NONLINEAR EXTENSIONS TO NEW CAUSALITY AND A NARMAX MODEL SELECTION ALGORITHM FOR CAUSALITY ANALYSIS

By

Pedro da Cunha Nariyoshi

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Electrical Engineering – Doctor of Philosophy

2021

ABSTRACT

NONLINEAR EXTENSIONS TO NEW CAUSALITY AND A NARMAX MODEL SELECTION ALGORITHM FOR CAUSALITY ANALYSIS

By

Pedro da Cunha Nariyoshi

Although the concept of causality is intuitive, an universally accepted objective measure to quantify causal relationships does not exist. In complex systems where the internal mechanism is not well understood, it is helpful to estimate how different parts of the system are related. In the context of time-series data, Granger Causality (GC) has long been used as a way to quantify such relationships, having been successfully been applied in fields as diverse as econometrics and neurology. Multiple Granger-like and extensions to GC have also been proposed. A recent measure developed to address limitations of GC, New Causality (NC), offers several advantages over GC, such as normalization and better proportionality with respect to internal mechanisms. However, NC is limited in scope by its seminal definition being based on parametric linear models. In this work, a critical analysis of NC is presented, NC is extended to a wide range of nonlinear models and finally, enhancements to a method of estimating nonlinear models for use with NC are reported.

A critical analysis is conducted to study the relationship between NC values and model estimation errors. It is shown that NC is much more sensitive to overfitting in comparison to GC. Although the variance of NC estimates is reduced by applying regularization techniques, NC estimates are also prone to bias. In this work, diverse case-studies are presented showing the behavior of NC estimation in the presence of regularization. A mathematical study of the sources of bias in the estimates is given.

For systems that cannot be modeled well by linear models, the seminal definition of NC performs poorly. This works gives examples in which nonlinear observation models cause NC values obtained with the seminal definition to behave contrary to intuitive expectations. A nonlinear extension of NC to all linear-in-parameters models is then developed and shown to

address these limitations. The extension reduces to the seminal definition of NC for linear models and offers a flexible weighting mechanism to distribute contributions among nonlinear terms. The nonlinear extension is applied to a range of synthetic data and real EEG data with promising results.

The sensitivity of NC to parameter estimation errors demands that special care be taken when using NC with nonlinear models. As a complement to nonlinear NC, enhancements to a algorithm for nonlinear parametric model estimation are presented. The algorithm combines a genetic search element for regressor selection with a set-theoretic optimal bounded ellipsoid algorithm for parameter estimation. The enhancements to the genetic search make use of sparsity and information theoretic measures to reduce the computational cost of the algorithm. Significant reductions are shown and direction for further improvements of the algorithm are given. The main contributions of this work are providing a method for estimating causal relationships between signals using nonlinear estimated models, and a framework for estimating the relationships using an enhanced algorithm for model structure search and parameter estimation.

This thesis is dedicated to God, the Creator, Redeemer and Giver of Life.

Praise God from whom all blessings flow
Praise Him all creatures here below
Praise Him above ye heavenly hosts
Praise Father, Son, and Holy Ghost.
– Thomas Ken, “Morning Hymn”

ACKNOWLEDGEMENTS

Throughout my graduate experience, I have come to fully appreciate that the completion of this thesis was a joint effort and would not have been possible with the help of others. I would like to show my heartfelt gratefulness here.

I would like to thank my doctoral advisor, Dr. John Deller. He has been a great advisor, mentor and friend during the duration of my program. I thank him for his patience, good humor, and unwavering support for my work through all the challenges I have faced. He has set an example of excellence in teaching, research and mentorship for me.

I would like to thank my wife, Shihua Liu. She deserves at least half of the credit for the completion of this dissertation. A most excellent mother, wife and friend. Words cannot express how grateful I am for your companionship and support. I would also like to thank my children, Alice, Samuel and Natalia, who fill my life with glee everyday and with whom there is never a dull moment. It has been an immense pleasure and privilege to have you in my life.

The members of my committee have been exceedingly patient and gracious with me and the numerous detours I have taken until I was able to have some traction in my research. I have known Dr. Goodman the longest and who I have never seen without a smile and positive attitude. I thank him for all the meetings shared, papers reviewed, the constant optimism, advice given and kindness received. I also thank Dr. Aviyente and Dr. Punch, who I have known first as excellent and passionate instructors and later had the pleasure to join my doctoral committee.

My family has been constant source of encouragement and unconditional love. I would like to especially acknowledge my mother, Maria Alice da Cunha Nariyoshi, who unfortunately was not able to live to see her grandchildren and my graduation. She has always encouraged me to strive forward and persevere. To my brother, João Fernando da Cunha Nariyoshi and father, Fernando Massanori Nariyoshi, who have given me much love and encouragement over the years, I convey here my utter gratefulness and love. I also must thank my parents-in-law, Yongheng Liu and Sufan Long, who have welcomed me into their family with open arms and have given me the honor

to marry their precious daughter. I also thank my extended family for so much love received throughout the years. Especially, I thank my grandparents, Ignácio Adonias da Cunha and Alice Eliza da Cunha, and my aunts Nanci and Alice Nariyoshi.

Many friends, new and old, have accompanied and assisted me also deserve my thanks. I would like to thank Fan Bin, Yiqun Yang and Xiaofeng Zhao with whom I shared countless lunches and lively conversation, and Blair Fleet and Jinyao Yan, who were great lab mates. I thank my good friends Danilo Luvizotto, Adelle Araújo, and Nicole Torelli who remained close, even though we lived so far. I would be remiss if I didn't also mention Shichen Zhang, Qianwei Jiang, Li Jie, He Qiong, Sichao Wang, Rebeca Gutierrez, Xiaoxing Han, Yu Cheng, Ifwat Ghazali, Abhinav Gaur, Thássyo Pinto and Anselmo Pontes. I am deeply thankful for each one of you.

I thank Dr. McGough for supporting me and serving as my advisor for the first 2 years of my program. I would also like to thank Dr. Radha, Dr. Wierzba, Dr. Mason and Dr. Chakrapani, who have mentored and supervised me during my teaching assignments. Through these experiences, I have continually grown to love teaching. I am also especially grateful to Dr. Katy Colbry for her kindness and encouragement to all engineering graduate students. And I cannot neglect also thanking Dr. Hogan, Dr. Balasubramanian, Dr. Rothwell, Dr. L. Udpa, Dr. S. Udpa, Dr. Papapolymerou and all members of the ECE office and support staff.

From the University of São Paulo, many professors also served as an inspiration to pursue a PhD. Specifically, I would like to thank Dr. Denise Consonni, who sparked my love for Electrical Engineering and teaching. I would also like to thank Dr. Vítor H. Nascimento, who advised me during my undergraduate research work.I also thank Dr. Magno T. M. Silva, Dr. Cristiano Panázio and Dr. Marco Alayo who great instructors and mentors. I would also like to thank all my friends in the classes of 2009, 2010 and 2011.

I would like to thank Nathanael Fawcett, my pastor during my youth and who I dearly love and am grateful to. I thank my brothers in sisters in Christ from Intervarsity, Aliança Bíblica Universitária, and the global Church of Christ. Specially, I would like to thank the Bauers, Bielers, Cogans, Fosters, Jeffries, and Starks for being mentors, friends and walking together with us.

TABLE OF CONTENTS

LIST OF TABLES x

LIST OF FIGURES xi

LIST OF ALGORITHMS xiv

KEY TO ABBREVIATIONS xv

CHAPTER 1 INTRODUCTION 1
- 1.1 General statement 1
- 1.2 Research objectives 7
- 1.3 Critical analysis of the study 8
- 1.4 Structure of the dissertation 10
- 1.5 Summary and contributions 11

CHAPTER 2 BACKGROUND METHODS 12
- 2.1 Overview 12
- 2.2 Modeling 12
  - 2.2.1 Generalized observation model 14
  - 2.2.2 Estimation model 15
  - 2.2.3 Least squares estimation 19
  - 2.2.4 ARMAX models 20
  - 2.2.5 ARX models with non-white error sequences 20
  - 2.2.6 NARMAX and modified NARX models 21
  - 2.2.7 LASSO regression 22
- 2.3 Set-membership optimum bounded ellipsoid algorithms 23
- 2.4 NARMAX model estimation and the EvolOBE method 26
  - 2.4.1 Genetic encoding and algorithm overview 28
- 2.5 Causality analysis 30
  - 2.5.1 Humean concept of causality 31
  - 2.5.2 Granger causality 34
  - 2.5.3 Spectral Granger causality 36
  - 2.5.4 Conditional Granger causality 38
  - 2.5.5 New causality 39
  - 2.5.6 Spectral new causality 40

CHAPTER 3 A CRITICAL ANALYSIS OF NEW CAUSALITY 42
- 3.1 Overview 42
- 3.2 Problematic aspects of models in NC literature 43
  - 3.2.1 Model 1 43
  - 3.2.2 Model 2 44
  - 3.2.3 Model 3 45

3.2.4 Model 4 . . . 45
3.2.5 Model 5 . . . 46
3.2.6 Model 6 . . . 48
3.2.7 Model 7 . . . 50
3.2.8 Discussion of models 1-7 . . . 51
3.3 Analysis of NC robustness to parameter errors through case studies . . . 52
3.3.1 Discussion . . . 62
3.3.2 NC and GC fluctuation . . . 62
3.3.3 Regression conditioning and over-fitting . . . 64
3.3.4 Comparing NC and GC . . . 66
3.4 Bias in NC estimates . . . 67
3.4.1 Case 1: $\Delta \boldsymbol{b}^T \boldsymbol{Z} \approx \boldsymbol{0}$ . . . 70
3.4.2 Case 2: $\Delta \boldsymbol{a}^T \boldsymbol{X} \approx \boldsymbol{0}$ . . . 71
3.4.3 Case 3: $\Delta \boldsymbol{a}^T \boldsymbol{X} + \Delta \boldsymbol{b}^T \boldsymbol{Z} \approx \boldsymbol{0}$ . . . 72
3.4.4 Case 4: Regularization . . . 72
3.4.4.1 Extending case 3 . . . 73
3.4.4.2 Discussion of bias . . . 75
3.5 Conclusion . . . 76

CHAPTER 4 A NONLINEAR EXTENSION TO NEW CAUSALITY . . . 77
4.1 Overview . . . 77
4.2 Motivation . . . 77
4.3 Choice of NARMAX models . . . 79
4.4 A nonlinear extension to NC for a restricted set of models . . . 80
4.5 A comprehensive NNC definition . . . 81
4.5.1 Form 1: λ^1 - create a new category for nonlinear cross-terms . . . 83
4.5.2 Form 2: λ^2 - weight regressor functions equally across regressor signals . . . 84
4.5.3 Form 3: λ^3 - weight regressor functions across regressor signals according to an application (model) dependent criterion . . . 86
4.5.4 Spectral nonlinear new causality . . . 87
4.6 Discussion and analysis through example models . . . 87
4.7 Application: EEG data . . . 96
4.8 Discussion of λ functions and preprocessing . . . 102
4.9 Conclusions . . . 104

CHAPTER 5 IMPROVEMENTS TO THE EvolOBE METHOD FOR NONLINEAR CAUSALITY ANALYSIS . . . 106
5.1 Overview . . . 106
5.2 Model form . . . 107
5.3 Identification strategy . . . 108
5.3.1 NSGA-II . . . 109
5.3.2 Asymmetric mutation operator . . . 110
5.3.3 Reduced surrogate crossover . . . 112
5.3.4 Linkage tree crossover . . . 112
5.4 Results of AM and RSX . . . 114

5.5 Results of LTX 117
5.6 Application to NNC analysis 123
5.7 Discussion and conclusions 125

CHAPTER 6 CONCLUSION 128
6.1 Overview 128
6.2 Contributions 130
6.3 Future Work 130

APPENDICES 132
APPENDIX A DERIVATION OF CLOSED-FORM EXPRESSIONS FOR GC AND NC FOR FIRST-ORDER BIJOINTLY REGRESSIVE OBSERVATION MODELS 133
APPENDIX B LISTINGS FOR ALGORITHMS 141

BIBLIOGRAPHY 146

LIST OF TABLES

Table 3.1: Theoretically evaluated GC and NC measures for the observation model in Eq. (3.19) 54

Table 3.2: Theoretically evaluated NC measures for the observation model in Eq. (3.20) . . 60

Table 4.1: Linear $\mathrm{NC}_{x_j \to x_k}$ values for the model of Eq. (4.1) 78

Table 4.2: Nonlinear $\mathrm{NC}_{x_j \to x_k}$ values for the model of Eq. (4.1) 81

Table 4.3: $\mathrm{NC}_{x_j \to x_k}$ values for the model in Eq. (4.17) 89

Table 4.4: $\mathrm{NC}_{y_j \to y_k}$ values for the model of Eq. (4.18) 90

Table 4.5: $\mathrm{NC}_{x_j \to x_k}$ values for the model of Eq. (4.20) 92

Table 4.6: $\mathrm{NC}_{x_j \to x_1}$ values for the model of Eq. (4.21) with 10dB SNR 94

Table 4.7: $\mathrm{NC}_{x_j \to x_1}$ values for the model of Eq. (4.21) with 50dB SNR 94

Table 4.8: $\mathrm{NC}_{x_j \to x_1}$ values for the model of Eq. (4.23) with 10dB SNR 95

Table 4.9: $\mathrm{NC}_{x_j \to x_1}$ values for the model of Eq. (4.23) with 50dB SNR 95

Table 4.10: GC, NC, NNC, and SNNC results on whether to accept x_{Fp1} causes x_{O2} 99

Table 4.11: GC, NC, NNC, and SNNC results on whether to reject x_{O2} causes x_{Fp1} 99

Table 4.12: GC, NC, NNC, and SNNC results on whether to accept x_{Fp1} causes x_{P4} 101

Table 4.13: GC, NC, NNC, and SNNC results on whether to reject x_{P4} causes x_{Fp1} 101

Table 5.1: Fitted parameters for different methods 117

LIST OF FIGURES

Figure 2.1: Geometric illustration of OBE algorithms 25

Figure 2.2: NSGA-II algorithm summary 29

Figure 2.3: Different explanations for large $GC_{x\to z}$ 38

Figure 3.1: Distribution of the $NC_{1\to 2}$ estimates as a function of λ for $M = 1$. 55

Figure 3.2: Distribution of the $GC_{1\to 2}$ estimates as a function of λ for $M = 1$. 55

Figure 3.3: Distribution of the $GC_{1\to 2}$ estimates as a function of λ for $M = 2$. 56

Figure 3.4: Distribution of the $NC_{1\to 2}$ estimates as a function of λ for $M = 2$. 56

Figure 3.5: Distribution of the $NC_{1\to 2}$ estimates as a function of λ for $M = 5$. 57

Figure 3.6: Distribution of the $NC_{1\to 2}$ estimates as a function of λ for $M = 6$. 58

Figure 3.7: Distribution of the $GC_{1\to 2}$ estimates as a function of λ for $M = 5$. 58

Figure 3.8: Distribution of the $GC_{1\to 2}$ estimates as a function of λ for $M = 6$. 59

Figure 3.9: Distribution of the $NC_{1\to 2}$ estimates as a function of λ for $M = 5$ and $N = 1024$. 59

Figure 3.10: Distribution of the $NC_{1\to 1}$ estimates as a function of λ for the model shown in Eq. (3.20) and $M = 4$ 61

Figure 3.11: Distribution of the $NC_{1\to 1}$ estimates as a function of λ for $M = 6$ 61

Figure 3.12: Distribution of the $NC_{1\to 1}$ estimates as a function of λ for $M = 1$ for the model shown in Eq. (3.20) 62

Figure 3.13: $NC_{1\to 1}$ vs $NC_{0,1\to 1}$ histogram plots as a function of λ 63

Figure 3.14: $NC_{1\to 1}$ vs $NC_{0,1\to 1}$ histogram plots as a function of λ 66

Figure 3.15: Distribution of the $GC_{2\to 1}$ estimates as a function of λ for the model shown in Eq. (3.20) and $M = 6$ 67

Figure 3.16: Distribution of the $GC_{1\to 2}$ estimates as a function of λ for the model shown in Eq. (3.20) and $M = 6$ 68

Figure 3.17: NC estimates for different values of NC_0 and β 71

Figure 3.18: Estimated probability density function of NC using exact and approximate expressions 74

Figure 3.19: Estimated probability density function of NC split into two cases 75

Figure 4.1: NC values for the model of Eq. (4.1) 79

Figure 4.2: NC and NNC values for the model of Eq. (4.1) 82

Figure 4.3: 10-20 International System Electrode Location Diagram 96

Figure 4.4: Spectrum of the Fp1 channel of the EEG recording 97

Figure 4.5: Average of $SNNC_{Fp1\rightarrow O2}$ values of subject 1 98

Figure 4.6: Receiver operating characteristic curves for the unfiltered tests 100

Figure 4.7: Receiver operating characteristic curves for 13.5Hz 101

Figure 4.8: Receiver operating characteristic curves for the unfiltered tests 102

Figure 4.9: Receiver operating characteristic curves for 13.5Hz 102

Figure 5.1: NSGA-II algorithm summary 110

Figure 5.2: Linkage tree example 113

Figure 5.3: Estimated pareto front 115

Figure 5.4: Histogram vs. Fitted distribution 116

Figure 5.5: Generations to arrive at the desired model 116

Figure 5.6: Estimated regressor functions present in best models 119

Figure 5.7: Estimated pareto-front for 15dB SNR 119

Figure 5.8: Histogram of required evaluations for RSX and LTX 120

Figure 5.9: Fitted PDF for the required evaluations 121

Figure 5.10: CDF for the required number of evaluations to find the desired solution 121

Figure 5.11: Close-up for fewer than 15000 evaluations 122

Figure 5.12: CDF for the required number of evaluations for 15dB SNR 123

Figure 5.13: Comparison between estimated pareto fronts for different SNR values 123

Figure 5.14: NNC values for the final candidate model set for 10dB SNR 124

Figure 5.15: NNC values for the final candidate model set for 50dB SNR 125

LIST OF ALGORITHMS

Algorithm B.1: Unified Optimum Bounded Ellipsoid Algorithm 142

Algorithm B.2: UOBE Recursion 142

Algorithm B.3: Automatic Bounds Estimation 144

Algorithm B.4: Generate Linkage Tree 144

Algorithm B.5: Linkage Tree Crossover 145

Algorithm B.6: Linkage Tree Crossover for multi-objective problems 145

KEY TO ABBREVIATIONS

AIC Akaike Information Criterion

AR Autoregressive

ARMAX Autoregressive Moving Average with Exogenous input

ARX Autoregressive with exogenous input

BIC Bayesian Information Criterion]

EvolOBE Evolved Optimum Bounded Ellipsoid

GC Granger Causality

LSE Least Squares Estimation

LTGA Linkage Tree Genetic Algorithm

LTI Linear and Time Invariant

LTIiP Linear and Time Invariant in Parameters

LTX Linkage Tree Crossover

MSE Mean Square Error

NSGA Non-dominated Sorting Genetic Algorithm

OBE Optimum Bounding Ellipsoid

RMSE Root Mean Square Error

RSX Reduced Surrogate Crossover

WRLS Weighted Recursive Least Squares

CHAPTER 1

INTRODUCTION

1.1 General statement

The concept of causation and consequence is at the foundation of the scientific method. Although causality is an intuitively simple concept, *action A causes event B to occur*, an universally accepted definition of causality has long eluded scientists and philosophers. Understanding causal relationships is an essential step in the analysis of complex systems. Despite significant theoretical and heuristic advances in the topic, quantifying and tracking causality strength and assessing the causal link between two dependent quantities or events is still an active field of research.

The scientific approach to establishing these relationships is by creating falsifiable hypotheses (*e.g.*, "A causes B" or "A does not cause B") and subsequently testing which hypothesis provides the most satisfactory answer. The analysis often starts by taking measurements or observations of quantities that are relevant (or at least possibly relevant) to the question. In the context of signal processing, these measurements are referred to as signals. Signals are frequently classified as ***inputs*** and ***outputs***, which are somewhat analogous to causes and effects. The ***system*** is the underlying entity that processes the quantities from which input signals are measured into the quantities from which the output signals are measured. When signals are measured sequentially in constant time intervals, the resulting sequence is called a time series. The mathematical representation of how the inputs and outputs are related is called a ***model***. The models are constructed given the available data, the particular hypothesis being considered and any *a priori* knowledge available about the system being studied. Many causality analysis methods involve the creation of models and measuring intrinsic characteristics of one or more models and statistical properties of the data.

At the heart of the scientific method, Occam's razor has been used as an heuristic tool to evaluate different explanations of observed phenomena. Also known as *lex parsimoniæ* (law of

parsimony), it states that "Entities are not to be multiplied without necessity," or, in other words, for different explanations of a phenomenon, the simplest (satisfactorily accurate) explanation is to be preferred. With regard to model construction, this has been re-expressed (somewhat amusingly) by Box [31]:

> Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model. However, cunningly chosen parsimonious models often do provide remarkably useful approximations. ... [T]here is no need to ask the question "Is the model true?". If "truth" is to be the "whole truth" the answer must be "No". The only question of interest is "Is the model illuminating and useful?".

For causality analysis, it is often expedient to disregard many fundamental aspects of a system in order to produce a model that provides better intuition of the relationships between potential inputs and outputs (or causes and effects) [104]. For instance, one need not know the line frequency or voltage to assert that a light switch controls a lamp, even though these are fundamental design parameters for the internal function of the circuit. For more complex system, determining what aspects to consider or ignore in constructing a model is not straightforward [138].

While the problem of model structure selection and validation cannot be universally solved, it is possible to employ general principles to find useful models. Models with higher complexity may potentially better represent the system being observed, but may also be prohibitively expensive or require large amounts of data to be accurately computed. Ljung summarizes the problem with [128, pg. 494]:

> The compromise between parsimony and flexibility is at the heart of the identification problem. How shall we obtain a good fit to data with few parameters? The answer usually is to use *a priori* knowledge about the system, intuition, and ingenuity. These facts stress that identification can hardly be brought into a fully automated procedure. The answer usually is to use *a priori* knowledge about the system, intuition, and ingenuity. ... A general advice is to "try simple things first."

Besides challenges of properly modeling systems, quantifying causal relationships represents an additional non-trivial problem. Causality can only be inferred (but not determined) from time-series records (and only under certain conditions [154]). The most widely used method for

assessing causality in the context of signals and systems - the context of this work - is known as Granger Causality (GC) [76, 77]. Borrowing from Hume's study of causality [103], GC focuses on evaluating how well past information about a signal or event A can predict the current state of second signal or event B. The method has received several extensions, such as conditional GC (CGC) [72] and spectral GC (SGC) [71], as well as similar spectral methods such as partial directed coherence (PDC) [12, 169, 173], the relative power contribution (RPC, also referred to as Akaike Causality) [3, 208] and the directed transfer function (DTF) [57, 108, 176].

In addition to transfer function and model based approaches, alternative methods abound for inferring connectivity between time-series records [80, 156]. Phase analysis methods have shown promise in inferring connectivity, such as the the phase-locking value (PLV) [91, 120], the phase slope index (PSI) [86, 150] and phase-syncrony [9, 10, 119]. More recently, phase-amplitude coupling methods have been applied with promising results [140]. Information theory based methods, such as directionality index (DI) [126, 170], Mutual Information (MI) and Transfer Entropy (TE) [196] also have been employed, but in general require more data for estimating probability distributions [114] and cannot capture quickly time-varying characteristics, such as functional connectivity microstates in the brain [61]. The present work focuses on model based approaches rather than phase analysis and information theoretic approaches.

A more recent causality analysis method, New Causality (NC) [95], uses a different approach, relying on the internal structure and states of a multivariate autoregressive model (MVAR) to estimate causality strength. The use of the internal structure presupposes that the models appropriately represent the mechanisms being studied. This assumption is not necessarily correct in complex models; nonetheless, NC possesses several desirable characteristics, such as the production of a normalized value for which the sum of all the NC values contributing to a particular "effect" signal adds to unity. Relative to GC, NC allows easier comparisons among different systems, because the measured causality strength increases with increasing NC values, whereas GC might produce "small values" even when signals have a strong causal link [94] or not depend on relevant model parameters [100]. Moreover, in the tests with real and surrogate data in [94, 95, 98–100],

NC is superior to GC in the indication of causality strength. However, as shown in [148], NC is more sensitive to model parameter overfitting than GC, requiring more accurate model parameter estimation to produce meaningful results. Further, the seminal formulation of NC is restricted to linear MVAR models. One of the central contributions of the present work is the extension of NC to the far more general ***nonlinear autoregressive moving average with exogenous input*** (NARMAX) models [22] while retaining all the advantageous properties of NC.

One principle often used to obtain "useful" models is to find the simplest models that provides good explanatory power. Since model simplicity and high explanatory power are often conflicting objectives, system identification algorithms involve a solution that provides the "best" balance/trade-off between the two objectives.

Evaluating the complexity[1] of a model is not simple, specially when distinct classes of models must be compared, such as ones generated by artificial neural networks (ANN) or genetic programming and linear models. Within the same class of models, however, there are often methods of *quantifying* complexity. Particularly, for linear models, many approaches to compare model complexity exist, such as the l_0 norm of the parameter space [193] and the model orders for autoregressive (AR), moving average (MA), and autoregressive moving average (ARMA) models.

For parametric models, accuracy is usually optimized using a mean square error (MSE) criterion, although in some cases other measures, such as total least squares [133] or the l_∞ norm (also known as max-norm) [56, 84, 185], might be preferable.

Parametric models frequently employ a form of regularization of the parameter space to balance model complexity and prediction error. This is achieved by adding a regularization term to the cost function, which assigns a penalty to solutions with higher order or larger norm of the parameter space. Regularization may be regarded as a Bayesian approach to model estimation, in which prior information (*i.e.*, assumptions) is used in the formulation of the models [111]. If the model is assumed to be sparse, the l_0 norm of the parameters can be used [174]. Total variation has been applied for image denoising, assuming the noise-free image is smooth [29, 171]. As long

[1] The term "complexity" is used here in a customary (non-technical) sense

as the assumptions about the formulation of the models is close enough to reality, regularization can greatly aid parameter estimation [166].

For ***linear and time-invariant*** (LTI) models, a vast variety of methods and literature are available which are based on well-established theories, such as Fourier transforms [164]. Linear and time-invariant models possess a number of properties that make them amenable to analysis [152] and can be completely characterized (within the constraints of short-term processing) by the impulse response of the model or, equivalently, the system function. The advantages of LTI modeling are such that it is sometimes desirable to linearize nonlinear models so that LTI techniques may be applied [151].

However, LTI models are increasingly deemed insufficient for system analysis and design in the 21st century [25, 39]. For linear time-*varying* systems, adaptive methods exist, such as Least Mean Squares (LMS) [205], Recursive Least Squares (RLS) [158] and derivatives, such as the Normalized Least Mean Squares (NLMS) [172] and Set-membership Weighted Recursive Least Squares [52]. However, addressing nonlinearity in models is an ongoing problem, for which the development of a concise universal methodology is unlikely.

