to corresponding outlier that is more/less than 1.5 times of upper/lower quartile; the central rectangle spans the upper quartile to the lower quartile; the segment inside the rectangle indicates the median; whiskers above and below the box refer to the maximum and minimum. A close look at Fig. 7.40 and Fig. 7.41 suggests that the time domain GC test cannot detect any regulatory link in the presence of noise, while the results for filtered series are significant and more consistent especially for those time classes after 14(1). Comparing the p-values illustrated in Fig. 7.40 and Fig. 7.41, it is evident that the length of the series and level of intensities have more effect on the result of the noisy data than the filtered one as the p-values in Fig. 7.40 are getting more insignificant for the final subclasses of time class 14, where there is a decreasing pattern for these two parameters in the expression profiles. Likewise, for the frequency domain GC test, the links have been detected for all the filtered series, whilst there is no regulatory relationship detected for non-filtered ones.

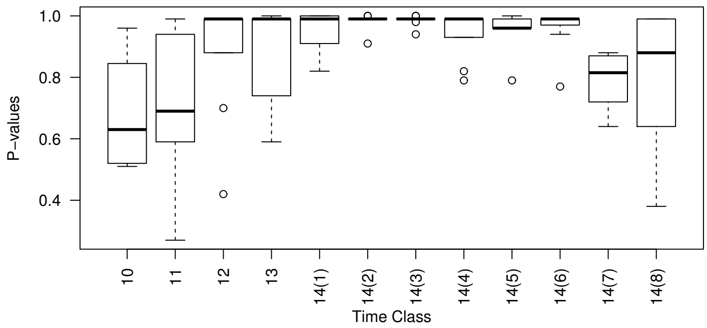


Fig. 7.40 Box Plots of Time Domain GC Test P-values for Noisy Series.

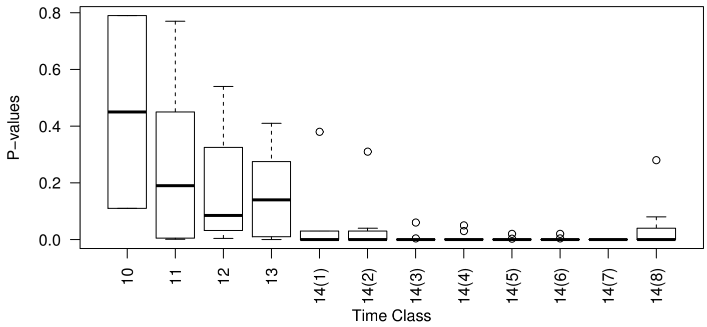


Fig. 7.41 Box Plots of Time Domain GC Test P-values for Filtered Series.

Table 7.6 and Table 7.7 present the results of the conducted analysis to detect the regulatory link between Bcd and Kr profiles and Cad and Kr profiles respectively. As can be seen, reducing the noise level is an essential step in detecting the regulatory link using the time domain and frequency domain tests. Similar to the results reported in Table 7.5, CCM method can again efficiently identify the regulatory relationship even in the presence of noise.

Table 7.6 Causality Tests Results for Noisy and Filtered Bcd on Kr Profiles.

| Time Class | **Time Domain GC** | | | | **Frequency Domain GC** | | **CCM** | |
|---|---|---|---|---|---|---|---|---|
| | Noisy Series | | Filtered Series | | Noisy Series | Filtered Series | Noisy Series | Filtered Series |
| | YES/NO | p-value | YES/NO | p-value | YES/NO | YES/NO | YES/NO | YES/NO |
| 12 | NO | 0.71 | NO | 0.15 | NO | YES | YES | YES |
| 13 | NO | 0.66 | YES | 0.04 | NO | YES | YES | YES |
| 14(1) | NO | 0.89 | YES | 0.03 | NO | YES | YES | YES |
| 14(2) | NO | 0.93 | YES | 0.01 | NO | YES | YES | YES |
| 14(3) | NO | 0.97 | YES | 0.01 | NO | YES | YES | YES |
| 14(4) | NO | 0.94 | YES | 0.00 | NO | YES | YES | YES |
| 14(5) | NO | 0.95 | YES | 0.00 | NO | YES | YES | YES |
| 14(6) | NO | 0.92 | YES | 0.00 | NO | YES | YES | YES |
| 14(7) | NO | 0.81 | YES | 0.00 | NO | YES | YES | YES |

Note: "Yes" stands for the detected regulatory link and "No" means the regulatory link could not be detected by the adopted test.

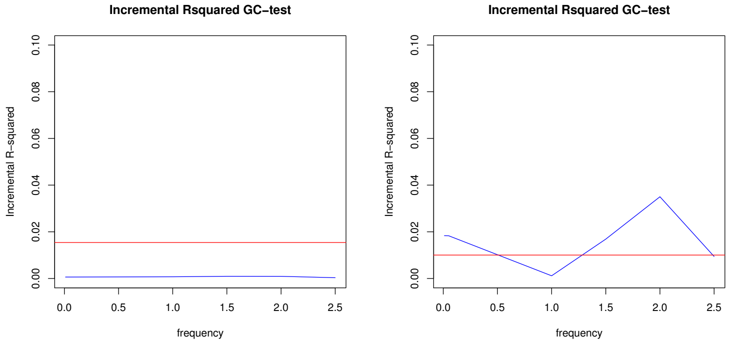
Table 7.7 Causality Tests Results for Noisy and Filtered Cad on Kr Profiles.

| Time Class | **Time Domain GC** | | | | **Frequency Domain GC** | | **CCM** | |
|---|---|---|---|---|---|---|---|---|
| | Noisy Series | | Filtered Series | | Noisy Series | Filtered Series | Noisy Series | Filtered Series |
| | YES/NO | p-value | YES/NO | p-value | YES/NO | YES/NO | YES/NO | YES/NO |
| 12 | NO | 0.39 | NO | 0.25 | NO | YES | YES | YES |
| 13 | NO | 0.78 | NO | 0.11 | NO | YES | YES | YES |
| 14(1) | NO | 0.84 | YES | 0.05 | NO | YES | YES | YES |
| 14(2) | NO | 0.89 | YES | 0.03 | NO | YES | YES | YES |
| 14(3) | NO | 0.94 | YES | 0.01 | NO | YES | YES | YES |
| 14(4) | NO | 0.91 | YES | 0.01 | NO | YES | YES | YES |
| 14(5) | NO | 0.87 | YES | 0.00 | NO | YES | YES | YES |
| 14(6) | NO | 0.82 | YES | 0.00 | NO | YES | YES | YES |
| 14(7) | NO | 0.75 | YES | 0.00 | NO | YES | YES | YES |

Note: "Yes" stands for the detected regulatory link and "No" means the regulatory link could not be detected by the adopted test.

Fig. 7.42, Fig. 7.44 and Fig. 7.46 depict an example of the results obtained by frequency domain GC test for Bcd–Cad, Bcd–Kr and Cad–Kr profile pairs respectively

[6]. As the component of each frequency is considered separately for identifying possible causal link, the impacts of relatively less information are significantly reduced. However, there are overwhelming evidences of filtered series showing minor differences between the test statistics and the 5% critical value, which consequently indicate causality.

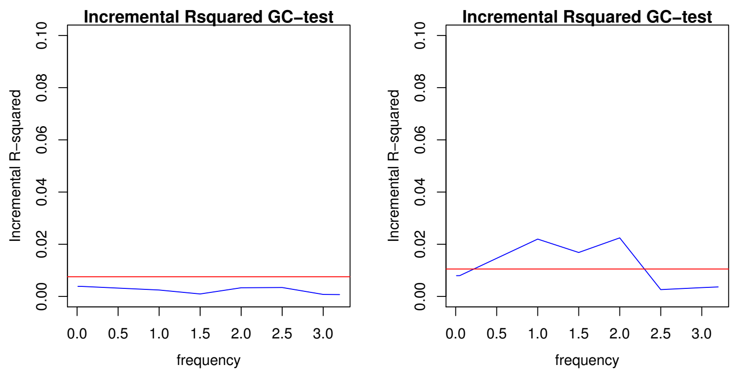


(a) Noisy bcd on cad (b) Filtered bcd on cad

Note: The blue line represents the statistic test of each specific frequency, and the red line represents the 5% critical value for all the frequencies.

Fig. 7.42 Example Frequency Domain GC Test Results for Noisy and Filtered Bcd and Cad (TC 11).

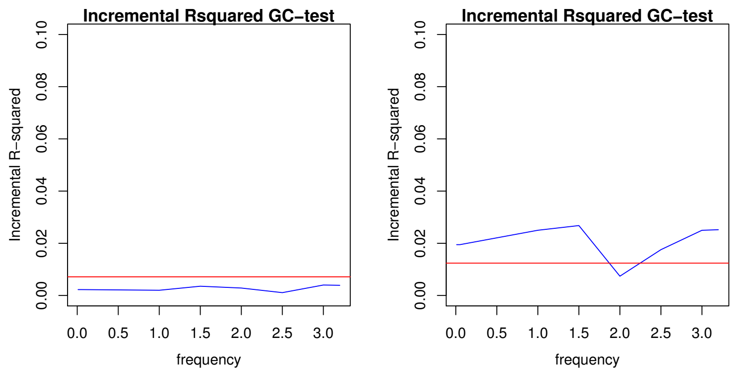
[6] The frequency domain GC test results for all considered pairs of filtered genes related to all different time classes can be found in Appendix D.5



(a) Noisy bcd on kr (b) Filtered bcd on kr

Note: The blue line represents the statistic test of each specific frequency, and the red line represents the 5% critical value for all the frequencies.

Fig. 7.44 Example Frequency Domain GC Test Results for Noisy and Filtered Bcd and Kr (TC 12).



(a) Noisy-t12-cad on kr (b) Filtered-t12-cad on kr

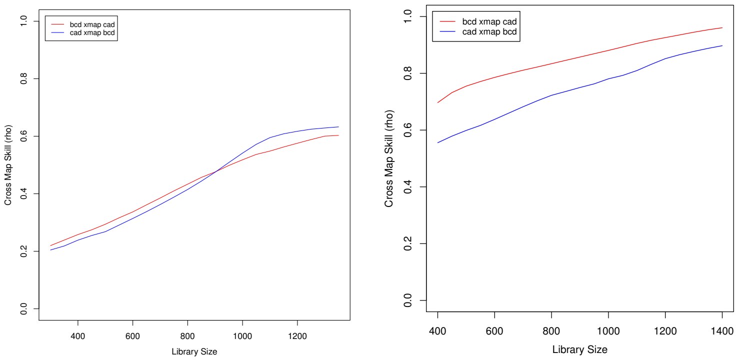
Note: The blue line represents the statistic test of each specific frequency, and the red line represents the 5% critical value for all the frequencies.

Fig. 7.46 Example Frequency Domain GC Test Results for Noisy and Filtered Cad and Kr (TC 12).

Fig. 7.48, Fig. 7.50 and Fig. 7.52 represent the examples of the CCM test result for Bcd–Cad, Bcd–Kr and Cad–Kr before and after filtering the profiles [7]. The results of CCM reflect close relationships between Bcd and Cad with and without

[7] The CCM test results for all considered pairs of filtered genes related to all different time classes can be found in Appendix D.6.

filtering, whilst Bcd shows more significant relationship with Kr comparing to Cad for both original and filtered data. The cross-mapping abilities of Bcd and Cad on Kr are fairly similar, however, Kr clearly indicates higher reconstruction ability on Bcd comparing to Cad. In more details regarding the relationship between Bcd and Cad, considering the average reconstruction ability represented by $\rho$, it is suggested that CCM is not affected by the smaller length of the series related to the initial time. However, the increasing pattern of the average level of cross-mapping ability up to time class 14(3), which follows by a decreasing trend for the rest of the subclasses, indicates less accuracy of the results for higher time classes. The approximate average value of $\rho$ over 0.5 for noisy series indicates significant cross-mapping (or reconstruction) ability to identify the causal links. Correspondingly, a average is found to be approximately over 0.8, which reflects stronger causal links detected between Bcd and Cad after filtering. In terms of the relationships between Bcd and Kr, the filtered series can be beneficial for identifying slightly more significant causality relationship and both original and filtered series indicate stronger cross-mapping ability from Kr to Bcd, which means that Bcd shows much stronger causal effect on Kr than the other way around. Regarding the Cad and Kr, the causality relationship identified are less significant comparing to the other pairs studied, and it still confirms the stronger performance of the filtered series with a average about 0.4 comparing to the average of 0.2 for original series.

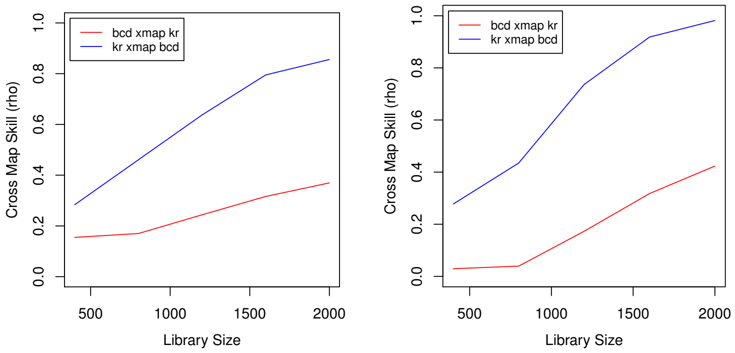


(a) Noisy-t14(8)-ccm

(b) Filtered-t14(8)-ccm

Note: The red line indicates the reconstruction ability of Bcd crossmap Cad, while the blue line represents the performance of Cad on crossmapping Bcd.

Fig. 7.48 Example CCM Test Results for Noisy and Filtered Bcd and Cad (TC 14(8)).

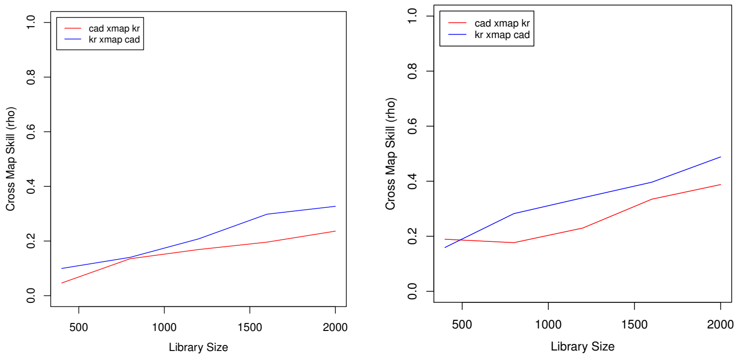


(a) Noisy-t14(7)-ccm

(b) Filtered-t14(7)-ccm

Note: The red line indicates the reconstruction ability of Bcd crossmap Kr, while the blue line represents the performance of Kr on crossmapping Bcd.

Fig. 7.50 Example CCM Test Results for Noisy and Filtered Bcd and Kr (TC 14(7)).



(a) Noisy-t14(5)-ccm (b) Filtered-t14(5)-ccm

Note: The red line indicates the reconstruction ability of Cad crossmap Kr, while the blue line represents the performance of Kr on crossmapping Cad.

Fig. 7.52 Example CCM Test Results for Noisy and Filtered Cad and Kr (TC 14(5)).

### 7.3.3 Discussion

This section further extend the CCM implementation on gene regulatory role detection in Chapter 6 through the SSA-CCM hybrid approach that considers the data preprocessing step prior to the causality analysis so to eliminate the possible influences of complex noise. This research is the first attempt of incorporating gene regulatory role detection with the advanced subspace-based causality detection techniques. Satisfying performances are achieved by the original data alone with CCM, and further improvements are established by introducing the SSA-CCM hybrid approach. In general, it proves that the existing complex noise can influent the causality analysis outcome, even the advanced subspace-based technique shows relatively low index that reflecting level of causality. It is also confirmed that noise filtering will absolutely contribute on assisting the causality analyses in gene regulatory role study, even the empirical approaches can benefit from the filtered series so to conduct relatively improved conclusions. The data preprocessing is evident necessary procedure for causality analysis in gene regulatory role study, and the SSA-CCM hybrid causality test is proved the reliable and well-thought-out approach for causality analysis in complex systems like GRN.

This application further challenges the SSA-CCM hybrid approach and the significant performances again assure the capability of this hybrid approach. Broadly, this research contributes on the current literature of causality analysis of data contain complex noise as well as the collaboration of quantitative causality analysis technique and the studies of GRN. Nevertheless, this can be easily adapted to the other pairs of genes and is also applicable to a wider range of GRNs to infer the regulatory interactions presented among the genes of that network.

## 7.4 Conclusion

This chapter introduces the SSA-CCM hybrid causality test through the implementations and validations by two representative case studies. The performance of subspace-based causality detection techniques introduced in previous chapters have shown remarkable performances and achieved satisfying outcomes comparing to the empirical approaches in a diverse range of applications. In order to further complete this research, this chapter considers the effects of the existing trend of the climate change data and the complex noise of the gene expression profiles as the focuses to conduct more comprehensive causality analysis. By introducing the data preprocessing procedure, CCM is significantly improved further and even the empirical approaches are significantly more applicable. The complex data are decomposed and/or filtered to achieve the causality analysis that is more accurate and less influenced by insignificant components. By overcoming the data and model complexity, possible existing nonlinearity and restrictions of parametric approaches, along with its progressive features as a reduced form, data driven, straight forward approach, there is no doubt that the SSA-CCM hybrid causality test can be incorporated with more broad range of subjects and bring valuable contributions in causality analysis of complex systems.

# Chapter 8

# Conclusion and Future Research

This thesis aims at the theoretical advancements of causality analysis methods by incorporating the advanced subspace-based techniques. It achieves a few novel developments of the quantitative methods in relation to causality analysis along with critical evaluations by simulations as well as a diverse range of real data applications. It is also the initial research that comprehensively extend the subspace-based techniques to different aspects of causality analysis, including similarity measure, association measure and causality measure. Moreover, for the first time, a few advanced, relatively new subspace-based techniques are adopted independently or as combination so to contribute on the theoretical literature of causality analysis as well as the corresponding subspace-based techniques. This chapter concludes the significant achievements of the whole research in section 8.1, followed by the critical summary of research challenges as well as the proposed directions of future research in section 8.2.

## 8.1 Discussion

In general, this research has managed to achieve many aspects of contributions, from the pioneer incorporation of the most advanced subspace-based techniques, to the theoretical and practical advancements of the novel causality analysis methods. The main contributions are summarized and listed in details as follows.

Firstly, it is important to address the significant role of the subspace-based techniques adopted in this research. As relatively new nonparametric techniques for time series analysis, these subspace-based techniques, including SVD, SSA and C-CM, have shown remarkable performances in various range of subjects as detailedly reviewed in **Chapter 2**. As the first attempt to my knowledge, this research focuses on these advanced techniques and explores their potentials on causality analysis by bring their advantages of data decomposition, nonlinearity applicability, no restrictions on modelling to the construction of novel causality analysis methods. These advanced features are the common inheritances that the advancements proposed by this thesis share with the subspace-based techniques. Moreover, these advanced features are the key reasons that the novel causality analysis methods can accomplish the progresses to contribute to the existing literature.

The most distinct contribution that this thesis offers is that the systematical advancements of all adopted subspace-based techniques regarding the significant aspect of causality analysis. Not only does this research extend these novel techniques to a multivariate system, it also addresses the crucial and challenging question of causality. These techniques are all developed to respectively construct novel causality analysis methods that are proved robust, reliable, as well as superior at overcoming the shortages of the empirical causality analysis approaches.

Another contribution that focuses on the causality aspect is that this research considers the philosophy of causes by Aristotle (384-322 B.C.) as the fundamental philosophy, for which, this thesis has incorporated the formal cause with the similarity measure and association measure respectively, whilst the efficient cause is extended as the theoretical basis of the causality measure. To my knowledge, this is the first research that brings the different aspects of causality to the theoretical advancements of causality analysis methods, not to mention, through the subspace-based techniques that no previous studies have utilized.

More specifically, regarding the specific advancements that this thesis brings, the novel similarity measure in **Chapter 3** overcomes one of the most crucial difficulties in similarity measure that the different type of features are not comparable by introducing eigenvalue distribution as the "formal" criterion of evaluating similarity for the first time. The novel mutual association measure in **Chapter 4** outperforms the empirical linear or nonlinear approaches by its sensitivity and improved capability on nonlinear association or complex association. The SSA causality test in **Chapter 5** makes innovative modification on the linear model based causality analysis approach through the incorporation of SSA and MSSA forecasting performance. More importantly, this thesis emphasises its data driven approach orientation and addresses the implementations with real data comparisons. **Chapter 6** adopts the advanced CCM technique and comprehensively evaluates its performance through a diverse range of applications for the first time. It is also acting as a benchmark comparison for the novel SSA-CCM hybrid method introduced in **Chapter 7**. Specifically, SSA contributes to the data preprocessing procedure due to its full-featured data processing superiority before the causality analysis is conducted by CCM. This study is the first time this hybrid approach is comprehensively proposed, and the satisfying performances evident by the applications have shown its significant potential and contribution as an advanced causality analysis technique to the existing literature.

It is of note that another contribution that this research emphasises is that the straight forward approach of problem solving and the orientation of reduced form, data driven quantitative methods. Causality is a broad research subject, while answering the question of causality can be extremely complicated. Instead of trying to reconstruct the complex system by some restricted models, this research aims to understand causality by proposing sufficient quantitative tools that only require two key variables. The data driven aspect also significantly contributes to the causality analysis study as it allows researchers to dive into the data itself and even identify the particular components that determine causality regardless of linear or nonlinear features. The remarkable performance on nonlinearity and complex systems cannot be achieved without its advantages on considering the data as a whole without overlooking any information contained by the data.

The last contribution this thesis achieved is that the diverse range of applications involved in this research, including climate change study, gene regulatory role detection, oil-tourism and oil-stock market studies. It is the first attempt of extending subspace-based causality analysis techniques to the applications in these areas that also produces consistent, solid, satisfying outcomes comparing to empirical approaches respectively. These real data applications also represent data from complex systems that may also involve one or combination of long time span trend, complex noise, etc. Therefore, the proposed methods can be easily applied to other relative research areas or data that have the similar feature or experience difficulties with empirical approaches.

## 8.2 Future Challenges and Directions of Research

Despite the research achievements that are summarized in the above section, one cannot overlook the fact that this research is a first attempt in many aspects. This thesis considers three subspace-based techniques and proposes the corresponding advancements in terms of similarity, association and causality measures through a few novel causality analysis methods. The wide scope that this research covers along with the high volume of quantitative methods this thesis develops are making a significant part of the contribution, whilst this also indicates that the depth of each measure/technique/novel method can be exploited further, which will be the main direction of future research. Following the discussion sections of each chapter respectively, here in this section, some identified challenges and proposals of future research are summarized as follows.

The simulations for the similarity measure in **Chapter 3** can be extended to involve more types of series, even combinations of series with different levels of variations or complex noise. The short series have relatively less information to form the criterion of similarity measure, which is the main reason that longer series show much better performances. This is due to the feature of the method that reduced the dimension of information in the first place. Further research can focus on the improvement of performance in short series. However, as the rapid advancements of technology and information, large volume of data are more frequently encountered nowadays for tremendous amount of subjects (more information can be found in the published works of Hassani et al. (2016) and Hassani et al. (2017a)), this method has shown potentials on the capability of working with long and complex time series. Future research can also investigate on the efficiency of obtaining population information as benchmark so to significantly speed up the calculation.

Similar to the similarity measure, the evaluation by simulation for the mutual association measure in **Chapter 4** can be expanded to more nonlinear patterns or combinations of complex patterns. In terms of the SSA causality test in **Chapter 5**, the minor difference outcomes are the most important concern. Future research can work on the approach to identify a shared ground of identical window length or number of eigenvalues between SSA and MSSA operations instead of the current version that selects the corresponding optimal performances for every step. As a relatively new subspace-based technique, CCM adopted in **Chapter 6** has satisfying sensitivity and capability in terms of casuality detection, the possible directions of future research involve the theoretical advancement for panel data, improving the current cross map skill measure index.

Finally, more applications can be explored and this applies for all proposed methods. The similarity measure can be further validated by time series classification applications, even implementations like image/text/chemical/gene expression recognition and classification. The mutual association measure can be easily applied to any groups of time series, even series with different lengths. Therefore, it stands as an alternative correlation measure with no restrictions of linearity or nonlinearity, any applications that encounter the correlation analysis can be adopted for further research. Regarding the SSA causality test, CCM test and SSA-CCM hybrid causality test, which successfully overcome the data and model complexity, nonlinearity and model restrictions, a much broader range of subjects can surely be applied. Some selected topics for future research include rain fall and sea surface temperature study, other pairs of genes in significant GRNs, air pollution and meteorological/economical factors, neural signal interaction, etc.

# References

Abdi, H. (2007). The kendall rank correlation coefficient. *Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA*, pages 508–510.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.

Alexandrov, T., Bianconcini, S., Dagum, E. B., Maass, P., and McElroy, T. S. (2012). A review of some modern approaches to the problem of trend extraction. *Econometric Reviews*, 31(6):593–624.

Alharbi, N., Ghodsi, Z., and Hassani, H. (2016). Noise correction in gene expression data: a new approach based on subspace method. *Mathematical Methods in the Applied Sciences*.

Alharbi, N. and Hassani, H. (2016). A new approach for selecting the number of the eigenvalues in singular spectrum analysis. *Journal of the Franklin Institute*, 353(1):1–16.

Alter, O., Brown, P. O., and Botstein, D. (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97(18):10101–10106.

Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics*, pages 193–212.

Antonakakis, N., Dragouni, M., and Filis, G. (2015). How strong is the linkage between tourism and economic growth in europe? *Economic Modelling*, 44:142–155.

Arun, K. S., Huang, T. S., and Blostein, S. D. (1987). Least-squares fitting of two 3-d point sets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 9(5):698–700.

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1):1–22.

Baird-Titus, J. M., Clark-Baldwin, K., Dave, V., Caperelli, C. A., Ma, J., and Rance, M. (2006). The solution structure of the native k50 bicoid homeodomain bound to the consensus taatcc dna-binding site. *Journal of Molecular Biology*, 356(5):1137–1151.

Balasubramaniyan, R., Hüllermeier, E., Weskamp, N., and Kämper, J. (2005). Clustering of gene expression data using a local shape-based similarity measure. *Bioinformatics*, 21(7):1069–1077.

Bao, P. and Ma, X. (2005). Image adaptive watermarking using wavelet domain singular value decomposition. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(1):96–102.

Barnard, J. M. and Downs, G. M. (1992). Clustering of chemical structures on the basis of two-dimensional similarity measures. *Journal of Chemical Information and Computer Sciences*, 32(6):644–649.

Becken, S. (2008). Developing indicators for managing tourism in the face of peak oil. *Tourism Management*, 29(4):695–705.

Becken, S. (2011). A critical review of tourism and oil. *Annals of Tourism Research*, 38(2):359–379.

Becken, S. and Lennox, J. (2012). Implications of a long-term increase in oil prices for tourism. *Tourism Management*, 33(1):133–142.

Berleth, T., Burri, M., Thoma, G., Bopp, D., Richstein, S., Frigerio, G., Noll, M., and Nüsslein-Volhard, C. (1988). The role of localization of bicoid rna in organizing the anterior pattern of the drosophila embryo. *The EMBO Journal*, 7(6):1749.

Bhattacharyya, A. (1946). On a measure of divergence between two multinomial populations. *Sankhyā: the Indian Journal of Statistics*, pages 401–406.

Bieler, J., Pozzorini, C., and Naef, F. (2011). Whole-embryo modeling of early segmentation in drosophila identifies robust and fragile expression domains. *Biophysical journal*, 101(2):287–296.

Boigelot, D. (2011). *An example of the correlation of x and y for various distributions of (x,y) pairs*, original online commons of correlation examples edition.

Breitung, J. and Candelon, B. (2006). Testing for short-and long-run causality: A frequency-domain approach. *Journal of Econometrics*, 132(2):363–378.

Brookshire, E. and Weaver, T. (2015). Long-term decline in grassland productivity driven by increasing dryness. *Nature Communications*, 6.

Broomhead, S. and King, G. P. (1986). Extracting qualitative dynamics from experimental data. *Physica D: Nonlinear Phenomena*, 20(2-3):217–236.

Carbó, R., Leyda, L., and Arnau, M. (1980). How similar is a molecule to another? an electron density measure of similarity between two molecular structures. *International Journal of Quantum Chemistry*, 17(6):1185–1189.

Casdagli, M., Eubank, S., Farmer, J. D., and Gibson, J. (1991). State space reconstruction in the presence of noise. *Physica D: Nonlinear Phenomena*, 51(1-3):52–98.

CEIC (2015). *Technical report*, ceic database edition.

Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density functions. *City*, 1(2):1.

Chamberlain, G. (1982). The general equivalence of granger and sims causality. *Econometrica: Journal of the Econometric Society*, pages 569–581.

Chandra, D. S. (2002). Digital image watermarking using singular value decomposition. In *Circuits and Systems, 2002. MWSCAS-2002. The 2002 45th Midwest Symposium on*, volume 3, pages III–III. IEEE.

Chapman, N. (2012). *Correlation analysis in chemistry: recent advances*. Springer Science & Business Media.

Chen, Z., Cai, J., Gao, B., Xu, B., Dai, S., He, B., and Xie, X. (2017). Detecting the causality influence of individual meteorological factors on local pm2. 5 concentration in the jing-jin-ji region. *Scientific Reports*, 7.

Ciner, C. (2011). Eurocurrency interest rate linkages: A frequency domain analysis. *International Review of Economics & Finance*, 20(4):498–505.

Clark, A. T., Ye, H., Isbell, F., Deyle, E. R., Cowles, J., Tilman, G. D., and Sugihara, G. (2015). Spatial convergent cross mapping to detect causal relationships from short time series. *Ecology*, 96(5):1174–1181.

Clark, M. (2013). A comparison of correlation measures. *Center for Social Research, University of Notre Dame*, page 4.

Copf, T., Schröder, R., and Averof, M. (2004). Ancestral role of caudal genes in axis elongation and segmentation. *Proceedings of the National Academy of Sciences of the United States of America*, 101(51):17711–17715.

Croux, C. and Reusens, P. (2013). Do stock prices contain predictive power for the future economic activity? a granger causality analysis in the frequency domain. *Journal of Macroeconomics*, 35:93–103.

Danilov, D. and Zhigljavsky, A. (1997). *Principal components of time series: the 'Caterpillar' method.* St. Petersburg: University of St. Petersburg.

Daub, C. O., Steuer, R., Selbig, J., and Kloska, S. (2004). Estimating mutual information using b-spline functions–an improved similarity measure for analysing gene expression data. *BMC Bioinformatics*, 5(1):118.

Davidson, E. and Levin, M. (2005). Gene regulatory networks.

Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227.

Demirel, H., Ozcinar, C., and Anbarjafari, G. (2010). Satellite image contrast enhancement using discrete wavelet transform and singular value decomposition. *IEEE Geoscience and Remote Sensing Letters*, 7(2):333–337.

Deyle, E. R., Fogarty, M., Hsieh, C.-h., Kaufman, L., MacCall, A. D., Munch, S. B., Perretti, C. T., Ye, H., and Sugihara, G. (2013). Predicting climate effects on pacific sardine. *Proceedings of the National Academy of Sciences*, 110(16):6430–6435.

Deyle, E. R. and Sugihara, G. (2011). Generalized theorems for nonlinear state space reconstruction. *PLoS One*, 6(3):e18295.

Dionisio, A., Menezes, R., and Mendes, D. A. (2004). Mutual information: a measure of dependency for nonlinear time series. *Physica A: Statistical Mechanics and its Applications*, 344(1):326–329.

Dixon, P. A., Milicich, M. J., and Sugihara, G. (1999). Episodic fluctuations in larval supply. *Science*, 283(5407):1528–1530.

Dost, F. (2015). A non-linear causal network of marketing channel system structure. *Journal of Retailing and Consumer Services*, 23:49–57.

Drineas, P., Frieze, A., Kannan, R., Vempala, S., and Vinay, V. (2004). Clustering large graphs via the singular value decomposition. *Machine Learning*, 56(1):9–33.

Eckart, C. and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, pages 1–26.

EIA (2016). *U.S. Energy Information Administration*.

Eichler, A., Olivier, S., Henderson, K., Laube, A., Beer, J., Papina, T., Gäggeler, H. W., and Schwikowski, M. (2009). Temperature response in the altai region lags solar forcing. *Geophysical Research Letters*, 36(1).

Elsner, J. and Tsonis, A. (1996). *Singular Spectrum Analysis: A New Tool in Time Series Analysis*. Springer Science & Business Media.

Falcon, A. (2015). Aristotle on causality. In Edward, N. Z., editor, *The Stanford Encyclopedia of Philosophy*. Stanford University.

Fan, B., Guo, L., Li, N., Chen, J., Lin, H., Zhang, X., Shen, M., Rao, Y., Wang, C., and Ma, L. (2014). Earlier vegetation green-up has reduced spring dust storms. *Scientific Reports*, 4:6749.

Feige, E. L. and Pearce, D. K. (1979). The casual causal relationship between money and income: some caveats for time series analysis. *The Review of Economics and Statistics*, pages 521–533.

Filis, G., Degiannakis, S., and Floros, C. (2011). Dynamic correlation between stock market and oil prices: The case of oil-importing and oil-exporting countries. *International Review of Financial Analysis*, 20(3):152–164.

Fraedrich, K. (1986). Estimating the dimensions of weather and climate attractors. *Journal of the Atmospheric Sciences*, 43(5):419–432.

FRED (2015). *Technical report*, federal reserve economic data edition.

George, K. W., Chen, A., Jain, A., Batth, T. S., Baidoo, E. E., Wang, G., Adams, P. D., Petzold, C. J., Keasling, J. D., and Lee, T. S. (2014). Correlation analysis of targeted proteins and metabolites to assess and engineer microbial isopentenol production. *Biotechnology and Bioengineering*, 111(8):1648–1658.

Geweke, J. (1982). Measurement of linear dependence and feedback between multiple time series. *Journal of the American Statistical Association*, 77(378):304–313.

Ghodsi, M., Alharbi, N., and Hassani, H. (2015a). The empirical distribution of the singular values of a random hankel matrix. *Fluctuation and Noise Letters*, 14(03):1550027.

Ghodsi, M., Hassani, H., and Sanei, S. (2010). Extracting fetal heart signal from noisy maternal ecg by singular spectrum analysis. *Journal of Statistics and its Interface, Special Issue on the Application of SSA*, 3(3):399–411.

Ghodsi, Z., Hassani, H., and McGhee, K. (2015b). Mathematical approaches in studying bicoid gene. *Quantitative Biology*, 3(4):182–192.

Ghodsi, Z., Silva, E. S., and Hassani, H. (2015c). Bicoid signal extraction with a selection of parametric and nonparametric signal processing techniques. *Genomics, Proteomics & Bioinformatics*, 13(3):183–191.

GISS (2015). *Goddard Institute for Space Studies*, database edition.

Goh, C. (2012). Exploring impact of climate on tourism demand. *Annals of Tourism Research*, 39(4):1859–1883.

Golub, G. H. and Reinsch, C. (1970). Singular value decomposition and least squares solutions. *Numerische Mathematik*, 14(5):403–420.

Golyandina, N., Nekrutkin, V., and Zhigljavsky, A. A. (2001). *Analysis of Time Series Structure: SSA and Related Techniques*. CRC Press.

Golyandina, N. and Osipov, E. (2007). The "caterpillar"-ssa method for analysis of time series with missing values. *Journal of Statistical planning and Inference*, 137(8):2642–2653.

Golyandina, N. and Zhigljavsky, A. (2013). *Singular Spectrum Analysis for time series*. Springer Science & Business Media.

Golyandina, N. E., Holloway, D. M., Lopes, F. J., Spirov, A. V., Spirova, E. N., and Usevich, K. D. (2012). Measuring gene expression noise in early drosophila embryos: nucleus-to-nucleus variability. *Procedia computer science*, 9:373–382.

Gorfine, M., Heller, R., and Heller, Y. (2012). Comment on detecting novel associations in large data sets. *Unpublished (available at http://emotion. technion. ac. il/ gorfinm/files/science6. pdf on 11 Nov. 2012)*.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438.

Gupta, R., Gil-Alana, L. A., and Yaya, O. S. (2015). Do sunspot numbers cause global temperatures? evidence from a frequency domain causality test. *Applied Economics*, 47(8):798–808.

Hajian, S. and Movahed, M. S. (2010). Multifractal detrended cross-correlation analysis of sunspot numbers and river flow fluctuations. *Physica A: Statistical Mechanics and its Applications*, 389(21):4942–4957.

Hamilton, J. D. (1996). This is what happened to the oil price-macroeconomy relationship. *Journal of Monetary Economics*, 38(2):215–220.

Hassani, H. (2007). Singular spectrum analysis: methodology and comparison. *Journal of Data Science*, 5(2):239–257.

Hassani, H. (2010). Singular spectrum analysis based on the minimum variance estimator. *Nonlinear Analysis: Real World Applications*, 11(3):2065–2077.

Hassani, H., Alharbi, N., and Ghodsi, M. (2014). A short note on the pattern of the singular values of a scaled random hankel matrix. *International Journal of Applied Mathematics*, 27(3):237–243.

Hassani, H., Alharbi, N., and Ghodsi, M. (2015a). A study on the empirical distribution of the scaled hankel matrix eigenvalues. *Journal of Advanced Research*, 6(6):925–929.

Hassani, H., Dionisio, A., and Ghodsi, M. (2010a). The effect of noise reduction in measuring the linear and nonlinear dependency of financial markets. *Nonlinear Analysis: Real World Applications*, 11(1):492–502.

Hassani, H. and Ghodsi, Z. (2014). Pattern recognition of gene expression with singular spectrum analysis. *Medical Sciences*, 2(3):127–139.

Hassani, H., Heravi, S., Brown, G., and Ayoubkhani, D. (2013a). Forecasting before, during, and after recession with singular spectrum analysis. *Journal of Applied Statistics*, 40(10):2290–2302.

Hassani, H., Heravi, S., and Zhigljavsky, A. (2009). Forecasting european industrial production with singular spectrum analysis. *International Journal of Forecasting*, 25(1):103–118.

Hassani, H., Heravi, S., and Zhigljavsky, A. (2013b). Forecasting uk industrial production with multivariate singular spectrum analysis. *Journal of Forecasting*, 32(5):395–408.

Hassani, H., Huang, X., and Ghodsi, M. (2017a). Big data and causality. *Annals of Data Science*, pages 1–24.

Hassani, H., Huang, X., Silva, E. S., and Ghodsi, M. (2016). A review of data mining applications in crime. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 9(3):139–154.

Hassani, H. and Mahmoudvand, R. (2013). Multivariate singular spectrum analysis: A general view and new vector forecasting approach. *International Journal of Energy and Statistics*, 1(01):55–83.

Hassani, H., Mahmoudvand, R., and Zokaei, M. (2011). Separability and window length in singular spectrum analysis. *Comptes Rendus Mathematique*, 349(17-18):987–990.

Hassani, H. and Silva, E. S. (2015). A kolmogorov-smirnov based test for comparing the predictive accuracy of two sets of forecasts. *Econometrics*, 3(3):590–609.

Hassani, H., Silva, E. S., Antonakakis, N., Filis, G., and Gupta, R. (2017b). Forecasting accuracy evaluation of tourist arrivals. *Annals of Tourism Research*, 63:112–127.

Hassani, H., Silva, E. S., Gupta, R., and Segnon, M. K. (2015b). Forecasting the price of gold. *Applied Economics*, 47(39):4141–4152.

Hassani, H., Soofi, A. S., and Zhigljavsky, A. (2013c). Predicting inflation dynamics with singular spectrum analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(3):743–760.

Hassani, H., Soofi, A. S., and Zhigljavsky, A. A. (2010b). Predicting daily exchange rate with singular spectrum analysis. *Nonlinear Analysis: Real World Applications*, 11(3):2023–2034.

Hassani, H. and Thomakos, D. (2010). A review on singular spectrum analysis for economic and financial time series. *Statistics and its Interface*, 3(3):377–397.

Hassani, H., Webster, A., Silva, E. S., and Heravi, S. (2015c). Forecasting us tourist arrivals using optimal singular spectrum analysis. *Tourism Management*, 46:322–335.

Hassani, H. and Zhigljavsky, A. (2009). Singular spectrum analysis: methodology and application to economics data. *Journal of Systems Science and Complexity*, 22(3):372–394.

Hassani, H., Zhigljavsky, A., Patterson, K., and Soofi, A. (2010c). A comprehensive causality test based on the singular spectrum analysis. *Causality in Science*, pages 379–406.

Haugh, L. D. (1976). Checking the independence of two covariance-stationary time series: a univariate residual cross-correlation approach. *Journal of the American Statistical Association*, 71(354):378–385.