While *ad-hoc* techniques, such as nonlinear state-space models, have been successful in applications like neural connectivity analysis [68] and stock market volatility [190], they require relatively intimate understanding of the systems being modeled and do not generalize well outside their application domains.

Artificial neural network models are also very powerful and have fostered advances in prediction [47], classification [123] and even complex gameplay [183, 197]. The universal approximation theorem states that ANNs can potentially represent any continuous function on compact subsets [48, 49, 129], although the learnability of the parameters is not addressed by the theorem. In spite of the performance, the black-box nature of ANNs remains one of the largest criticisms [58] and a barrier to interpretability. Recent developments seek to address some of these criticisms by providing methods of interpreting ANN models [130, 181].

Finally, ***linear and time-invariant in parameters*** (LTIiP) models, which are most concisely

expressed using NARMAX models [22] (of which the Volterra series [198] and the Hammerstein models are special cases [145]) have shown to be very powerful in many applications, from epidemiology [165] and microbial growth [210] to human physiology [116] and aerospace engineering [34]. A significant advantage of LTIiP methods is that they allow the use of the wealth of powerful and well understood LTI methods of system identification in identifying nonlinear models. Additionally, NARMAX models enable sparse, interpretable and transparent modeling [202], all of which are characteristics desirable in causality analysis.

While NARMAX models provide a concise but flexible parsimonious model paradigm [117], NARMAX models also introduce a new set of challenges. Unlike ARMAX models, which only allow time-shift operators to be applied to the regressor signals, NARMAX models allow the application of other operators - generally called ***regressor functions***. Depending on the model order and class of regressor functions, the number of such functions may be very large. Also, it is often the case that regressors are highly correlated, leading to slow and inaccurate convergence [11]. Although many methods exist, the selection of the subset of the regressor functions is an unresolved issue in system identification for over-parameterized models [2, 22, 25, 27, 81, 117, 118, 201, 203, 214, 219].

A recent method for NARMAX model identification [214], henceforth called ***evolved OBE*** (EvolOBE),[2] has shown promise in developing accurate, sparse and interpretable results. This method searches for a family of NARMAX model structures that maximize accuracy while minimizing the number of regressors. The method involves a hybrid approach which uses a genetic algorithm to select regressor functions, while employing a set-membership based optimum bounding ellipsoid algorithm [52] to estimate the parameters values. A significant advantage of such a method is that does not require any assumptions about stationarity or distributional characteristics of the model disturbances. The capability of identifying simple nonlinear models with good accuracy with unbiased parameters under complex noise conditions makes this algorithm compelling for use in complex nonlinear systems, avoiding overfitting and maintaining good interpretability of model structure.

[2] In [209], this algorithm is called OBE with evolved regressor signals (OBE-ERS).

1.2 Research objectives

Causality analysis is often employed to gain insight about systems whose internal properties are unknown. Granger causality possesses a intuitive interpretation, if the inclusion of past values of a signal x improve the prediction of the current value of a second signal y compared to predicting y using only past values of y itself, then this improvement can be used as evidence that x causes y. While conceptually simple, it can be difficult to map GC values to information about the systems which relate x and y [95], as GC is designed to measure effect, not mechanism [19].[3] On the other hand, NC draws directly from the mechanism (of the model) and thus is complementary to GC, providing new insight into the models. However, the literature on NC is limited in comparison to the wealth of methods for estimating and applying GC. Additionally, most of the studies of NC have assumed that the observational and the estimated models are equivalent, with little discussion on the validity of that assumption and the consequences to the analysis results.

A deeper characterization of the robustness of NC to model order and parameter uncertainty is required to increase understanding and confidence in the use of NC [148]. Although GC was only defined for MVAR models in its seminal form [76, 77], nonlinear extensions exist [7, 13, 66, 132]. The seminal definition of NC is also restricted to MVAR models, so the extension of NC to NARMAX models developed in this work will allow NC to be useful in a much wider range of applications.

To improve upon NC and address some of its drawbacks, this work takes a two pronged approach: first, an extension of NC to a more comprehensive set of linear and nonlinear models is developed [146]; second, the framework for nonlinear system identification found in [214] is explored and improved in the search for "useful" models. The present work also includes the implementation and discussion of state-of-the-art methods for improved search speed and accuracy [149, 192, 217].

Thus, the research objectives of this study are to:

1. Characterize the behavior of NC under model order and parameter uncertainty.

[3] The authors of [95] dispute this claim in [99].

2. Extend the formulation of NC to enable application to LTIiP nonlinear models.

3. Improve model structure search for LTIiP nonlinear models through use of enhanced genetic algorithms.

4. Apply the model structure search algorithms to causality analysis using sets of simulated and real data.

1.3 Critical analysis of the study

In the same vein as Box's remark, it would not be expected that causality would be discriminable from time-series records alone. While all techniques discussed in this work could potentially be applied to any multivariate time-series data, *a priori* information should be used to first evaluate if the hypothesis of causality is plausible and whether all relevant factors have been considered.[4] A machine cannot correct operator mistakes because "it cannot think for itself" [137]. Thus, causality measures must represent only a part of causality analysis, because such measures are unable to differentiate between alleged causality and deficient experimental design. New causality is under the same restrictions and is prone to produce misleading results if incorrect or incomplete data are used.

Holland and Durbin [92] have also argued that only one cause can be observed at a time, what they referred to as the ***fundamental problem of causal inference***. That is, supposing it is desired to know if intervention A (*e.g.*, medication) will cause B (*e.g.*, reduction of a particular symptom) on a particular patient C. If it is chosen to do A, one can measure the outcome of A given C (*e.g.*, giving the medicine to C), but not the outcome of not doing A on C (*e.g.*, not giving the medicine to C), and vice-versa. Therefore, one must either take a *statistical* approach of testing different interventions over a large population (*e.g.*, giving the medicine to people similar to C reduced the symptom on 80% of them, when given a placebo, the symptom was reduced in 40% of

[4] Cliff stated this fact as "these programs are not magic. They cannot tell the user about what is not there." [46] Cartwright argues that one cannot get knowledge of causes from equations and associations alone [36], but instead, old causal knowledge must be used to extract new causal knowledge.

them) or an approach they call *scientific*, which requires the assumptions of homogeneity (*e.g.*, the outcome of an intervention in the past would be the same in the present) so that different outcomes can be compared (*e.g.*, the sentence "symptoms are reduced every time C takes the medicine" assumes that the effect of A on C is time-invariant even if C might change over time). Additionally, they assert that causes can only be interventions that are imposed (not voluntary) and are not attributes (*e.g.*, one cannot state that a car is fast *because* it is a *Ferrari*, since it would be impossible to measure the speed of the *same* car if it were made by Ford, because it would not be same car after all. Instead, one could only say that cars made by *Ferrari* are usually faster than cars made by Ford, without establishing a causal relationship). Their conclusions were summarized in the motto: "no causation without manipulation." However, Pearl argues in [154] that, while manipulation is simply one way to test the workings of mechanisms, it is by no means necessary for causal determination. Humans can confidently say that the moon causes tides (even if we cannot observe the effects of the lack of a moon) or that the genetic code of a raven causes it to be black (even without manipulating its DNA).

As will be discussed in Sec. 2.5, Hume believes humans to be unable to assert causation. Thus he devises a framework through which causation can be *inferred*. Granger causality builds upon Hume's work, creating a formal measure for causal inference. Granger causality is closely linked with the concept of TE, which measures transferred information rather than how two signals are interconnected. In fact, GC and TE are equivalent for normally distributed signals [14]. The differences between transferred information (and therefore GC) and causal effects are sometimes subtle but not negligible [127].

Similarly to the seminal definition of NC, the nonlinear extension of NC [146] fundamentally relies on the quality[5] of the estimated models being used. As shown in [148], even when the data are generated by a parametric model of the same class as the estimated models, the NC measure values depend heavily on the accuracy of the parameter estimates, whereas GC was

[5]Quality in this context refers to the ability of a model to sufficiently represent the internal dynamics of a system. This is in contrast to many predictive models, whose design is based on the ability to predict the output of a system given a set of inputs, often without regard to the actual internal dynamics of the system.

shown to be much more robust to parameter estimation errors. The use of robust parametric model estimation methods mitigates this uncertainty somewhat, but careful selection and examination of the estimated models remains essential in evaluating causality using NC.

Causality analysis studies generally focus on systems with complex behaviors and/or unknown internal mechanisms. The goal is often to gain some insight into the functioning of a system, without necessarily fully comprehending internal interaction. This poses a problem for the evaluation of novel causality analysis tools, as most real datasets do not possess a "ground truth" for validation. Synthetic datasets offer several advantages, the foremost for causality analysis being the presence of ground truth. The knowledge of internal parameters also allows decoupling the quality of the causality measure from the model estimation aspect of the measure. On the other hand, while the ability to tune models to exhibit different behaviors is often desirable, as one can test the measure under different scenarios, the use of synthetic datasets can also (accidentally or intentionally) produce misleading results [147]. This work utilizes a set of real and synthetic datasets to show performance on a variety of problems, showing interesting results in a number of applications, but makes no claim of supremacy, rather presenting the nonlinear extension of NC as an additional and useful tool in an signal processing practitioner. Just as any powerful analysis tool, care must be taken in its application and the interpretation of the results. Again, a machine cannot correct operator mistakes regardless of how powerful the machine and how smart the operator may be.

1.4 Structure of the dissertation

This dissertation begins with the background methods chapter, in which an overview of modeling and modeling philosophy are given. This is foundational for the discussion which follows. The background material is followed by a short review of existing causality analysis tools. Finally, the model identification framework is laid out, with discussion of the particular techniques implemented. The model development is followed by a series of studies. First, a critical analysis of NC is given, which discusses models used in the literature, the robustness of NC under model

uncertainty, and derivations of sources of bias in NC estimation. Second, nonlinear extensions to NC are developed, with application examples using synthetic and real data. Third, enhancements to the EvolOBE method, where the method is tested against simulated data and the results are evaluated against observational models, also the GC and NC values obtained under the evolutionary algorithm are compared to the values obtained using the observational models. The studies are followed by the conclusion chapter, where a summary and a discussion of the results is given.

1.5 Summary and contributions

The concept of causality is integral to the scientific method. However, concisely defining and quantifying causality relationships is an elusive task. Many methods of evaluating causality have been created, with GC being the most prominent. However, since GC is designed to measure effect, not mechanism, NC can be used in conjunction to obtain more insight into the systems being studied. This work expands on NC by extending it to a wide range of nonlinear models and, thus, its applicability to a wider set of problems, and by doing a deeper critical analysis of NC, as portrayed in existing literature, and its behavior under model structure and parameter uncertainty. Additionally, this work also includes improvements to the EvolOBE method, which are applied to the nonlinear extension to NC. These results will drive the field forward to a more comprehensive set of causality analysis tools that include nonlinear NC.

CHAPTER 2

BACKGROUND METHODS

2.1 Overview

This chapter includes an overview of some of the methods used in this work. A large portion of Sec. 2.2 is quoted directly from [147–149] with a few modifications for improved flow and clarity.

2.2 Modeling

Before delving into the topic of causality analysis, it is important to make a distinction between systems and models. The time-series literature tends to be somewhat cavalier in the formulation of parametric time-series models. Widespread understanding of the fundamental modeling concepts allows a certain lack of precision in model notation. In particular, it is not uncommon to use the same modeling notation for the putative observation model and the estimation model. The ***observation model***, ordinarily one of the standard time-series models [32] with a white-noise or more strongly-independent disturbances is assumed to generate the observed sequence. Accordingly, its parameters are unknown, but the model is posed for theoretical analysis. The ***estimation model*** (or ***estimated***, following model identification) is the parametric model resulting from the model identification process. Although the observation model and the estimated model are naturally similar in form, the two models which may have quite different parameter values and accompanying disturbances. Since this distinction is important in the causality analysis approaches studied in this work, this section is dedicated to a clear explanation of the intricacies of models, including the establishment of a clear convention for model nomenclature and notation.

To simplify this task, the discussion will be restricted to a class of models that are linear, time-invariant and causal (over the interval of observation). This restriction simplifies he task of modeling *signals* – the model therefore representing a discrete time system of which only the output is observable. Moreover, the intention to use conventional ***least-square-error*** (LSE)

estimation of model parameters (in keeping with existing literature to which this work refers) prescribes that the natural choice of signal observation model is – at least in the case of model involving a single signal – the standard time-series model known as the ***autoregressive*** (AR) ***model***, often denoted AR($\mathcal{M}$) to indicate that the model has $\mathcal{M}$ parameters.

The AR($\mathcal{M}$) observation model for a signal sequence x is given by

$$x[n] = \sum_{m=1}^{\mathcal{M}} a^{m*} x[n-m] + \eta[n] \doteq (\boldsymbol{a}^*)^T \boldsymbol{x}[n] + \eta[n], \qquad n \in \mathbb{Z}, \tag{2.1}$$

in which, by convention, η is a discrete-time white noise process, and in which we have defined the Cartesian $\mathcal{M}$-vectors,

$$\begin{aligned} \boldsymbol{a} &\doteq \begin{bmatrix} a^{1*} & a^{2*} & \cdots & a^{\mathcal{M}*} \end{bmatrix}^T, \\ \boldsymbol{x}[n] &\doteq \begin{bmatrix} x[n-1] & x[n-2] & \cdots & x[n-\mathcal{M}] \end{bmatrix}^T. \end{aligned} \tag{2.2}$$

The parameter values, a^{m*}, include the superscript symbol "$*$" to indicate the "true" parameters – that is, the parameters associated with the observation model. The estimation of these parameters is discussed in a more general context below.

Let us digress momentarily to comment on a terminology issue. Some authors might choose to refer to the model of form (2.1) as a "generative model" (or "synthesis model") referring its assumed role in "generating" or "synthesizing" the sequence x. For the reporting of future research extending the present developments, the authors prefer to reserve the term generative model to refer to an unconstrained (and generally unknowable) operator, say $\mathbb{H}$, across normed vector spaces that is "used by nature" to *exactly* (without error at any level of precision) produce the signal x from the input η, say $x = \mathbb{H}\eta$. We will therefore deliberately use the term "observation model" when referring to Eq. (2.1) and related extensions.

It remains to specify the models used in estimation (following some further consideration of the observation model). The issue we are addressing by taking extra care in defining what each model refers to is necessitated by the following matter: it is not unusual for an author (across many fields) to, for example, implicitly use model (2.1) – with parameters a_i, rather than a^{m*} – then to refer to the estimated parameters with the same notation $a_1, \ldots, a_{\mathcal{M}}$, thus creating ambiguity in the

meaning of the parameter symbol notation. Less frequently, but all too commonly, the sequence name η may also be used to indicate the error sequence in the estimated model (in the AR case, the residual in the linear prediction of $x[n]$ using $x[n-1], \ldots, x[n-\mathcal{M}]$), thereby creating further ambiguity. Whereas such practices are generally accepted and lead to no adverse issues for the experienced practitioner, it is critical to clearly distinguish the various models used in the present discussion.

2.2.1 Generalized observation model

Before addressing the estimation models, we need to enhance the AR model of Eq. (2.1) for the present purposes. One can approach the required modification in several ways. Equation (2.1) represents a model for a single signal generated by passing uncorrelated noise through a linear filter. Causality analysis is generally concerned with multiple signals, say $x_1, x_2, \ldots, x_{N_s}$, where $N_s \geq 2$ denotes the number of such signals, and the possibility that any of the signals $\left\{ x_j \right\}_{j=1}^{N_s}$ may contribute to (may "cause") the generation of x_p for a given $1 \leq p \leq N_s$. The inclusion of linear combinations of samples from further signals on the right side of Eq. (2.1) makes it improper to refer to the model as "autoregressive." The augmented model (in the "careful" notation suggested above), assuming, for convenience, that, for every p, x_p has a linear dependency on $\mathcal{M}$ past values of each of the signals including itself, takes the form

$$\begin{aligned} x_p[n] &= \sum_{m=1}^{\mathcal{M}} a_{pp}^{m*} x_p[n-m] + \left(\sum_{\substack{q=1 \\ q \neq p}}^{N_s} \sum_{m=1}^{\mathcal{M}} a_{pq}^{m*} x_q[n-m] \right) + \eta_p[n] \\ &\doteq (\boldsymbol{a}_p^*)^T \boldsymbol{x}[n] + \eta_p[n] \,, \end{aligned} \tag{2.3}$$

where η_p continues to denote a scalar white-noise excitation for p and the vectors $\boldsymbol{a}_p^*$ and $\boldsymbol{x}[n]$ are extended in the natural way relative to Eq. (2.1):

$$\begin{aligned} \boldsymbol{a}_p^* &= \begin{bmatrix} a_{p1}^{1*} & a_{p1}^{2*} & \cdots & a_{p1}^{\mathcal{M}*} & a_{p2}^{1*} & \cdots & a_{p2}^{\mathcal{M}*} & \cdots & \cdots & a_{pN_s}^{1*} & \cdots & a_{pN_s}^{\mathcal{M}*} \end{bmatrix}^T \quad \text{and} \\ \boldsymbol{x}[n] &= \begin{bmatrix} x_1[n-1] & \cdots & x_1[n-\mathcal{M}] & \cdots & \cdots & x_{N_s}[n-1] & \cdots & x_{N_s}[n-\mathcal{M}] \end{bmatrix}^T \end{aligned} \tag{2.4}$$

with $\boldsymbol{a}_p^*$ and $\boldsymbol{x}[n]$ both vectors in $\mathbb{R}^{M_{\boldsymbol{a}^*}}$ where $M_{\boldsymbol{a}^*}$ is the number of parameters used in modeling signal x_p,

$$M_{\boldsymbol{a}^*} \doteq N_s\mathcal{M} = \dim\{\boldsymbol{a}_p^*\}. \tag{2.5}$$

Although this is not customary in the current literature on causality modeling, the most conventional way to refer to such a model (for each p) would be as an ***autoregressive model with exogenous inputs*** (ARX). One can also view this model as representing a multiple-input, single-output (MISO), discrete-time system (if the disturbance η_p is viewed as an excitation), but with the caution that it is only recursive in the signal x_p, with x_j, $\forall j \neq p$ serving as exogenous inputs for each p. Models accounting for multiple outputs are sometimes referred to as ***jointly regressive*** models [100, 109] or ***multivariate autoregressive*** (MVAR) models [33, 108, 204].

An important special case of the observation model of Eq. (2.3) occurs for $N_s = 2$ which appears in problems in which the causality effects between two signals are analyzed. In this case, the estimation model can be written as two explicit equations,

$$\begin{aligned} x_1[n] &= \sum_{m=1}^{\mathcal{M}} a_{11}^{m*} x_1[n-m] + \sum_{m=1}^{\mathcal{M}} a_{12}^{m*} x_2[n-m] + \eta_1[n] \\ &\doteq \boldsymbol{a}_{11}^{*T}\boldsymbol{x}_1[n] + \boldsymbol{a}_{12}^{*T}\boldsymbol{x}_2[n] + \eta_1[n], \\ x_2[n] &= \sum_{m=1}^{\mathcal{M}} a_{22}^{m*} x_2[n-m] + \sum_{m=1}^{\mathcal{M}} a_{21}^{m*} x_1[n-m] + \eta_2[n] \\ &\doteq \boldsymbol{a}_{22}^{*T}\boldsymbol{x}_2[n] + \boldsymbol{a}_{21}^{*T}\boldsymbol{x}_1[n] + \eta_2[n]. \end{aligned} \tag{2.6}$$

These equations can be formulated as the more general model of Eq. (2.3). For example, for $N_s = 2$,

$$\begin{aligned} \boldsymbol{a}_p^* &= \begin{bmatrix} \boldsymbol{a}_{p1}^{*T} & \boldsymbol{a}_{p2}^{*T} \end{bmatrix}^T \quad \text{and} \\ \boldsymbol{x}[n] &= \begin{bmatrix} \boldsymbol{x}_1^T[n] & \boldsymbol{x}_2^T[n] \end{bmatrix}^T. \end{aligned} \tag{2.7}$$

2.2.2 Estimation model

Turning to the estimation model, it is customary in the linear modeling case – and consistent with ***minimum-mean-squared-error*** (MMSE) estimation theory – to take the form of the noise-free

observation model as the basis of the estimation model. For the general observation model of Eq. (2.3), the estimation model for signal x_p becomes

$$\hat{x}_p[n] = \sum_{m=1}^{M} a_{pp}^{m} x_p[n-m] + \left(\sum_{\substack{q=1 \\ q \neq p}}^{N_s} \sum_{m=1}^{M} a_{pq}^{m} x_q[n-m] \right) \doteq \boldsymbol{a}_p^T \boldsymbol{x}[n], \tag{2.8}$$

where M is the model order. It is to be observed that the "$*$" superscripts do not appear on the notation for the parameter estimates. This is a deliberate effort to distinguish a "true" coefficient in the observation model, say a_{pq}^{m*}, from the symbolic representation of the corresponding parameter to be determined in the estimation model. It will be our custom to refer to estimation model of (2.8) as the ***estimated model*** when we wish to stress that the parameters have taken values determined by an optimization procedure over observed data [161].

Note that M itself is a parameter of the model, which must also be predetermined. While theoretically any model with $M \geq \mathcal{M}$ could perfectly represent the observation model, the parameter estimators become less accurate as M increases. An example of the distributional characteristics of the parameter estimates will be given in Sec. 2.2.3 for jointly normally distributed signals. Many methods of comparing models with different M values exist, such as Akaike Information Criterion (AIC) [5], Final Prediction Error (FPE) [4], Minimum Description Length (MDL) [168], Bayesian information criterion (BIC) [174], and other hybrid methods [62].

It is further noteworthy that, whereas the observation model is AR or ARX in the signal x_p – that is, it is recursive in the signal x_p – the estimation model is purely "feedforward" in producing an output as a linear combination of past values of x_p, and of some subset of the remaining $N_s - 1$ signals, at time n. Such a model does not correspond to any conventional (Box-Jenkins-type) time-series model, but, in the parlance of signal processing, corresponds to a MISO discrete-time system [32]. Note also the absence of any noise in the estimated model process.

Associated with an estimated model for signal x_p is an error sequence, say ϵ_p, with value at time n given by

$$\epsilon_p[n] \doteq x_p[n] - \hat{x}_p[n]\,. \tag{2.9}$$

By subtracting Eq. (2.8) from Eq. (2.3), we see that this error contains components due to inaccura-

cies in the estimated coefficients, as well as the disturbance sequence η_p,

$$\epsilon_p[n] = (\boldsymbol{a}_p^* - \boldsymbol{a}_p)^T \boldsymbol{x}[n] + \eta_p[n]. \tag{2.10}$$

A slight abuse of notation is used here, where $\boldsymbol{a}_p^*$ and $\boldsymbol{a}_p$ are zero-padded to account for the missing elements (when $M \neq \mathcal{M}$) and $\boldsymbol{x}[n]$ is similarly adjusted to account for any missing elements. For example, suppose $M > \mathcal{M}$, then $\boldsymbol{a}_p^*$ is padded with $M - \mathcal{M}$ zeros in the locations that correspond to $x_q[n - M - 1] \cdots x_q[n - \mathcal{M}]$ for all $q \in \{1, \cdots, N_s\}$.

When the parameters are correctly identified in the estimation model, so that $\boldsymbol{a} = \boldsymbol{a}^*$, then the estimation error is equivalent to the white-noise disturbance of the observation model at each n, $\epsilon_p[n] = \eta_p[n]$. This is known to be the case for the MMSE estimate of the parameters of such a linear model [153], assuming that the model order of the estimated model is greater or equal to that of the observation model. The LSE solution asymptotically approaches the MMSE solution as the number of observations increase.

In practice, of course, the parameter estimates $\boldsymbol{a}$ must be determined from finite data records of the signals $\left\{ x_j \right\}_{j=1}^{N_s}$. Without loss of generality, we may assume that each of the signals is observed on the time indices, $n = 1, 2, \ldots, N - 1$, observation $x_p[N]$ is additionally available, and the parameters are sought with which to model the signal x_p on the interval $n = 1, \ldots, N$. Let $\boldsymbol{a}[N]$ denote the vector of parameter estimates obtained on this interval, and let $\left\{ \epsilon_p\big(n \mid N \big) \right\}_{n=1}^{N}$ be the corresponding error sequence associated with the estimated model with parameters $\boldsymbol{a}[N]$. The assumption of "small errors" (i.e., $\sigma_{\eta_p}^2 \ll \sigma_{\hat{x}_p}^2$) is often used to justify the use of (LSE) estimation of the parameters on the finite interval. In fact, in the present context, the lack of correlation in the sequence $\eta_p[n]$ leads to an unbiased LSE estimate, $\boldsymbol{a}[N]$, for finite N, and asymptotic convergence in mean square to $\boldsymbol{a}^*$.

The observations on the given time range comprise a set of N equations in $M_{\boldsymbol{a}} \doteq \dim\{\boldsymbol{a}\}$

unknown parameters (maximally $M_{\boldsymbol{a}} = MN_s$), which, written in vector-matrix form as,

$$\underbrace{\begin{bmatrix} \hat{x}_p[1] \\ \hat{x}_p[2] \\ \vdots \\ \hat{x}_p[N] \end{bmatrix}}_{\doteq \hat{\boldsymbol{x}}_p[N]} = \underbrace{\begin{bmatrix} x_1[0] & x_1[1] & \cdots & x_1[N-1] \\ \vdots & \vdots & \ddots & \vdots \\ x_1[-M+1] & x_1[-M+2] & \cdots & x_1[N-M] \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ x_{N_s}[0] & x_{N_s}[1] & \cdots & x_{N_s}[N-1] \\ \vdots & \vdots & \ddots & \vdots \\ x_{N_s}[-M+1] & x_{N_s}[-M+2] & \cdots & x_{N_s}[N-M] \end{bmatrix}^T}_{\doteq \boldsymbol{X}^T[N]} \underbrace{\begin{bmatrix} a_{p1}^1[N] \\ \vdots \\ a_{p1}^M[N] \\ \vdots \\ \vdots \\ a_{pN_s}^1[N] \\ \vdots \\ a_{pN_s}^M[N] \end{bmatrix}}_{\doteq \boldsymbol{a}[N]}, \tag{2.11}$$

where $\hat{\boldsymbol{x}}_p[N] \in \mathbb{R}^N$, $\boldsymbol{X}[N] \in \mathbb{R}^{N \times M_a}$, and $\boldsymbol{a}[N] \in \mathbb{R}^{M_a}$. In these terms, the LSE estimate is the solution to

$$\boldsymbol{X}^T[N]\boldsymbol{X}[N]\boldsymbol{a}_p[N] = \boldsymbol{X}^T[N]\hat{\boldsymbol{x}}_p[N]. \tag{2.13}$$

The error sequence may be added to the estimated model for signal x_p to create a model that exactly produces the original signal:

$$x_p[n] = \boldsymbol{a}_p^T \boldsymbol{x}[n] + \epsilon_p[n], \tag{2.14}$$

or, if we wish to emphasize the short-term temporal nature of the estimated parameters in the model,

$$x_p[n] = \boldsymbol{a}_p^T[N]\boldsymbol{x}[n] + \epsilon_p\big(n \mid N\big). \tag{2.15}$$

Although this model theoretically produces the exact signal x_p over the interval $n = 1, \ldots, N$, it is generally very different from the observation model of Eq. (2.3). We refer to Eq. (2.15) as the ***error-augmented estimated model***. As noted near Eq. (2.10), the estimation error sequence ϵ_p is dependent upon the misadjustment in the parameter values relative to the presumed true values of the observation model, $\boldsymbol{a}_p^* - \boldsymbol{a}_p[N]$, as well as the disturbance sequence in the observation, η_p. Not discussed above is the fact that the error sequence is also dependent upon the short-term

estimation of the parameters (*i.e.*, the duration N). The error sequence is therefore a key indicator of the quality of the model and we will see this sequence play an important role in causality analysis.

2.2.3 Least squares estimation

An error sequence accounts for both the disturbance sequence and errors in parameter estimation [Eq. (2.10)]. Under the assumption of "small errors", minimizing the error sequence therefore approximately minimizes the parameter estimation error. The well-known solution to the normal equations, Eq. (2.13), is given by [74]

$$\boldsymbol{a}_p = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{x}_p, \tag{2.16}$$

in which $(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is the pseudoinverse of $\boldsymbol{X}$.

Assuming that the disturbance is i.i.d. zero mean Gaussian random process with variance $\boldsymbol{\sigma}_{\boldsymbol{\eta}}^2$, that the regressors are bounded and the covariance matrix of the regressors Σ_X exists and is non-singular and that the observation model is BIBO[1] stable, the solution is distributed as

$$\boldsymbol{a}_p \sim \mathcal{N}_{M_a}\left(\boldsymbol{a}_0,\ \boldsymbol{\sigma}_{\boldsymbol{\eta}}^2\Sigma_X^{-1}/(N - M_a)\right), \tag{2.17}$$

where $\mathcal{N}_{M_a}(\boldsymbol{\mu}, \Sigma)$ is a multivariate normal distribution of dimension M_a with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, Σ_X^{-1} is the inverse of the covariance matrix of the regressors, and N is the number of time samples.

For sets of regressors with ill-conditioned covariance matrices, the variance of $\boldsymbol{a}_p$ can be very large. As NC depends directly on the accuracy of the model parameters, it is prone to misleading results for small N (compared to the largest element of the vector $\Sigma_X^{-1}\sigma_\eta^2$).

If the i.i.d. Gaussian assumption is not satisfied, ill-conditioned regressor matrices will still cause the parameter estimates to be have potentially large variance, although the parameters may

[1] Bounded-input-bounded-output (BIBO) stability is a form of system stability linking the output of a system to its inputs. A discrete time signal $x[n]$ is called bounded if there exists a $B > 0 \in \mathbb{R}$ such that for every $n \in \mathbb{Z}$ $|x[n]| < B$. A system is called BIBO stable if and only if, given any bounded input, the output is also guaranteed to be bounded [151].

not be normally distributed. Special attention must be taken in the case of NC, to assure that the covariance matrix is well conditioned or regularization must be applied to reduce errors in the NC measure estimation.

2.2.4 ARMAX models

The most comprehensive way to represent LTI models is using the ARMAX representation. This representation encompasses AR, MA, models with exogenous inputs and any combination thereof. Starting with the error augmented model of Eq. (2.14), expanded to highlight the AR, MA and exogenous inputs,

$$\begin{aligned} x_p[n] &= \overbrace{\underbrace{\sum_{m=1}^{M_{\mathrm{AR}}} a_{pp}^{m} x_p[n-m]}_{\text{autoregressive}} + \underbrace{\left(\sum_{\substack{q=1 \\ q\neq p}}^{N_s} \sum_{m=1}^{M_{\mathrm{X}}} a_{pq}^{i} x_q[n-m]\right)}_{\text{exogenous input}} + \underbrace{\sum_{m=1}^{M_{\mathrm{MA}}} a_{p\epsilon}^{m} \epsilon_p[n-m]}_{\text{moving average}}}^{\hat{x}_p[n]} + \underbrace{\epsilon_p[n]}_{\substack{\text{error} \\ \text{sequence}}}, \\ &= \boldsymbol{a}_p^T \boldsymbol{x}[n] + \epsilon_p[n], \end{aligned} \tag{2.18}$$

where M_{AR} is the model order for the AR term, M_{MA} is the model order for the MA term and M_{X} is the model order for the exogenous input term. Note that the model orders and generally unknown (unless predicated on *a priori* knowledge of the system being modeled) and must be estimated prior to the parameter estimation, additionally ϵ_p must also be estimated. The Box-Jenkins method [32] is the standard approach to iteratively identify ARMAX model structures.

2.2.5 ARX models with non-white error sequences

Digressing momentarily into ARMAX modeling, note that ARMAX models of Eq. (2.18) can be expressed in terms of the sum of ARX model and a colored noise term

$$x_p[n] = \underbrace{\sum_{m=1}^{M_{\mathrm{AR}}} a_{pp}^{m} x_p[n-m] + \left(\sum_{\substack{q=1 \\ q\neq p}}^{N_s} \sum_{m=1}^{M_{\mathrm{X}}} a_{pq}^{i} x_q[n-m]\right)}_{\text{ARX model}} + \underbrace{\epsilon_p'[n]}_{\substack{\text{colored} \\ \text{error} \\ \text{sequence}}}, \tag{2.19}$$

where

$$\epsilon'[n] = \sum_{m=1}^{M_{\text{MA}}} a_{p\epsilon}^{m} \epsilon_p[n-m] + \epsilon_p[n], \tag{2.20}$$

such that

$$\max_{n\in[1,N]} \left|\epsilon_p'[n]\right| \leq \left(1 + \sum_{m=1}^{M_{\text{MA}}} \left|a_{p\epsilon}^{m}\right|\right) \max_{n\in[1,N]} \left|\epsilon_p[n]\right|, \tag{2.21}$$

which shows that if ϵ_p is bounded, ϵ_p' will also be bounded. These characteristics will be exploited in Sec. 2.4.

2.2.6 NARMAX and modified NARX models

The LTIiP class of models extend traditional LTI models by allowing nonlinear transformations of the model inputs and past outputs, while allowing the use of many of the classical modeling, prediction and estimation techniques with well-understood and well-tested convergence characteristics. LTIiP models have shown to be a viable alternative to highly nonlinear in parameter models [41], with excellent results in many applications, from epidemiology [165] and microbial growth [210] to human physiology [116]. The most comprehensive representation of LTIiP models is the ***nonlinear ARMAX*** (NARMAX) [39, 122], which are expressed as

$$\begin{aligned} x_p[n] &= \sum_{k=1}^{K} a_{pk} \varphi_{pk}\left(x_1\big|_{n-M}^{n-1}, x_2\big|_{n-M}^{n-1}, \dots, x_{N_s}\big|_{n-M}^{n-1}, \epsilon_p\big|_{n-M}^{n-1}\right) + \epsilon_p[n] \\ &\doteq \boldsymbol{a}_p^T \boldsymbol{\varphi}_p[n] + \epsilon_p[n] \end{aligned} \tag{2.22}$$

where K is the number of regressor functions, φ_{qp} is the q^{th} regressor function, $x_r\big|_{n-M}^{n-1}$ represents the set of all available samples of signal x_r from time $n-M$ until time $n-1$ and a_{pk} is the parameter weight associated with φ_{pk}. Here, the argument of the φ_{qp} is included to reinforce the fact that the regressor functions may depend on any combination of the regressor signals (including the error). Common regressor function families include radial basis functions [40], wavelets [26], and polynomials [6, 8, 24, 82].

In [202], Wei uses the ***linear in parameters nonlinear in variables*** (LIP-NIV) terminology to describe NARMAX models. However, this implies that the models are inherently nonlinear

in variables, which would exclude ARMAX models from the category. Instead, this work will maintain the usage of the LTIiP terminology to highlight that traditional LTI models are a subset of NARMAX models.

The modeling power of NARMAX models comes at the cost increased complexity in estimating parameters. Due to the large number of highly correlated regressors, slow convergence, overfitting and inaccurate parameter estimates are common challenges faced when estimating model parameters [11].

The estimation of parameters that depend on past values of the error sequence in linear ARMAX models (MA portion) is considerably more complex than for the parameters associated autoregressive and exogenous inputs portions of the model. While there are methods for estimating MA parameters [63, 204], and iterative approaches exist for NARMAX models, many approaches focus on NARX models [26, 40]. Additionally, the interpretability of terms that depend on the error sequence have lower interpretability and are often not included in final predictive model [200], as these noise terms are not useful for model prediction but are only used to reduce bias in model estimation [200, 202]. A small modification to NARMAX models simplify parameter estimation is

$$\begin{aligned} x_p[n] &= \sum_{k=1}^{K} a_{pk}\varphi_{pk}\left(x_1\big|_{n-M}^{n-1}, x_2\big|_{n-M}^{n-1}, \ldots, x_{N_s}\big|_{n-M}^{n-1}\right) + \sum_{k=1}^{K_\epsilon} b_{pk}\phi_{pk}\left(\epsilon_p\big|_{n-M}^{n-1}\right) + \epsilon_p[n] \\ &\doteq \boldsymbol{a}_p^T\boldsymbol{\varphi}_p[n] + \epsilon_p'[n] \end{aligned} \tag{2.23}$$

where

$$\epsilon_p'[n] = \sum_{k=1}^{K_\epsilon} b_{pk}\phi_{pk}\left(\epsilon_p\big|_{n-M}^{n-1}\right) + \epsilon_p[n], \tag{2.24}$$

so that regressor functions may depend on either the regressor signals or past values of the error sequence. This restriction to NARMAX models is equivalent to NARX models with colored noise.

2.2.7 LASSO regression

The ***least absolute shrinkage and selection operator*** (LASSO) [193] is an extension to traditional least squares estimation, in which an l_1 norm regularization is employed to encourage sparsity in

the parameters. LASSO regression is equivalent to finding a parameter vector $\boldsymbol{a}$ that satisfies

$$\underset{\boldsymbol{a}\in\mathbb{R}^{Ma}}{\arg\min}\left\{\|x-\hat{x}(\boldsymbol{a})\|_2^2+\lambda\|\boldsymbol{a}\|_1\right\}, \tag{2.25}$$

in which x the signal being modeled and $\hat{x}(\boldsymbol{a})$ is the prediction of x based on the parameter vector $\boldsymbol{a}$, and λ a the regularization factor.

Unlike the l_0 norm, the l_1 norm allows the use of efficient gradient-based optimization techniques [70], while being more effective at encouraging sparsity in the parameter space than, for example, Tikhonov (l_2 norm) regularization.

2.3 Set-membership optimum bounded ellipsoid algorithms

All parameter estimation strategies share a similar goal: finding the optimum parameter estimates given a limited amount of data. The optimality criterion differs between algorithms, for example, the smallest prediction error for LSE or a compromise between prediction error and sparsity of parameters [Eq. (2.25)] for LASSO. Set-membership estimation approaches aim at providing the set of parameters that are consistent with the observed data and the model.

Starting with a putative NARMAX observation model of the form

$$\begin{aligned} x[n] &= \sum_{k=1}^{K} a_{pk}^{*}\varphi_{pk}^{*}\left(x_1\big|_{n-M}^{n-1}, x_2\big|_{n-M}^{n-1}, \dots, x_{N_s}\big|_{n-M}^{n-1}, \eta_p{}_{n-M}^{n-1}\right)+\epsilon_p[n] \\ &\quad (\boldsymbol{a}_p^{*})^T\boldsymbol{\varphi}_p^{*}+\eta[n], \end{aligned} \tag{2.26}$$

where K is the number of regressor functions and M is the model order for which there exists a sequence of positive numbers $\gamma[n]$, such that

$$|\eta[n]|^2 < \gamma[n]. \tag{2.27}$$

For a estimation model of the form

$$\hat{x}[n] = \boldsymbol{a}_p^T\boldsymbol{\varphi}_p[n]+\epsilon[n], \tag{2.28}$$

the sequence $\gamma[n]$ imposes the constraint at each time n,

$$\left|x[n]-\boldsymbol{a}_p^T\boldsymbol{\varphi}_p[n]\right| < \gamma[n], \tag{2.29}$$

or, equivalently,

$$\begin{aligned} \boldsymbol{a}_p^T \boldsymbol{\varphi}_p[n] &< x[n] + \gamma[n] \\ \boldsymbol{a}_p^T \boldsymbol{\varphi}_p[n] &> x[n] - \gamma[n] \end{aligned} \tag{2.30}$$

which define a hyperstrip (region between the two parallel hyperplanes) in which the set of valid parameters - known as ***feasibility set*** - must lie. The intersection of any set of K or more hyperstrips defined by linearly independent observations forms a convex polytope of dimension K. If at time n, the polytope defined by the intersection of all previous hyperstrips is not fully contained within the hyperstrip defined by Eq. (2.29), the feasibility set is refined. This is akin to faceting a gem, where each new refinement potentially adds up to two flat facets to the polytope.[2]

Although the polytope defined by the feasibility set has finite dimension, there is no limit to the number of facets. The evaluation of the intersection of hyperstrips becomes increasingly complex as the number of considered time samples increases. Optimum bounded ellipsoid algorithms provide a computationally efficient approximation to the polytope by evaluating a hyperellipsoid that bounds the polytope [52]. Compared with the polytope, the unfaceted nature of the hyperellipsoid is more akin to a cabochon (a polished unfaceted gem). A geometric illustration for a bidimensional parameter space is shown in Fig. 2.1. In Fig. 2.1, the x-axis represents the value of θ_1, the y-axis represents the value of θ_2. ω_2 is strip defined by $\boldsymbol{\varphi}_p[2]$ and $x[2]$, likewise, ω_3 is strip defined by $\boldsymbol{\varphi}_p[3]$ and $x[3]$. Ω_3 is the intersection between ω_2 and ω_3 and Θ_3 is an ellipsoid that bounds Ω_3. Note that Θ_3 and Ω_3 are both completely contained within the strip ω_4, so no refinement occurs at time $n = 4$.

The ability to reject samples that do not reduce the feasibility set is a significant advantage of OBE algorithms. While the recursion for OBE algorithms is very similar to a ***weighted recursive least squares*** (WRLS), thus $\mathcal{O}\left(K^2\right)$ complexity per time sample processed, typically only a small fraction of samples provides refinement to the ellipsoid [54]. Despite the increased computational efficiency in comparison to WRLS, OBE algorithms produced guaranteed bounds for the feasibility

[2] Any facets completely located outside the hyperstrip are removed from the polytope, so the number of facets does not necessarily monotonically increase, even though the volume of the polytope monotonically decreases with every refinement.

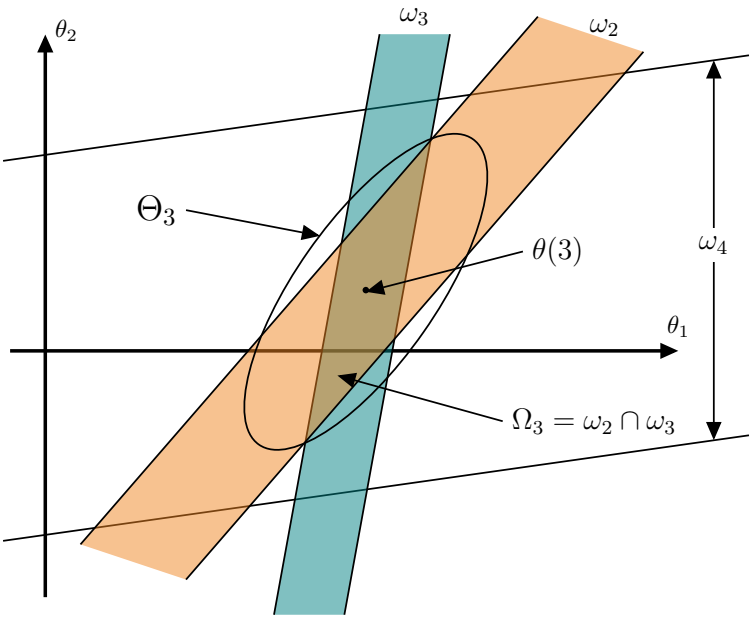


Figure 2.1: Geometric illustration of OBE algorithms

set.

The ellipsoid can be succinctly defined by the centroid and a matrix containing the principal axes of the ellipsoid. In the case of OBE algortithms, the feasibility set at time n is defined as

$$\Theta \doteq \left\{ \boldsymbol{\theta} \in \mathbb{R}^K : (\boldsymbol{\theta} - \boldsymbol{\theta}_c)^T \frac{\boldsymbol{C}}{\kappa} (\boldsymbol{\theta} - \boldsymbol{\theta}_c) < 1 \right\}, \tag{2.31}$$

where $\boldsymbol{\theta}_c$ is the centroid of the ellipsoid, $\boldsymbol{C}$ is the sample covariance matrix of $\boldsymbol{\varphi}_p$ (thus a positive semidefinite matrix), and κ is a positive scalar, such that $\frac{\boldsymbol{C}}{\kappa}$ define the principal axes of the ellipsoid.

Another advantage of SM estimation is that it requires fewer assumptions about the distributional characteristics of the noise term. The only requirement for the employment of SM algorithms is that the noise be bounded over the observed sequence. Methods for estimating the bounds have been developed with proven conditions under which convergence is guaranteed [125].

The recursion steps for OBE algorithms can be found in Alg. B.1 of the appendix with a short overview on the difference between variants and enhancements to the algorithm.

2.4 NARMAX model estimation and the EvolOBE method

While NARMAX models are often able to represent many complex interactions with few terms, the parameters associated with such terms must still be estimated. As the number of regressor functions increases, parameter estimation is very likely to become an ill-conditioned problem. Thus, traditional regression methods such as LSE do not generally produce good results when coupled with nonlinear models with a large number of regressor functions. The number of candidate regressor functions often is very large. For example, for polynomial regressor functions, the number of such functions grows factorially with the polynomial order. An exhaustive search of all possible subsets is computationally prohibitive for most practical applications. Thus, finding optimal subsets of the regressor functions becomes fundamental to properly estimate models. This section contains an overview of a family of NARMAX model estimation algorithms that is particularly suited for causality analysis.

The poor conditioning of a large set of regressor functions is in large part due to the fact that many regressor functions will be highly correlated with one another (*e.g.*, x and x^3 have a correlation coefficient of 0.77 for x normally distributed with zero mean and unity variance). Additionally, many commonly used sets of regressor functions form overcomplete systems, which creates null spaces in the regressor space.

Many techniques have been developed specifically for nonlinear model selection and parameter estimation [22, 25, 27, 81, 118, 201, 203, 214]. These tend to fall within three categories: stepwise search algorithms, bridge regression and evolutionary search.

Stepwise search algorithms iteratively add or remove candidate regressor functions from the model until a criterion is reached. Since the number of possible "paths" grows factorially with the number of regressor functions, most employ greedy approaches, where the regressor which most reduces the prediction error of the NARMAX model is chosen and/or the prune the regressor functions which least increase the prediction error when removed. Matching pursuit [131], Forward-Regression Orthogonal Least Squares (FROLS) [25] and Least Angle Regression (LARS) [65] are prominent examples. Stepwise approaches suffer from shortcomings in practice.

Particularly, autoregressive terms are typically included first in the search, especially for systems with dynamics well below the sampling frequencies [23]. This is true regardless of how important those terms are in the final model. Once the initial autoregressive terms are selected, the remaining prediction error is often small enough that the choice of regressors is sensitive to noise in the data [157].

Bridge regression methods [67] add a penalty to the cost function proportional to the ℓ_ρ-norm of the parameters[3]. Bridge methods can be used independently or combined with stepwise methods. Ridge regression [90] (also known as Tikhonov regularization [194]) uses ℓ_2-norm and possesses closed-form solution and can improve conditioning in ill-posed problems, but do not generate sparse solutions. LASSO regression [193] (also know as basis pursuit [42]) use the ℓ_1-norm and are effective ways of finding sparse solutions.

However, existing model structure and parameter estimation methods suffer (to differing degrees) from slow or inaccurate convergence of the parameters [11], high computational cost [138] and often produce inaccurate model structures [201].

Evolutionary search is well suited for regressor selection with many examples in the literature [115, 121, 163, 187]. While more computationally expensive than bridge or step-wise methods, it is able to find global optima within the search space at an acceptable computational cost (sometimes even comparable to gradient-based approaches [139]). The EvolOBE method [210–217] differs from these approaches by combining the evolutionary search with set-theoretic OBE algorithms. The OBE class of parameter estimation algorithms possesses several desirable characteristics that make it particularly suited for the problem of estimating parameters for models of the form given by Eq. (2.23), for example, no necessity to make assumptions about the stationarity and distributional characteristics of the noise, and efficient computation of parameters.

Earlier variants of the algorithm used more traditional methods of evaluating model fitness, such as AIC and FPE, but later variants use a bi-objective evolutionary search [149, 217] that produces a set of models with the best compromise between predictive power and complexity.

[3] ρ is most often set such that $0 \leq \rho \leq 2$ [70]

This obviates the choice of hyperparameters or assumptions to regulate the trade-off between the two objectives and allows a wider search and greater population diversity [195] as solutions that have high fitness for different objectives can more easily coexist and coevolve.

2.4.1 Genetic encoding and algorithm overview

In the EvolOBE method, models are treated as chromosomes. The LTIiP model is the phenotype of a chromosome, a binary sequence in which each bit indicates the presence or absence of a particular gene. Each gene codes for a particular regressor function in the model. The algorithm starts with a random population of chromosomes. The parameter sets result from the set-membership processing of the data and the genetic makeup of each chromosome. Unlike other estimation methods, the set-membership algorithms provide sets of feasible parameter vectors rather than a single point estimate. Measurable set properties are then used to assign fitness values to each chromosome, and the fitness value is used in the genetic algorithm selection process to evolve the population toward better solutions (e.g. [167]). This framework simultaneously addresses selection of the model structure and the parameter estimation.

To reduce the computational complexity of this process, the search space of regressor models must be controlled, and the candidate and final models must use the fewest regressors that are consistent with an objective of prediction-error minimization, Since these objectives are conflicting, a multi-objective optimization approach is desired. For this work, the Non-dominated Sorting Genetic Algorithm - II (NSGA-II) [51] approach is adopted, since it generates set solutions (ideally the Pareto-front), providing the best solution for a given number of regressors and allowing the model with the best trade-off to be chosen.

NSGA-II is a standard algorithm for solving multiobjective optimization problems. It requires a small number of parameters and is able to obtain solution sets with good spread. The basic NSGA-II algorithm is shown in Fig. 2.2. An initial random population of size N is generated and evaluated according to the two objectives: prediction accuracy and number of regressors. The population is then sorted, the best half is selected as parents, which go through selection,

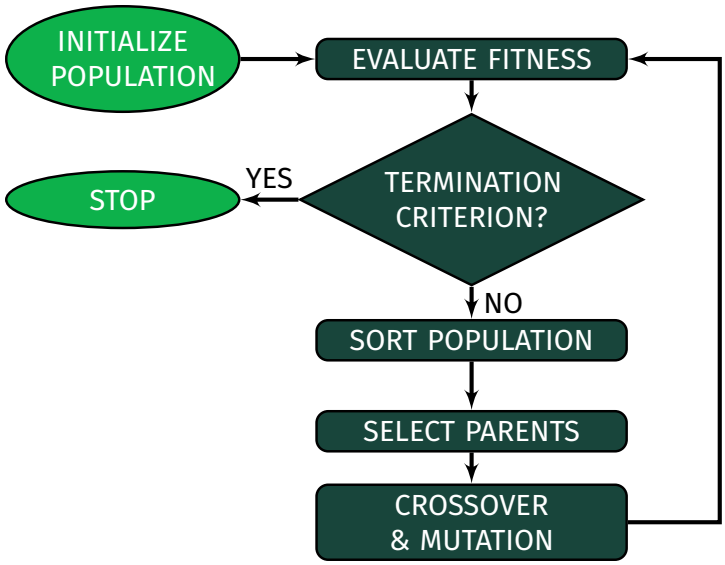


Figure 2.2: NSGA-II algorithm summary

mutation and crossover to generate a new population of children. The parents and children of this generation become the parents of the following generation. The cycle is repeated until the termination criterion is reached.

In the seminal EvolOBE paper [214], Yan *et al.* used binary tournament, bit-wise mutation and single-point crossover.Later variants of the EvolOBE algorithm use different mutation and cross-over algorithms tailored for discovery of sparse models which provide faster convergence [149].

The sorting of the population occurs at two tiers. First, the population is sorted by ***fronts***, each front is formed by a set of solutions has higher optimality than all other members of the set (this is called being ***non-dominated***). The population is sorted such that the members of the first front is placed higher in the set of solutions, followed by the subsequent fronts sequentially until the entire population has been sorted. Within a front, the population is sorted by the sum of the edge lengths of the cuboid formed by the two surrounding solutions within the front (in the bi-objective case), this is known as the ***crowding distance***. The elements with larger crowding distance are placed higher within their respective fronts. As a consequence of how the crowding distance is computed, the edge solutions are always ranked higher, as they do only have a single solution surrounding them (infinite crowding distance).

2.5 Causality analysis

Philosophers and scientists have vigorously debated the meaning of causality and no universally accepted definition exists. In [103], philosopher David Hume argued that the human mind is not able to fully assert true causality, only to observe events occurring in succession. Nevertheless, Hume proposes conditions for a relationship to be called causal. While, not universally accepted, this work will use Hume's definition of causality, because it is testable and quantifiable. Like Box [31], our intent is not to find "true" models, but rather gain insight and understanding of the systems being studied. Nevertheless, a "true" model may be posed in some cases for theoretical analysis.

The most widely known method of assessing causality strength is GC. It was first postulated by Norbert Wiener that if the inclusion of a regressor could improve the prediction of a regressand, then the relationship between the regressor and regressand could be assumed "causal" [206]. Granger used this idea to give a formal definition of causality and feedback in the context of AR models [76]. Granger Causality relies on Hume's work [103], which focused on epistemological causality (focusing on what can can be learned and known), rather than ontological (how things are). Hume posed certain conditions under which causality can be ascertained. These conditions are discussed in Sec. 2.5.2 and connected with the definition of GC. While Granger himself distinguished GC from "true causality" [77], GC performs well in a number of applications, from econometrics [60, 87] to neurology [175].

While causality analysis often involves the use of predictive models, there is no guarantee that the predictive models internally represent the systems that they are modeling. This is closely related to the distinction between correlation and causation (known as the *cum hoc ergo propter hoc* fallacy). Although precedence (when coupled high correlation) may seem like a good indicator of causality, it also cannot be equated with causation (known as the *post hoc* fallacy). For example, many people brush their teeth before going to sleep; however, brushing teeth does not cause sleep. Granger himself has highlighted the distinction between "true" causality and GC [77].