Helmert, F. R. (1876). Über die wahrscheinlichkeit der potenzsummen der beobachtungsfehler. *Z. Math. u. Phys*, 21:192–218.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, pages 293–325.

Holloway, D. M., Harrison, L. G., Kosman, D., Vanario-Alonso, C. E., and Spirov, A. V. (2006). Analysis of pattern precision shows that drosophila segmentation develops substantial independence from gradients of maternal gene products. *Developmental Dynamics*, 235(11):2949–2960.

Holmes, J. M. and Hutton, P. A. (1990). On the casual relationship between government expenditures and national income. *The Review of Economics and Statistics*, pages 87–95.

Hsiao, C. (1979). Autoregressive modeling of canadian money and income data. *Journal of the American Statistical Association*, 74(367):553–560.

Huang, A. (2008). Similarity measures for text document clustering. In *Proceedings of the Sixth New Zealand Computer Science Research Student Conference (NZCSRSC2008), Christchurch, New Zealand*, pages 49–56.

Hung, W.-L. and Yang, M.-S. (2004). Similarity measures of intuitionistic fuzzy sets based on hausdorff distance. *Pattern Recognition Letters*, 25(14):1603–1611.

Ineson, S., Scaife, A. A., Knight, J. R., Manners, J. C., Dunstone, N. J., Gray, L. J., and Haigh, J. D. (2011). Solar forcing of winter climate variability in the northern hemisphere. *Nature Geoscience*, 4(11):753–757.

Jarvis, R. A. and Patrick, E. A. (1973). Clustering using a similarity measure based on shared near neighbors. *IEEE Transactions on Computers*, 100(11):1025–1034.

Jha, S. K. and Yadava, R. (2011). Denoising by singular value decomposition and its application to electronic nose data processing. *IEEE Sensors Journal*, 11(1):35–44.

Jung, H. and Park, C. (2011). Stock market reaction to oil price shocks. *Journal of Economic Theory and Econometrics*, 22:1–29.

Kalman, D. (1996). A singularly valuable decomposition: the svd of a matrix. *The College Mathematics Journal*, 27(1):2–23.

Kanjilal, P. P., Palit, S., and Saha, G. (1997). Fetal ecg extraction from single-channel maternal ecg using singular value decomposition. *IEEE Transactions on Biomedical Engineering*, 44(1):51–59.

Karlebach, G. and Shamir, R. (2008). Modelling and analysis of gene regulatory networks. *Nature Reviews Molecular Cell Biology*, 9(10):770–780.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1/2):81–93.

Kilian, L. and Park, C. (2009). The impact of oil price shocks on the us stock market. *International Economic Review*, 50(4):1267–1287.

Kirchgässner, G., Wolters, J., and Hassler, U. (2012). *Introduction to modern time series analysis*. Springer Science & Business Media.

Klema, V. and Laub, A. (1980). The singular value decomposition: Its computation and some applications. *IEEE Transactions on Automatic Control*, 25(2):164–176.

Kolmogorov, A. N. (1933). *Sulla determinazione empirica di una legge di distribuzione*. na.

Konstantinides, K., Natarajan, B., and Yovanof, G. S. (1997). Noise estimation and filtering using block-based singular value decomposition. *IEEE Transactions on Image Processing*, 6(3):479–483.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Lai, C.-C. and Tsai, C.-C. (2010). Digital image watermarking using discrete wavelet transform and singular value decomposition. *IEEE Transactions on Instrumentation and Measurement*, 59(11):3060–3063.

Lawson, C. L. and Hanson, R. J. (1974). Solving least squares problems. *Prentice-Hall Series in Automatic Computation, Englewood Cliffs: Prentice-Hall, 1974.*

Le Bihan, N. and Mars, J. (2004). Singular value decomposition of quaternion matrices: a new tool for vector-sensor signal processing. *Signal Processing*, 84(7):1177–1199.

Lean, J. and Rind, D. (1998). Climate forcing by changing solar radiation. *Journal of Climate*, 11(12):3069–3094.

Lewis, E. B. (1978). A gene complex controlling segmentation in drosophila. In *Genes, Development and Cancer*, pages 205–217. Springer.

Lin, D. et al. (1998). An information-theoretic definition of similarity. In *ICML*, volume 98, pages 296–304. Citeseer.

Liu, S. and Jack, J. (1992). Regulatory interactions and role in cell type specification of the malpighian tubules by the cut, krüppel, and caudal genes of drosophila. *Developmental Biology*, 150(1):133–143.

Liu, Y. and Bahadori, M. T. (2012). A survey on granger causality: A computational view. *University of Southern California*, pages 1–13.

Lockwood, M. (2012). Solar influence on global and regional climates. *Surveys in Geophysics*, 33(3-4):503–534.

Lopes, F. J., Spirov, A. V., and Bisch, P. M. (2012). The role of bicoid cooperative binding in the patterning of sharp borders in drosophila melanogaster. *Developmental Biology*, 370(2):165–172.

Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. (2003). Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283.

Luo, C., Zheng, X., and Zeng, D. (2014). Causal inference in social media using convergent cross mapping. In *Intelligence and Security Informatics Conference (JISIC), 2014 IEEE Joint*, pages 260–263. IEEE.

Lütkepohl, H. (1982). Non-causality due to omitted variables. *Journal of Econometrics*, 19(2-3):367–378.

Malinvaud, E. (1980). Statistical methods of econometrics.

Margolis, E., Oot, A., and Fredricks, D. N. (2016). Human microbiome dynamics: Causality detection with convergent cross mapping. In *Open Forum Infectious Diseases, 3(1)*, page 2224. Oxford University Press.

Massidda, C. and Mattana, P. (2013). A svecm analysis of the relationship between international tourism arrivals, gdp and trade in italy. *Journal of Travel Research*, 52(1):93–105.

McBride, J. C., Zhao, X., Munro, N. B., Jicha, G. A., Schmitt, F. A., Kryscio, R. J., Smith, C. D., and Jiang, Y. (2015). Sugihara causality analysis of scalp eeg for detection of early alzheimer's disease. *NeuroImage: Clinical*, 7:258–265.

Menezes, R., Dionísio, A., and Hassani, H. (2012). On the globalization of stock markets: An application of vector error correction model, mutual information and singular spectrum analysis to the g7 countries. *The Quarterly Review of Economics and Finance*, 52(4):369–384.

Miranian, A., Abdollahzade, M., and Hassani, H. (2013). Day-ahead electricity price analysis and forecasting by singular spectrum analysis. *IET Generation, Transmission & Distribution*, 7(4):337–346.

Mitchell, H. B. (2003). On the dengfeng–chuntian similarity measure and its application to pattern recognition. *Pattern Recognition Letters*, 24(16):3101–3104.

Myasnikova, E., Surkova, S., Panok, L., Samsonova, M., and Reinitz, J. (2009). Estimation of errors introduced by confocal imaging into the data on segmentation gene expression in drosophila. *Bioinformatics*, 25(3):346–352.

Nelson, C. R. (1973). *Applied time series analysis for managerial forecasting*. Holden-Day San Francisco.

Niessing, D., Blanke, S., and Jäckle, H. (2002). Bicoid associates with the 5-cap-bound complex of caudal mrna and represses translation. *Genes & Development*, 16(19):2576–2582.

Nikolova, N. and Jaworska, J. (2003). Approaches to measure chemical similarity–a review. *Molecular Informatics*, 22(9-10):1006–1026.

Patterson, K., Hassani, H., Heravi, S., and Zhigljavsky, A. (2011). Multivariate singular spectrum analysis for forecasting revisions to real-time data. *Journal of Applied Statistics*, 38(10):2183–2211.

Pearson, K. (1895). Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242.

Pearson, K. (1900). X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Penney, G. P., Weese, J., Little, J. A., Desmedt, P., Hill, D. L., et al. (1998). A comparison of similarity measures for use in 2-d-3-d medical image registration. *IEEE Transactions on Medical Imaging*, 17(4):586–595.

Pérez-Rodríguez, J. V., Ledesma-Rodríguez, F., and Santana-Gallego, M. (2015). Testing dependence between gdp and tourism's growth rates. *Tourism Management*, 48:268–282.

Pierce, D. A. (1977). Relationships—and the lack thereof—between economic time series, with special reference to money and interest rates. *Journal of the American Statistical Association*, 72(357):11–22.

Pierce, D. A. and Haugh, L. D. (1977). Causality in temporal systems: Characterization and a survey. *Journal of Econometrics*, 5(3):265–293.

Pisarev, A., Poustelnikova, E., Samsonova, M., and Reinitz, J. (2009). Flyex, the quantitative atlas on segmentation gene expression at cellular resolution. *Nucleic Acids Research*, 37(suppl 1):D560–D566.

Poustelnikova, E., Pisarev, A., Blagov, M., Samsonova, M., and Reinitz, J. (2004). A database for management of gene expression data in situ. *Bioinformatics*, 20(14):2212–2221.

Rajwade, A., Rangarajan, A., and Banerjee, A. (2013). Image denoising using the higher order singular value decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):849–862.

Reshef, D. N., Reshef, Y. A., Finucane, H. K., Grossman, S. R., McVean, G., Turnbaugh, P. J., Lander, E. S., Mitzenmacher, M., and Sabeti, P. C. (2011). Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.

Roche, A., Malandain, G., Pennec, X., and Ayache, N. (1998). The correlation ratio as a new similarity measure for multimodal image registration. *Medical Image Computing and Computer-Assisted Interventation—MICCAI'98*, pages 1115–1124.

Rodrıguez-Aragón, L. J. and Zhigljavsky, A. (2010). Singular spectrum analysis for image processing. *Statistics and its Interface*, 3(3):419–426.

Sahami, M. and Heilman, T. D. (2006). A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th International Conference on World Wide Web*, pages 377–386. AcM.

Sanei, S. and Hassani, H. (2015). *Singular spectrum analysis of biomedical signals*. CRC Press.

Savas, B. and Eldén, L. (2007). Handwritten digit classification using higher order singular value decomposition. *Pattern Recognition*, 40(3):993–1003.

Scafetta, N. (2009). Empirical analysis of the solar contribution to global mean air surface temperature change. *Journal of Atmospheric and Solar-Terrestrial Physics*, 71(17):1916–1923.

Scafetta, N. (2014). Global temperatures and sunspot numbers. are they related? yes, but non linearly. a reply to gil-alana et al.(2014). *Physica A: statistical Mechanics and its Applications*, 413:329–342.

Scafetta, N. and West, B. J. (2005). Estimated solar contribution to the global surface warming using the acrim tsi satellite composite. *Geophysical Research Letters*, 32(18).

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Schwert, G. W. (1979). Tests of causality: The message in the innovations. In *Carnegie-Rochester Conference Series on Public Policy*, volume 10, pages 55–96. Elsevier.

Serra, J. and Arcos, J. L. (2014). An empirical evaluation of similarity measures for time series classification. *Knowledge-Based Systems*, 67:305–314.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3-4):591–611.

Shlens, J. (2014). A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*.

SIDC (2015). *Solar Influences Data Analysis Centre*, database edition.

Silva, E. S., Ghodsi, Z., Ghodsi, M., Heravi, S., and Hassani, H. (2017). Cross country relations in european tourist arrivals. *Annals of Tourism Research*, 63:151–168.

Silva, E. S. and Hassani, H. (2015). On the use of singular spectrum analysis for forecasting us trade before, during and after the 2008 recession. *International Economics*, 141:34–49.

Simon, H. A. (1954). Spurious correlation: a causal interpretation. *Journal of the American Statistical Association*, 49(267):467–479.

Simon, N. and Tibshirani, R. (2014). Comment on" detecting novel associations in large data sets" by reshef et al, science dec 16, 2011. *arXiv preprint arXiv:1401.7645*.

Sims, C. A. (1972). Money, income, and causality. *The American Economic Review*, 62(4):540–552.

Sims, C. A. (1980). Macroeconomics and reality. *Econometrica: Journal of the Econometric Society*, pages 1–48.

Sokal Robert, R. and James, R. F. (1981). *Biometry. The principles and practice of statistics in biological research(2nd edition)*. WH Freeman and Company.

Spearman, C. (1904). " general intelligence," objectively determined and measured. *The American Journal of Psychology*, 15(2):201–292.

Stewart, G. W. (1993). On the early history of the singular value decomposition. *SIAM Review*, 35(4):551–566.

Sugihara, G. (1994). Nonlinear forecasting for the classification of natural time series. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 348(1688):477–495.

Sugihara, G., Grenfell, B., May, R. M., Chesson, P., Platt, H., and Williamson, M. (1990). Distinguishing error from chaos in ecological time series [and discussion]. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 330(1257):235–251.

Sugihara, G., May, R., Ye, H., Hsieh, C.-h., Deyle, E., Fogarty, M., and Munch, S. (2012). Detecting causality in complex ecosystems. *science*, 338(6106):496–500.

Sugihara, G. and May, R. M. (1990). Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, 344:6268.

Surkova, S., Kosman, D., Kozlov, K., Myasnikova, E., Samsonova, A. A., Spirov, A., Vanario-Alonso, C. E., Samsonova, M., Reinitz, J., et al. (2008). Characterization of the drosophila segment determination morphome. *Developmental Biology*, 313(2):844–862.

Székely, G. J., Rizzo, M. L., Bakirov, N. K., et al. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794.

Székely, G. J., Rizzo, M. L., et al. (2009). Brownian distance covariance. *The Annals of Applied Statistics*, 3(4):1236–1265.

Takens, F. (1981). Detecting strange attractors in turbulence. In *Dynamical Systems and Turbulence, Warwick 1980*, pages 366–381. Springer.

Tang, C. F. and Abosedra, S. (2016). Tourism and growth in lebanon: new evidence from bootstrap simulation and rolling causality approaches. *Empirical Economics*, 50(2):679–696.

Tang, C. F. and Tan, E. C. (2013). How stable is the tourism-led growth hypothesis in malaysia? evidence from disaggregated tourism markets. *Tourism Management*, 37:52–57.

Tang, C. F. and Tan, E. C. (2015). Does tourism effectively stimulate malaysia's economic growth? *Tourism Management*, 46:158–163.

Thompson, R. C. and Therianos, S. (1972). The eigenvalues of complementary principal submatrices of a positive definite matrix. *Canad. J. Math*, 24(4):658–667.

Tsonis, A. A., Deyle, E. R., May, R. M., Sugihara, G., Swanson, K., Verbeten, J. D., and Wang, G. (2015). Dynamical evidence for causality between galactic cosmic rays and interannual variation in global temperature. *Proceedings of the National Academy of Sciences*, 112(11):3253–3256.

Tsui, W. H. K. and Fung, M. K. Y. (2016). Causality between business travel and trade volumes: Empirical evidence from hong kong. *Tourism Management*, 52:395–404.

Van Der Veen, A.-J., Deprettere, E. F., and Swindlehurst, A. L. (1993). Subspace-based signal analysis using singular value decomposition. *Proceedings of the IEEE*, 81(9):1277–1308.

Van Loan, C. F. (1976). Generalizing the singular value decomposition. *SIAM Journal on Numerical Analysis*, 13(1):76–83.

Vautard, R. and Ghil, M. (1989). Singular spectrum analysis in nonlinear dynamics, with applications to paleoclimatic time series. *Physica D: Nonlinear Phenomena*, 35(3):395–424.

Vautard, R., Yiou, P., and Ghil, M. (1992a). Singular-spectrum analysis: A toolkit for short, noisy chaotic signals. *Physica D: Nonlinear Phenomena*, 58(1):95–126.

Vautard, R., Yiou, P., and Ghil, M. (1992b). Singular-spectrum analysis: A toolkit for short, noisy chaotic signals. *Physica D: Nonlinear Phenomena*, 58(1):95–126.

Wall, M. E., Rechtsteiner, A., and Rocha, L. M. (2003). Singular value decomposition and principal component analysis. In *A Practical Approach to Microarray Data Analysis*, pages 91–109. Springer.

Wallace, J. M., Smith, C., and Bretherton, C. S. (1992). Singular value decomposition of wintertime sea surface temperature and 500-mb height anomalies. *Journal of Climate*, 5(6):561–576.

Wold, H. (1954). Causality and econometrics. *Econometrica: Journal of the Econometric Society*, pages 162–177.

Yang, C., Duraiswami, R., and Davis, L. (2005). Efficient mean-shift tracking via a new similarity measure. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE computer society conference on*, volume 1, pages 176–183. IEEE.

Ye, H., Deyle, E. R., Gilarranz, L. J., and Sugihara, G. (2015). Distinguishing time-delayed causal interactions using convergent cross mapping. *Scientific Reports*, 5.

Zhang, X., Xu, Z., Jia, N., Yang, W., Feng, Q., Chen, W., and Feng, Y. (2015). Denoising of 3d magnetic resonance images by using higher-order singular value decomposition. *Medical Image Analysis*, 19(1):75–86.

# Appendix A

# Forecasting Algorithms of SSA/MSSA

Here in this Appendix, all detailed forecasting algorithms of SSA/MSSA are listed as in Golyandina et al. (2001), Hassani and Mahmoudvand (2013) and Sanei and Hassani (2015). Note that these are built above the fundamental knowledge of SSA and MSSA that has been comprehensively introduced in Chapter 2 section 2.2.

## A.1 SSA

### A.1.1 Recurrent SSA Forecasting

1. Donate a time series $Y_N = (y_1, ..., y_N)$ with lengh of $N$.

2. Set the window length $L$.

3. Consider the linear space $\mathfrak{L}_r \subset \mathbf{R}^L$ of dimension $r < L$, where assume that $e_L \notin \mathfrak{L}_r$, where $e_L = (0, 0, \ldots, 1) \in \mathbf{R}^L$.

4. Obtain the trajectory matrix $\mathbf{X} = [X_1, ..., X_K]$ of $Y_N$.

5. Extract the orthonormal basises $U_i (i = 1,...,r)$ from the SVD of $\mathbf{X}$.

6. Compute matrix $\widehat{\mathbf{X}} = [\widehat{X}_1 : \ldots : \widehat{X}_K] = \sum_{i=1}^{r} U_i U_i^T \mathbf{X}$. The vector $\widehat{X}_i$ is the orthogonal projection of $X_i$ onto the space $\mathfrak{L}_r$.

7. Construct the Hankelized matrix $\widetilde{\mathbf{X}} = \mathscr{H}\widehat{\mathbf{X}} = [\widetilde{X}_1 : \ldots : \widetilde{X}_K]$, where $\mathscr{H}$ is a Hankel operator..

8. Set $v^2 = \pi_1^2 + \ldots + \pi_r^2$, where $\pi_i$ is the last component of the vector $U_i$ $(i = 1,\ldots,r)$. Since $e_L \notin \mathfrak{L}_r$, so that $\mathfrak{L}_r$ is not a vertical space, $v^2 < 1$.

9. Determine vector $A = (\alpha_1,\ldots,\alpha_{L-1}) = \frac{1}{1-v^2}\sum_{i=1}^{r} \pi_i U_i^{\nabla}$, where $U^{\nabla} \in \mathbf{R}^{L-1}$ is the vector consisting of the first $L-1$ components of the vector $U \in \mathbf{R}^L$.

10. Obtain the time series $Y_{N+h} = (y_1,\ldots,y_{N+h})$ by

$$y_i = \begin{cases} \widetilde{y}_i & \text{for } i = 1,\ldots,N \\ \sum_{j=1}^{L-1} \alpha_j y_{i-j} & \text{for } i = N+1,\ldots,N+h \end{cases} \tag{A.1}$$

where $\widetilde{y}_i$ $(i = 1,\ldots,N)$ are the reconstructed series; $y_{N+1},\ldots,y_{N+h}$ are the $h$-step ahead recurrent forecasts.

### A.1.2 Vector SSA Forecasting

Consider the following matrix:

$$\Pi = \mathbf{V}^{\nabla}(\mathbf{V}^{\nabla})^T + (1-v^2)AA^T, \tag{A.2}$$

where $\mathbf{V}^{\nabla} = [U_1^{\nabla},...,U_r^{\nabla}]$. Consider the linear operator $\theta^{(v)} : \mathfrak{L}_r \mapsto \mathbf{R}^L$, where

$$\theta^{(v)}U = \begin{pmatrix} \Pi U^{\nabla} \\ A^T U^{\nabla} \end{pmatrix}. \tag{A.3}$$

Therefore, vector $Z_i$ is obtained following:

$$Z_i = \begin{cases} \widetilde{X}_i & \text{for } i = 1, \ldots, K \\ \theta^{(v)} Z_{i-1} & \text{for } i = K+1, \ldots, K+h+L-1 \end{cases} \tag{A.4}$$

where $\widetilde{X}_i$ indicate the reconstructed columns of the trajectory matrix after grouping and filtering the noise components. Then, through diagonal averaging on the constructed matrix $\mathbf{Z} = [Z_1, ..., Z_{K+h+L-1}]$, a new series $y_1, ..., y_{N+h+L-1}$ is obtained, where $y_{N+1}, ..., y_{N+h}$ indicate the $h$-step ahead vector predictions.