Of particular interest is NC, which was developed to address limitations of GC in measuring

causal mechanisms and which has shown useful results in a number of applications [95, 96, 98–100, 105, 112, 220]. It has been pointed out that GC measures causal effect rather than mechanism [19] and NC measures a fundamentally different (although related) quantity.[4] New Causality is better suited as a complement for GC (and other causality measure tools) rather than a replacement.

2.5.1 Humean concept of causality

Hume claims that the relationship between cause and effect cannot be established simply by reasoning, but instead requires an assumption of "uniformity of nature," *i.e.*, that certain natural laws and processes do not change overtime [102]. Although unprovable by means of observation alone, "uniformity of nature" serves as a first principle through which causation can be judged. While Hume believes that "nothing is more evident than that the human mind cannot form such an idea of two objects as to conceive any connection between them" [103, Sec. XIV], he studies causality within the context of what can be understood through experience.

In [103, Sec. XV], Hume postulates the following set of rules by which to judge causes and effects (quoted verbatim here, other than use of modern spelling):

1. The cause and effect must be contiguous in space and time.
2. The cause must be prior to the effect.
3. There must be a constant union between the cause and effect. It is chiefly this quality that constitutes the relation.
4. The same cause always produces the same effect, and the same effect never arises but from the same cause. This principle we derive from experience, and is the source of most of our philosophical reasonings. For when by any clear experiment we have discovered the causes or effects of any phenomenon, we immediately extend our observation to every phenomenon of the same kind,

[4] The claim is disputed by the authors of [99]. Nonetheless, the author tends to agree with [19].

without waiting for that constant repetition, from which the first idea of this relation is derived.

5. There is another principle, which hangs upon this, namely that where several different objects produce the same effect, it must be by means of some quality, which we discover to be common among them. For as like effects imply like causes, we must always ascribe the causation to the circumstance, wherein we discover the resemblance.

6. The following principle is founded on the same reason. The difference in the effects of two resembling objects must proceed from that particular, in which they differ. For as like causes always produce like effects, when in any instance we find our expectation to be disappointed, we must conclude that this irregularity proceeds from some difference in the causes.

7. When any object increases or diminishes with the increase or diminution of its cause, it is to be regarded as a compounded effect, derived from the union of the several different effects, which arise from the several different parts of the cause. The absence or presence of one part of the cause is here supposed to be always attended with the absence or presence of a proportionable part of the effect. This constant conjunction sufficiently proves, that the one part is the cause of the other. We must, however, beware not to draw such a conclusion from a few experiments. A certain degree of heat gives pleasure; if you diminish that heat, the pleasure diminishes; but it does not follow, that if you augment it beyond a certain degree, the pleasure will likewise augment, for we find that it degenerates into pain.

8. The eighth and last rule I shall take notice of is, that an object, which exists for any time in its full perfection without any effect, is not the sole cause of that effect, but requires to be assisted by some other principle, which may forward its influence and operation. For as like effects necessarily follow from like causes,

> and in a contiguous time and place, their separation for a moment shows, that these causes are not complete ones.

A discussion of the philosophical implications of "uniformity of nature" assumption lies outside the scope of this work. Here, systems will be assumed to vary slowly enough that a time-invariant model adequately represents the system dynamics over "short" periods of time in which analysis takes place. Similarly, item 7 implies some proportionality in the causal relationship, where an increase in the cause will proportionally affect the effect. Hume, however, does not exclude the possibility of nonlinearity in the relationship. One must not indiscriminately assume an affine relationship between cause and effect exists even if the observations (under a limited range) closely follow an affine relationship. Therefore, as discussed in Sec. 2.2.2, it is important to remember the distinction between models and the systems they represent.

Additionally, Hume's items 1 and 3 cannot be derived from samples of signals alone, but must be evaluated separately. Note that Hume's concepts of contiguity and union in time and space are loosely defined. Even if internally to the systems, causes and effects might be contiguous, often these states are unobtainable. Additionally, discrete time data collected from a finite number of sensors implies these requirements will never be fully satisfied without additional assumptions (*e.g.*, limited bandwidth). For the purposes of this work, it will be assumed that signals satisfy these requirements. Time-series data are unable to provide information regarding items 1 and 3, which must be evaluated using *a priori* information.

Hume's item 6 states that if two outcomes are different, then the causes must also be different. When some causes cannot be measured or estimated, the outcomes will also not be estimable. The error augmented and observation models [Eq. (2.1) and Eq. (2.14), respectively] account for this by including an unknown disturbance sequence. That is, even if the parameters of the observation model were to be known, discrepancies (however small) are still expected in the prediction. Nevertheless, a disturbance sequence with small variance suggests (but does not guarantee) that most of the "causes" are being accounted for.

What remains for analysis are Hume's items 2, 4 and 5. Item 4 states that if A causes B, then A

must co-occur with B. In the domain of continuous random variables, this is roughly equivalent to the concept of dependence (or correlation for linear models). Item 5 states that if A causes B, and C also causes B, there must be a common element between A and C. Uncovering such mechanisms is helpful when analyzing systems, but the existence of a common factor between A and C does not aid in the decision on whether A and/or C cause B. Finally item 2 requires event A to precede B in order to establish causality of B by A. Although this requirement is intuitive, careful examination is required to ascertain whether A truly precedes B. This is particularly evident in systems that exhibit predictable, periodic, or quasiperiodic behavior. Apparent "noncausal" behavior can be attributed to predictive learning. For instance, rooster crows do not cause the sun to rise, instead, roosters possess the ability to predict sunrise times due to its quasiperiodicity using an internal circadian clock [180] (and also using other cues such as light and even social rank [179]). Nonetheless, few would object to the statement that "the rooster crows just *before* the break of dawn."

2.5.2 Granger causality

By combining item 4 (correlation) and item 2 (precedence), GC assesses the causality strength using the relative increase in predictive power gained by including a second signal into an estimation model. This is done by comparing an estimated ARX model (joint model) over an estimated AR model (disjoint model) where the exogenous input is formed of past samples of the causing signal being studied.[5] The increase in predictive power is used as evidence of causality.[6]

Suppose that stochastic signals x_1 and x_2 are sampled. It is then possible to create predictive models for x_1 varying the presence or absence of x_2. A model that only uses past values of x_1 can

[5] When the current sample of this signal is used, the increase in predictive power is called instantaneous GC. Instantaneous GC violates precedence and therefore weakens the case for calling it "causality."

[6] To highlight the distinction between "true causality" and GC, some authors choose to use the Granger-cause (A Granger-causes B) jargon, however, keeping with Hume's notion of "obtainable causality" and for brevity's sake, this work will refrain from using the term, while acknowledging the distinction between epistemological and ontological causality.

be written as

$$\begin{aligned} x_1[n] &= \varphi_1(x_1|_{n-M}^{n-1}) + \epsilon[n], \\ &= \varphi_1(x_1[n-1], x_1[n-2], x_1[n-3], \cdots, x_1[n-M]) + \epsilon[n], \end{aligned} \tag{2.32}$$

where $\varphi_1(x_1|_{n-M}^{n-1})$ is a function of past values of x_1 from time $n-M$ to time $n-1$ inclusive and ϵ is the error sequence. If φ_1 is a linear function, this predictive model reduces to an AR observation model. A second predictive model using past values of both x_1 and x_2 can be written as

$$x_1[n] = \varphi_2(x_1|_{n-M}^{n-1}, x_2|_{n-M}^{n-1}) + \epsilon'[n], \tag{2.33}$$

where $\varphi_2(x_1|_{n-M}^{n-1}, x_2|_{n-M}^{n-1})$ is a function of past values of x_1 and x_2 from time $n-M$ to time inclusive $n-1$, and ϵ' is the error sequence. If φ_2 is a linear function, this predictive model reduces to an ARX observation model, where x_2 is the exogenous input. Note that both predictive models must have their topology and parameters estimated (in the case of parametric models).

Although in many applications the signals being analyzed are of the same nature (*e.g.,* two EEG channels, two stocks, etc) and minimally processed (*e.g.,* filtering applied for removing volume conduction, line noise, EMG interference, etc), GC can analyzed distinct quantities like the effect of phase from one channel into amplitude of a second channel [141].

The GC value in the contrast represented by [Eqs. (2.32) and (2.33)] is defined as

$$\mathrm{GC}_{2\to 1} = \ln(\sigma_\epsilon^2/\sigma_{\epsilon'}^2), \tag{2.34}$$

where σ_ϵ^2 is the sample variance of the error sequence of the estimated model where x_2 is absent and $\sigma_{\epsilon'}^2$ is the sample variance of the error sequence of the model with x_2 as exogenous input. Since, in general, one of the rational objectives of model estimation is minimizing the residual error, the inclusion of x_2 in Eq. (2.33) assures that $\sigma_{\epsilon'}^2 \leq \sigma_\epsilon^2$ and thus GC≥ 0. In order to evaluate the hypothesis of whether x_2 causes x_1, a statistical significance test, such as an F-test [184], is conducted on the GC statistic.

It is noteworthy that, in general, $\varphi_1(x_1|_{n-M}^{n-1}) \neq \varphi_2(x_1|_{n-M}^{n-1}, \mathbf{0})$ unless $x_2|_{n-M}^{n-1} \doteq \mathbf{0}$; that is, the model estimation method employed for obtaining φ_1 and φ_2 will attempt to fit the data, so φ_1 will adapt

to the absence of x_2. If $x_2\big|_{n-M}^{n-1}$ can be predicted well by $x_1\big|_{n-M}^{n-1}$, then σ_ϵ^2 might not be significantly larger than $\sigma_{\epsilon'}^2$, even if the contribution of x_2 to φ_2 is large [23, 157].

The simplicity of GC allows it to be easily applied to a wide range of problems with good results, *e.g.* [33, 69, 177]. However, since it is designed to measure causal effect, GC value does not fully consider the internal states of the underlying observation model, only the outputs of the model. Further, it has been claimed that GC values are difficult to compare across observation models, as GC values are not normalized and obtaining a threshold for statistical significance is not straightforward [95].

The use of two independently estimated models is vulnerable to resulting bias and larger variance [16, 43]. More recent methods have been developed to derive GC values from a single full regression using factorization of the spectral density matrix [16, 17, 59]. Nevertheless, conceptually, these methods still stem from the comparison of the predictive power of two models.

Although authors have pointed out apparent limitations of GC, [79, 94, 95, 97, 100, 135, 188], GC is a well established methodology for analyzing causal relationships [33]. Additionally, for normally distributed signals, GC has been shown to be equivalent to TE (save by a scaling factor) [14], but can be evaluated reliably with fewer samples. Barrett and Barnett acknowledge in [19] that "GC is not a perfect measure for all stochastic time series: if the true process is not a straightforward multivariate autoregressive process with white-noise residuals, then it becomes only an approximate measure of causal influence. In each real-world scenario, discretion is required in deciding if confounds such as non-linearity and correlations in the noise are mild enough for the measure to remain applicable." While TE is applicable to other models, other authors have also pointed out that causal effects and transferred information [127].

2.5.3 Spectral Granger causality

Spectral GC is the frequency domain decomposition of GC introduced by Geweke [71]. Spectral GC uses the power spectral density (PSD) function to assess GC at particular frequencies. Suppose

there is a pair of signals x_1 and x_2 that can be modeled by

$$\begin{aligned} x_1[n] &= \boldsymbol{a}_{12}^T \boldsymbol{x}_2[n] + \boldsymbol{a}_{11}^T \boldsymbol{x}_1[n] + \epsilon_1[n] \\ x_2[n] &= \boldsymbol{a}_{22}^T \boldsymbol{x}_2[n] + \boldsymbol{a}_{21}^T \boldsymbol{x}_1[n] + \epsilon_2[n], \end{aligned} \tag{2.35}$$

where ϵ_1 and ϵ_2 are assumed to be sampled from white and mutually uncorrelated random processes. Applying the discrete time Fourier transform (DTFT) yields

$$\begin{aligned} X_1(f) &= \boldsymbol{A}_{12}(f) X_2(f) + \boldsymbol{A}_{11}(f) X_1(f) + E_1(f) \\ X_2(f) &= \boldsymbol{A}_{22}(f) X_2(f) + \boldsymbol{A}_{21}(f) X_1(f) + E_2(f), \end{aligned} \tag{2.36}$$

where $\boldsymbol{A}_{12}(f)$, $\boldsymbol{A}_{11}(f)$, $\boldsymbol{A}_{22}(f)$, and $\boldsymbol{A}_{21}(f)$ are the DTFTs of $\boldsymbol{a}_{12}$, $\boldsymbol{a}_{11}$, $\boldsymbol{a}_{22}$, and $\boldsymbol{a}_{21}$ respectively, $X_1(f)$ and $X_2(f)$ and the DTFTs of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ respectively and $E_1(f)$ and $E_2(f)$ are the DTFT of samples of ϵ_1 and ϵ_2 respectively. Through manipulation, Eq. (2.36) can be rewritten as

$$\begin{bmatrix} E_1(f) \\ E_2(f) \end{bmatrix} = \begin{bmatrix} \boldsymbol{B}_{11}(f) & \boldsymbol{B}_{12}(f) \\ \boldsymbol{B}_{21}(f) & \boldsymbol{B}_{22}(f) \end{bmatrix} \begin{bmatrix} X_1(f) \\ X_2(f) \end{bmatrix}. \tag{2.37}$$

As long as $\boldsymbol{B}_{11}(f)\boldsymbol{B}_{22}(f) \neq \boldsymbol{B}_{12}(f)\boldsymbol{B}_{21}(f)$ for any $f \in [-0.5, 0.5]$, Eq. (2.37) can be inverted yielding

$$\begin{bmatrix} X_1(f) \\ X_2(f) \end{bmatrix} = \begin{bmatrix} \boldsymbol{C}_{11}(f) & \boldsymbol{C}_{12}(f) \\ \boldsymbol{C}_{21}(f) & \boldsymbol{C}_{22}(f) \end{bmatrix} \begin{bmatrix} E_1(f) \\ E_2(f) \end{bmatrix}. \tag{2.38}$$

Under these circumstances, the spectral density of x_1 can be written as

$$|X_1(f)|^2 = |\boldsymbol{C}_{11}(f)E_1(f)|^2 + |\boldsymbol{C}_{12}(f)E_2(f)|^2. \tag{2.39}$$

Using Eq. (2.39), SGC is defined as

$$\mathrm{SGC}_{x_2 \to x_1} = \ln\left(\frac{|X_1(f)|^2}{|\boldsymbol{C}_{11}(f)E_1(f)|^2}\right), \tag{2.40}$$

or, equivalently

$$\mathrm{SGC}_{x_2 \to x_1} = \ln\left(1 + \frac{|\boldsymbol{C}_{12}(f)E_2(f)|^2}{|\boldsymbol{C}_{11}(f)E_1(f)|^2}\right). \tag{2.41}$$

This means that the $\mathrm{SGC}_{x_2 \to x_1}$ is proportional to the ratio between the "contribution" of $E_2(f)$ (originating from x_2) and $E_1(f)$ (originating from x_1). As the contribution of $E_2(f)$ to x_1 increases, so does the $\mathrm{SGC}_{x_2 \to x_1}$.

It is important to note that, due to the matrix inversion in Eq. (2.38), the relationship between the parameters in vectors $\boldsymbol{a}_{11}$, $\boldsymbol{a}_{12}$, $\boldsymbol{a}_{21}$ and $\boldsymbol{a}_{22}$ [from Eq. (2.35)] and the functions $\boldsymbol{C}_{11}(f)$ and $\boldsymbol{C}_{12}(f)$ is not straightforward and is model order dependent. An example of the nontrivial relationship between the parameters and GC is shown in Appendix A.

Spectral GC is particularly helpful when the frequency bands of interest are well known or concentrated into relatively narrowband peaks [35]. Another noteworthy characteristic of SGC is that it is (at least theoretically) filtering invariant, that is, the SGC values do not change when the signals are filtered by an invertible filter [15]. In fact, prefiltering the data has been recommended against unless the noise can be very well characterized (*e.g.*, 50Hz/60Hz mains hum) [16].

2.5.4 Conditional Granger causality

Geweke also developed an extension to GC for MVAR models [72]. When analyzing more than two signals, traditional GC is unable to differentiate chains of causal relationships. For example, suppose x, y and z are signals that can be represented by a MVAR model. If both $\mathrm{GC}_{x\to z}$ and $\mathrm{GC}_{y\to z}$ are large, GC cannot distinguish between the model A in Fig. 2.3a from the model B in Fig. 2.3b. Conditional GC solves the ambiguity by evaluating the improvement in the prediction conditioned to other signal or set of signals.

(a) Model A: x causes z directly (b) Model B: x causes z indirectly though z

Figure 2.3: Different explanations for large $\mathrm{GC}_{x\to z}$

In other words, conditional GC compares the variance of the error sequence associated with the model for z predicted using past values of y and z - $\sigma^2_{\epsilon_{z|y}}$ - to the the variance of the prediction error of signal z given past values of x, y and z - $\sigma^2_{\epsilon_{z|x,y}}$. Similarly to Eq. (2.34), conditional GC is defined as

$$\mathrm{GC}_{x\to z|y} = \ln(\sigma^2_{\epsilon_{z|y}}/\sigma^2_{\epsilon_{z|x,y}}), \tag{2.42}$$

In model A, $\mathrm{GC}_{x\to z|y}$ remains large, while in model B, $\mathrm{GC}_{x\to z|y}$ will be small (ideally 0). Thus, a large $\mathrm{GC}_{x\to z|y}$ would indicate model A is more likely than model B.

2.5.5 New causality

Instead of focusing on predictive power as a measure of causality, the NC measure relies on the internal structure of a parametric model and upon evaluating the proportion of the energy of each contribution [which is formally defined in Eq. (2.44)] to infer causation. By making use of the models, NC is able to more proportionately represent the strength of internal mechanisms of the observation model. Also, unlike GC which requires the careful selection of conditioning sets beforehand (otherwise potentially leading to false conclusions [184]), NC foregoes the use of two models and derives its value from a single MVAR model.

Suppose an estimated model is generated for time-series data using an error augmented model in Eq. (2.14), which can be expanded and grouped by regressor signal as

$$x_p[n] = \sum_{h=1}^{N_s}\sum_{m=1}^{M} a_{ph}^m x_h[n-m] + \epsilon_p[n]. \tag{2.43}$$

Under this model, we define the contribution from x_q into x_p as

$$c_{pq}[n] = \sum_{m=1}^{M} a_{pq}^m x_q[n-m] \tag{2.44}$$

such that the NC measure is defined as

$$\mathrm{NC}_{x_q\to x_p} = \frac{\sum_{n=M}^{N}\left(c_{pq}[n]\right)^2}{\sum_{h=1}^{N_s}\sum_{n=M}^{N}\left(c_{ph}[n]\right)^2 + \sum_{n=M}^{N}\epsilon_p^2[n]}, \tag{2.45}$$

or, equivalently,

$$\mathrm{NC}_{x_q\to x_p} = \frac{\sum_{n=M}^{N}\left(\sum_{m=1}^{M} a_{pq}^m x_q[n-m]\right)^2}{\sum_{h=1}^{N_s}\sum_{n=M}^{N}\left(\sum_{m=1}^{M} a_{ph}^m x_h[n-m]\right)^2 + \sum_{n=M}^{N}\epsilon_p^2[n]}, \tag{2.46}$$

where $\mathrm{NC}_{x_q\to x_p}$ is the NC value of x_q into x_p, N is the number of observed time samples of x_p and x_h, M is the model order and N_s is the number of signals compared. When comparing two signals,

the equation reduces to

$$\mathrm{NC}_{x_2 \to x_1} = \frac{\sum_{n=M}^{N} \left(\sum_{m=1}^{M} a_{12}^{i} x_2[n-m] \right)^2}{\sum_{n=M}^{N} \left[\left(\sum_{m=1}^{M} a_{12}^{i} x_2[n-m] \right)^2 + \left(\sum_{m=1}^{M} a_{11}^{i} x_1[n-m] \right)^2 + \epsilon_1^2[n] \right]}. \tag{2.47}$$

2.5.6 Spectral new causality

One characteristic shared by many causality analysis tools is the ability to spectrally decompose the measure to analyze particular frequency bands. The spectral extension of new causality, henceforth referred as ***Spectral New Causality*** (SNC),[7] proceeds rather intuitively from the seminal definition. First, the contributions are defined in the frequency domain

$$C_{pq}(f) = \mathcal{F}\{c_{pq}[n]\} = \sum_{n=m}^{N} c_{pq}[n] e^{-j2\pi f n} \tag{2.48}$$

where $\mathcal{F}$ is the DTFT operator, which is shown on the right hand side. The SNC is then defined as

$$\mathrm{SNC}_{x_q \to x_p} = \frac{|C_{pq}(f)|^2}{\sum_{h=1}^{N_s} \int_{-0.5}^{0.5} |C_{ph}(f)|^2 df + (N-m)\sigma_{\epsilon_p}^2}, \tag{2.49}$$

where $\sigma_{\epsilon_p}^2$ is the sample variance of ϵ_p. Note also that the denominator has been modified for consistency, but it can be shown using Parseval's theorem that the value of the denominator is equivalent to the denominator in Eq. (2.46).

In [95], SNC is defined using the power spectrum of the regressors signals, but the definition using contributions is equivalent and greatly simplifies the derivations in Ch. 4. Also note that in [95], the integrals in the denominator are erroneously omitted. One characteristic shared between SGC and SNC is that the integral of $\mathrm{SNC}_{x_q \to x_p}(f)$ over one period of the DTFT (*e.g.*. from -0.5 to 0.5) yield the GC and NC values respectively.

The expression for SNC is conceptually similar to RPC [3]. The difference lies in that RPC uses the power contribution of the innovation sequence of a signal (ϵ_q) instead of the signals (x_q). One

[7] The spectral extension is called "new spectral causality" in [95], which is confusing, as it is the spectral extension to NC, rather than a new definition of spectral causality (which does not exist).

advantage of RPC is that the denominator is model invariant, whereas in NC the squared sum of the elements in the denominator depend on the model parameter estimates. This occurs because ϵ_p and ϵ_q are assumed to be mutually uncorrelated for all $p \neq q$, whereas x_p and x_q are (in general) correlated. This can lead to the presence of bias in the NC estimates (further explored in Sec. 3.4).

CHAPTER 3

A CRITICAL ANALYSIS OF NEW CAUSALITY

3.1 Overview

In the causality analysis literature, the distinctions among systems, observation models, and estimated models is often blurred. Models are often taken at face value without further discussion on the validity of the model, order and parameter estimates. In this chapter, some of the observation models used in NC literature [94, 95, 100] are discussed. Then, two case studies are done in order to evaluate the robustness of NC and GC to model order and parameter estimation errors. Finally, four scenarios for bias in NC estimation are explored.

From the perspective of the equivalence of GC to TE (measuring transferred information), GC will not measure causal contributions from signals that follow predictable patterns (*e.g.*, slow changing signals, periodic or quasiperiodic signals). While it is true that the GC values estimated using data from some of these observation models may defy intuition on causal strength, signals with high temporal correlation will also require a large number of epochs to produce accurate parameter estimates. Sec. 3.2 discusses the challenge some of the models in NC literature pose to parameter estimation and also the plausibility of some of models.

Some of the observation models used in the literature to showcase the advantages of NC over GC are severely ill-posed. Although it has been shown that NC can more proportionally represent the causal mechanisms than GC [95, 100], the NC values can only improve upon the inference from GC values if the the estimated models correctly mimic the internal dynamics of the observation models. In Sec. 3.3, particular examples are shown of how NC estimates are susceptible to errors in the parameter estimation. In summary, the NC value is as good (or useful) as the model used. On the other hand, GC is generally more robust to parameter estimation errors.

During the investigation of the robustness of NC estimates to model estimation errors, bias was observed in the estimates. This led to the study reported in Sec. 3.4, in which, a mathematical

approach is used to predict likely the sources of bias in NC estimates.

A significant portion of this chapter is quoted directly from the author's work in [147, 148] with a few modifications for improved flow and clarity.

3.2 Problematic aspects of models in NC literature

With the assessment of causality strength in mind, several observation models previously used in comparisons between GC and NC will be re-examined.

3.2.1 Model 1

A principal example observation model studied by Hu *et al.* [95, Eq. (14)] is re-examined. The observation model is compared to a second observation model [95, Eq. (15)], to argue that GC does not reflect the "real strength of causality," the observation models share the same GC value, in spite of their differences. This model is ill-posed in a way that produces relatively small GC estimates. However, the ill-posedness also presents a challenge for NC, as NC depends on the parameter estimates (further discussion on the effect of parameter estimate errors on NC is given in Sec. 3.3). The observation model from [95, Eq. (14)] is expressed as

$$\begin{aligned} x_1[n] &= 0.8x_1[n-1] - 0.8x_2[n-1] + \eta_1[n], \\ x_2[n] &= \qquad\qquad + 0.8x_2[n-1] + \eta_2[n], \end{aligned} \tag{3.1}$$

in which η_1 and η_2 are white noise processes of variances 0.005 and unity, respectively. It is noteworthy that $\sigma^2_{\eta_2} = 200\sigma^2_{\eta_1}$. In [95], it is claimed that the GC value does not reflect the apparent real causal interaction between x_1 and x_2. Although the low variance of η_1 of [Eq. (3.1)] aids in the estimation of the parameters associated with $x_1[n]$, it also can cause the covariance matrix of the regressors to be ill-conditioned. Because of the small $\sigma^2_{\eta_1}$, for any $\ell \in \mathbb{Z}$, one can write

$$x_1[n-\ell] \approx 0.8x_1[n-\ell-1] - 0.8x_2[n-\ell-1], \tag{3.2}$$

so regressors $x_1[n-\ell]$, $x_1[n-\ell-1]$ and $x_2[n-\ell-1]$ are approximately linearly dependent. The linear dependance can also be characterized as a null space in the regressor matrix, in which

variations in the parameters have little effect on the residual error. When combined with the relatively large variance found in x_2, one can write

$$\begin{aligned} x_2[n] \approx 0.8x_2[n-1] + \eta_2[n] + \cdots + \\ \sum_{\ell=1}^{M-1} \beta_\ell \Big[x_1[n-\ell] - 0.8x_1[n-\ell-1] + 0.8x_2[n-\ell-1] \Big], \end{aligned} \tag{3.3}$$

in which the β_ℓ are scalars that represent errors in the estimated parameters in the direction given by the parameters in the brackets. A large variance on the parameter estimates is expected in light of Eq. (2.17), because the covariance matrix is ill-conditioned.

Because Eq. (3.2) contains both x_1 and x_2 terms and the way NC is computed, the estimates of $\mathrm{NC}_{1\to2}$ and $\mathrm{NC}_{2\to2}$ will be biased towards 0.5, which is particularly problematic for $\mathrm{NC}_{1\to2}$, since ideally $\mathrm{NC}_{1\to2} = 0$. A full treatment for the presence of bias in ill-posed problems is given Sec. 3.4.3.

3.2.2 Model 2

The observation model studied by Hu *et al.* in [95, Eq. (15)] is used in conjunction with the observation model in Eq. (3.1) to compare GC and NC. This model is ill-posed as well and has an unrealistic structure which also produces small GC values. The observation model is given by

$$\begin{aligned} x_1[n] &= -0.8x_2[n-1] + \eta_1[n], \\ x_2[n] &= +0.8x_2[n-1] + \eta_2[n], \end{aligned} \tag{3.4}$$

where η_1 and η_2 are white noise processes of variances 0.01 and unity, respectively. Note that $x_1[n]$ does not depend on previous samples of itself. Due to the small variance of η_1 relative to η_2, this is also an ill-posed problem, as one can deduce from Eq. (3.4) that

$$x_1[n-\ell] + 0.8x_2[n-\ell-1] \approx 0, \tag{3.5}$$

for any $\ell \in \mathbb{Z}$. Following an argument similar to Eq. (3.3), as $\sigma^2_{\eta_1}$ is much smaller than $\sigma^2_{\eta_2}$, the estimated model is likely to contain contributions from x_1 into x_2, in the form of $x_1[n-\ell]+0.8x_2[n-\ell-1]$, which are absent in the observation model.

3.2.3 Model 3

Hu *et al.* [95, Eq. 24] use the following observation model to argue that GC underrepresents causality strength

$$\begin{aligned} x_1[n] &= -0.99x_2[n-1] + \eta_1[n], \\ x_2[n] &= 0.99x_1[n-1] + 0.1x_2[n-1] + \eta_2[n], \end{aligned} \tag{3.6}$$

where η_1 and η_2 are white noise processes of variances unity and 0.1, respectively.

In this observation model, the asymptotic value for $\mathrm{GC}_{2\to1}$ is 0.093 (the derivation of the expression is given in Sec. A.2.2 of Appendix A), meaning that the power of the residual error of the prediction of $x_1[n]$ is reduced by less than 10% by including previous samples of x_2 relative to using only past samples of x_1. It is claimed in [95] that the GC cannot identify the causal relationship between the two signals, as the theoretical value for $\mathrm{NC}_{2\to1}$ is 0.96, which indicates that current value of x_1 can be almost fully explained by first delayed value of x_2.

The small GC value is a result of the particular conditions in this observation model. The relatively large $\sigma^2_{\eta_1}$ means that a larger portion of the signal cannot be explained by previous values of either x_1 or x_2. Therefore, the theoretical minimum variance of the residual of x_1 is relatively large. Additionally, since $\sigma^2_{\eta_2}$ is relatively small, the previous values of x_2 can be well predicted by previous values of x_1, so the reduction of residual error by considering x_2 is small.

If $\sigma^2_{\eta_1}$ is made equal to $\sigma^2_{\eta_2}$, $\mathrm{GC}_{1\to2} = 0.67$, and $\mathrm{GC}_{2\to1} = 0.70$, meaning that the contributions from x_1 to x_2 is similar, but smaller, than the contribution of x_2 to x_1. The power of residual error is reduced by about half in both cases. The NC values also indicate that the strengths of the contributions are similar to each other with $\mathrm{NC}_{1\to2} = 0.96$ and $\mathrm{NC}_{2\to1} = 0.98$.

3.2.4 Model 4

In Hu *et al.* in [95, Eq. (25)], the following observation model is used to further argue that GC does not represent causality strength. The observation model is given by

$$\begin{aligned} x_1[n] &= -0.99x_2[n-1] + \eta_1[n], \\ x_2[n] &= 0.1x_2[n-1] + \eta_2[n], \end{aligned} \tag{3.7}$$

where η_1 and η_2 are white noise processes of variances unity and 0.1, respectively. The GC value from x_2 into x_1 is 0.092, which is claimed to be too small, given that $x_1[n]$ is clearly caused by $x_2[n-1]$.

However, upon closer inspection, it becomes clear that the contribution of x_2 into x_1 is indeed small. The variances of $x_1[n]$ and $x_2[n]$ are 1.099 and 0.101 respectively. So the contribution of η_1 to x_1 is about 10 times larger than that of x_2. Even under perfect estimation conditions, the residual can only be reduced by approximately 9%, so, although x_2 represents the only measurable contribution to x_1, the contribution is significantly smaller than that of η_1, as GC correctly indicates.

3.2.5 Model 5

Model 5 is presented in the paper by Hu *et. al.* [100, Eq. (25)] to support a claim that GC possesses a "fatal drawback" that makes it unsuitable in some scenarios. This, of course, is only true if a similar observation model is plausible in any practical application, otherwise the analysis should have limited bearing on judging GC. The model is given by

$$\begin{aligned} x_1[n] &= \quad x_2[n-1] + \eta_1[n], \\ x_2[n] &= -0.9x_1[n-1] + \eta_2[n], \end{aligned} \tag{3.8}$$

where η_1 and η_2 are white noise processes of unity variance.

In this model, at every timestep, x_1 and x_2 exchange values with one another. The current value of x_1 depends solely on the delayed sample of x_2 and a white noise process. Similarly, x_2 depends solely on the first delayed sample of x_1 and a white noise process. The variances of x_1 and x_2 are 10.53 and 9.53 respectively, which are significantly larger than that of η_1 and η_2. Therefore, the contribution of x_1 into x_2 and that of x_2 into x_1 are indeed relatively large. The asymptotic values for GC are $\mathrm{GC}_{1\to 2} = 0.26$ and $\mathrm{GC}_{2\to 1} = 0.30$, whereas the theoretically-evaluated NC values are $\mathrm{NC}_{1\to 2} = 0.90$ and $\mathrm{NC}_{2\to 1} = 0.89$. The NC values clearly indicate how strongly x_1 and x_2 are coupled, whereas the GC values are relatively low, representing a potential of reduction of the variance of the error of only 23% and 26% for $\mathrm{GC}_{1\to 2}$ and $\mathrm{GC}_{2\to 1}$ respectively.

However, a bigger question is: "Under what conditions would a similar observation model occur in nature?" The propagation delay between the two signals is *exactly* one time sample, which can only be achieved if the sampling rate is *designed* this way or by faulty delay embedding. This means the observation model would not be so cleanly representable if the sampling rate were even slightly different. Additionally, (according to the model equations) the signals do not depend directly on previous samples of themselves, however, assuming that the continuous-time counterparts of x_1 is differentiable, we have

$$x_1(t + \Delta t) \approx x_1(t) + \Delta t \cdot \frac{dx_1(t)}{dt}, \tag{3.9}$$

for small enough Δt, where $x_1(t)$ is the continuous time signal from which $x_1[n]$ is sampled (*i.e.*, $x_1(nT_s) = x_1[n]$, where T_s is the sampling period and the sampling rate $f_s = 1/T_s$). Consequently, for small enough T_s,

$$x_1[n + 1] \approx x[n] + T_s \cdot \left.\frac{dX_1(t)}{dt}\right|_{t=nT_s}, \tag{3.10}$$

where d/dt denotes the derivative in time. For any sufficiently high sampling rate, the signals *should* be at least correlated to previous samples of themselves. A similar analysis is done in [79], but in the context of GC. Another example of insufficient sampling rate is given in Sec. 3.2.7.

The observation model in Eq. (3.8) can be written as

$$\begin{aligned} x_1[n] &= -0.9x_1[n-2] + \eta_3[n], \\ x_2[n] &= -0.9x_2[n-2] + \eta_4[n], \end{aligned} \tag{3.11}$$

where the residual errors $\eta_3[n] = \eta_1[n] + \eta_2[n-1]$, and $\eta_4[n] = 0.9\eta_1[n-1] + \eta_2[n]$ are independent white Gaussian processes of variances 2 and 1.81 respectively.

The small difference in variance of the residuals explains the low GC values. The model given in Eq. (3.11) might have slightly larger power in the residual error, but it also does not require inter-channel contributions. So the model given in Eq. (3.11) is arguably simpler than Eq. (3.8), with a minimal increase of predictive error. Additionally, notice that $x_1[n]$ and $x_2[n]$ are uncorrelated, further strengthening the case for the model given in Eq. (3.11).

The claim that GC incorrectly represents the causal relationship requires knowledge that the model of Eq. (3.8) correctly models signals x_1 and x_2. While this can be argued in a computational simulation, such a strong claim cannot be made about a complex problem where the underlying mechanism cannot be easily explained. Thus, while an interesting mental exercise concerning GC, a case cannot be made that the occurrence of such cases is significant enough to warrant the term "fatal flaw."

The choice of sampling frequency is also important for modeling and causality inference in real applications. A discussion of the relationship of regression and sampling rates is given in [23] in the context of nonlinear models. Existing literature of the effects of insufficient sampling in GC estimation in the context of econometrics is found in [136] and in the context of neurophysiological processes in [18].

3.2.6 Model 6

In [94], Hu *et al.* provide two example observation models in which GC values are zero, even though there are clearly causal relationships between the two signals. These example highlight that GC measures transferred information, rather than "causal" influence. In this case, x_2 is a periodic signal, and therefore no new entropy (information) is added to x_2 beyond the first period. The first example model is found in [94, Eq. (10)],

$$\begin{aligned} x_1[n] &= -0.99x_2[n-1] + \eta_1[n], \\ x_2[n] &= -x_2[n-2], \end{aligned} \tag{3.12}$$

where η_1 is a white noise process of unity variance. Note that this model does not contain a η_2 term, effectively making $\sigma_{\eta_2} = 0$.

It can be shown that x_2 can be expressed as a periodic signal with a period of exactly four samples, repeating $x_2[0]$, $x_2[1]$, $-x_2[0]$, and $-x_2[1]$ indefinitely. The values of these samples depend on the initial conditions of x_2. Similarly to Model 5, the observation model seems to be sampled in a way that synchronizes with x_2, such that the period of x_2 is *exactly* four samples. Due to the

lack of external driving forces, x_2 is stable, however, it is noteworthy that the observation model contains a pole on the unit circle.

Since x_2 is not stochastic, it can be estimated using past values of x_1 as

$$\hat{x}_2[n] \approx -\frac{\sum_{\ell=0}^{N_\ell} x_1[n-3-4\ell] - \sum_{k=0}^{N_k} x_1[n-1-4k]}{0.99(N_\ell + N_k)}, \tag{3.13}$$

for any $N_\ell > 0$ and $N_k > 0$. For large enough $N_\ell + N_k$ (*e.g.* $N_\ell + N_k \gg \sqrt{\sigma_{x_2}^2/\sigma_{\eta_1}^2}$), past values of x_1 can predict the value x_2, so that the first line of Eq. (3.12) can be rewritten as

$$x_1[n] = \lim_{\substack{N_\ell \to \infty \\ N_k \to \infty}} \frac{\sum_{\ell=0}^{N_\ell} x_1[n-4-4\ell] - \sum_{k=0}^{N_k} x_1[n-2-4k]}{N_\ell + N_k} + \eta_1[n]. \tag{3.14}$$

In this case, the GC value tends to zero, as Eq. (3.14) does not contain any x_2 terms and yet has the same residual as Eq. (3.13). However, since AR models must have finite order, the GC is never zero. Often a maximum order is imposed in the regression algorithm to avoid overfitting, which would bound GC away from zero.

However, because x_2 is deterministic, it cannot be discerned whether $x_2[n-1]$ truly causes $x_1[n]$. Since Eq. (3.12) can be rewritten as

$$\begin{aligned} x_1[n] &= -0.99x_2[n-1-4\ell_1] + \eta_1[n], \\ x_2[n] &= -x_2[n-2-4\ell_2], \end{aligned} \tag{3.15}$$

for any $\ell_1, \ell_2 \in \mathbb{Z}^+$. Thus, it is impossible to discern $x_2[n-1]$ from any $x_2[n-1-4\ell]$ where[1] $\ell \in \mathbb{Z}^+$. In terms of parameter estimation, this ambiguity causes the regressor matrix to become singular. In this case, additional assumptions (*e.g.* sparsity in parameters, bias towards smaller weights or maximum allowable model order) are necessary to estimate the model parameters correctly. This is especially true if it is desired to also concurrently estimate the propagation delay between x_2 and its effect on x_1.

[1] in fact any convex combination of $x_2[n-1-4\ell]$ terms would be indistinguishable.

3.2.7 Model 7

The second example given by Hu *et al.* in [94, Eq. (13)] shows another instance in which GC is allegedly zero:

$$\begin{aligned} x_1[n] &= -0.99x_2[n-1] + \eta_1[n] \\ x_2[n] &= \eta_1[n], \end{aligned} \tag{3.16}$$

where η_1 is a white noise process of unity variance. Note that the equations for both x_1 and x_2 have $\eta_1[n]$ instead of separate $\eta_1[n]$ and $\eta_2[n]$.

Although $x_2[n]$ is not linearly predictable by any strictly causal model, $x_2[n-1]$ is predictable given past samples of x_1. Hu *et al.* state that for any realization of Eq. (3.16), it is possible to rewrite the equation for $x_1[n]$ as

$$x_1[n] = \lim_{M\to\infty} \sum_{j=1}^{M} a_j x_1[n-j] + \eta_1[n]. \tag{3.17}$$

for some $\{a_j\}_{j=1}^{\infty}$.However, this is only true as $M \longrightarrow \infty$. For $M = 1$, the GC value from x_2 into x_1 is 0.4, decreasing monotonically as M increases. Although this value is arguably low given the mechanism in Eq. (3.16), the NC value of x_2 into x_1 is = 0.5, which also underrepresents the causal relationship.

The authors suggest in [94] that there is a instantaneous causality relationship from x_2 into x_1. To illustrate this, the first line of Eq. (3.16) can be rewritten as

$$x_1[n] = x_2[n] - 0.99x_2[n-1]. \tag{3.18}$$

In this case, the $\mathrm{NC}_{2\to1} = 1$, which implies that x_1 can be fully explained by x_2. In this case, $\mathrm{GC}_{2\to1} \longrightarrow \infty$, which also implies that x_1 can be fully explained by x_2.

While η_1 and η_2 are assumed white in AR models, x_1 and x_2 are not. The whiteness of the residuals implies that each new samples of η_1 and η_2 provide innovation to the observation model that is independent of any previous sample. If that were not true, previous samples of η_1 and η_2 could be used to predict the current values of η_1 and η_2, and, in turn, the current values of x_1 and x_2. Since most regression techniques aim at minimizing the residual error, η is usually to be assumed white. However, in this model, x_2 is white. The implication is that x_2 changes

unpredictably and that no previous values of x_2 can be used to predict its current value. This seems to indicate that the system is being sampled insufficiently and that we cannot determine whether x_2 is being aliased.

3.2.8 Discussion of models 1-7

The basis of AR modeling is that previous samples of a signal provide information about the expected current sample. When the signal changes slowly, the previous sample often provides a good estimate of the current sample. By considering two samples, one can estimate the derivative of the signal and use it to improve the estimate. Assuming no overfitting occurs, the inclusion of more regressors will further improve the estimate. The same argument can be expressed in the spectral domain. The estimation filter attempts to match the spectrum of the signal, where filters with larger orders can better match the desired spectrum.

In Models 2 through 7 (Sec. 3.2.1 through Sec. 3.2.7), at least one of the regressands does not depend on previous samples of itself. In a physical system, that would imply that either the quantity being estimated has no inertia (not continuous) or that the system is being sampled too slowly. However, if the signals are assumed to be continuous in time and sampled above the Nyquist frequency, it would be expected that, the difference between the current sample of signal should be constrained by the previous samples. While such systems exist, one can argue that they are degenerate cases of more practical observation models. In particular, neural systems have been shown to be strongly dependent on internal states [68].

While some models shown in [94, 95, 100] highlight alleged drawbacks of GC, they represent only a small restrictive class observation models, which are not representative of the performance of GC in most problems, or may even pose difficulty for NC, as parameter estimation could be adversely affected by ill-posed problems. In comparison, the models used as examples in the MVGC toolbox [16] contain models designed to mimic particular realistic scenarios (*e.g.* 5-node networks, 9-node networks, non-stationary linear models, etc). Although not a comprehensive list, these models provide better means of analyzing and comparing causality tools.

In many of the analyses of experimental data presented in [95, 98, 99], NC outperforms GC in showing the causality mechanism strength, however, some simulations in [94, 95, 100] show extreme cases in which GC will predictably underrepresent the causality mechanism and might be considered degenerate observation models. Although it is important to highlight instances where GC does not perform well, these are far from being "fatal drawbacks."

In [19], Barrett and Barnett assert that the claim of GC not capturing "how strongly one time-series influences another" could be considered "radical." However, it was conceded that "GC is not a perfect measure for all stochastic time series: if the true process is not a straightforward multivariate AR process with white-noise residuals, then it becomes only an approximate measure of causal influence." For different applications, other methods are available such as conditional GC [14, 44], spectral GC [71] and other methods such as partial directed coherence (PDC) [12], relative power contribution (RPC) [3], directed transfer function (DTF) [93], and phase slope index (PSI) [150]. NC is a new addition to that list, which has shown promising results and its strengths and weaknesses will likely be explored in the following years.

When comparing two techniques, it is important that the observation models chosen represent the strengths as well as the limitations of both techniques. A possible remedy to this challenge is to create a set of benchmark problems or datasets that represent a variety realistic scenarios. For instance, the multivariate GC toolbox (MVGC) [16] provides a small set of example models representing realistic scenarios. This set could be expanded to account more scenarios and serve as a benchmark set, which would allow a fairer and comprehensive comparison between causality analysis tools.

3.3 Analysis of NC robustness to parameter errors through case studies

To empirically evaluate the robustness of NC measures to model parameter estimation error and overfitting under model uncertainty, one of the primary example observation model used in [95,

Eq. 14] is re-examined. The observation model is expressed as

$$\begin{aligned} x_1[n] &= 0.8x_1[n-1] \quad - 0.8x_2[n-1] + \eta_1[n] \\ x_2[n] &= \qquad\qquad\qquad + 0.8x_2[n-1] + \eta_2[n] \end{aligned} \tag{3.19}$$

where η_1 and η_2 are white noise processes of variances 0.005 and unity, respectively.

Three scenarios are observed in the present study, each using a different values for M of the regressors. To study the statistical properties of the NC and GC estimates under the effects of overfitting and regularization, 65536 simulations were run in each scenario. In each simulation, the number of time samples $N = 256$, and model parameters were estimated using LASSO regression. A wide range of regularization parameters was used ($\lambda \in [10^{-7}, 10^{1}]$). In the following figures, the value of λ is shown in the x-axis and the NC and GC values are shown in the y-axis. For each value of λ, the probability distribution of NC and GC values were estimated from the histogram taken at that λ value and are shown in a color plot, where yellow represents higher probability density and blue represents lower probability density.

Although λ must also be estimated, several techniques exist to find appropriate values. The most common method is using a resampling method, such as bootstrapping and cross-validation [21, 110]. k-fold cross-validation splits the dataset into k subsets, then for each subset, evaluating the prediction error of that subset using the estimated model obtained by using the union of the other $k - 1$ subsets to estimate the parameters. The λ value that produces the smallest average of variance of prediction errors is chosen. Alternatively, thresholding can be used to evaluate parameter significance, with the number of significant parameters used in conjunction with a method such as AIC or BIC to compare models [124]. In this work, the wide range of λ values was chosen to showcase how NC and GC estimation react to differing levels of regularization.

For the observation model in Eq. (3.19), the theoretically evaluated NC values [see Eq. (A.22) in Appendix A for the exact expression] are shown in Table 3.1. These NC measures indicate that x_1 strongly dictates its own behavior, that x_1 and x_2 together can very accurately predict x_1 (in other words, $\mathrm{NC}_{\eta_1 \to x_1}$ is small), x_1 does not contribute to x_2 [as Eq. (3.19) indicates] and x_2 contributes to its own behavior strongly, but a significant portion cannot be explained by either x_1 or x_2 (in

other words, $\mathrm{NC}_{\eta_2 \to x_2}$ is relatively large). Although there are other factors at play, the relatively smaller value of $\mathrm{NC}_{\eta_1 \to x_1}$ compared to $\mathrm{NC}_{\eta_2 \to x_2}$ is expected as the variance of η_1 is much smaller than the variance of η_2. The theoretical values for the GC measures are also shown in Table 3.1, where $\mathrm{GC}_{2 \to 1}$ shows a range, due to GC being model order dependent. These values indicate that the presence of x_1 does not improve the prediction of x_2 (x_1 does not cause x_2), but the presence of x_2 reduces the residual prediction error energy from 14 to 140 times (x_2 does cause x_1).

Table 3.1: Theoretically evaluated GC and NC measures for the observation model in Eq. (3.19)

$\mathrm{NC}_{1 \to 1}$	0.89
$\mathrm{NC}_{2 \to 1}$	0.11
$\mathrm{NC}_{1 \to 2}$	0
$\mathrm{NC}_{2 \to 2}$	0.64
$\mathrm{GC}_{2 \to 1}$	[4.86,5.38]
$\mathrm{GC}_{1 \to 2}$	0

Of particular interest in these tests is $\mathrm{NC}_{1 \to 2}$, which should indicate that x_2 is not caused by x_1. Similarly, $\mathrm{GC}_{1 \to 2}$ is expected to be small, indicating that x_1 does not cause x_2. Ideally, $\mathrm{NC}_{1 \to 2} = 0$ and $\mathrm{GC}_{1 \to 2} = 0$, but, in practice the values will be greater than zero as a consequence of the statistical variance of the estimator in conjunction with the data properties.[2] Nevertheless, significance thresholds T_{NC} and T_{GC} may be set such that if $\mathrm{NC}_{1 \to 2} < T_{\mathrm{NC}}$ or $\mathrm{NC}_{1 \to 2} < T_{\mathrm{GC}}$, x_1 is assumed not to cause x_2. The significance thresholds can be obtained through a number of techniques, such as block resampling [159], stationary bootstrap [160] or trial shuffling [37, 196].

In the studies with $M = 1$ (exact model order), NC and GC using LASSO regression produce good results, despite a small value of N ($N = 256$). In Fig. 3.1, the distribution of estimated $\mathrm{NC}_{1 \to 2}$ values is shown for different regularization parameters. Even for λ as low as 10^{-7}, most simulations yield NC measures close to the theoretical NC values. Fig. 3.2 shows the distribution of $\mathrm{GC}_{1 \to 2}$ values under the same circumstances. Similarly to the $\mathrm{NC}_{1 \to 2}$, little variation is seen over the range of λ values.

To evaluate the effects of order overestimation, simulations for the observation model of

[2] If the estimated $a_{21}^i = 0$ for $i = 1, 2, \ldots, M$, then $\mathrm{NC}_{1 \to 2} = 0$ and $\mathrm{GC}_{1 \to 2} = 0$, but at least in terms of LSE, the probability of exact equivalence is nil.

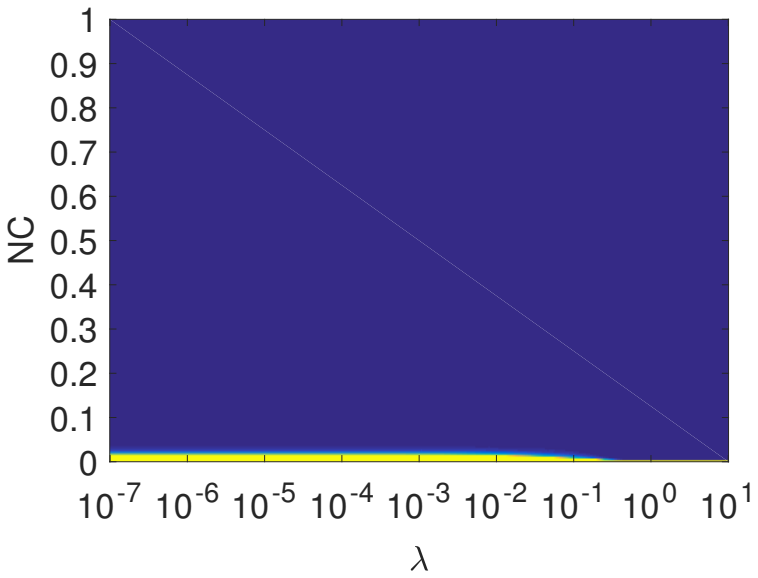


Figure 3.1: Distribution of the $NC_{1\to 2}$ estimates as a function of λ for $M = 1$.

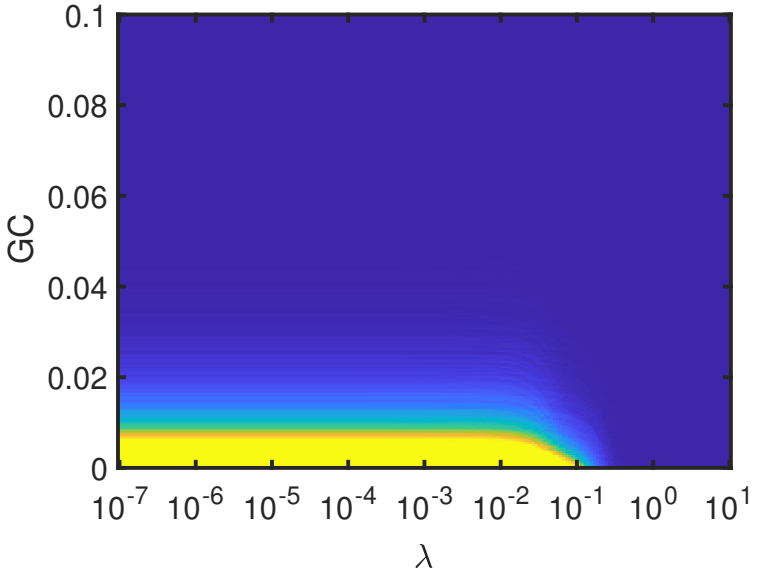


Figure 3.2: Distribution of the $GC_{1\to 2}$ estimates as a function of λ for $M = 1$.

Eq. (3.19) were run under the same conditions except for a larger M. With $M = 2$, until enough regularization is applied ($\lambda \geq 10^{-3}$), the simulation does not yield satisfactory results for NC, as shown in Fig. 3.4, where there is a large spread of values for $NC_{1\to 2}$. The $GC_{1\to 2}$ estimate for $M = 2$ (shown in Fig. 3.3) has higher variance than the estimate for $M = 1$, but is robust to the overfitting and regularization, having a consistent value for lowest values of λ tested and only changing when excessive regularization is applied ($\lambda > 10^{-2}$), in other words, when the parameter

estimates are substantially biased towards zero.

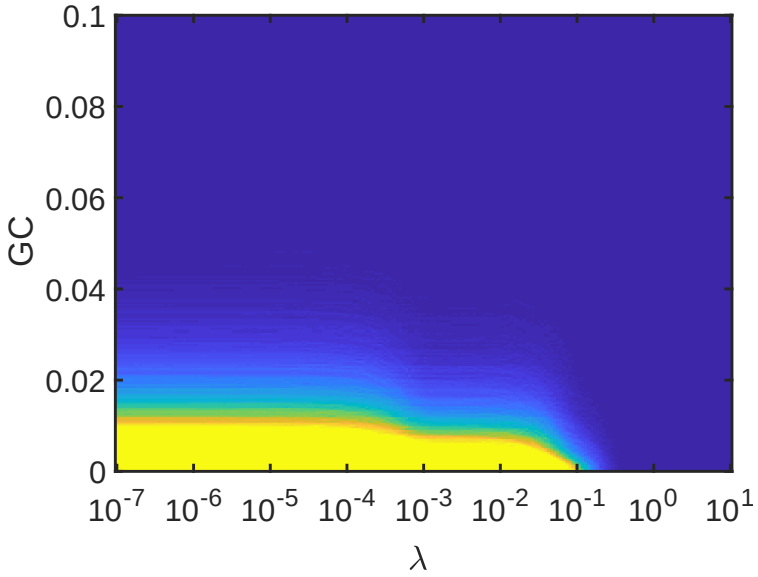


Figure 3.3: Distribution of the $GC_{1\to 2}$ estimates as a function of λ for $M = 2$.