## A.2 MSSA

### A.2.1 Vertical MSSA

#### Recurrent VMSSA Forecasting

Let us have $M$ series $Y_{N_i}^{(i)} = (y_1^{(i)}, \ldots, y_{N_i}^{(i)})$ and corresponding window length $L_i$, $1 < L_i < N_i/2, i = 1, \ldots, M$. Thus, the $h$-step ahead VMSSA-R forecasting algorithm is as follows (Hassani and Mahmoudvand, 2013; Sanei and Hassani, 2015).

1. Construct the trajectory matrix $\mathbf{X}^{(i)} = [X_1^{(i)}, \ldots, X_K^{(i)}] = (x_{mn})_{m,n=1}^{L_i,K}$ for each single series $Y_{N_i}^{(i)}$ $(i = 1, \ldots, M)$ respectively, where $K$ is a fixed value for all series.

2. Construct the block trajectory matrix:

$$\mathbf{X}_V = \begin{bmatrix} \mathbf{X}^{(1)} \\ \vdots \\ \mathbf{X}^{(M)} \end{bmatrix}.$$

3. Let $\mathbf{U}_{V_j} = (U_j^{(1)}, \dots, U_j^{(M)})^T$ be the $j^{th}$ eigenvector of $\mathbf{X}_V \mathbf{X}_V^T$, where $U_j^{(i)}$ with length $L_i$ corresponds to the series $Y_{N_i}^{(i)}$ $(i = 1, \dots, M)$.

4. Consider $\widehat{\mathbf{X}}_V = [\widehat{X}_1 : \dots : \widehat{X}_K] = \sum_{i=1}^{r} U_{V_i} U_{V_i}^T \mathbf{X}_V$ as the reconstructed matrix achieved from $r$ eigentriples:
$$\widehat{\mathbf{X}}_V = \begin{bmatrix} \widehat{\mathbf{X}}^{(1)} \\ \vdots \\ \widehat{\mathbf{X}}^{(M)} \end{bmatrix}.$$

5. Consider matrix $\widetilde{\mathbf{X}}^{(i)} = \mathscr{H}\widehat{\mathbf{X}}^{(i)}$ $(i = 1, \dots, M)$ as the result of the Hankelization procedure of the matrix $\widehat{\mathbf{X}}^{(i)}$ obtained from the previous step, where $\mathscr{H}$ is a Hankel operator.

6. Donate $U_j^{(i)\triangledown}$ the vector of the first $L_i - 1$ components of the vector $U_j^{(i)}$ and $\pi_j^{(i)}$ is the last component of the vector $U_j^{(i)}$ $(i = 1, \dots, M)$.

7. Select the number of $r$ eigentriples for the reconstruction.

8. Define matrix $\mathbf{U}^{\triangledown M} = \left(U_1^{\triangledown M}, \dots, U_r^{\triangledown M}\right)$, where
$$U_j^{\triangledown M} = \begin{bmatrix} U_j^{(1)\triangledown} \\ \vdots \\ U_j^{(M)\triangledown} \end{bmatrix}.$$

9. Donate matrix $\mathbf{W}$ as follows:
$$\mathbf{W} = \begin{bmatrix} \pi_1^{(1)} & \pi_2^{(1)} & \cdots & \pi_r^{(1)} \\ \pi_1^{(2)} & \pi_2^{(2)} & \cdots & \pi_r^{(2)} \\ \vdots & \vdots & \cdots & \vdots \\ \pi_1^{(M)} & \pi_2^{(M)} & \cdots & \pi_r^{(M)} \end{bmatrix}.$$

10. If the matrix $\left(\mathbf{I}_{M\times M}-\mathbf{W}\mathbf{W}^{T}\right)^{-1}$ exists and $r \leq L_{sum}-M$, then the $h$-step ahead VMSSA forecasts are achieved by:

$$\begin{cases} \left[\tilde{y}_{j_1}^{(1)},\ldots,\tilde{y}_{j_M}^{(M)}\right], & j_i = 1,\ldots,N_i \\ \\ \left(\mathbf{I}_{M\times M}-\mathbf{W}\mathbf{W}^{T}\right)^{-1}\mathbf{W}\mathbf{U}^{\nabla M^{T}}\mathbf{Z}_h\,, & j_i = N_i+1,\ldots,N_i+h, \end{cases}$$

where, $\mathbf{Z}_h = \left[Z_h^{(1)},\ldots,Z_h^{(M)}\right]^T$ and $Z_h^{(i)} = \left[\hat{y}_{N_i-L_i+h+1}^{(i)},\ldots,\hat{y}_{N_i+h-1}^{(i)}\right]$ $(i = 1,\ldots,M)$.

**Vector VMSSA Forecasting**

On top of the recurrent VMSSA algorithms, consider the matrix:

$$\Pi = \mathbf{U}^{\nabla}\mathbf{U}^{\nabla^{T}} + \mathscr{R}\left(\mathbf{I}_{M\times M}-\mathbf{W}\mathbf{W}^{T}\right)\mathscr{R}^{T}, \tag{A.5}$$

where, $\mathscr{R} = \mathbf{U}^{\nabla}\mathbf{W}^{T}\left(\mathbf{I}_{M\times M}-\mathbf{W}\mathbf{W}^{T}\right)^{-1}$. The vector VMSSA forecasts algorithms are listed below by following Hassani and Mahmoudvand (2013).

1. Define vectors $Z_i$ as follows:

$$Z_i = \begin{cases} \widetilde{X}_i & \text{for } i = 1,\ldots,k \\ \mathscr{P}^{(v)}Z_{i-1} & \text{for } i = k+1,\ldots,k+h+L_{\max}-1, \end{cases} \tag{A.6}$$

   where, $L_{\max} = \max\{L_1,\ldots,L_M\}$.

2. Construct the matrix $\mathbf{Z} = [Z_1 : \ldots : Z_{K+h+L_{\max}-1}]$. The corresponding Hankelization is conducted to obtain $\hat{y}_1^{(i)},\ldots,\hat{y}_{N+h+L_{\max}}^{(i)}$ $(i = 1,\ldots,M)$.

3. The numbers $\hat{y}_{N_i+1}^{(i)},\ldots,\hat{y}_{N_i+h}^{(i)}$ $(i = 1,\ldots,M)$ represent the $h$ step ahead vector VMSSA predictions.

### A.2.2 Horizontal MSSA

**Recurrent HMSSA Forecasting**

1. Construct the trajectory matrix $\mathbf{X}^{(i)} = [X_1^{(i)}, \ldots, X_K^{(i)}] = (x_{mn})_{m,n=1}^{L,K_i}$ for each series $Y_{N_i}^{(i)}$ $(i = 1, \ldots, M)$ respectively with a fixed value of $L$.

2. Obtain the block trajectory matrix:

$$\mathbf{X}_H = \left[ \begin{array}{cccc} \mathbf{X}^{(1)} : & \mathbf{X}^{(2)} : & \cdots & : \mathbf{X}^{(M)} \end{array} \right].$$

3. Donate vector $U_{H_j} = (u_{1_j}, \ldots, u_{L_j})^T$, with length $L$, be the $j^{th}$ eigenvector of $\mathbf{X}_H \mathbf{X}_H^T$.

4. Determine $\widehat{\mathbf{X}}_H = \sum_{i=1}^{r} U_{H_i} U_{H_i}^T \mathbf{X}_H$ as the reconstructed matrix obtained using $r$ eigentriples that:

$$\mathbf{X}_H = \left[ \begin{array}{cccc} \widehat{\mathbf{X}}^{(1)} : & \widehat{\mathbf{X}}^{(2)} : & \cdots & : \widehat{\mathbf{X}}^{(M)} \end{array} \right].$$

5. Donate matrix $\widetilde{\mathbf{X}}^{(i)} = \mathscr{H} \widehat{\mathbf{X}}^{(i)}$ $(i = 1, \ldots, M)$ as the Hankelization product of $\widehat{\mathbf{X}}^{(i)}$ obtained from the previous step.

6. Let $U_{H_j}^{\triangledown}$ denotes the vector of the first $L-1$ coordinates of the eigenvectors $U_{H_j}$, and $\pi_{H_j}$ indicates the last coordinate of the eigenvectors $U_{H_j}$ $(j = 1, \ldots, r)$.

7. Define $\upsilon^2 = \sum\limits_{j=1}^{r} \pi_{H_j}^2$.

8. Construct the linear coefficients vector $\mathscr{R}$ by:

$$\mathscr{R} = \frac{1}{1-\upsilon^2} \sum_{j=1}^{r} \pi_{Hj} U_{Hj}^{\triangledown}. \tag{A.7}$$

9. If $\upsilon^2 < 1$, then the $h$-step ahead recurrent HMSSA forecasts exist and can be calculated by:

$$\left[\hat{y}_{j_1}^{(1)},\ldots,\hat{y}_{j_M}^{(M)}\right]^T = \begin{cases} \left[\tilde{y}_{j_1}^{(1)},\ldots,\tilde{y}_{j_M}^{(M)}\right], & j_i = 1,\ldots,N_i, \\ \\ \mathscr{R}^T \mathbf{Z}_h, & j_i = N_i+1,\ldots,N_i+h, \end{cases} \tag{A.8}$$

where $\mathbf{Z}_h = \left[Z_h^{(1)},\ldots,Z_h^{(M)}\right]^T$ and $Z_h^{(i)} = \left[\hat{y}_{N_i-L+h+1}^{(i)},\ldots,\hat{y}_{N_i+h-1}^{(i)}\right]$ $(i=1,\ldots,M)$.

### Vector HMSSA Forecasting

Following the items (1)-(7) of recurrent HMSSA forecasting algorithms, donate the following matrix

$$\Pi = \mathbf{U}^{\triangledown}\mathbf{U}^{\triangledown T} + (1-v^2)RR^T, \tag{A.9}$$

where $\mathbf{U}^{\triangledown} = [U_1^{\triangledown},\ldots,U_r^{\triangledown}]$. Thus, consider the linear operator $\mathscr{P}^{(v)} : \mathfrak{L}_r \mapsto \mathbb{R}^L$ that

$$\mathscr{P}^{(v)}Y = \begin{pmatrix} \Pi Y_{\triangle} \\ R^T Y_{\triangle} \end{pmatrix}, \; Y \in \mathfrak{L}_r, \tag{A.10}$$

and $Y_{\triangle}$ is vector of last $L-1$ elements of $Y$.

Define vector $Z_j^{(i)}$ $(i=1,\ldots,M)$ as follows:

$$Z_j^{(i)} = \begin{cases} \widetilde{X}_j^{(i)} & \text{for } j = 1,\ldots,k_i \\ \mathscr{P}^{(v)} Z_{j-1}^{(i)} & \text{for } j = k_i+1,\ldots,k_i+h+L-1 \end{cases} \tag{A.11}$$

where $\widetilde{X}_j^{(i)}$ indicate the reconstructed columns of trajectory matrix of the $i^{\text{th}}$ series after grouping and leaving noise components. Through diagonal averaging of the constructed matrix $\mathbf{Z}^{(i)} = [Z_1^{(i)},\ldots,Z_{k_i+h+L-1}^{(i)}]$, a new series $\hat{y}_1^{(i)},\ldots,\hat{y}_{N_i+h+L-1}^{(i)}$ is achieved, where $\hat{y}^{(i)}_{N_i+1},\ldots,\hat{y}^{(i)}_{N_i+h}$ indicate the *h*-step ahead of vector HMSSA predictions.

# Appendix B

# Comparison of Eigenvalue Distributions by Horizontal/Vertical Techniques

Following the newly proposed similarity measure in Chapter 3, which is firstly built on a fundamental multivariate system with the benchmark series as the dominant role and uses corresponding eigenvalue distribution information as the similarity criterion. Section 3.2.1 illustrates the detailed process of extracting the eigenvalue distribution information from a two series system, while this Appendix is provided to clarify the comparison between horizontal form and vertical form embedding and decomposition technique. Similarly, both with and without the premise of multivariate system scenarios are considered and compared respectively as follows.

## B.1 With Premise of Multivariate System

In order to evaluate the differences between horizontal and vertical forms, the comparisons are set between the Hankel matrix made by a random variable $X$ with itself and the matrix formed by $X$ and a very similar random variable $Y$. Consequently, the expected results should be that these two Hankel matrices are sharing the similar distribution of eigenvalues.

The comparisons of results are listed as follows by different types of simulated variables respectively. Note that the blue line represents the singular values from the matrix formed by a random variable $X$ with itself, and the red line refers to the singular values of the Hankel matrix structured by $X$ and a very similar random variable $Y$. Identically, as stated in section 3.2.2, the default window length is set as about 1/10 of the time series length, for which the default length for all tested series is 1000 and default window length is 100, unless it has been specified and noted.

According to the introduction above, the expecting result is that showing similar distributions, the singular values of a Hankel matrix formed by random white noise series $X$ and itself show significantly similar distribution with the singular values of Hankel matrix formed by two random white noise series $X$ and $Y$ when the according matrices are structured in horizontal form. In terms of the vertically formed matrices, the distributions of two groups of singular values vary greatly especially for the second half of eigenvalues.

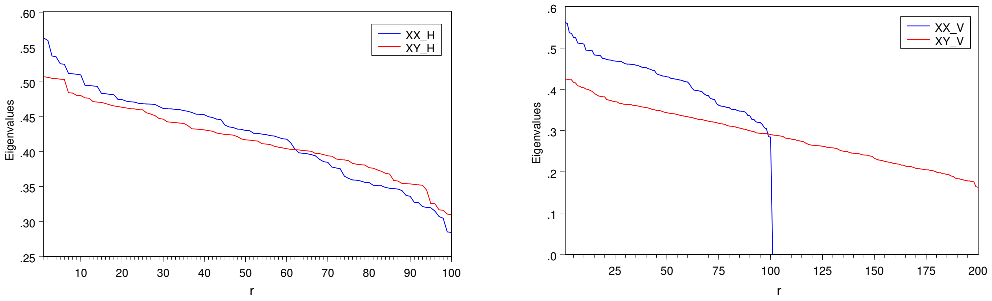


Fig. B.1 Eigenvalues of Two White Noise by Horizontal (left) and Vertical (right) SVD

For random uniform distributed series, the empirical distribution of eigenvalues differ by the specified range of values. For the [0,1] uniform distribution group by vertical decomposition, the differences start to be significant only for the second half of eigenvalues, whilst in terms of the [-1,1] uniform distribution group, significant differences between singular values can be found since the first component. However, in general, the horizontally formed matrices still provide much similar distributions of singular values than the vertically formed group.

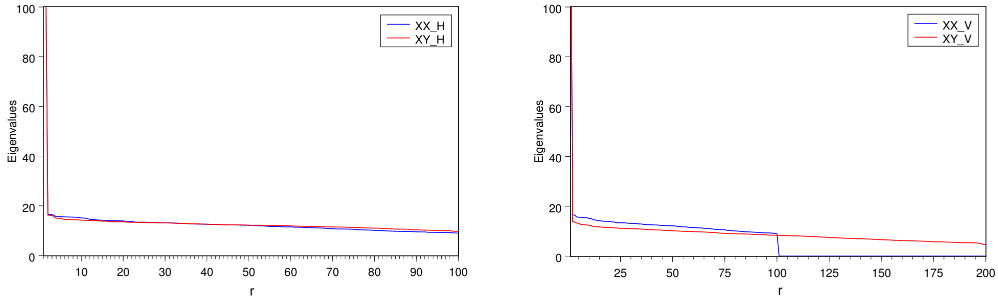


Fig. B.2 Eigenvalues of Two Uniform Distributions [0,1] by Horizontal (left) and Vertical (right) SVD

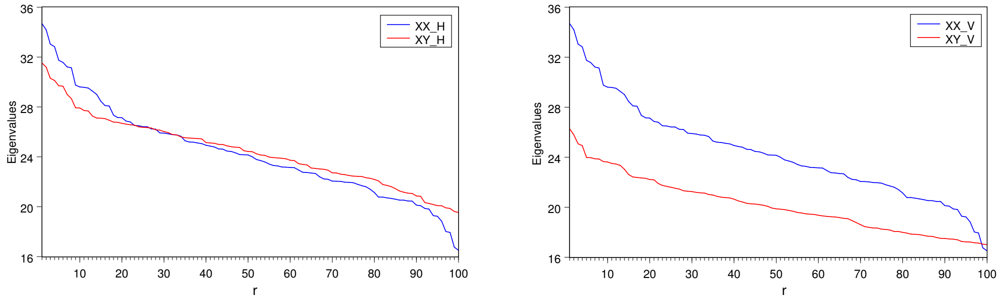


Fig. B.3 Eigenvalues of Two Uniform Distributions [-1,1] by Horizontal (left) and Vertical (right) SVD

Considering the natural character of eigenvalue distribution of sine waves that only contain two significant eigenvalues (reflecting information of trend and cycle respectively), the figures above are partially enlarged at the comparable parts. Also, since the differences between distributions of eigenvalues for sine waves can only be reflected when the window length is relatively small, the second group of figures are singular values by specifying window length $L$ as 2, in which the minimum window length is considered and the most significant difference of eigenvalues is shown between horizontal and vertical decomposition. There are only minor differences between the horizontally formed and the vertically formed groups when the window length L is set 100 as default. However, the distributions of eigenvalues vary greatly when the window length is correspondingly small. Therefore, again in short, the horizontally formed matrices can provide the results as expected.

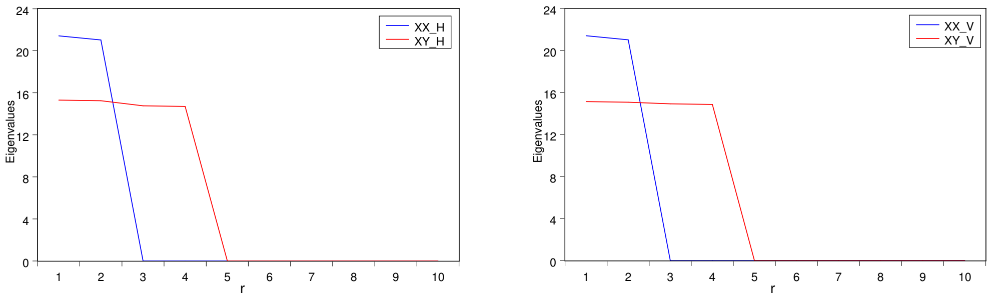


Fig. B.4 Eigenvalues of Two Sine Waves [-1,1] by Horizontal (left) and Vertical (right) SVD

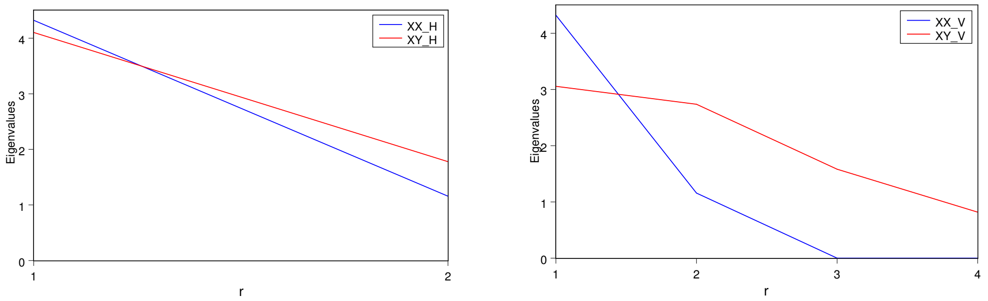


Fig. B.5 Eigenvalues of Two Sine Waves [-1,1] by Horizontal (left) and Vertical (right) SVD (L=2)

The differences between the horizontally formed and the vertically formed groups of exponential distribution series are relatively significant just as previous scenarios. The comparison shows that horizontally formed matrices can better provide the similar distribution of eigenvalues as simultaneously expected.

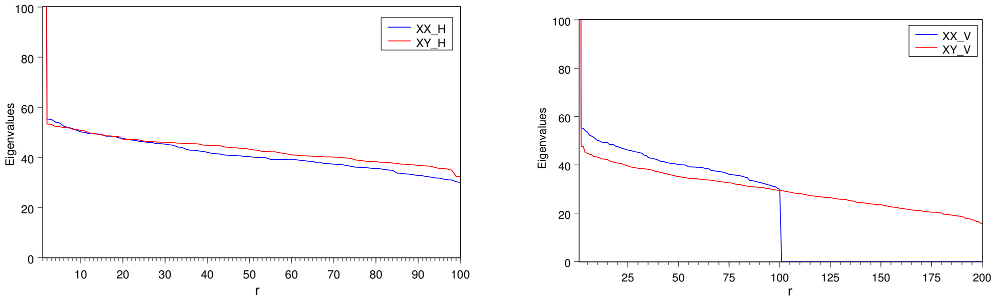


Fig. B.6 Eigenvalues of Two Exponential Distribution by Horizontal (left) and Vertical (right) SVD

In summary, according to the results of different types of simulated series and scenarios considered, horizontal form of matrices can always present the results as expected in the first place, therefor horizontal SVD will be the suitable procedure to produce comparable singular values before examining the distribution of eigenvalues for the proposed novel similarity measure.

## B.2 Without Premise of Multivariate System

The previous statement of this new proposed similarity measure is built on a fundamental multivariate system of the benchmark series. Instead, here only the $X$ and $Y$ are considered separately. Consistent with the previous statements, the same group of different types of series are generated for evaluation. Regarding all simulated series for each type of distribution, it is expected to have similar eigenvalues distributions between $X$ and $Y$ formed matrices. Similarly, the default number of observation for simulated series is 1000 and default window length for embedding process is 100.

It is worth to be noted that this time two series are considered without premise of a multivariate system, instead, two system of each series with itself are formed separately, the corresponding horizontal and vertical form matrices for one series are symmetric, hence the first 100 eigenvalues are identical while the second half of eigenvalues for vertical case are nearly identical to 0. Eigenvalues for several different types of generated series are shown in the following figures, which confirm our expected results.

As the eigenvalues by horizontal form matrix are identical to the first half of eigenvalues of vertical form matrix, the 100 eigenvalues will be employed for measuring similarity by comparing its distribution with the eigenvalues from the other series accordingly.

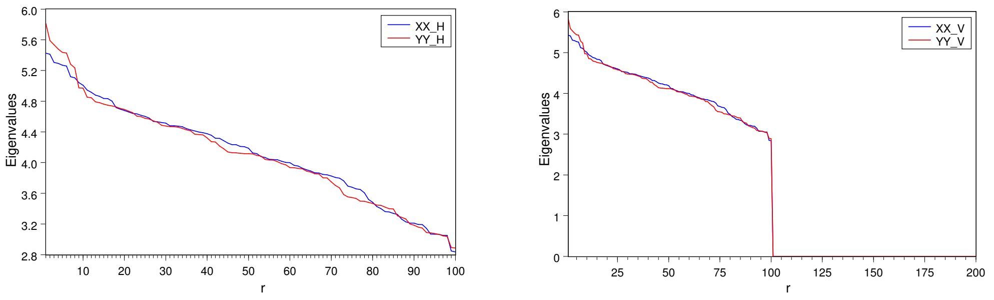


Fig. B.7 Eigenvalues of Two White Noise by Horizontal (left) and Vertical (right) SVD

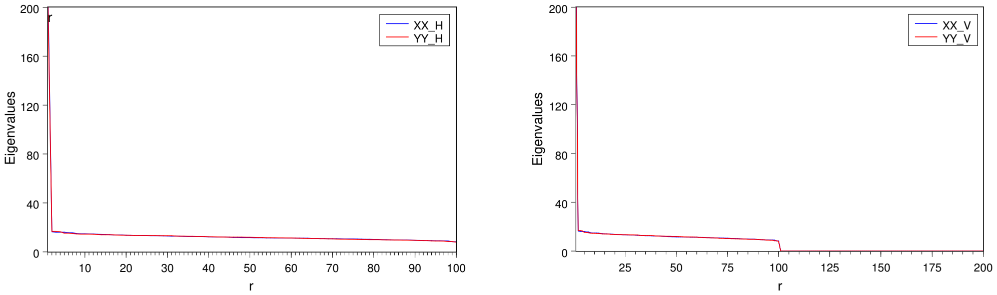


Fig. B.8 Eigenvalues of Two Uniform Distributions [0,1] by Horizontal (left) and Vertical (right) SVD

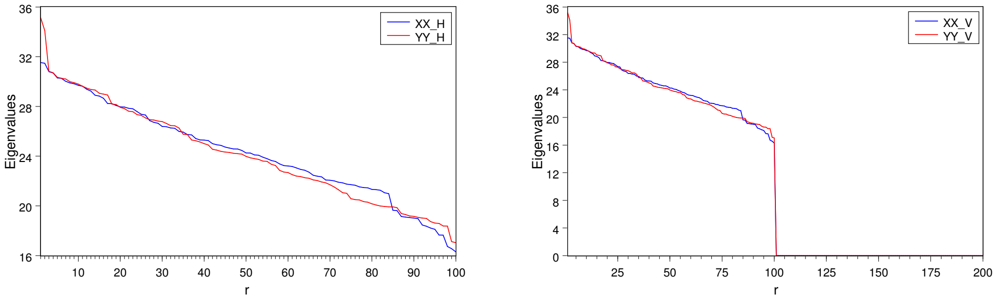


Fig. B.9 Eigenvalues of Two Uniform Distributions [-1,1] by Horizontal (left) and Vertical (right) SVD

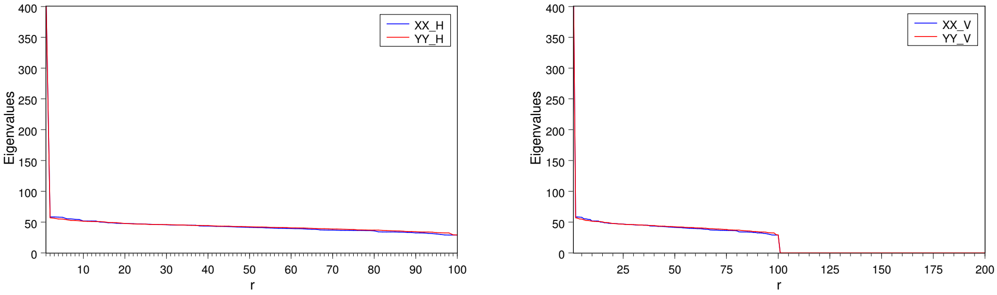


Fig. B.10 Eigenvalues of Two Exponential Distribution by Horizontal (left) and Vertical (right) SVD

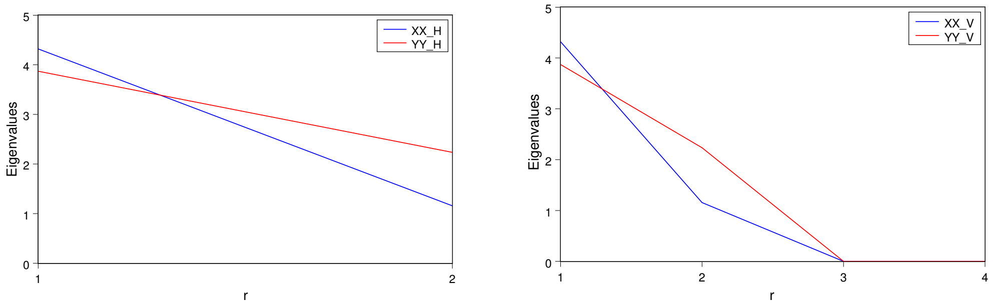


Fig. B.11 Eigenvalues of Two Sine Waves [-1,1] by Horizontal (left) and Vertical (right) SVD (L=2)

# Appendix C

# Development of Causality Analysis by Linear Models

Wold (1954) stated that the concept of causality is indispensable and fundamental to all science. The explorations and investigations of the "why" question have started and consistently insisted since the very beginning prior to Aristotle. The approaches of causality analysis vary greatly when the causes are considered from different aspects. In this Appendix, as an extension of the review of Granger casuality approach in Chapter 5, section 5.2, it is aimed to providing a brief review of the developments of Granger causality approach alongside with corresponding statistical techniques as follows.

## C.1 Simon's Approach

Simon (1954) proposed the "genuine" correlation as a causal interpretation, in which he stated the necessity of concluding additional variables and equations to create a wider system. Therefore, in order to answer the question whether including a third variable will affect the relationship between two variables, a three variable system was introduced by Simon (1954) as follows:

$$\begin{aligned} x+a_{12}y+a_{13}z &= u_1 \\ a_{21}x+y+a_{23}z &= u_2 \\ a_{31}x+a_{32}y+z &= u_3 \end{aligned} \tag{C.1}$$

where $u$ refers to the error term, therefor $A = \|a_{ij}\|$ indicates the coefficient matrix of the system. Consequently, if some of the elements of the coefficient matrix are zero, equivalently, not all the variables directly influence all the others. By making assumptions of time precedence and non-correlation of the error terms, the parameters estimated by sufficient assumptions reflect whether the original variable are causally related, therefore, determine the "genuine" correlation and interpret the causal relationship. More specifically, one example from Simon (1954) stated that if one assumes $a_{31} = a_{32} = a_{21}$, the significant results by examining the system equations indicate that $y$ is causally dependent on $z$ and $x$ is causally dependent on $y$ and $z$. Therefore, for $a_{12} \neq 0$ and the assumption of $x$ and $y$ were correlated, one can then lead to the conclusion of genuine correlation. However, it has to be noted that the "wider system" by Simon (1954) is assumed on the basis of linearity and variables included are measured from their respective means.

## C.2 Wold's Approach

Wold (1954) proposed the "causal chain" to specify a recursive structure for a system of simultaneous equation. It is briefly introduced below by following the example from Wold (1954): Consider the analysis is based on the relationship between demand ($d$) and price ($p$) $d_t = D(p_t)$, where $t$ is the trend or time index. Additionally, it is already assumed that the relationship between supply ($s$) and price is realistic and far complex to be specified. Therefor the estimates can be built on previous development accordingly as follows:

$$\begin{aligned} s_t &= S(d_{t-1}, s_{t-1}, p_{t-1}, d_{t-2}, \cdots) + u_t \\ p_t &= P(d_{t-1}, s_{t-1}, p_{t-1}, \cdots) + v_t \end{aligned} \tag{C.2}$$

Wold (1954) stated that each of the relations above allows a causal interpretation and it will provide the same precise results as the original model of:

$$\begin{aligned} d_t &= D(p_t) \\ s_t &= S(p_{t-1}) \\ p_t &= p_{t-1} + \gamma(d_{t-1} - s_{t-1}) \end{aligned} \tag{C.3}$$

Malinvaud (1980) proposed that there is a natural analogue in a dynamic system to Wold's "causal chain" form for a static econometric model. Sims (1972) stated that this analogue turns out to be exactly a model in which causation is unidirectional according to the criterion developed later by Granger (1969); Wold's form is in general not testable in a static context as any multivariate set of data with a specified list of endogenous variables can be fit by a recursive model. According to Sims (1972), the dynamic analogue is easily testable: If and only if causality runs one way from current and past values of some list of exogenous variables to a given endogenous variable, then in a regression of the endogenous variable on past, current, and future values of the exogenous variables, the future value of the exogenous variables should have zero coefficients.

## C.3 Granger's Approach

Granger (1969) focused on the incremental predictability for answering the question of definition of causality and he proposed the statistical approach, Granger causality test, that is the most general and significant method for testing the causality relationship between two variables in the linear regression model. Granger suggested ”causality” is tastable using of simple regression or correlation techniques in two-variable models. In addition, a simple Granger Causality, the instantaneous Granger Causality and a feedback model are also discussed, where current as well as past values of $x$ are used to predict $y_t$. If $y$ is related to current or lagged $x$, but not future $x$, $x$ is exogenous relative to $y$ (Schwert, 1979). Details of Granger's approach to test causality are listed as follows, which primarily follows Granger (1969):

Assume there are two stationary stochastic time series $X$ and $Y$, let $I_t$ be the set of all the information in the universe accumulated since time $t-1$, so the $I_t - Y_t$ will denote all the information apart from series $Y_t$ and let $\sigma^2$ be the corresponding forecast error. $\overline{X_t}$ represents the set of past values $\{X_{t-j}, j = 1,2,...,\infty\}$ and $\overline{\overline{X_t}}$ represents the set of past and present values $\{X_{t-j}, j = 0,1,2,...,\infty\}$. Thus $\sigma^2(X|\overline{I_t})$ will be the prediction of $X$, using all the information from the past, and $\sigma^2(X|\overline{I_t - Y_t})$ will be the prediction of $X$, using all the information from the past apart from the series $Y$.

- ***Simple Granger Causality***

  If the forecast error of $X$ based on all the information $I$ is smaller than the forecast error of $X$ based on the past information apart from series $Y$, which is denoted as $\sigma^2(X|I_t) < \sigma^2(X|\overline{I_t - Y_t})$, then $Y$ is causing $X$. Granger (1969) stated as ”if we are better able to predict $X$ using all available information than if the information apart from $Y$ had been used, we say that $Y$ is causing $X$”.

- ***Feedback Model***

  If the Simple Granger Causality from $Y$ to $X$ is donated as $Y \Rightarrow X$, then the feedback indicates the situation that when $X$ is causing $Y$ and also $Y$ is causing $X$, which can be represented as $X \Leftrightarrow Y$, also can be denoted as following: If $\sigma^2(X|\overline{I}) = \sigma^2(X|\overline{I-Y})$ and $\sigma^2(Y|\overline{I}) = \sigma^2(Y|\overline{I-X})$, then we say $X \Leftrightarrow Y$.

- ***Instantaneous Granger Causality***

  The instantaneous causality is indicated if better forecast of current value of $X$ can be conducted when the present value of $Y$ is considered together than only considering the set of all the past information. It can be donated as: if $\sigma^2(X|\bar{I},\bar{\bar{Y}})$, the instantaneous causality of $Y_t \Rightarrow X_t$ is occurring.

Another significant definition proposed is the ”causality lag” by Granger (1969). In which, the least value of $k$ such that $\sigma^2(X|I-Y(k)) < \sigma^2(X|I-Y(k+1))$ is defined as (integer) causality lag $m$, which also indicates that the values $Y_{t-j}, j = 0,1,...,m-1$ can provide no additional help in improving the forecast of $X_t$. The regression formulation of Granger causality states that a variable $X$ is the cause of another variable $Y$ if the past values of $X$ are helpful in predicting the future value of $Y$, two regressions are considered as follows:

$$\begin{aligned} Y(t) &= \sum_{l=1}^{L} \pi_l Y(t\text{-}l) + \varepsilon_1, \\ Y(t) &= \sum_{l=1}^{L} \pi_l Y(t\text{-}l) + \sum_{l=1}^{L} \gamma_l X(t\text{-}l) + \varepsilon_2, \end{aligned} \tag{C.4}$$

where $L$ is the maximal time lag, $\pi$ and $\gamma$ are vectors of coefficients, $\varepsilon$ is the prediction error term (Liu and Bahadori, 2012). If the second is a significantly better model than the first one, one determines that time series $X$ Granger causes time series $Y$ (Liu and Bahadori, 2012).

## C.4 Sims Test

Sims (1972, 1980) provided the vector auto-regressive (VAR) processes to answer the question that whether some specific time series are generated independently of the other time series considered. Assume two variables $X$ and $Y$, Sims (1972) proposed the estimate model as:

$$Y_t = a + b_{-k}X_{t-k} + \cdots + b_{-1}X_{t-1} + b_0X_t + b_1X_{t+1} + \cdots + b_mX_{t+m} + u_t \tag{C.5}$$

where $a$ and $b$ are corresponding parameters; $k$ and $m$ are positive integers; $t$ refers to time index; and $u_t$ is the disturbance term.

Therefore, the null hypothesis is $H_0 : b_1 = \cdots = b_m = 0$. Next, calculate the sum of square error of the model by ordinary least squares (OLS) under the null hypothesis and note as $SSE_h$; whilst again compute the sum of square error of the model without the null hypothesis condition by OLS and note it as $SSE_a$. The $F$ statistic then can be calculated by:

$$F_{sim} = (\frac{SSE_h - SSE_a}{k+1})/(\frac{SSE_a}{n-k-m-2}) \tag{C.6}$$

where $n$ is the number of observations. In accordance of the corresponding $F$ statistic, the null hypothesis then can be determined to be rejected or not accordingly. Consequently, the conclusion can be summarized whether $X$ and $Y$ have causal relationship or not.

According to Sims (1972), the proposed approach is equivalent to the approach of Granger. However, Chamberlain (1982) extends the Granger and Sims approaches by using conditional independence instead of linear predictors, it indicated these two approaches varies and non-causality is stronger than strict exogeneity. One drawback of this test is that there might be a third variable $Z$ causing $Y$, but $Z$ might be contemporaneously correlated with $X$. In such circumstances the test would still show $X$ causing $Y$ leading to a false conclusion.

## C.5 Haugh-Pierce Test

Haugh (1976) firstly stated that causal relations between two time series can also be characterised by the residuals of their univariate auto-regressive moving average models. Then in (Pierce and Haugh, 1977), the Haugh-Pierce test was made popular, by which employs the estimated residuals of the univariate models for $X$ and $Y$. Briefly, this proposed two step procedure of causality test is introduced mainly following (Kirchgässner et al., 2012; Schwert, 1979).

First, the evaluated variables are transformed to $x$ and $y$ by using logarithm or differentiating to have constant unconditional mean and variance over the sample period. Followed by the estimations of two univariate autoregressive moving average (ARMA) models for the transformed variables below (Nelson, 1973):

$$\begin{aligned} \Phi_y(L)y_t &= a + \theta_y(L)a_{yt} \\ \Phi_x(L)x_t &= b + \theta_x(L)b_{xt} \end{aligned} \tag{C.7}$$

where $\Phi(L)$ are finite autoregressive polynomials in the lag operator for transformed variables $x$ and $y$ respectively; whilst $\theta(L)$ refers to finite moving average polynomials scenario. Then, the second step is examining cross-correlations between the residuals of the univariate ARMA models. Based on Kirchgässner et al. (2012), assume the corresponding cross-correlations as $\rho_{ab}(t)$, then the following statistics is computed:

$$S = T \cdot \sum_{t=t_1}^{t_2} \hat{\rho}_{ab}(t) \tag{C.8}$$

Consequently, if the null hypothesis $H_0 : \rho_{ab}(t0 = 0$ is reject, for $t_1 = 1 \wedge t_2 \geq 1$, the causal relation from $x$ to $y$ can be examined; reversely, for $t_1 \leq -1 \wedge t_2 = -1$, the causal relationship from $y$ to $x$ can be tested.

Schwert (1979) mentioned that the power of this procedure, which use correlations, is smaller than the power of the Granger approach which uses regressions. What is more, spurious independence might occur (Pierce, 1977), sometimes caused by omitted variables (Lütkepohl, 1982). Feige and Pearce (1979) stated that this test provided deep insight of information, but it might only be a first step to analyse causal relations between time series; on the other hand, information on the relations between two time series, which is contained in cross-correlations, can be useful even if no formal test is applied.

## C.6 Hsiao Test

Hsiao (1979) developed another procedure to estimate bivariate models and interpret causal relationships. It is similar to the Granger approach and is also built on autoregressive representation. Specifically, Hsiao (1979) suggested that the lag length-s should be determined with an information criterion – Akaike Criterion (Akaike, 1974) or Schwarz Criterion (Schwarz et al., 1978).

Assume the two evaluated variables in the bivariate model are $X$ and $Y$, the procedure of Hsiao test of $X$ "Hsiao causes" $Y$ starts from determining the optimal lag $l_y$ of the univariate autoregressive (AR) of $Y$. By fixing $l_y$, the optimal lag $l_x$ of $X$ in regression of explaining $Y$ is defined. Next step is fixing $l_x$ and re-evaluating the optimal lag of dependent variable $Y$, note as $l_{\bar{y}}$. Therefore, by comparing the information criterion in the processes of defining $l_{\bar{y}}$ and $l_y$, it can be concluded that $X$ has significant impact on $Y$ when the last information criterion is larger than the value in the beginning. Reversely, the test of causal impact from $Y$ to $X$ can be reproduced by exchanging the positions of $X$ and $Y$.

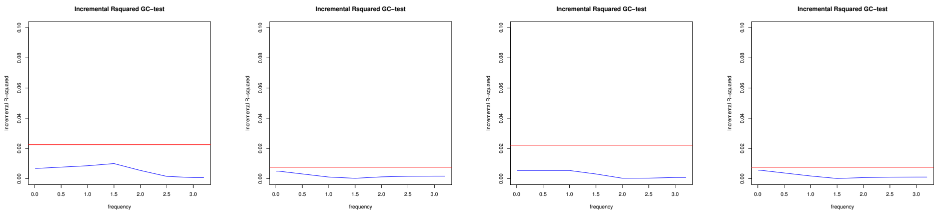
Kirchgässner et al. (2012) stated that the Hsiao test only can capture the simple causal relations between the two variables. Whilst, the correlation between the residuals can reflect the possible instantaneous relation. It also has to be noticed that the optimal lag length of the dependent variable and the conditioning variables must be determined before the optimal lag length of the other independent variable is fixed.