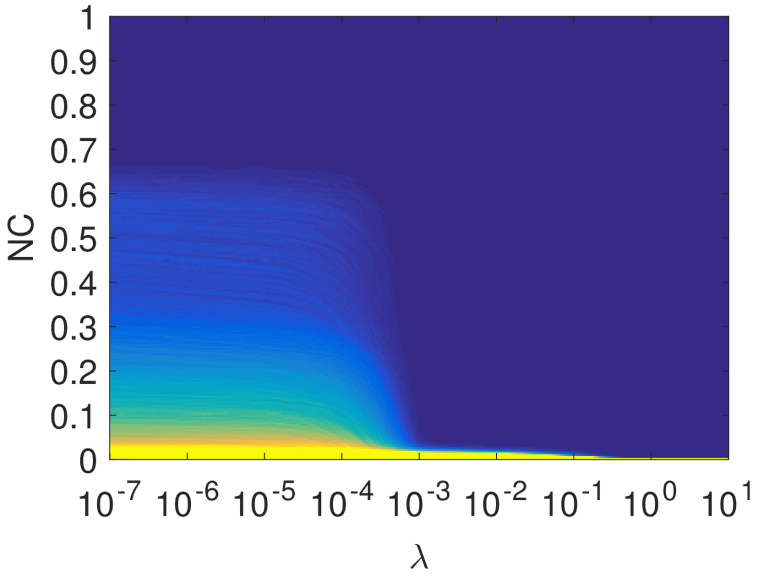


Figure 3.4: Distribution of the $NC_{1\to 2}$ estimates as a function of λ for $M = 2$.

Because of the large correlations between x_1, x_2 and their delayed samples, the covariance matrix of the regressors is ill-conditioned. When M is increased to five, the estimate of $NC_{1\to 2}$ not only has large variance, but also tends to bifurcate and cluster around two values (approximately 0.35 and 0.55) when $\lambda < 10^{-3}$. Neither of these is the theoretically correct value, as shown in

Fig. 3.5. This tendency is strengthened as the mismatch between model order and regressor order increases, as shown in Fig. 3.6, where $M = 6$. Meanwhile, the GC estimates remain close to what was observed for $M = 2$, as shown in Fig. 3.7 for $M = 5$ and Fig. 3.8 for $M = 6$. While the extra regressors cause the GC estimates not to have any probability mass at GC $= 0$ for $\lambda < 10^{-3}$, the estimates remain close to zero, even for small amounts of regularization. This suggests that, in this test, GC is more robust to overfitting and model order overestimation, even when the parameters cannot be estimated accurately.

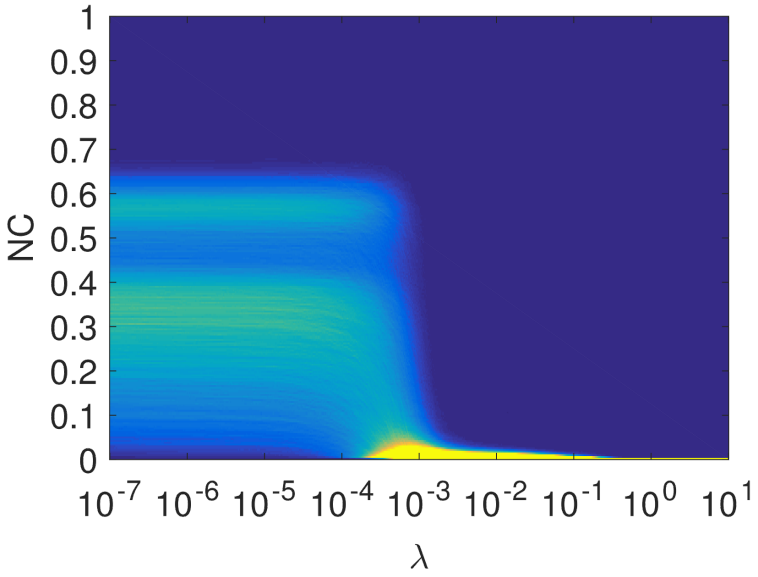


Figure 3.5: Distribution of the $NC_{1\to 2}$ estimates as a function of λ for $M = 5$.

Although it is possible to mitigate the need for regularization with larger sample sizes, as shown in Fig. 3.9, in which $N = 1024$ (instead of $N = 256$ in previous figures), it is sometimes necessary to infer the change in causality strength over short time intervals, so that the model parameters must be estimated over data blocks spanning the same short time intervals. Blindly increasing the sampling rate is often not advisable as it would adversely interfere with the models and conditioning of the regressor matrix [23]. Therefore, special care must be taken when estimating the model order and its parameters to avoid misleading NC values.

In order to study the NC performance in models with longer propagation delays between

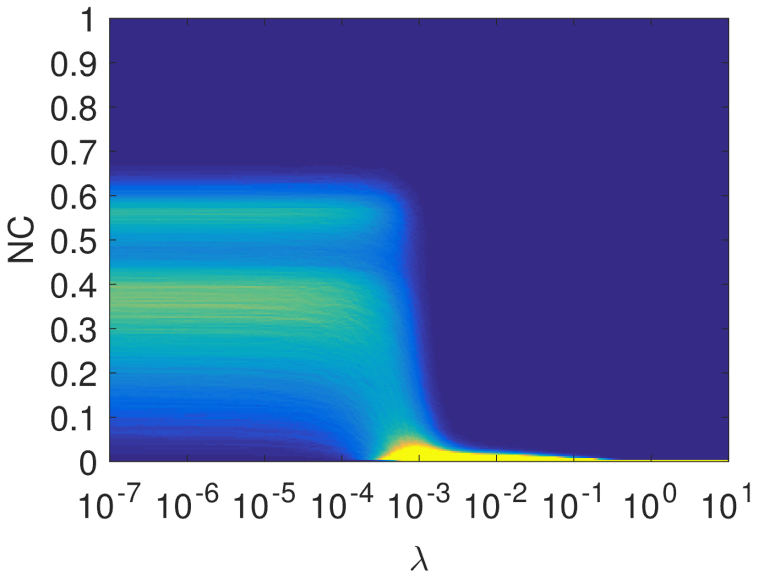


Figure 3.6: Distribution of the $NC_{1\to 2}$ estimates as a function of λ for $M = 6$.

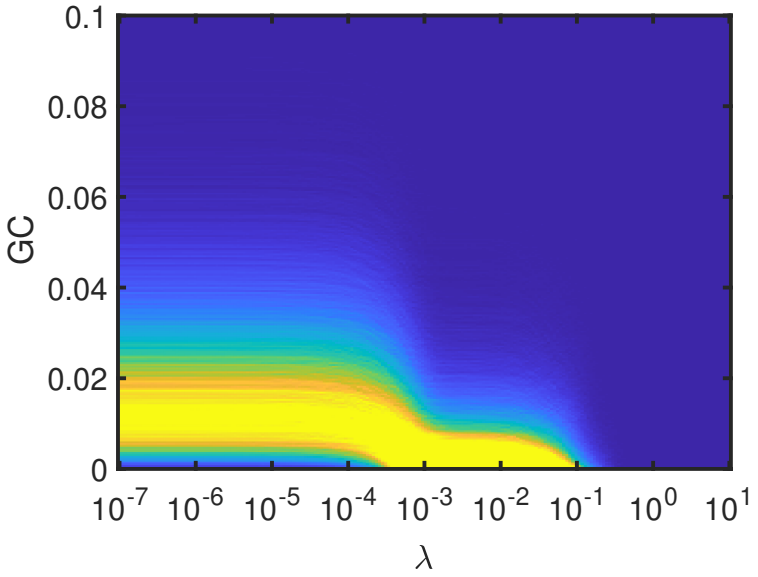


Figure 3.7: Distribution of the $GC_{1\to 2}$ estimates as a function of λ for $M = 5$.

channels, we propose a similar model here, with

$$\begin{aligned} x_1[n] &= \quad 0.6x_1[n-1] - 0.3x_2[n-4] + \eta_1[n] \\ x_2[n] &= -0.5x_1[n-1] + 0.6x_2[n-1] + \eta_2[n] \end{aligned} \tag{3.20}$$

where η_1 and η_2 are both white Gaussian noise processes of zero mean and variances unity and 0.005, respectively. Note that the contribution of x_2 into x_1 has a delay of four samples. In this case, the order of the joint ARX model is four, even though the contribution of x_2 into x_1 can be

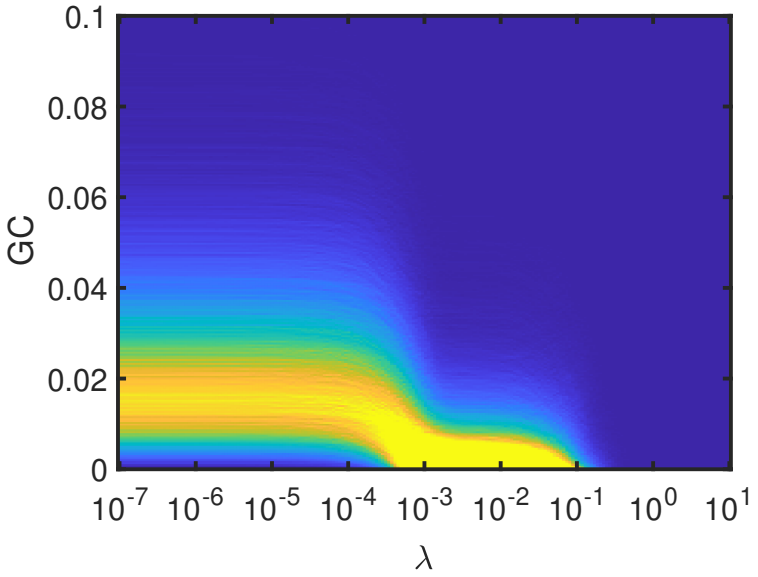


Figure 3.8: Distribution of the $GC_{1\to 2}$ estimates as a function of λ for $M = 6$.

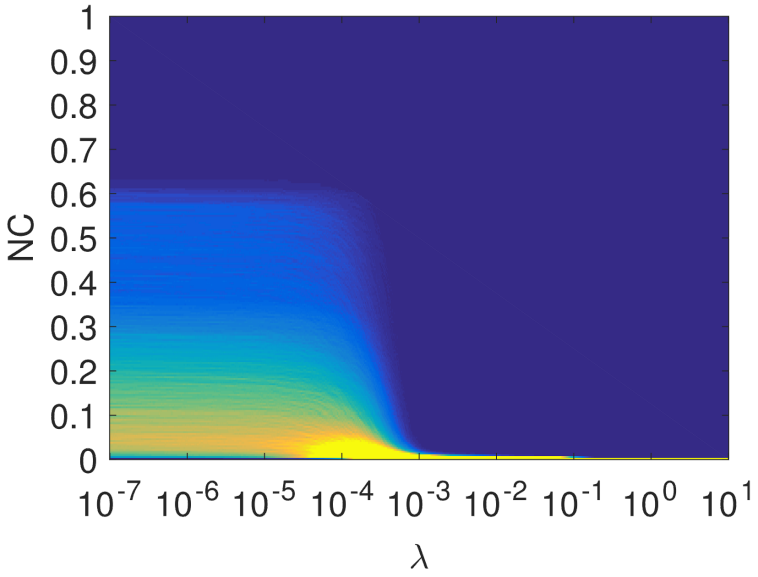


Figure 3.9: Distribution of the $NC_{1\to 2}$ estimates as a function of λ for $M = 5$ and $N = 1024$.

represented with a single regressor ($x_1[n-4]$). When assuming that the system can be modeled as a joint AR model and varying only the order of the model, either a large number of regressors is made available (when $M \geq 4$) or an insufficient number of terms is made available ($M < 4$). Therefore problems with larger propagation delays present a challenge for parameter estimation.

The theoretical NC values for Eq. (3.20) are shown in Table 3.2, which can be obtained by evaluating Eq. (A.23) with the expected values for the squared terms. This indicates that for x_1 the

contribution of previous values of x_1 is roughly twice that of the contributions of x_2. Also, the past values of x_1 and x_2 can be used to almost fully predict the current value of x_1. Additionally, the power of the contribution of previous values of x_2 is about three times that of the power of x_1 to x_2, but a portion of the current value of x_2 cannot be well predicted by past values of x_1 and x_2 (since $NC_{1\to2}$+$NC_{2\to2}$ = 0.77).

Table 3.2: Theoretically evaluated NC measures for the observation model in Eq. (3.20)

$NC_{1\to1}$	0.57
$NC_{2\to1}$	0.20
$NC_{1\to2}$	0.33
$NC_{2\to2}$	0.67

When the exact model order ($M = 4$) is used, results show that regularization is necessary when $N = 256$. Fig. 3.10 shows the $NC_{1\to1}$ values obtained using LASSO regression for different values of the regularization factor (λ). The behavior observed in Fig. 3.5 is present, despite the use of the exact model order. When enough regularization is applied, the NC values approach the theoretical values. The results are biased towards zero, partially due to tendency of the regularization to bias the value of the parameters towards zero, while also increasing the residual error. Additionally, a small bias is observed due to the nonlinear dependence on parameters in the definition of NC. Further analysis of biases in NC estimates is found in Sec. 3.4.

If the model order is overestimated at 6, the NC values will exhibit more variance, even when enough regularization is applied. Fig. 3.11 shows the probability density function of the NC values. This indicates that LASSO-assisted regression is not able to accurately estimate the model parameters, regardless of the choice of regularization factor.

When underestimating the model order as unity, in other words, attempting to predict $x_1[n]$ and $x_2[n]$ with only $x_1[n-1]$ and $x_2[n-1]$, the results were unexpectedly good, as shown in Fig. 3.12. When compared to Fig. 3.11, the results do not bifurcate and have lower variance. Although the average NC value estimate is lower than the theoretical value, they are comparable to the results obtained with the exact model order, while accepting a wider range of regularization factors. Part of the improvement comes from the fact that the autocorrelation of the signals is high and that

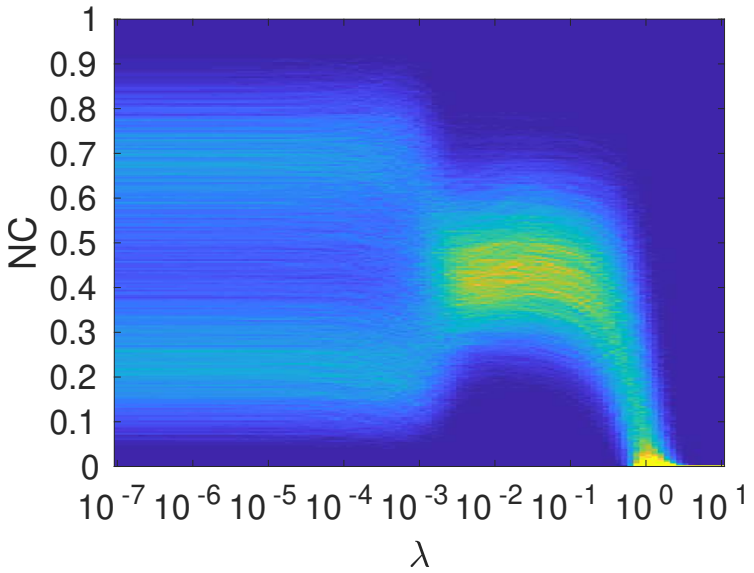


Figure 3.10: Distribution of the $NC_{1\to1}$ estimates as a function of λ for the model shown in Eq. (3.20) and $M = 4$

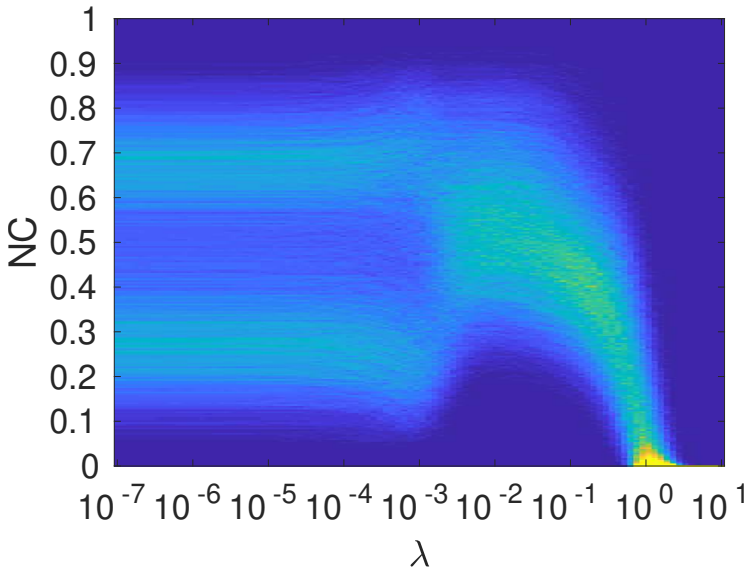


Figure 3.11: Distribution of the $NC_{1\to1}$ estimates as a function of λ for $M = 6$

$M = 1$ improves the posedness of the problem.

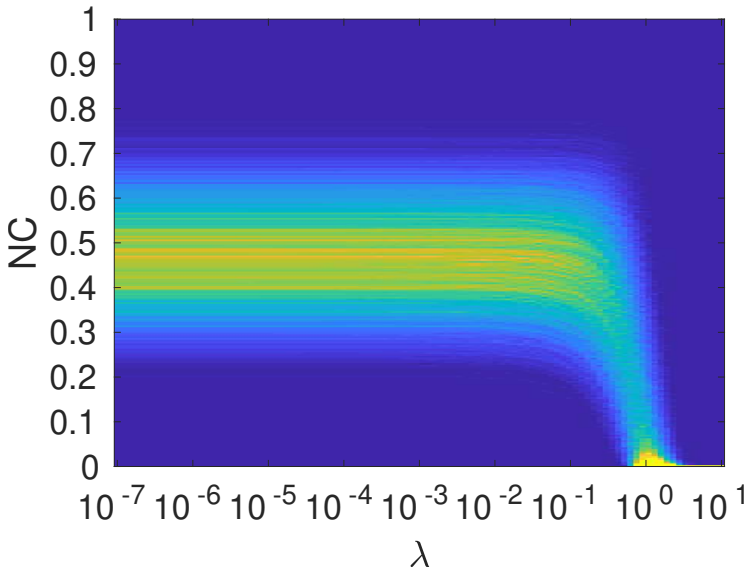


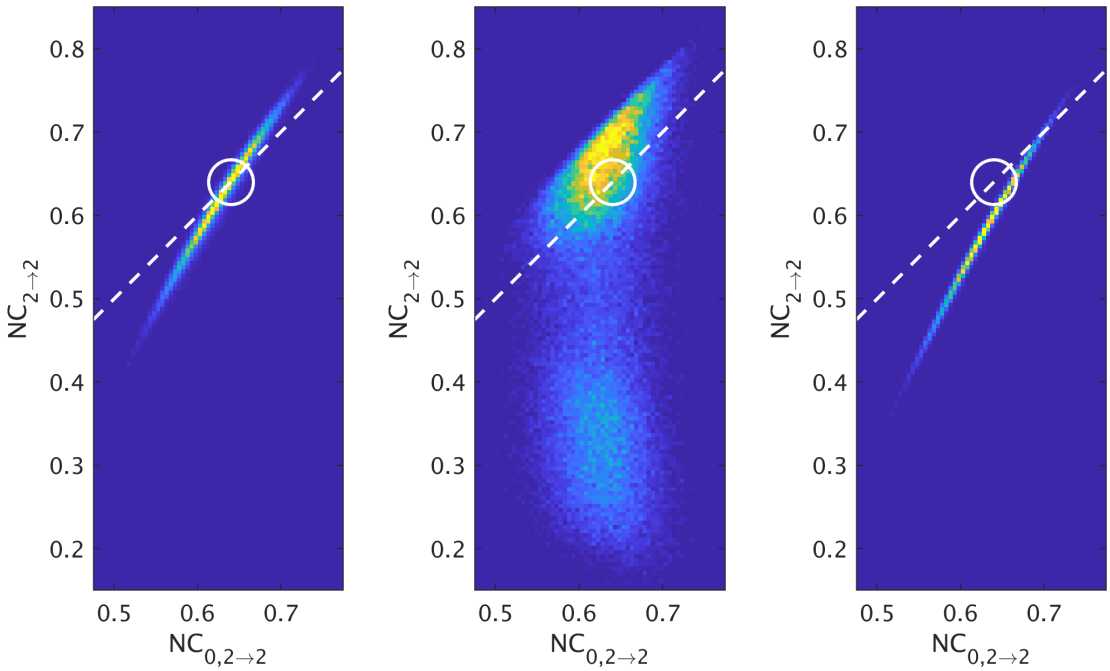
Figure 3.12: Distribution of the $NC_{1\to1}$ estimates as a function of λ for $M = 1$ for the model shown in Eq. (3.20)

3.3.1 Discussion

3.3.2 NC and GC fluctuation

For synthetic models, the NC and GC values can be calculated theoretically using the variances of η_1 and η_2 and the observation model parameters and the MMSE model parameter values for the disjoint model [see Eq. (A.15) in Sec. A.2.2 for the derivation of the disjoin model parameters]. However, even when the model parameters are known, NC and GC values estimated using their respective definitions also depend on the particular samples (practically, only sample variances can be obtained, which are used as an estimate of the variances) that are used to estimate the model parameters. To decouple the variation in the NC values due to parameter estimation errors and the variation due to sample variance, in this subsection, the NC values are calculated twice, once with the parameters obtained from LASSO regression and once with the true model parameters. The NC values obtained with the true model parameters will henceforth be called NC_0. The variation found in NC_0 values indicates how much variation is inherent in the short-term record, while the correlation between the NC_0 values and the NC values serve to assess the effect of parameter estimation on the variation of NC values.

The NC and NC_0 values are expected to be strongly correlated, as this would indicate that the NC value estimates are close to the theoretical values, and that the variation in the NC values originates mostly from sample variation, rather than noise or inaccurate assessment of causality strength. Fig. 3.13 shows the bidimensional histogram data for $NC_{2\to2}$ and $NC_{0,2\to2}$ for the model in Eq. (3.20) and $M = 4$ for differing regularization factors. On the x-axis are the bins for the $NC_{0,2\to2}$ values and in the y-axis are the bins for $NC_{2\to2}$ values. The dashed line region of the histogram that corresponds to $NC = NC_0$ and the circle is drawn centered at the theoretically calculated NC value. The NC values in Fig. 3.13a were obtained with $\lambda = 10^{-2}$, which shows a strong correlation between the NC and NC_0, although not perfect alignment with the dashed line.



(a) $\lambda = 10^{-2}$ (b) $\lambda = 10^{-4}$ (c) $\lambda = 10^{-1}$

Figure 3.13: $NC_{1\to1}$ vs $NC_{0,1\to1}$ histogram plots as a function of λ

If the regularization factor is not sufficiently large, however, NC estimates deviate from NC_0. For small regularization factors, the correlation is much lower between NC and NC_0. Fig. 3.13b shows the $NC_{2\to1}$ for $\lambda = 10^{-4}$, where some of the estimates are correctly located around the dashed line and circle. However, many of the estimates do not correlate well with NC_0, but rather are even weakly negatively correlated with it (the bottom peak).

For large regularization factors, the NC estimates tend to be biased towards zero, even if the correlation is high. A more formal treatment of the bias is given in Sec. 3.4.4. Note that this is expected as regularization introduces bias to the parameter estimator in exchange for lower variance of the estimates. Fig. 3.13c shows the histogram data for $\lambda = 10^{-1}$, which is only one order of magnitude larger than that of Fig. 3.13a, but the results are no longer centered on the dashed line.

3.3.3 Regression conditioning and over-fitting

The model of Eq. (3.19) is introduced by Hu *et al.* in [95], where it is claimed that the GC value does not reflect the real causal influence between x_1 and x_2. In the present work, it is shown that this system also poses problems for NC estimation. Although the low variance of η_1 of Eq. (3.19) aids in the estimation of the parameters associated with $x_1[n]$, it also can cause the regressors to be highly colinear, as one can write

$$x_1[n-l] \approx 0.8x_1[n-l-1] - 0.8x_2[n-l-1] \tag{3.21}$$

so regressors $x_1[n-l]$, $x_1[n-l-1]$ and $x_2[n-l-1]$ are nearly linearly dependent. This can also be characterized as a null space in the regressor matrix [$\boldsymbol{X}[N]$ in Eq. (2.11)], in which variations in the parameters have very little effect on the residual error, creating large variances in the parameter estimation [see Eq. (2.17) for the distribution of the parameters under LSE]. When combined with the relatively large variance found in x_2, one can write

$$\begin{gathered} x_2[n] \approx 0.8x_2[n-1] + \\ \sum_{l=1}^{P-1} \beta_l \Big[x_1[n-l] - 0.8x_1[n-l-1] + 0.8x_2[n-l-1] \Big] \end{gathered} \tag{3.22}$$

in which the β_l are scalars that govern the deviation of the predicted parameters in the direction given by the parameters in the brackets.