# Appendix D

# Detailed Results of Frequency Domain GC Test and CCM Causality Test

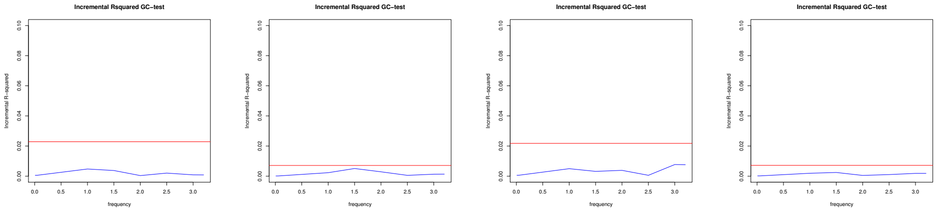
## D.1 Frequency Domain GC Test Results of Oil Prices and Tourist Arrivals

Note that having greater test statistics (blue) than the corresponding 5% critical values (red) indicates possible causal links within corresponding frequency range. Also, the optimal lag-structures are maintained for all tests.



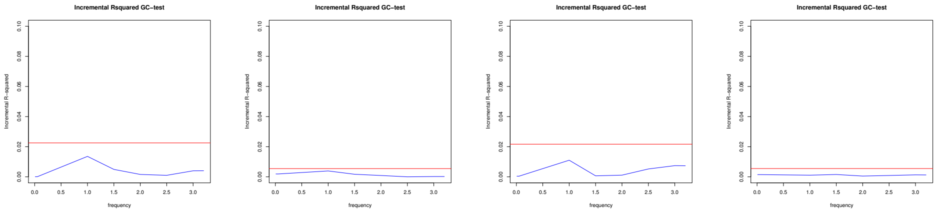
(a) TA→BRT (b) TA←BRT (c) TA→WTI (d) TA←WTI

Fig. D.1 Frequency Causality Results for Austria Tourists Arrivals and Oil Prices.



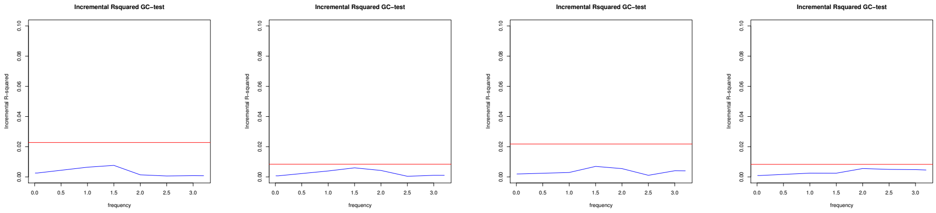
(a) TA→BRT (b) TA←BRT (c) TA→WTI (d) TA←WTI

Fig. D.2 Frequency Causality Results for Germany Tourists Arrivals and Oil Prices.



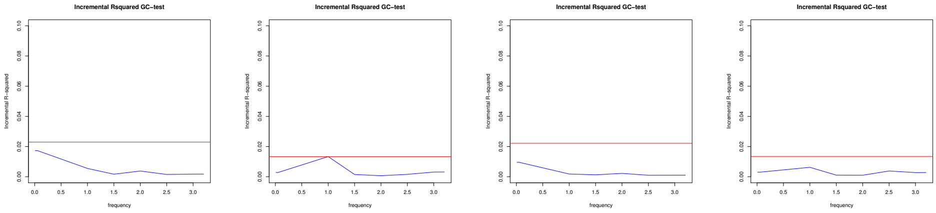
(a) TA→BRT (b) TA←BRT (c) TA→WTI (d) TA←WTI

Fig. D.3 Frequency Causality Results for Greece Tourists Arrivals and Oil Prices.



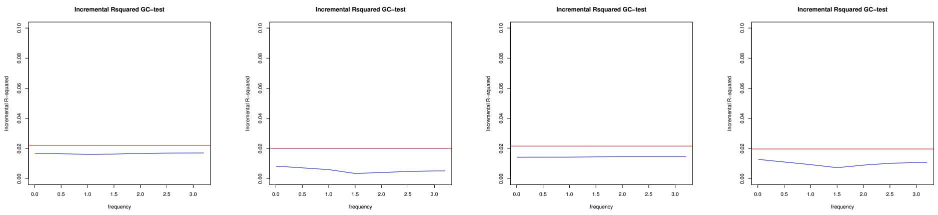
(a) TA→BRT (b) TA←BRT (c) TA→WTI (d) TA←WTI

Fig. D.4 Frequency Causality Results for Italy Tourists Arrivals and Oil Prices.



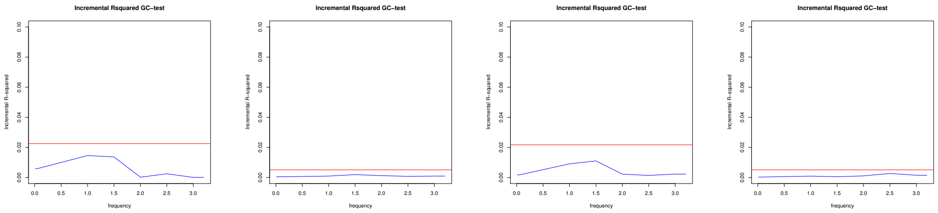
(a) TA→BRT (b) TA←BRT (c) TA→WTI (d) TA←WTI

Fig. D.5 Frequency Causality Results for Netherland Tourists Arrivals and Oil Prices.



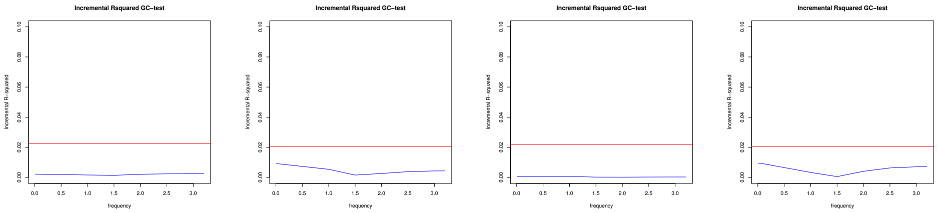
(a) TA→BRT (b) TA←BRT (c) TA→WTI (d) TA←WTI

Fig. D.6 Frequency Causality Results for Portugal Tourists Arrivals and Oil Prices.



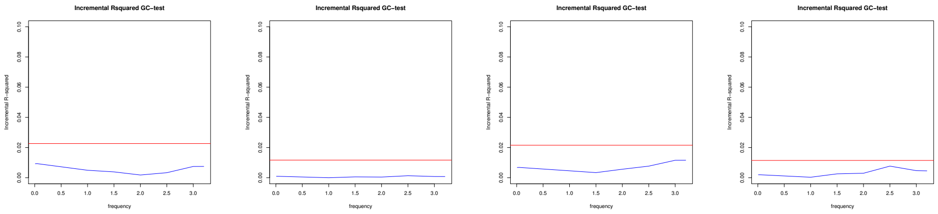
(a) TA→BRT (b) TA←BRT (c) TA→WTI (d) TA←WTI

Fig. D.7 Frequency Causality Results for Spain Tourists Arrivals and Oil Prices.



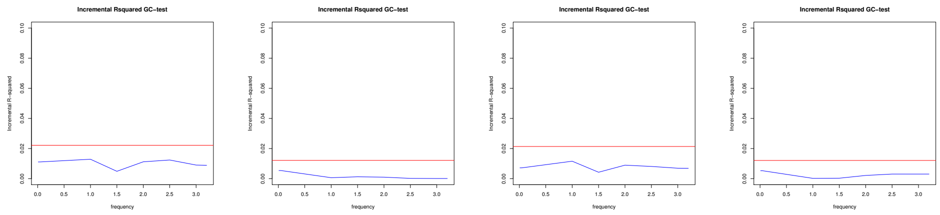
(a) TA→BRT (b) TA←BRT (c) TA→WTI (d) TA←WTI

Fig. D.8 Frequency Causality Results for Sweden Tourists Arrivals and Oil Prices.



(a) TA→BRT (b) TA←BRT (c) TA→WTI (d) TA←WTI

Fig. D.9 Frequency Causality Results for UK Tourists Arrivals and Oil Prices.

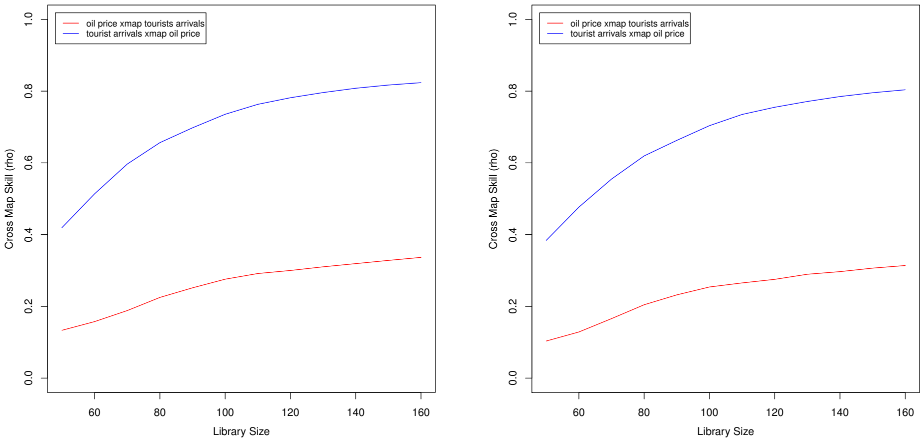


(a) TA→BRT (b) TA←BRT (c) TA→WTI (d) TA←WTI

Fig. D.10 Frequency Causality Results for US Tourists Arrivals and Oil Prices.

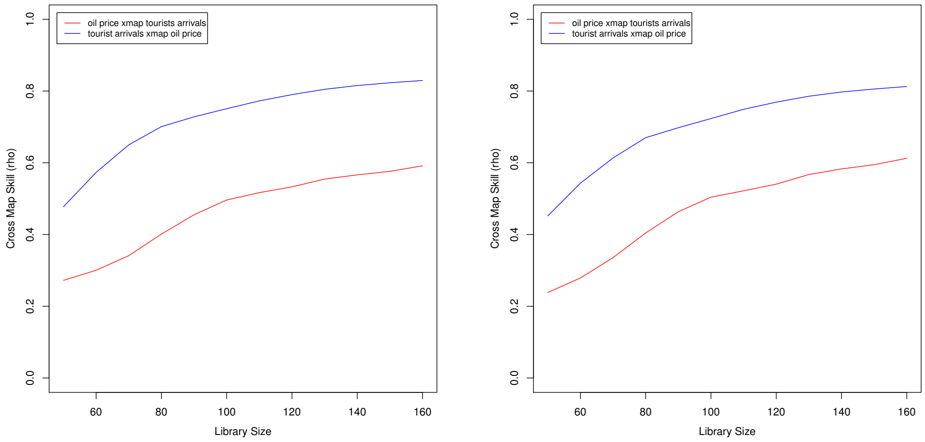
## D.2 CCM Test Results of Oil Prices and Tourist Arrivals

The cross map skill index reflects the reconstruction ability of the fact factor to the cause factor for both directions respectively. Here more specifically, the blue line above red line means significant cross map skill of tourist arrivals on oil price, which indicates causality from oil price to tourist arrivals.



(a) BRT (b) WTI

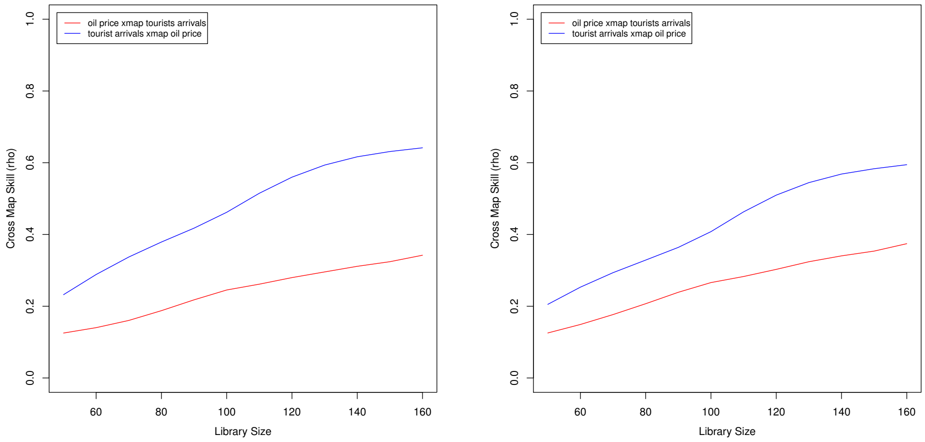
Fig. D.11 CCM Causality Results for Austria Tourists Arrivals and Oil Prices.



(a) BRT

(b) WTI

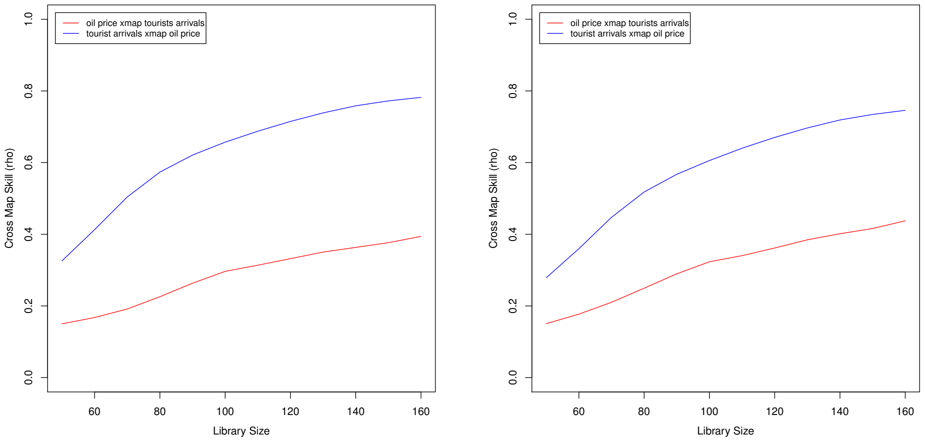
Fig. D.12 CCM Causality Results for Germany Tourists Arrivals and Oil Prices.



(a) BRT

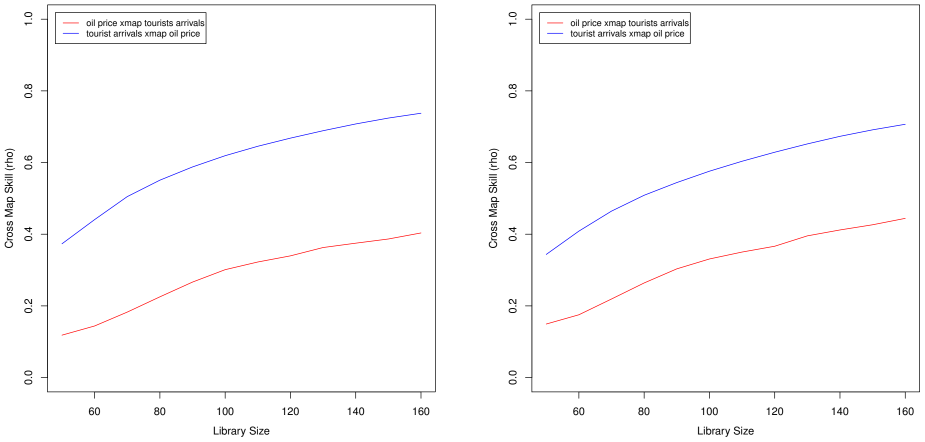
(b) WTI

Fig. D.13 CCM Causality Results for Greece Tourists Arrivals and Oil Prices.



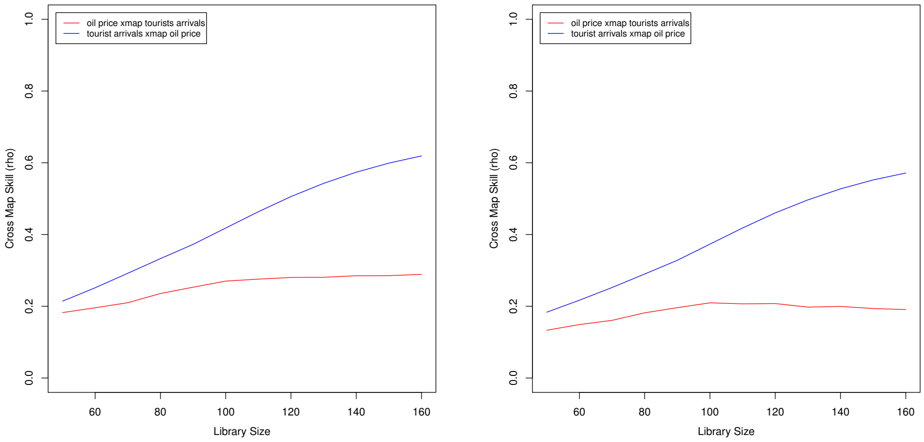
(a) BRT (b) WTI

Fig. D.14 CCM Causality Results for Italy Tourists Arrivals and Oil Prices.



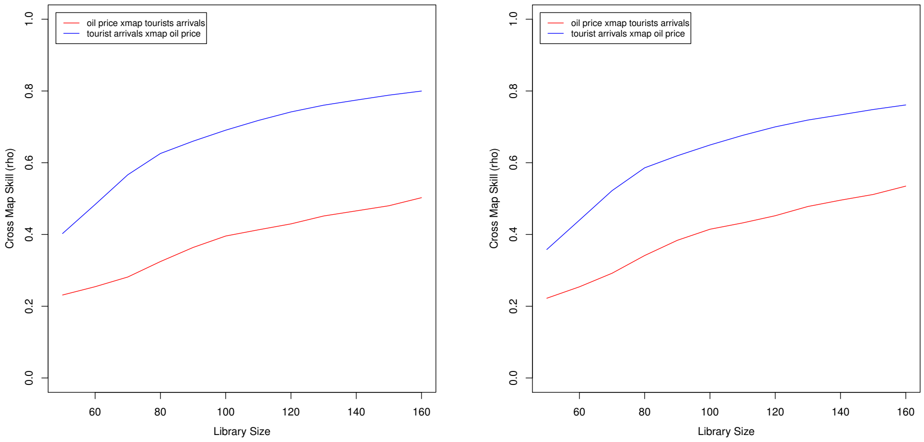
(a) BRT (b) WTI

Fig. D.15 CCM Causality Results for Netherland Tourists Arrivals and Oil Prices.



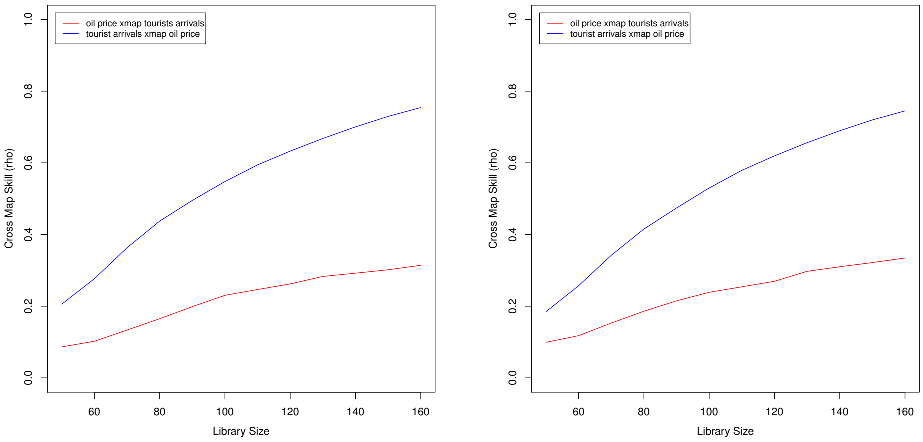
(a) BRT (b) WTI

Fig. D.16 CCM Causality Results for Portugal Tourists Arrivals and Oil Prices.



(a) BRT (b) WTI

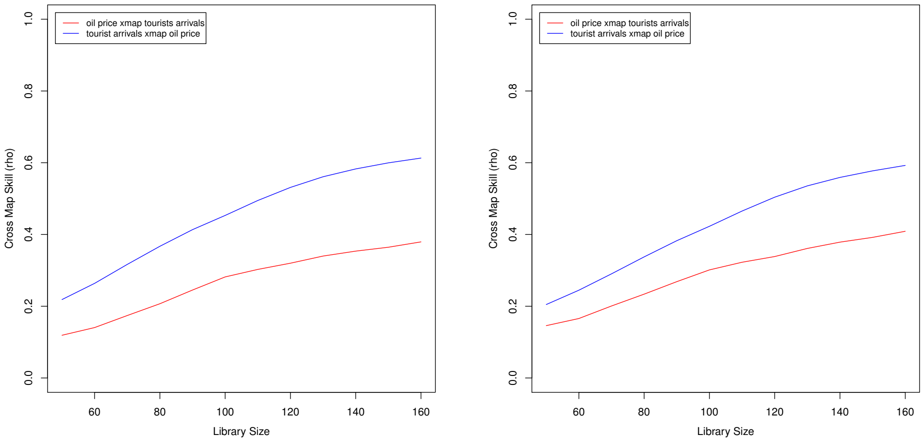
Fig. D.17 CCM Causality Results for Spain Tourists Arrivals and Oil Prices.



(a) BRT

(b) WTI

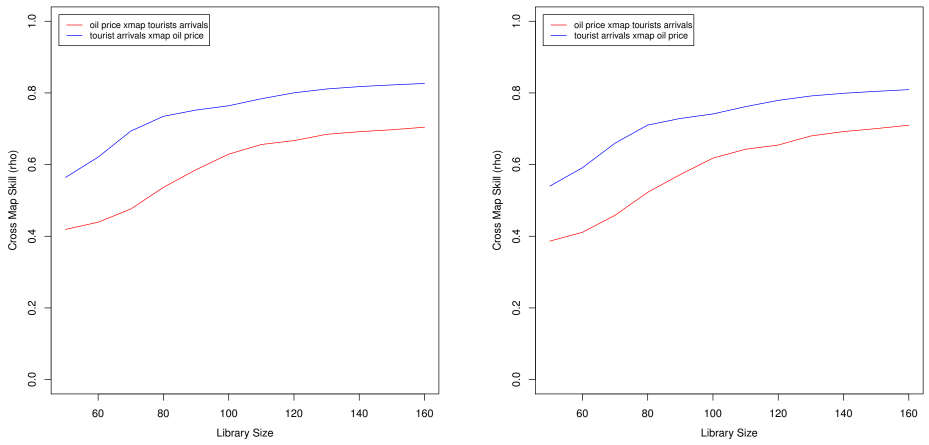
Fig. D.18 CCM Causality Results for Sweden Tourists Arrivals and Oil Prices.



(a) BRT

(b) WTI

Fig. D.19 CCM Causality Results for UK Tourists Arrivals and Oil Prices.



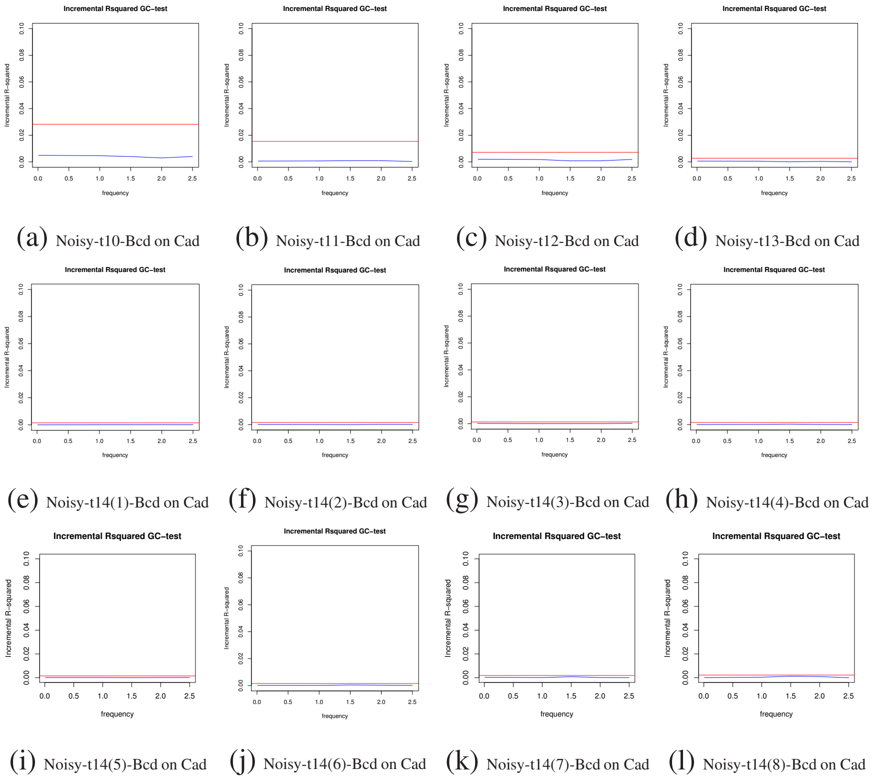
(a) BRT

(b) WTI

Fig. D.20 CCM Causality Results for US Tourists Arrivals and Oil Prices.

## D.3 Frequency Domain GC Test Results of Gene Profiles

It is of note that having greater test statistics (blue) than the corresponding 5% critical values (red) indicates possible causal links within corresponding frequency range. Also, the optimal lag-structures are maintained for all tests.

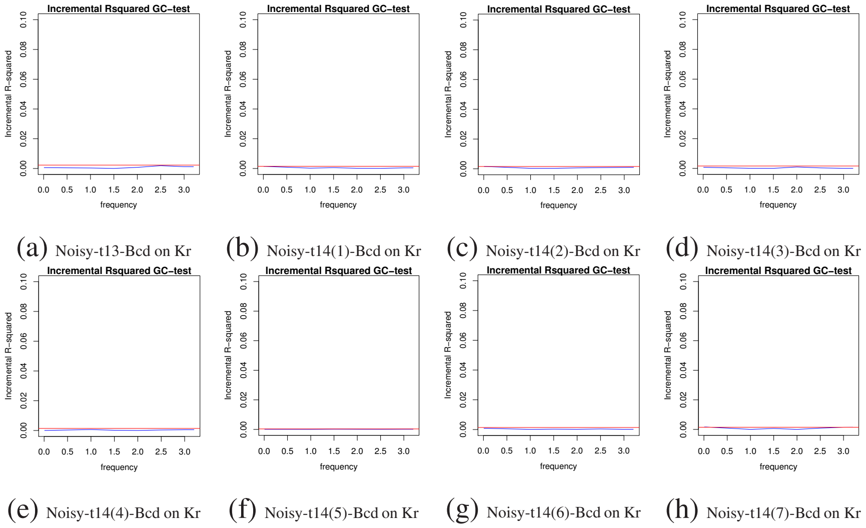


(a) Noisy-t10-Bcd on Cad (b) Noisy-t11-Bcd on Cad (c) Noisy-t12-Bcd on Cad (d) Noisy-t13-Bcd on Cad

(e) Noisy-t14(1)-Bcd on Cad (f) Noisy-t14(2)-Bcd on Cad (g) Noisy-t14(3)-Bcd on Cad (h) Noisy-t14(4)-Bcd on Cad

(i) Noisy-t14(5)-Bcd on Cad (j) Noisy-t14(6)-Bcd on Cad (k) Noisy-t14(7)-Bcd on Cad (l) Noisy-t14(8)-Bcd on Cad

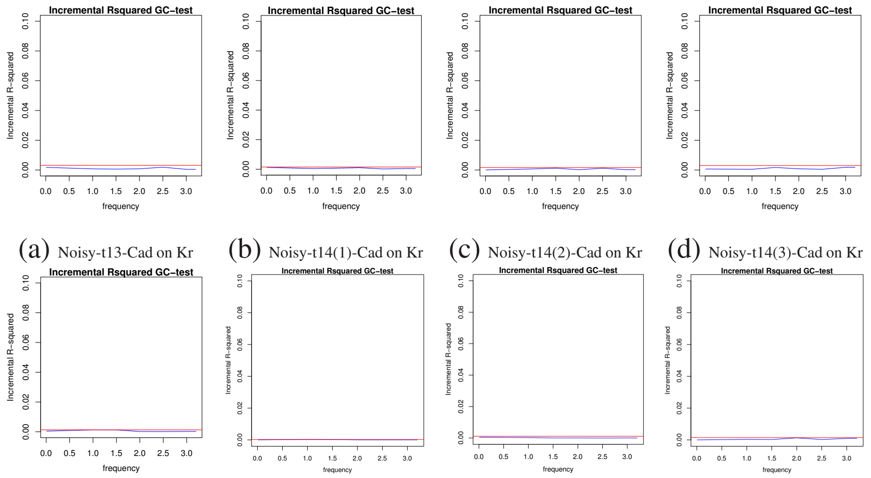
Fig. D.21 Frequency Domain GC Test Results for Bcd and Cad (Noisy Series).



(a) Noisy-t13-Bcd on Kr (b) Noisy-t14(1)-Bcd on Kr (c) Noisy-t14(2)-Bcd on Kr (d) Noisy-t14(3)-Bcd on Kr

(e) Noisy-t14(4)-Bcd on Kr (f) Noisy-t14(5)-Bcd on Kr (g) Noisy-t14(6)-Bcd on Kr (h) Noisy-t14(7)-Bcd on Kr

Fig. D.22 Frequency Domain GC Test Results for Bcd and Kr (Noisy Series).



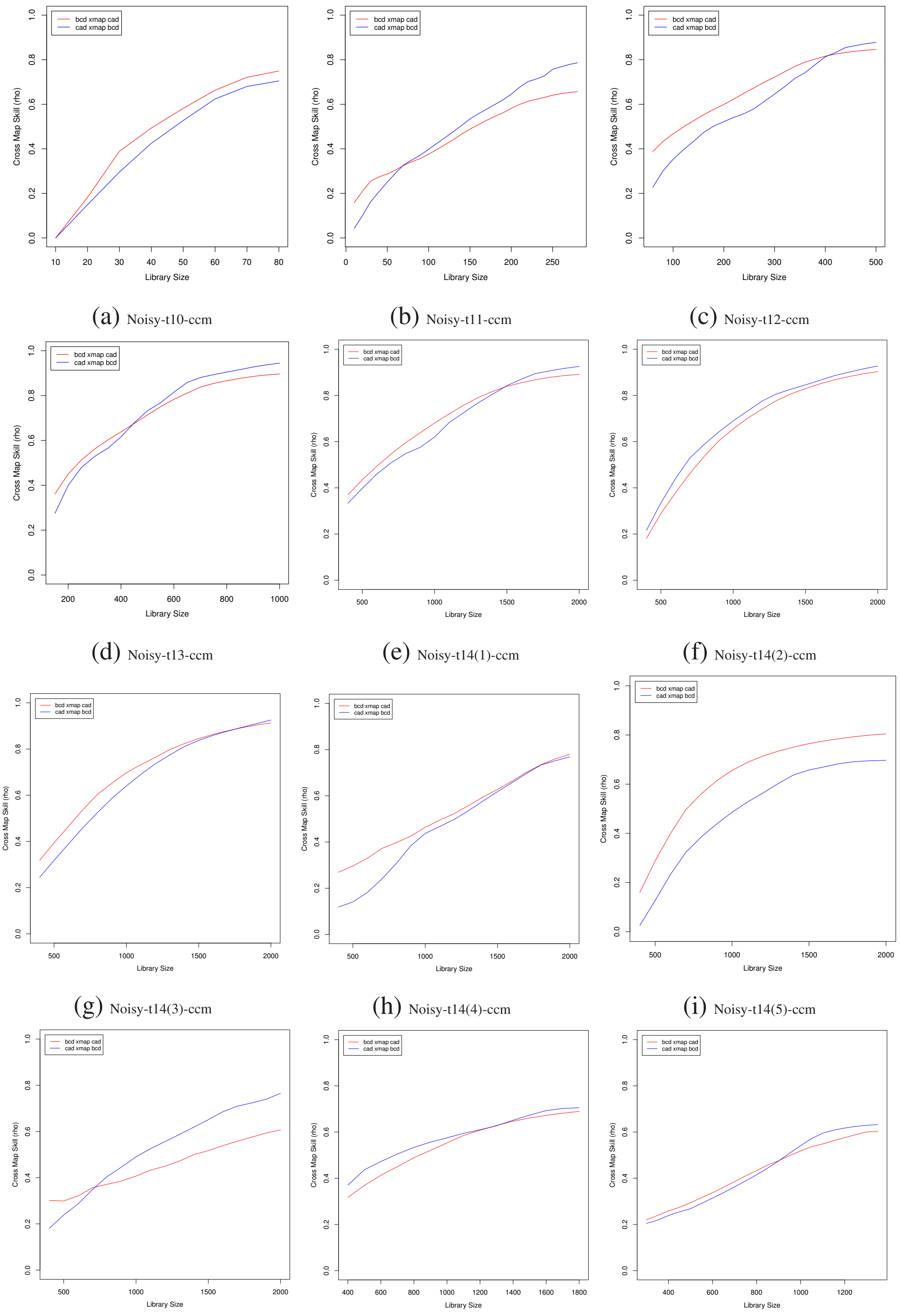
(a) Noisy-t13-Cad on Kr (b) Noisy-t14(1)-Cad on Kr (c) Noisy-t14(2)-Cad on Kr (d) Noisy-t14(3)-Cad on Kr

(e) Noisy-t14(4)-Cad on Kr (f) Noisy-t14(5)-Cad on Kr (g) Noisy-t14(6)-Cad on Kr (h) Noisy-t14(7)-Cad on Kr

Fig. D.23 Frequency Domain GC Test Results for Cad and Kr (Noisy Series).

## D.4 CCM Test Results of Gene Profiles

The cross map skill index reflects the reconstruction ability of the fact factor to the cause factor for both directions respectively. For instance, if the blue line represents the cross map skill index of *A* on *B* and the blue line lies above red line, it means significant cross map skill of *A* on *B*, which indicates causality from *B* to *A*.



(a) Noisy-t10-ccm

(b) Noisy-t11-ccm

(c) Noisy-t12-ccm

(d) Noisy-t13-ccm

(e) Noisy-t14(1)-ccm

(f) Noisy-t14(2)-ccm

(g) Noisy-t14(3)-ccm

(h) Noisy-t14(4)-ccm

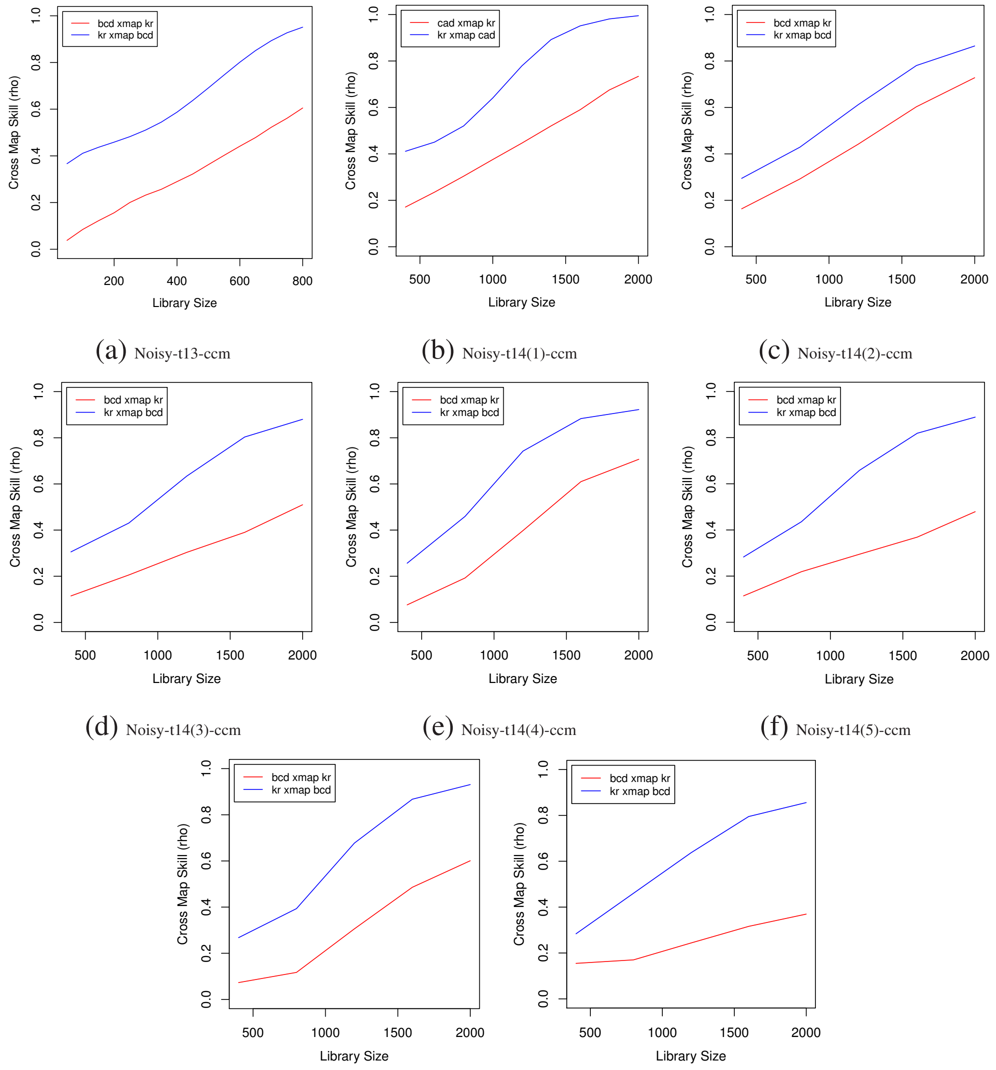
(i) Noisy-t14(5)-ccm

(j) Noisy-t14(6)-ccm

(k) Noisy-t14(7)-ccm

(l) Noisy-t14(8)-ccm

Fig. D.24 CCM Test Results for Bcd and Cad (Noisy Series).



(a) Noisy-t13-ccm

(b) Noisy-t14(1)-ccm

(c) Noisy-t14(2)-ccm

(d) Noisy-t14(3)-ccm

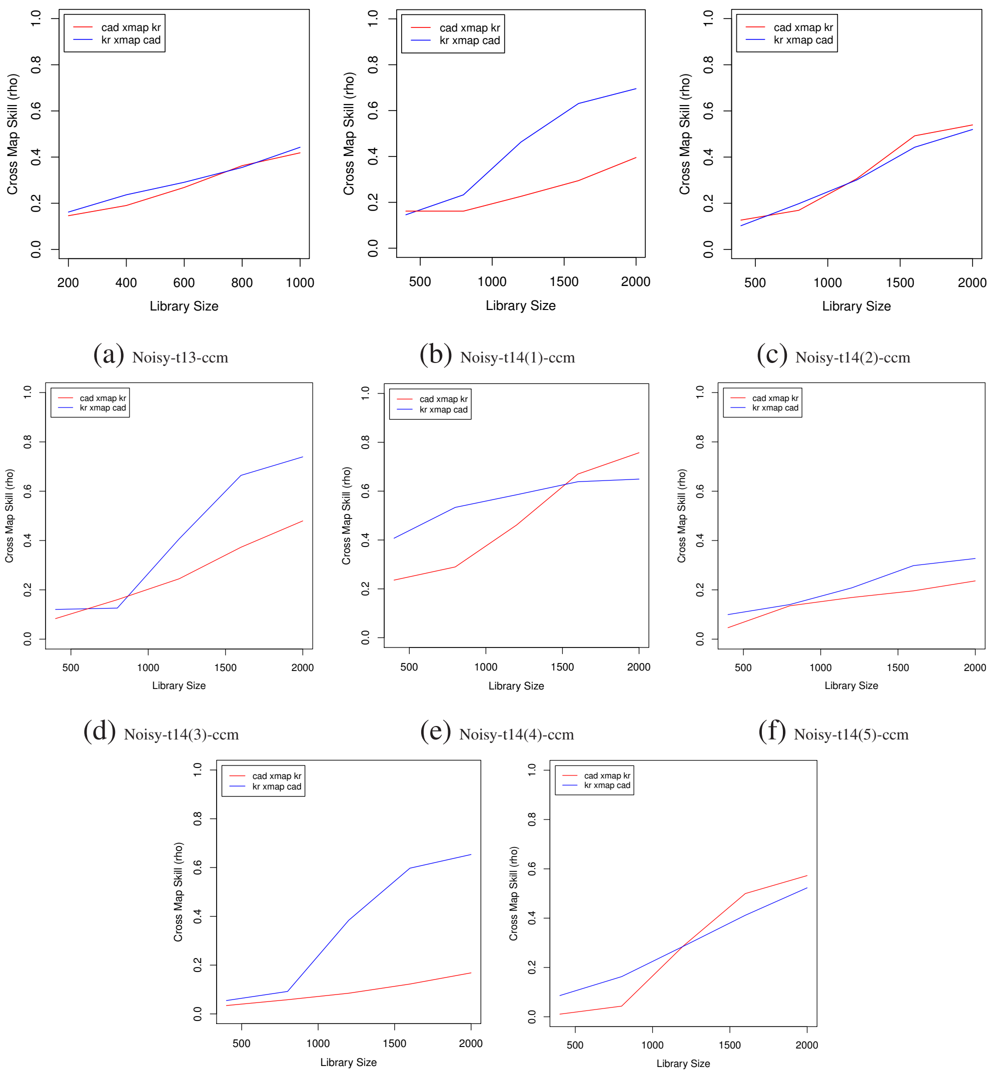
(e) Noisy-t14(4)-ccm

(f) Noisy-t14(5)-ccm

(g) Noisy-t14(6)-ccm

(h) Noisy-t14(7)-ccm

Fig. D.25 CCM Test Results for Bcd and Kr (Noisy Series).