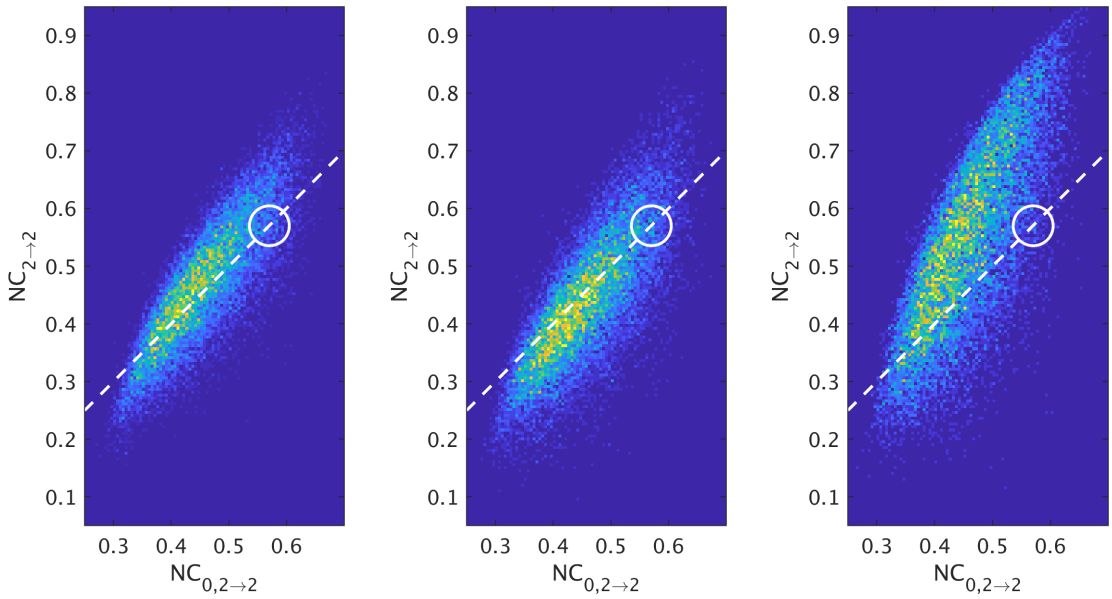
The increased model order also adversely affects the predictive error estimation, thus, GC analysis. Since the model order is also used to estimate the variance of the residual error, GC

values will change as the model order increases. The parameter estimation algorithm attempts to match the spectrum of the regressand, so an increase in model order yields improvements in the prediction using the disjoint model, even when using the joint observation model. The reduction in the residual error of the estimated disjoint model causes the GC value for large model orders to be lower than desired. This was analyzed for a particular model by Zhuo *et al.* in [220], where a series of backward recursive operations was used to expand the the AR model. A similar approach was taken by Grassmann in [79]. A general expansion and discussion is given in Sec. A.2.2. One important discussion presented in [95] is that for overestimated model orders, GC can be invariant to the model parameters,therefore model order estimation is an important aspect when estimating causality strength using GC as well.

While methods of estimating model orders exist, such as using the Akaike Information Criterion (AIC) [5] or Bayesian Information Criterion (BIC) [174], these criteria can only compare the quality of different models, but a different method must be used to generate the models, as exhaustive search of all possible combination of regressors is ordinarily prohibitive. Additionally, there are instances where AIC and BIC perform sub-optimally and misestimate the model order. The topic is still an active area of research, with new criteria still being developed [62].

observation models that contain terms with large delays, such as the observation model described in Eq. (3.20) (which has one element of order four), further complicate the proper regression model order selection. As shown in Fig. 3.11, there are instances in which LASSO regression is unable to accurately estimate the model parameters, regardless of the regularization factor.

Fig. 3.14 shows the histogram plots for three different values of model order (M), 1, 2 and 6. As expected, for $M = 6$, the variance of NC increases relative to $M = 4$. However, $M = 1$ performs comparably to $M = 4$, which is surprising, since $M = 1$ cannot fully model the observation model given in Eq. (3.20). In this particular case, it is preferable to underestimate the model order rather than overestimating it.



(a) $M = 1$ (b) $M = 4$ (c) $M = 6$

Figure 3.14: $NC_{1\to1}$ vs $NC_{0,1\to1}$ histogram plots as a function of λ

3.3.4 Comparing NC and GC

For the two models studied, depending on the estimated model order and regularization, the NC measure does not produce the expected results. However, this is not an issue with the measure itself, but rather with the estimation of the model parameters. To compare the sensitivity to model parameter uncertainty of NC to GC, the GC values were measured for the models described in Eqs. (3.19) and (3.20).

For the model in Eq. (3.19), the theoretical values for the GC measures are $GC_{1\to2} = 0$ and $GC_{2\to1} \in [4.86, 5.38]$ (depending on the model order). These values indicate that the presence of x_1 does not improve the prediction of x_2, inferring that x_1 does not cause x_2, but the presence of x_2 reduces the residual prediction error energy from 14 (for large M) to 140 times (for $M = 1$), inferring x_2 does cause x_1.

In contrast to NC, the GC measure is not as sensitive to errors in the model parameter estimates. $GC_{2\to1}$ remained relatively flat and close to the theoretical value for a wide variety of regularization factors. More importantly, $GC_{1\to2}$ is close to zero, even when very little regularization is applied, showing that there is no causal relationship from x_2 into x_1.

For the observation model of Eq. (3.20), the GC measures are also robust to model uncertainty. Figs. 3.15 and 3.16 shows the GC measures evaluated for $M = 6$. If excessive regularization ($\lambda \gg 10^{-2}$) is applied, the GC measure tends to 0, but for a wide range of values, the GC measures closely approximate the theoretical values. However, $\mathrm{GC}_{1\to 2}$ is very small, even though the contribution of x_1 to x_2 is significant. This is due to the large autocorrelation of x_2 and large variance of η_2, which leads x_1 not to reduce the variance of the residual significantly.

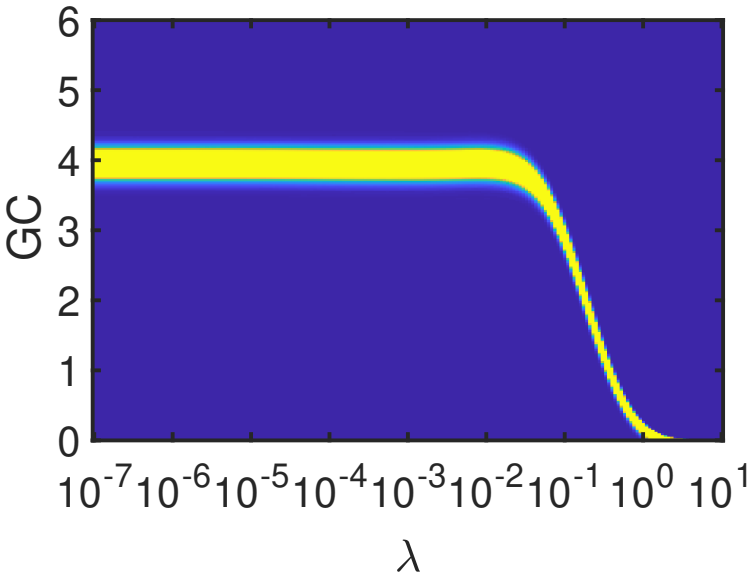


Figure 3.15: Distribution of the $\mathrm{GC}_{2\to 1}$ estimates as a function of λ for the model shown in Eq. (3.20) and $M = 6$

3.4 Bias in NC estimates

Although this work focuses on examining the variance of NC estimates observed two example models, a small bias was observed in the tested models. This subsection contains further investigation of the bias in the NC estimates.

The analysis will be constrained to the bivariate case, where a signal y can be expressed as the weighted sum of two signals x and z and a white noise process η. In order to increase clarity in the notation, this appendix will utilize different notation from the rest of paper. Instead of the $*$ superscript, the observation model parameters will have subscript 0, to avoid confusion between the superscripts and the transpose operator. Instead of x_1 and x_2, signals x and z are used, so that

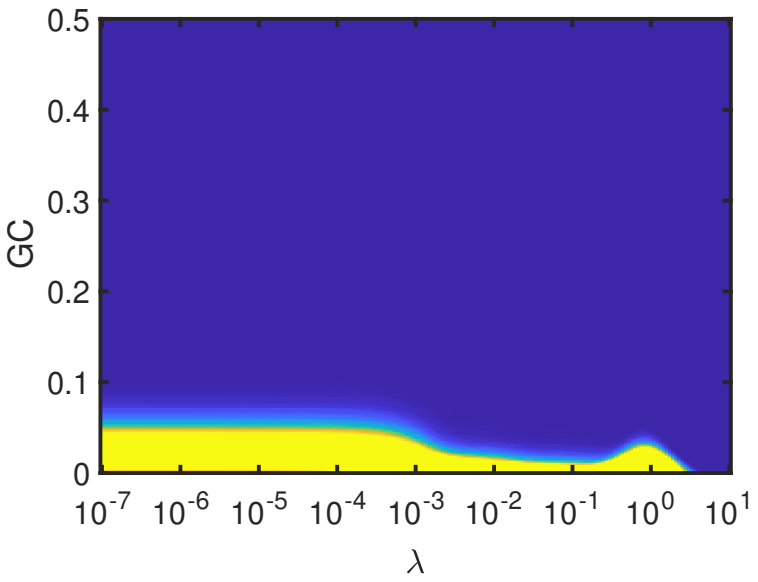


Figure 3.16: Distribution of the $GC_{1\to 2}$ estimates as a function of λ for the model shown in Eq. (3.20) and $M = 6$

no subscripts are needed and to make clear that the 0 subscript refers to the observation model, rather than being an index. The separation into x and z also serves to more clearly denote the difference between the parameter vectors associated with x and z, which will be called $\boldsymbol{a}_0$ and $\boldsymbol{b}_0$ respectively. The observation model, therefore, is written as

$$y[n] = \boldsymbol{a}_0^T \boldsymbol{x}[n] + \boldsymbol{b}_0^T \boldsymbol{z}[n] \ + \eta[n] \tag{3.23}$$

where $y[n]$ is the signal being modeled at time n, $\boldsymbol{x}[n]$ is the $M_a \times 1$ vector of regressors associated with signal x at time n (*i.e.* $\boldsymbol{x}[n] = \{x[n], x[n-1] \dots x[n-M_a+1]\}^T$) and parameterized by the $M_a \times 1$ vector $\boldsymbol{a}_0$, $\boldsymbol{z}[n]$ is the $M_b \times 1$ vector of regressors associated with signal z at time n (*i.e.* $\boldsymbol{z}[n] = \{z[n], z[n-1] \dots z[n-M_b+1]\}^T$) and parameterized by the $M_b \times 1$ vector $\boldsymbol{b}_0$ and η is a white noise process.

For further simplification of the calculations, Eq. (3.23) can be further condensed into a matrix form that contains all N time samples as

$$\boldsymbol{Y} = \boldsymbol{a}_0^T \boldsymbol{X} + \boldsymbol{b}_0^T \boldsymbol{Z} + \boldsymbol{\eta} \tag{3.24}$$

where $\boldsymbol{Y} = [y[n] y[n-1] \cdots y[n-N+1]]^T$ is a $1 \times N$ time-series vector of the regressand signal, $\boldsymbol{X}$ is the $M_a \times N$ regressor matrix where column j is $\boldsymbol{x}[n-j]$, $\boldsymbol{Z}$ is the $M_b \times N$ regressor matrix

where each column j is $\boldsymbol{z}[n-j]$ and $\boldsymbol{\eta} = [\eta[n]\eta[n-1]\cdots\eta[n-N+1]]^T$. Vectors $\boldsymbol{a}_0$ and $\boldsymbol{b}_0$ remain unchanged. Having defined these variables, $\mathrm{NC}_{0,x\to y}$ can be calculated

$$\mathrm{NC}_{0,x\to y} = \frac{||\boldsymbol{a}_0^T\boldsymbol{X}||^2}{||\boldsymbol{a}_0^T\boldsymbol{X}||^2 + ||\boldsymbol{b}_0^T\boldsymbol{Z}||^2 + ||\boldsymbol{\eta}||^2} \tag{3.25}$$

This value is the desired NC value calculated assuming perfect parameter estimates. In these examples, $\boldsymbol{\eta}$ are assumed to be non-zero for at least one element, to avoid degenerate cases where the denominator is zero. For greater clarity, the $x \longrightarrow y$ subscript will be dropped from the following expressions, but for all subsequent analyses, NC should be understood to be $\mathrm{NC}_{x\to y}$. Using an estimated model defined by

$$\boldsymbol{Y}_p = \boldsymbol{a}^T\boldsymbol{X} + \boldsymbol{b}^T\boldsymbol{Z} \tag{3.26}$$

allows the calculation of the residual

$$\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{Y}_p = \Delta\boldsymbol{a}^T\boldsymbol{X} + \Delta\boldsymbol{b}^T\boldsymbol{Z} + \boldsymbol{\eta} \tag{3.27}$$

where $\Delta\boldsymbol{a} = \boldsymbol{a}_0 - \boldsymbol{a}$ and $\Delta\boldsymbol{b} = \boldsymbol{b}_0 - \boldsymbol{b}$. $\boldsymbol{a}_0$ and $\boldsymbol{a}$ are assumed to be of the same size. Whenever the estimated order is not M_a, vectors $\boldsymbol{a}_0$ or $\boldsymbol{a}$ must be zero-padded such that their sizes match. The same applies to $\boldsymbol{b}_0$ and $\boldsymbol{b}$. This is similar to the approach in taken in Eq. (2.10) and is taken without loss of generality. After obtaining the estimated model parameters, the NC estimate is evaluated

$$\begin{aligned}\mathrm{NC} &= \frac{||\boldsymbol{a}^T\boldsymbol{X}||^2}{||\boldsymbol{a}^T\boldsymbol{X}||^2 + ||\boldsymbol{b}^T\boldsymbol{Z}||^2 + ||\Delta\boldsymbol{a}^T\boldsymbol{X} + \Delta\boldsymbol{b}^T\boldsymbol{Z} + \boldsymbol{\eta}||^2} \\ &= \frac{||\boldsymbol{a}_0^T\boldsymbol{X}||^2 - 2\boldsymbol{a}_0^T\boldsymbol{X}\boldsymbol{X}^T\Delta\boldsymbol{a} + ||\Delta\boldsymbol{a}^T\boldsymbol{X}||^2}{\begin{aligned}&||\boldsymbol{a}_0^T\boldsymbol{X}||^2 - 2\boldsymbol{a}_0^T\boldsymbol{X}\boldsymbol{X}^T\Delta\boldsymbol{a} + 2||\Delta\boldsymbol{a}^T\boldsymbol{X}||^2 + ||\boldsymbol{b}_0^T\boldsymbol{Z}||^2 - 2\boldsymbol{b}_0^T\boldsymbol{Z}\boldsymbol{Z}^T\Delta\boldsymbol{b} \\ &\quad + 2||\Delta\boldsymbol{b}^T\boldsymbol{Z}||^2 + 2\Delta\boldsymbol{a}^T\boldsymbol{X}\boldsymbol{Z}^T\Delta\boldsymbol{b} + 2(\Delta\boldsymbol{a}^T\boldsymbol{X} + \Delta\boldsymbol{b}^T\boldsymbol{Z})\boldsymbol{\eta}^T + ||\boldsymbol{\eta}||^2.\end{aligned}}\end{aligned} \tag{3.28}$$

Without further assumptions, this is as far as the expression can be simplified. In the interest of gaining further insight, a few additional assumptions are made. As $\boldsymbol{\eta}$ is assumed white, the term $(\Delta\boldsymbol{a}^T\boldsymbol{X} + \Delta\boldsymbol{b}^T\boldsymbol{Z})^T\boldsymbol{\eta}$ asymptotically approaches 0, so it will be assumed to be small enough to be disregarded in further analyses unless otherwise specified. In the next subsections, four distinct special cases will be analyzed that emulate some typical model estimation conditions.

3.4.1 Case 1: $\Delta \boldsymbol{b}^T \boldsymbol{Z} \approx \mathbf{0}$

The first case being analyzed is where $\Delta \boldsymbol{b}^T \boldsymbol{Z} \approx \mathbf{0}$. This encompasses both the case in which $\Delta \boldsymbol{b} \approx \mathbf{0}$ (where the estimate of $\boldsymbol{b} \approx \boldsymbol{b}_0$) and the case in which $\Delta \boldsymbol{b} \perp \boldsymbol{Z}$ (*e.g.* when $\boldsymbol{x}$ represents a FIR filter with zeros that coincide with the spectrum of $\boldsymbol{Z}$ or, in more algebraic terms, the vector $\Delta \boldsymbol{b}$ is inside the null-space of matrix $\boldsymbol{Z}$, due to its containing highly colinear terms).

First, two auxiliary variables are defined

$$\alpha = \frac{\boldsymbol{a}_0^T \boldsymbol{X} \boldsymbol{X}^T \Delta \boldsymbol{a}}{||\boldsymbol{a}_0^T \boldsymbol{X}||^2} \qquad\qquad \beta = \frac{||\Delta \boldsymbol{a}^T \boldsymbol{X}||^2}{||\boldsymbol{a}_0^T \boldsymbol{X}||^2} \tag{3.29}$$

where α represents the level of colinearity of $\Delta \boldsymbol{a}$ and $\boldsymbol{a}_0^T$ in the inner product defined by the matrix $\boldsymbol{X}\boldsymbol{X}^T$, which converges asymptotically to $(N-1)\Sigma_X$, where Σ_X is the covariance matrix of X. Geometrically, in the subspace defined by $\boldsymbol{X}$, $\alpha \boldsymbol{a}_0 X$ is the projection of $\Delta \boldsymbol{a}$ into $\boldsymbol{a}_0$ and β is the ratio between the square of the norm of $\Delta \boldsymbol{a}$ and the norm of $\boldsymbol{a}_0$. It has been assumed here that $||\boldsymbol{a}_0^T \boldsymbol{X}||^2 > 0$, whereas the case where $||\boldsymbol{a}_0^T \boldsymbol{X}||^2$ will be evaluated separately later in this subsection. It can be shown that $\beta \geq 0$ and $\alpha^2 < \beta$.

For $\Delta \boldsymbol{b}^T \boldsymbol{Z} = 0$, substituting α and β into Eq. (3.28) allows it to be rewritten as

$$\mathrm{NC} = \frac{(1 - 2\alpha + \beta)||\boldsymbol{a}_0^T \boldsymbol{X}||^2}{(1 - 2\alpha + 2\beta)||\boldsymbol{a}_0^T \boldsymbol{X}||^2 + ||\boldsymbol{b}_0^T \boldsymbol{Z}||^2 + ||\boldsymbol{\eta}||^2} \tag{3.30}$$

which can be manipulated using Eq. (3.25) into

$$\mathrm{NC} = \frac{(1 - 2\alpha + \beta)\mathrm{NC}_0}{1 + 2\mathrm{NC}_0(\beta - \alpha)} \tag{3.31}$$

Under these conditions, further assumptions are needed for further analysis. Under least squares estimation of parameters, the error in the estimated variables is expected to be uncorrelated with the variables being estimated (as long as $\boldsymbol{a}_0^T \boldsymbol{X}$ is uncorrelated with $\boldsymbol{\eta}$). Therefore, the case for $\alpha = 0$ will be explored first. In this case,

$$\mathrm{NC} = \frac{(1 + \beta)\mathrm{NC}_0}{1 + 2\mathrm{NC}_0\beta} \tag{3.32}$$

Notice how the numerator contains a $1 + \beta$ factor, while the denominator contains a $1 + 2\mathrm{NC}_0\beta$. This means that NC will be equal to NC_0 if and only if $\beta = 0$; otherwise, it will be slightly biased towards 0.5. Fig. 3.17 shows contour curves for different values of NC_0 and β.

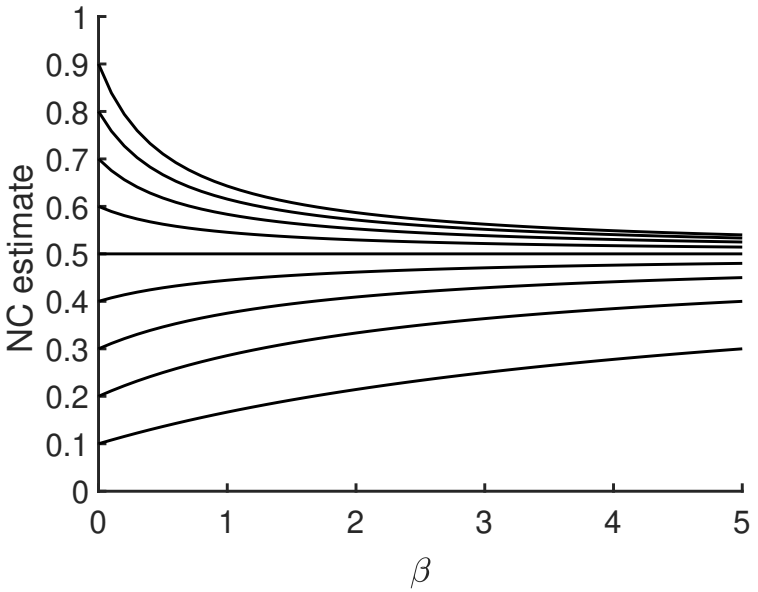


Figure 3.17: NC estimates for different values of NC_0 and β

For $||\boldsymbol{a}_0^T\boldsymbol{X}||^2 = 0$, the expression needs a small modification, but yields a similar expression

$$NC = \frac{||\Delta\boldsymbol{a}^T\boldsymbol{X}||^2}{2||\Delta\boldsymbol{a}^T\boldsymbol{X}||^2 + ||\boldsymbol{b}_0^T\boldsymbol{Z}||^2 + ||\boldsymbol{\eta}||^2} \tag{3.33}$$

which also biases NC towards 0.5. Therefore, for any $\Delta\boldsymbol{a}$ such that $||\Delta\boldsymbol{a}^T\boldsymbol{X}||^2 > 0$, NC is expected to be biased towards 0.5. Evidently, as long as the error in parameters is small (*i.e.* $||\Delta\boldsymbol{a}^T\boldsymbol{X}||^2 \ll ||\boldsymbol{a}_0^T\boldsymbol{X}||^2 + ||\boldsymbol{b}_0^T\boldsymbol{Z}||^2 + ||\boldsymbol{\eta}||^2$), this bias is also small.

3.4.2 Case 2: $\Delta\boldsymbol{a}^T\boldsymbol{X} \approx \boldsymbol{0}$

Similarly to Case 1, this analysis focuses varying only a single parameter vector ($\Delta\boldsymbol{b}$), while all terms related to the other parameter vectors ($\Delta\boldsymbol{a}$) are disregarded. Also similarly to Case 1, assuming $||\boldsymbol{a}_0^T\boldsymbol{X}||^2 > 0$, two new auxiliary variables are defined.

$$\gamma = \frac{\boldsymbol{b}_0^T\boldsymbol{Z}\boldsymbol{Z}^T\Delta\boldsymbol{b}}{||\boldsymbol{a}_0^T\boldsymbol{X}||^2} \qquad\qquad \delta = \frac{||\Delta\boldsymbol{b}^T\boldsymbol{Z}||^2}{||\boldsymbol{a}_0^T\boldsymbol{X}||^2} \tag{3.34}$$

Here, similarly to Eq. (3.29), $\delta > 0$ and $\gamma^2 < \delta$. By substituting Eq. (3.34) into Eq. (3.28), NC can be expressed as

$$NC = \frac{NC_0}{1 + 2NC_0(\delta - \gamma)} \tag{3.35}$$

As in Case 1, under least square assumptions (and assuming $\boldsymbol{b}_0^T\boldsymbol{Z}$ is uncorrelated with $\boldsymbol{\eta}$), γ can be assumed small. In this case, NC is biased towards 0. For $||\boldsymbol{a}_0^T\boldsymbol{X}||^2 = 0$, NC_0=0 regardless of η and $\boldsymbol{b}$. Therefore, there is no bias in NC if $||\boldsymbol{a}_0^T\boldsymbol{X}||^2 = 0$.

3.4.3 Case 3: $\Delta\boldsymbol{a}^T\boldsymbol{X} + \Delta\boldsymbol{b}^T\boldsymbol{Z} \approx \boldsymbol{0}$

The third case extends both previous cases. In the previous cases, it was assumed that the parameters associated with one of the regressors were equal to the observation model parameters, or that the error in the parameters was located in the null space of the regressor matrix (representing ill-posed problems). However, the parameter error due to the regressor matrix conditioning was constrained to a single signal. Case 3 deals with the case in which the parameter errors are spread across multiple signals. This is represented by

$$\Delta\boldsymbol{a}^T\boldsymbol{X} + \Delta\boldsymbol{b}^T\boldsymbol{Z} \approx 0 \tag{3.36}$$

Substituting Eq. (3.36) and variables α, β and γ into Eq. (3.28) yields

$$\text{NC} = \frac{(1 - 2\alpha + \beta)\text{NC}_0}{1 + 2\text{NC}_0(\beta - \alpha - \gamma)} \tag{3.37}$$

Note that under the assumption of small $|\gamma|$, this reduces to Eq. (3.32), being equivalent to Case 1 and, therefore, subject to the same biasing towards 0.5. Likewise, the expression for $||\boldsymbol{a}_0^T\boldsymbol{X}||^2 = 0$ is

$$\text{NC} = \frac{||\Delta\boldsymbol{a}^T\boldsymbol{X}||^2}{2||\Delta\boldsymbol{a}^T\boldsymbol{X}||^2 + ||\boldsymbol{b}_0^T\boldsymbol{Z}||^2 - 2\boldsymbol{b}_0^T\boldsymbol{Z}\boldsymbol{Z}^T\Delta\boldsymbol{b} + ||\boldsymbol{\eta}||^2} \tag{3.38}$$

3.4.4 Case 4: Regularization

In previous cases, α and γ were assumed to be small. If parameter estimates are obtained using LSE, this assumption is appropriate. However, if regularization methods are applied, the parameter estimates are expected to be biased towards zero. The bias towards zero can be modeled as

$$\Delta\boldsymbol{a}^T\boldsymbol{X} = \mu\boldsymbol{a_0}^T\boldsymbol{X} \qquad\qquad \Delta\boldsymbol{b}^T\boldsymbol{Z} = \mu\boldsymbol{b}_0^T\boldsymbol{Z} \tag{3.39}$$

for $0 \leq \mu < 1$. Effectively, this introduces a bias such that $||\boldsymbol{a}^T\boldsymbol{X}||^2 < ||\boldsymbol{a}_0^T\boldsymbol{X}||^2$ and $||\boldsymbol{b}^T\boldsymbol{Z}||^2 < ||\boldsymbol{b}_0^T\boldsymbol{Z}||^2$, and, therefore, $||\boldsymbol{\epsilon}||^2 > ||\boldsymbol{\eta}||^2$. Under these conditions, for any $0 < \mu < 1$ and $\mathrm{NC}_0 > 0$, the estimated NC value is expected to be lower than NC_0. By combining Eq. (3.39) and Eq. (3.28), the NC estimate becomes

$$\mathrm{NC} = \frac{\mathrm{NC}_0}{1 + \mathrm{NC}_0 \left(\frac{\mu^2}{(1-\mu)^2} \frac{||\boldsymbol{y}||^2}{||\boldsymbol{a_0}^T\boldsymbol{X}||^2} + 2\frac{\mu}{(1-\mu)} \frac{\boldsymbol{y}^T\boldsymbol{\eta}}{||\boldsymbol{a_0}^T\boldsymbol{X}||^2} \right)}. \tag{3.40}$$

As η is assumed to be white, it is uncorrelated with $\boldsymbol{X}$ and $\boldsymbol{Y}$, so it can be further assumed that $\boldsymbol{y}^T\boldsymbol{\eta} = ||\eta||^2$. Thus, Eq. (3.40) can be further manipulated into

$$\mathrm{NC} = \frac{\mathrm{NC}_0}{1 + \mathrm{NC}_0 \left(\frac{\mu^2}{(1-\mu)^2} \frac{||\boldsymbol{y}||^2}{||\boldsymbol{a_0}^T\boldsymbol{X}||^2} + 2\frac{\mu}{(1-\mu)} \frac{||\boldsymbol{\eta}||^2}{||\boldsymbol{a_0}^T\boldsymbol{X}||^2} \right)}. \tag{3.41}$$

Note that both terms in μ in the denominator are strictly positive for all $0 < \mu < 1$, therefore confirming that any regularization is bound to reduce the estimate of NC, particularly for data with low signal-to-noise ratio (*i.e.* $||\boldsymbol{a_0}^T\boldsymbol{X}||^2 + ||\boldsymbol{b_0}^T\boldsymbol{Z}||^2 \gg ||\boldsymbol{\eta}||^2$). Low signal-to-noise ratios already imply low NC_0 values, but Eq. (3.41) demonstrates that estimates of NC using regularization are expected to be even lower than NC_0.