(a) Noisy-t13-ccm (b) Noisy-t14(1)-ccm (c) Noisy-t14(2)-ccm

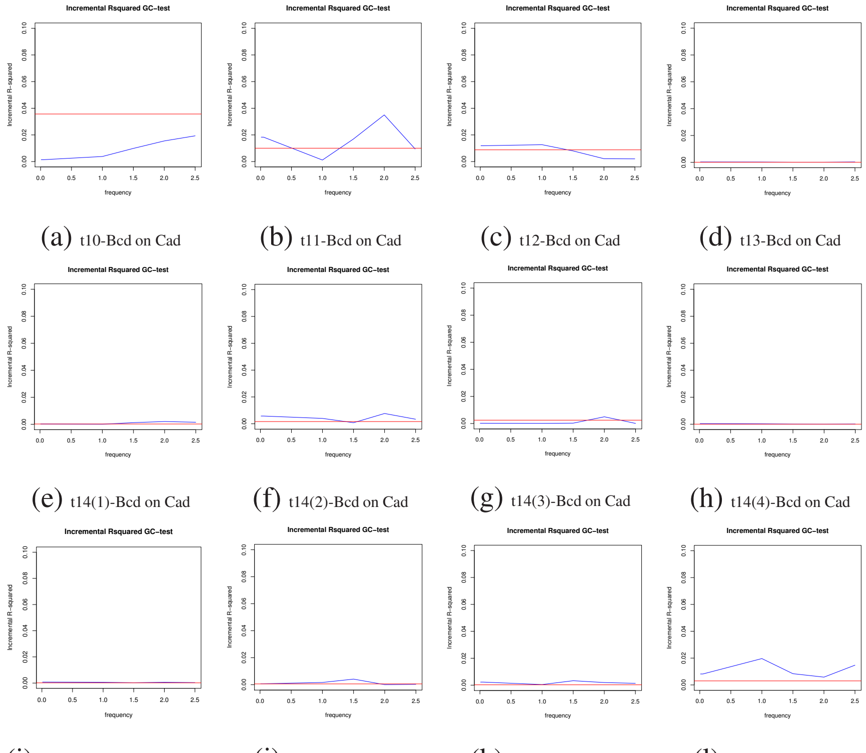
(d) Noisy-t14(3)-ccm (e) Noisy-t14(4)-ccm (f) Noisy-t14(5)-ccm

(g) Noisy-t14(6)-ccm (h) Noisy-t14(7)-ccm

Fig. D.26 CCM Test Results for Cad and Kr (Noisy Series).

## D.5 Frequency Domain GC Test Results of Filtered Gene Profiles

It is of note that having greater test statistics (blue) than the corresponding 5% critical values (red) indicates possible causal links within corresponding frequency range. Also, the optimal lag-structures are maintained for all tests.

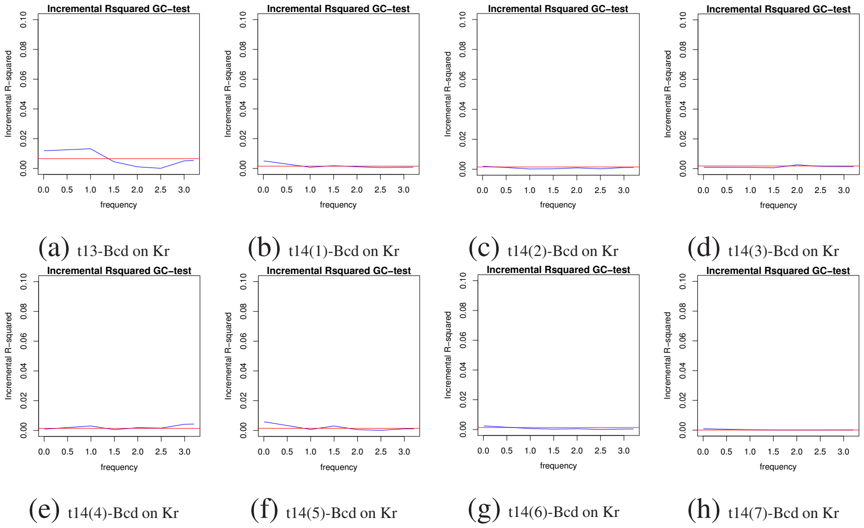


(a) t10-Bcd on Cad (b) t11-Bcd on Cad (c) t12-Bcd on Cad (d) t13-Bcd on Cad

(e) t14(1)-Bcd on Cad (f) t14(2)-Bcd on Cad (g) t14(3)-Bcd on Cad (h) t14(4)-Bcd on Cad

(i) Filtered-t14(5)-Bcd on Cad (j) t14(6)-Bcd on Cad (k) t14(7)-Bcd on Cad (l) t14(8)-Bcd on Cad

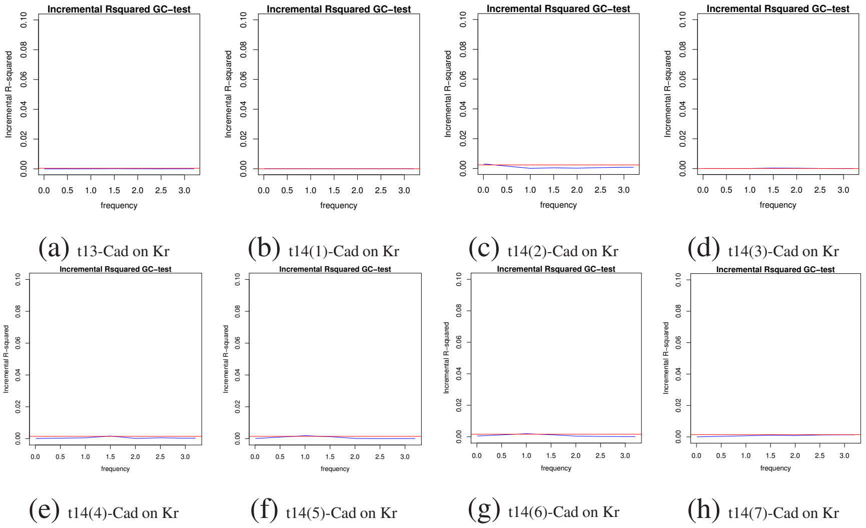
Fig. D.27 Frequency Domain GC Test Results for Bcd and Cad (Filtered Series).



(a) t13-Bcd on Kr (b) t14(1)-Bcd on Kr (c) t14(2)-Bcd on Kr (d) t14(3)-Bcd on Kr

(e) t14(4)-Bcd on Kr (f) t14(5)-Bcd on Kr (g) t14(6)-Bcd on Kr (h) t14(7)-Bcd on Kr

Fig. D.28 Frequency Domain GC Test Results for Bcd and Kr (Filtered Series).



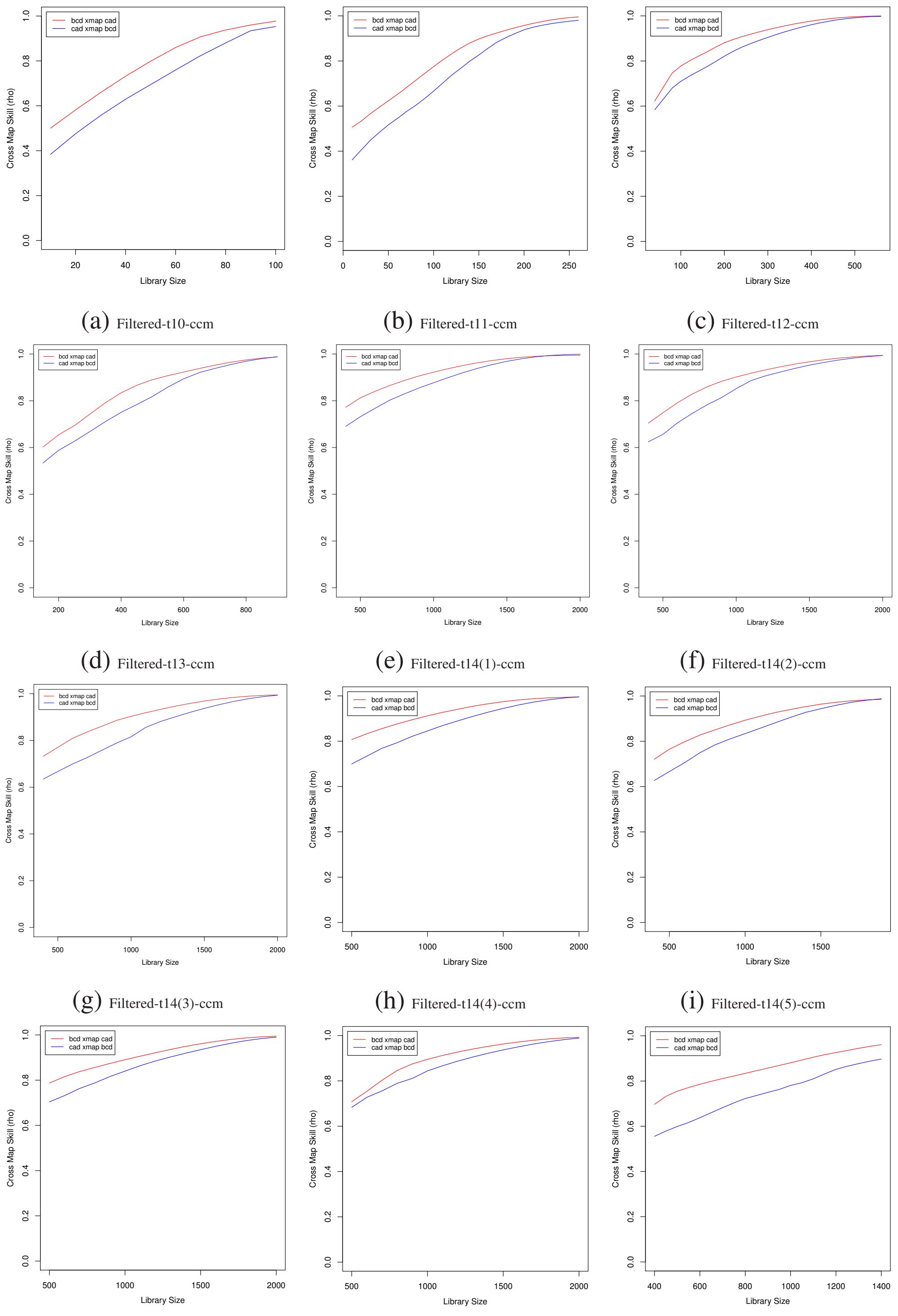
(a) t13-Cad on Kr (b) t14(1)-Cad on Kr (c) t14(2)-Cad on Kr (d) t14(3)-Cad on Kr

(e) t14(4)-Cad on Kr (f) t14(5)-Cad on Kr (g) t14(6)-Cad on Kr (h) t14(7)-Cad on Kr

Fig. D.29 Frequency Domain GC Test Results for Cad and Kr (Filtered Series).

## D.6 CCM Test Results of Filtered Gene Profiles

The cross map skill index reflects the reconstruction ability of the fact factor to the cause factor for both directions respectively. For instance, if the blue line represents the cross map skill index of *A* on *B* and the blue line lies above red line, it means significant cross map skill of *A* on *B*, which indicates causality from *B* to *A*.



(a) Filtered-t10-ccm

(b) Filtered-t11-ccm

(c) Filtered-t12-ccm

(d) Filtered-t13-ccm

(e) Filtered-t14(1)-ccm

(f) Filtered-t14(2)-ccm

(g) Filtered-t14(3)-ccm

(h) Filtered-t14(4)-ccm

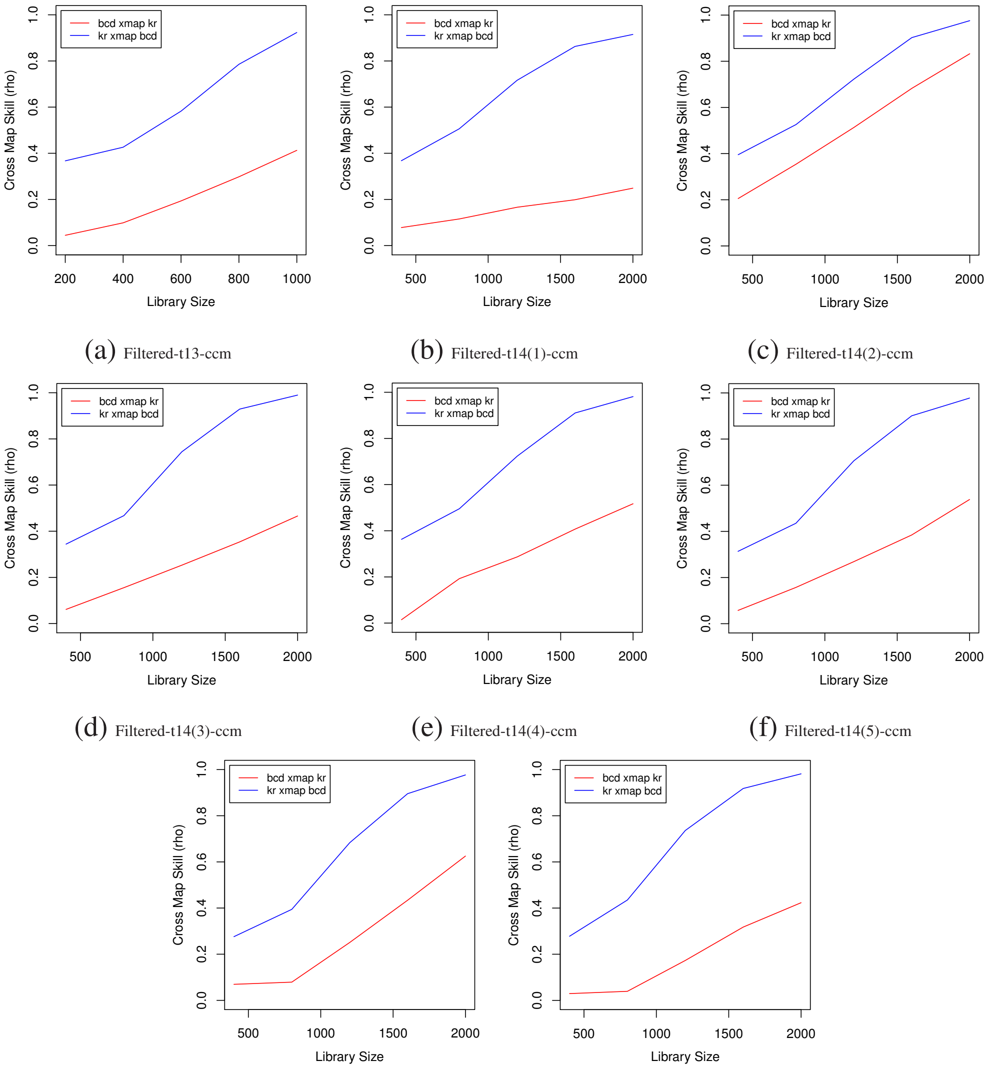
(i) Filtered-t14(5)-ccm

(j) Filtered-t14(6)-ccm

(k) Filtered-t14(7)-ccm

(l) Filtered-t14(8)-ccm

Fig. D.30 CCM Test Results for Bcd and Cad (Filtered Series).

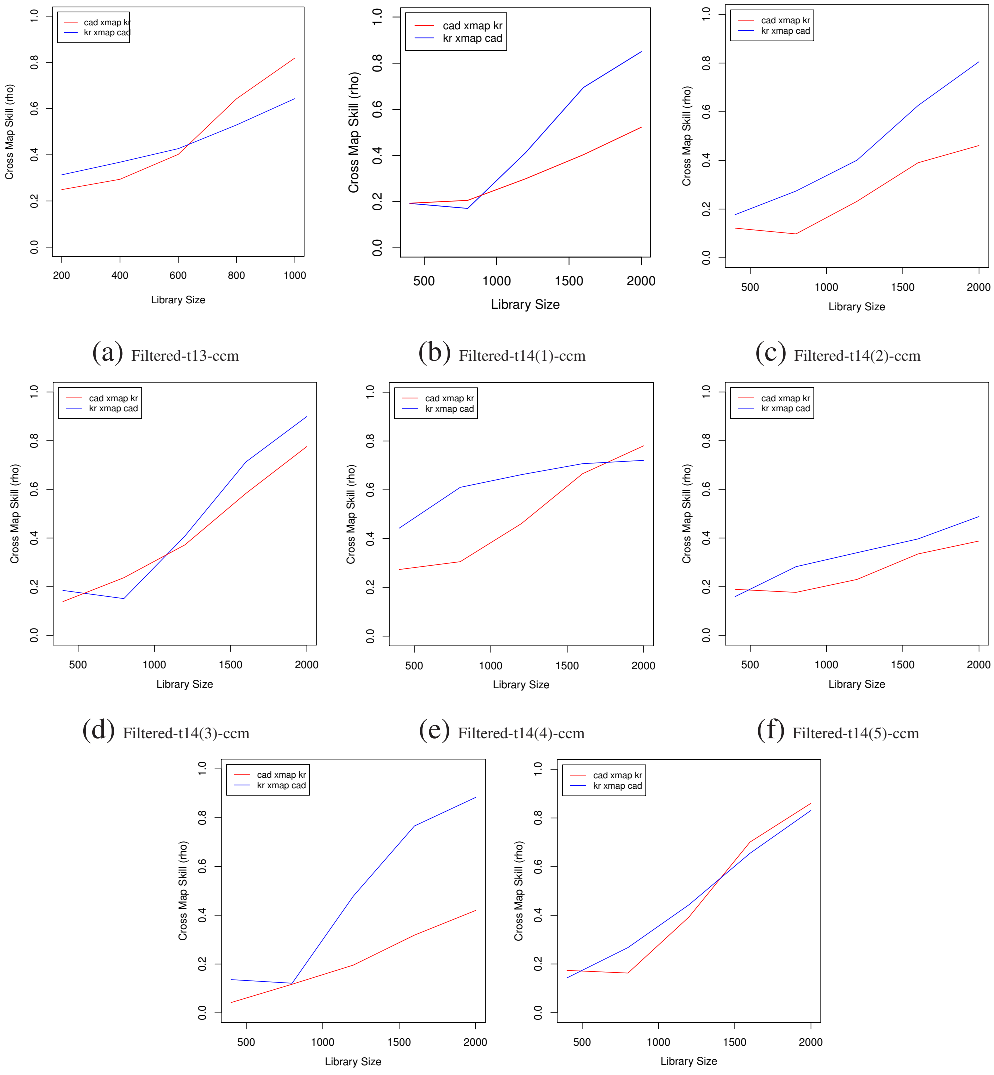


(a) Filtered-t13-ccm (b) Filtered-t14(1)-ccm (c) Filtered-t14(2)-ccm

(d) Filtered-t14(3)-ccm (e) Filtered-t14(4)-ccm (f) Filtered-t14(5)-ccm

(g) Filtered-t14(6)-ccm (h) Filtered-t14(7)-ccm

Fig. D.31 CCM Test Results for Bcd and Kr (Filtered Series).



(a) Filtered-t13-ccm

(b) Filtered-t14(1)-ccm

(c) Filtered-t14(2)-ccm

(d) Filtered-t14(3)-ccm

(e) Filtered-t14(4)-ccm

(f) Filtered-t14(5)-ccm

(g) Filtered-t14(6)-ccm

(h) Filtered-t14(7)-ccm

Fig. D.32 CCM Test Results for Cad and Kr (Filtered Series).