3.4.4.1 Extending case 3

The behavior observed in the simulations for the observation model described by Eq. (3.19) for $M \geq 2$, particularly the bifurcation observed for $M = 5$ and $M = 6$, requires further analysis. The bifurcation behavior cannot be explained fully by Eq. (3.37) alone, and particularly, some of the bifurcation points are for NC>0.5, which violates the assumption that $|\alpha| \ll 1$ and $|\gamma| \ll 1$. In order to model this behavior, the simplifying assumptions must be reconsidered.

In order to observe the bifurcation behavior in the solutions, it is necessary to assume some distributional characteristics of $\Delta\boldsymbol{a}^T\boldsymbol{X}$ and $\Delta\boldsymbol{b}^T\boldsymbol{Z}$. In this analysis, $\Delta\boldsymbol{a}^T\boldsymbol{X}$ and $\boldsymbol{a_0}^T\boldsymbol{X}$ will be assumed to be samples from a bivariate normal distribution. Note that these terms appear only as inner products in Eq. (3.28), so the distributional characteristics need only apply to the sum of all the time

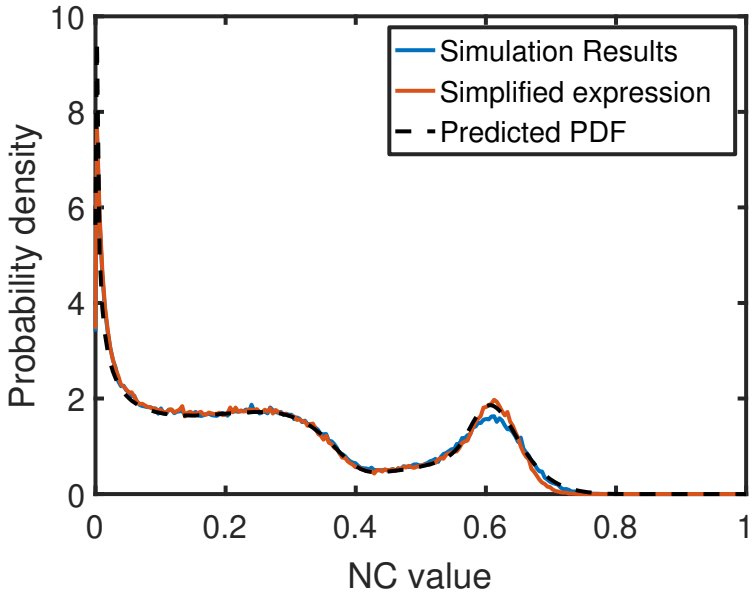


Figure 3.18: Estimated probability density function of NC using exact and approximate expressions

samples of each term; therefore, for sufficiently large N, the Gaussian assumption can be made under the central limit theorem, regardless of the distribution of the regressors and parameters.[3]

Obtaining the covariance matrix for $\Delta \boldsymbol{a}^T \boldsymbol{X}$ and $\boldsymbol{a_0}^T \boldsymbol{X}$ is not straightforward, as the distributional characteristics of $\boldsymbol{X}$ and $\boldsymbol{Z}$ depend on the observation model parameters (*i.e.* $\boldsymbol{a_0}$, $\boldsymbol{b_0}$ and the distributional characteristics of $\boldsymbol{\eta}$), which have interactions that are strongly coupled through a feedback loop (a solution to a bivariate second-order regressive model can be found in Appendix A and [147]). In a simulation study, however, these can be obtained empirically.

For the observation model of Eq. (3.19), a simulation was run using LSE to estimate the model parameters, under the same conditions as in Sec. 3.3. NC was estimated via Eq. (2.46), using the approximate form of Case 3 [Eq. (3.38)] and by calculating the mean and variances of the terms in the equation and estimating the probability density function using Monte Carlo simulation. The probability density functions were estimated using histograms and the results can be seen in Fig. 3.18. Note the excellent agreement between the exact and approximate expressions.

The peak seen around 0.6 occurs due to the large values of $\boldsymbol{a}_0^T \boldsymbol{X} \boldsymbol{X}^T \Delta \boldsymbol{a} + \boldsymbol{b}_0^T \boldsymbol{Z} \boldsymbol{Z}^T \Delta \boldsymbol{b}$. Fig. 3.19 was obtained by computing cases where $\|\boldsymbol{a}_0^T \boldsymbol{X}\|^2 + \|\boldsymbol{b}_0^T \boldsymbol{Z}\|^2 + \|\boldsymbol{\eta}\|^2$ is greater or smaller than

[3]While the classical central limit theorem requires i.i.d. samples, later developments prove convergence to Gaussian distributions under non-i.i.d. conditions [28, Theorem 27.5].

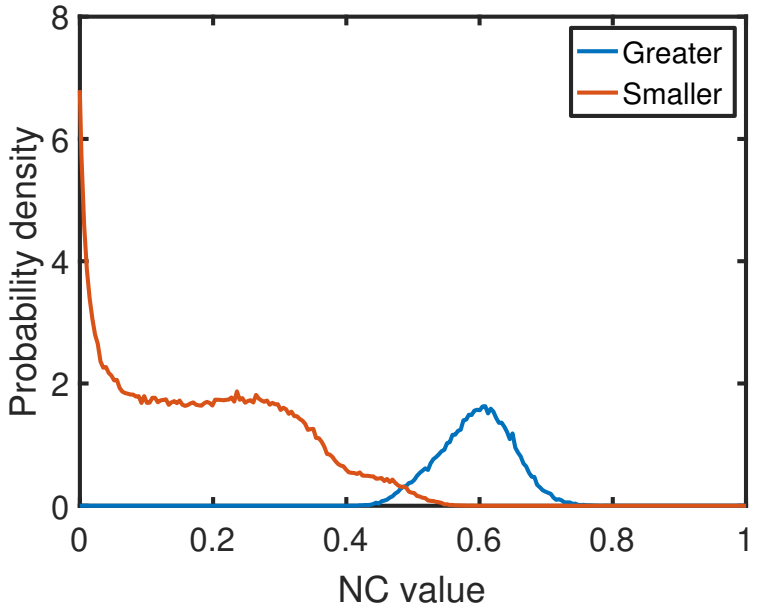


Figure 3.19: Estimated probability density function of NC split into two cases

$2\boldsymbol{a}_0^T\boldsymbol{X}\boldsymbol{X}^T\Delta\boldsymbol{a} + 2\boldsymbol{b}_0^T\boldsymbol{Z}\boldsymbol{Z}^T\Delta\boldsymbol{b}$ separately, and estimating their probability function. The blue line (unimodal function with mode close to 0.6) represent when the first term is greater than the second term, with the orange line representing the opposite case.

This behavior is exacerbated for larger values of M as the variances of $\Delta\boldsymbol{a}$ and $\Delta\boldsymbol{b}$ tend to increase with the increase of model order. Obviously, in most cases, the errors in the model parameters are expected to be small such that the case shown with the orange line never occurs and the NC estimate is close to NC_0.

3.4.4.2 Discussion of bias

These 4 cases demonstrate that due to the nonlinear nature of the NC calculation, NC estimates will often be biased towards a particular value. Although the studied cases represent particular conditions, combinations of one or more of these cases should represent a wide range of problems. This is not to say that NC is inherently flawed, but instead that special care must be taken when estimating NC. Particularly when the parameter estimates are close enough to the parameters of the observation model, the bias observed is small. Additionally, even a biased NC estimate can still provide helpful information for causality analysis.

3.5 Conclusion

The NC measure is an important development in causality analysis, addressing some limitations of GC. Particularly, it is designed to measure the causality mechanism, unlike GC, which measures the causal effect [19]. This ties to the relationship between GC and TE [14], since transferred information is indicator of causality, but whose differences must not be neglected [127]. As with many powerful analysis tools, proper care must be taken in order to avoid incorrect results. In this chapter, two examples are given where NC is shown to be more susceptible to model parameter estimation errors and overfitting than GC, particularly for ill-conditioned problems.

Although GC seems to be more robust to model parameter estimation, it still possesses many of the limitations described in [95, 99, 220]. Another advantage of NC is that is allows (pairwise) causality analysis for the entire model, while GC requires causality analysis to be done by considering one additional regressor signal at a time.

When estimating NC, a proper regression method must be applied to prevent overfitting. In this work, LASSO regression was used as an *ad-hoc* method of imposing sparsity in the model. For more complex systems, more sophisticated methods of obtaining model structure might be necessary. One recent method to obtain model structure is [149, 211, 217], which produces a family of models with differing levels of complexity and residual error, allowing easy trade-off selection. Future work will explore the performance of such methods for better model structure selection and causality measure estimation.

Although NC requires accurate parameter and model estimation, when these conditions are met, NC provides reliable results that in some cases have more powerful explanatory power than GC, more closely representing causality strength.

CHAPTER 4

A NONLINEAR EXTENSION TO NEW CAUSALITY

4.1 Overview

The seminal version of NC is defined for (linear) AR models [95]. While suitable in many applications, modern applications increasingly find linear and time-invariant (LTI) models to be insufficient [25, 39]. At the same time, as most generalizations, it is important to extend the applicability of a technique without losing its identifying characteristics. In this chapter, the definition of NC is extended to NARMAX models. The new definition, henceforth called nonlinear NC (NNC), not only maintains the same intuitive meaning, but identically reduces to the seminal definition when applied to ARMAX models.

The chapter starts with a motivating problem and the reasoning for the choice of NARMAX models. These are followed by the definition of the extension and examples of possible implementations. These are followed by application of this nonlinear extension into a series of progressively more complex synthetic models and discussions of the results. The technique is then applied to a EEG dataset and the results compared to GC and the seminal definition of NC. Finally, the results are summarized and discussed.

A significant portion of this chapter is quoted directly from [147] and [146] with a few modifications for improved flow and clarity.

4.2 Motivation

The ARMAX models used in the seminal formulation of NC of Eq. (2.46), contain only linear combinations of the regressors x_1, x_2, ..., x_{N_s} (and their time-delayed counterparts). observation models containing significant nonlinear terms, when modeled using ARMAX models, will less accurately predict the outputs of the model, and, more importantly, inadequately represent the underlying nature of the model. The following example illustrates one simple case where linear

NC is unable to represent causal strength.

Example A: Simple quadratic model

Consider the nonlinear model in Eq. (4.1) which contains a quadratic term, where η_1 and η_2 are samples from i.i.d. normally distributed processes with zero means and unity variances:

$$\begin{aligned} x_1[n] &= 0.53x_1[n-1] + 0.5x_2[n-1] + \alpha x_2^2[n-1] + \eta_1[n], \\ x_2[n] &= 0.5x_2[n-1] + \eta_2[n]. \end{aligned} \tag{4.1}$$

α is a coupling parameter that regulates the strength of the contribution of the quadratic term to x_1. Although the model is relatively simple, the seminal definition of NC has no mechanism to account for the quadratic term. A linear estimation model like that of Eq. (4.2) can be fit to predict the x_1, although at a reduced level of accuracy. As the effect of the quadratic term increases, an estimated ARX model of cannot represent the internal mechanism of the observation model and will produce an increasing prediction error variance. Consider the estimation model

$$x_1[n] = \sum_{i=1}^{M} a_{11}^i x_1[n-i] + \sum_{i=1}^{M} a_{12}^i x_2[n-i] + \epsilon_1[n]. \tag{4.2}$$

For $\alpha = 0.5$, the variance of x_1 is $\sigma_{x_1}^2 = 3.80$, the variance of x_2 is $\sigma_{x_2}^2 = 1.33$ and the variance of x_2^2 is $\sigma_{x_2^2}^2 = 3.55$. Using the model of Eq. (4.2), the optimum variance of the prediction error is $\sigma_{\epsilon_1}^2 = 1.95$ (opposed to the variance of η_1 which is $\sigma_{\eta_1}^2 = 1$). The evaluated NC value estimates are found in table 4.1.

Table 4.1: Linear $\mathrm{NC}_{x_j \to x_k}$ values for the model of Eq. (4.1)

		x_j	
		x_1	x_2
x_k	x_1	0.49	0.058
	x_2	0	0.25

Considering the variances of x_1 and x_2 (and x_2^2), notice that $\mathrm{NC}_{x_2 \to x_1}$ is small in comparison to $\mathrm{NC}_{x_1 \to x_1}$. Also note that the NC values for x_1 and x_2 add only to about 0.54, while the value which represents the contribution of the prediction error of the model is relatively large ($\mathrm{NC}_{\eta_1 \to x_1} = 0.46$).

For NC values over a range of α values, another undesirable result is observed. In Fig. 4.1, the NC values for this model are plotted for $\alpha \in [0, 1]$. Observe how, as α increases, $\mathrm{NC}_{x_2 \to x_1}$

decreases and $\mathrm{NC}_{x_1 \to x_1}$ increases. This is contrary to the intuition that the influence of x_2 over x_1 is increasing as α increases. This behavior stems from the fact that the model is using past values

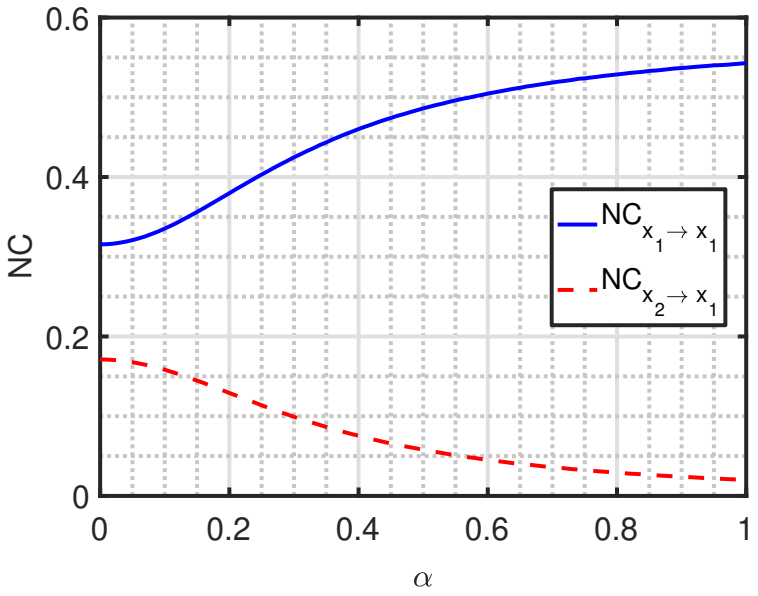


Figure 4.1: NC values for the model of Eq. (4.1)

of x_1 to estimate the value of $x_2^2[n-1]$, since (for $\alpha > 0$), $x_1[n-1]$ is correlated with $x_2^2[n-1]$, whereas $x_2[n-1]$ is not correlated to $x_2^2[n-1]$. As α increases, the correlation between $x_1[n]$ and $x_2^2[n-1]$ increases, and so does $\mathrm{NC}_{x_1 \to x_1}$.

Thus, it follows that the influence of x_2 on x_1 is not only underestimated when using the seminal definition of NC, but can also be negatively correlated. Moreover, it shows that, for observation models with significant nonlinear components, the seminal definition of NC is unable to properly assess the causal relationships between signals.

△ *End of Example A*

4.3 Choice of NARMAX models

ARMAX models are the most general representation of scalar linear systems. As shown in the previous section, the original definition of NC in terms of the parameters of a ARMAX model limits its use to systems that can be well-modeled by linear models. Since no canonical representation for all nonlinear models exists, a general nonlinear extension for NC is not possible. Particularly,

for many nonlinear models, it is not possible to decompose the model into a sum of "contributions" for each regressor [the concept of contribution will be later defined in Eq. (4.4)]. The seminal definition of NC requires the model to be decomposable into a sum of contributions, so a general extension of NC to all nonlinear models is infeasible.

Note that common LTI models comprise a subset of LTIiP models, so the extension of NC to NARMAX models subsumes the original NC development inherently. As long as certain conditions are met [discussed near Eq. (4.6)], the extension of NC developed in this work reduces to Eq. (2.46) for linear models.

The modeling power of NARMAX models comes at the cost of increased difficulty in estimating parameters. Due to the potentially large number of highly correlated regressors, overfitting and slow or inaccurate convergence are common challenges faced when estimating model parameters [11]. As a consequence, the quality of the models must be carefully evaluated, as NC values are dependent on accurate model structure and parameter estimation [148]. Nevertheless, many techniques have been developed specifically for nonlinear model selection and parameter estimation [22, 25, 27, 81, 118, 201, 203, 214].

4.4 A nonlinear extension to NC for a restricted set of models

A straightforward extension of NC to treat the LTIiP model occurs by grouping φ_p functions according to the regressor signal upon which they depend (*e.g.*, $x_q|_{n-M}^{n-1}$). A tentative expression for the nonlinear extension is as follows

$$\mathrm{NC}_{x_q \to x_p} = \frac{\left\| \sum_{k_q=1}^{K_q} a_{pk_q} \varphi_{k_q}^{q}\left(x_q|_{n-M}^{n-1} \right) \right\|_2^2}{\sum_{h=1}^{N_s} \left\| \sum_{k_h=1}^{K_h} a_{pk_h} \varphi_{pk_h}^{h}\left(x_h|_{n-M}^{n-1} \right) \right\|_2^2 + \left\| \eta_p[n] + \sum_{k_{\eta_p}=1}^{K_{\eta_p}} a_{pk_q} \varphi_{k_q}\left(\eta_p|_{n-M}^{n-1} \right) \right\|_2^2}, \tag{4.3}$$

where K_h is the number of regressor functions that depend exclusively on $x_h|_{n-M}^{n-1}$, $\varphi_{k_h}^{h}(x_h|_{n-M}^{n-1})$ is the k_h^{th} regressor function found in $\boldsymbol{\varphi}_p$ that depends exclusively on $x_h|_{n-M}^{n-1}$ and a_{pk_h} is the respective parameter associated with $\varphi_{k_h}^{h}(x_h|_{n-M}^{n-1})$. Note that this definition reduces identically to the seminal definition of NC for linear models.

This expression allows us to revisit the observation model of Eq. (4.1) and recompute the NC values with this definition.

Example B: Simple quadratic model revisited

The NC values shown in table 4.2 are computed using Eq. (4.3) for the model described by Eq. (4.1) for $\alpha = 0.5$. When using a quadratic NARMAX model, the NC values are more intuitive than the values found in Example A. $\mathrm{NC}_{x_2 \to x_1}$ is comparable to $\mathrm{NC}_{x_1 \to x_1}$, just as the contribution of x_2 to the current value of x_1 is comparable to the contributions of past values of x_1 to the current value of x_1.

Table 4.2: Nonlinear $\mathrm{NC}_{x_j \to x_k}$ values for the model of Eq. (4.1)

		x_j	
		x_1	x_2
x_k	x_1	0.32	0.37
	x_2	0	0.25

The NC and NNC values for the model of Eq. (4.1) are shown in Fig. 4.2 for varying values of α. At $\alpha = 0$, the NC and NNC values are equivalent, but as α increases, the values diverge significantly. As previously mentioned, the NC values follow a counterintuitive trend, with $\mathrm{NC}_{x_2 \to x_1}$ decreasing as α increases, whereas the NNC values follow a more intuitive trend. Notice that $\mathrm{NNC}_{x_1 \to x_1}$ remains almost constant over the range of α, where the small increase originates from the increased SNR, as the variance of x_1 increases with α whereas the variance of η_1 does not.

$\triangle$ *End of Example B*

4.5 A comprehensive NNC definition

Although the definition of Eq. (4.3) is intuitive, it cannot be used with general NARMAX models due to regressor functions that depend on multiple regressors (*e.g.*, $\varphi_k[n] = x_1[n-1]x_2[n-2]$). The presence of regressor functions that depend on more than one signal poses an additional challenge: how to best split the contribution across different signals? This question also appears in other causality related work [189]. Before answering this question, it is helpful to modify Eq. (4.3) to account for these terms. By including a weighting function, λ, to the contribution of each

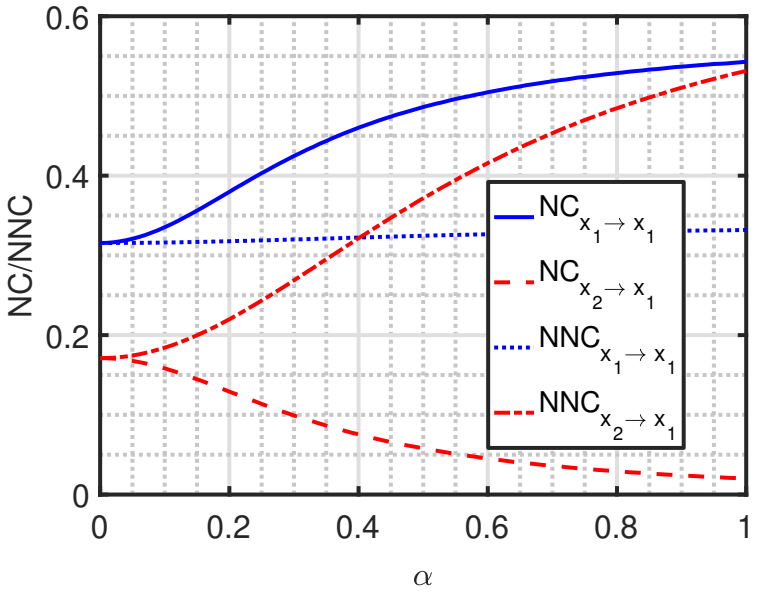


Figure 4.2: NC and NNC values for the model of Eq. (4.1)

regressor function, the NNC expression becomes

$$\mathrm{NC}_{x_q \to x_p} = \frac{\left\| \sum_{k=1}^{K} a_{pk} \varphi_k \lambda_{pq}(\varphi_k) \right\|_2^2}{\sum_{h=1}^{N_s} \left\| \sum_{k=1}^{K} a_{pk} \varphi_k \lambda_{ph}(\varphi_k) \right\|_2^2 + \left\| \eta_p[n] + \sum_{k=1}^{K} a_{pk} \varphi_k \lambda_{p\eta}(\varphi_k) \right\|_2^2} \tag{4.4}$$

where $\varphi_k = \varphi_k\left(x_1|_{n-M}^{n-1}, x_2|_{n-M}^{n-1}, \dots, x_{N_s}|_{n-M}^{n-1}, \eta_p{}_{n-M}^{n-1} \right)$[1] and $\lambda_{pq}(\varphi_k)$ is a function of φ_k associated with $x_q \longrightarrow x_p$ with the following properties

$$0 \leq \lambda_{pq}(\varphi_k) \leq 1 \tag{4.5a}$$

$$\lambda_{p\eta}(\varphi_k) + \sum_{q=1}^{N_\lambda} \lambda_{pq}(\varphi_k) = 1 \tag{4.5b}$$

Further, the following constraints are required so that the definition of NC for linear models remains as a special case:

$$\lambda_{pq}(\varphi_k) = \begin{cases} 1 & \text{if } \varphi_k \text{ is a function of } \textit{only } x_q|_{n-M}^{n-1}, \\ 0 & \text{if } \varphi_k \text{ does not depend on } x_q|_{n-M}^{n-1}. \end{cases} \tag{4.6}$$

[1]The arguments have been omitted for clarity, but the definition of the regressor functions remains the same as in Eq. (2.22), *i.e.*, can potentially depend on any set of the previous inputs and outputs.

Similarly to general nonlinear models, where no single concise representation is able to account for all cases, a single definition for the weighting function $\lambda_{pq}(\varphi_k)$ is impossible. Even when considering only LTIiP models, which can be concisely specified by the NARMAX representation, there is no canonical choice for set of regressor functions $\boldsymbol{\varphi}_k$. Instead, the NARMAX representation requires the choice of the proper function set for the particular problem. A similar challenge is present in this work. It is not possible to identify a unique definition of λ that would be appropriate for all applications of the method.

4.5.1 Form 1: λ^1 - create a new category for nonlinear cross-terms

The first form for λ discriminates terms that only depend on a single regressor signal from signals that depend on multiple regressors and assign them to separate categories. This way, the regressor functions $\varphi_j[n]$ and $\varphi_k[n]$ are joined if there is a $q \in 1, \ldots, N_s$ such that $\varphi_j[n]$ and $\varphi_k[n]$ can be expressed solely as functions of $x_q|_{n-M}^{n-1}$. This principle is used in the original definition of NC for the linear regressors, where past values of a signal as weighted and summed (*i.e.*, filtered) before the variance is estimated. The main distinction is that, in the linear NC definition, only time-shifting is used (as it is a linear transformation) and scaling the regressors would be absorbed in the parameter estimation. In other words, λ^1 can be defined as

$$\lambda_{pq}^{1}(\varphi_k) = \begin{cases} 1 & \text{if } \varphi_k \text{ is a function of } \textit{only } x_q|_{n-M}^{n-1} \\ 0 & \text{if } \varphi_k \text{ does not depend on } x_q|_{n-M}^{n-1} \text{ or depends on more than one regressor} \end{cases} \tag{4.7}$$

In order to satisfy Eq. (4.5b), a slight modification of the set of regressor signals must be made. Instead of $x_1, x_2, \ldots, x_{N_s}$, the set of regressor signals must be augmented by the set of all combinations of two or more signals (*e.g.*, $x_1 \cup x_2$, $x_1 \cup x_3$, $x_1 \cup x_2 \cup x_3$, etc.) For example, for a bivariate observation model of the form

$$\begin{aligned} x_1[n] = \; & a_1 x_1[n-1] + a_2 x_1[n-2] + a_3 x_2[n-1] + a_4 x_2[n-2] + \\ & a_5 x_1[n-1] x_2[n-1] + a_6 x_1[n-2] x_2[n-2] + \eta_1[n] \end{aligned} \tag{4.8}$$

the set of regressors is x_1, x_2 and $x_1 \cup x_2$, as all the regressor functions can be expressed as a function of x_1, x_2 or $x_1 \cup x_2$. Note that, regardless of time and polynomial order, this set covers all possibilities for bivariate nonlinear autoregressive with exogenous input (NARX) models,[2] therefore the inclusion of terms such as $x_1[n-1]x_2[n-2]$ or $x_1[n-2]x_2[n-1]$ does not alter the set.

To simplify notation, let us create a virtual regressor $x_3[n] = x_1[n]x_2[n]$, so the regressor set is x_1, x_2 and x_3. Assuming the values for a_k, for $k \in 1, 2, \ldots, 6$, have been accurately estimated, the expression for $\mathrm{NC}_{x_3 \to x_1}$ is therefore

$$\mathrm{NC}_{x_3 \to x_1} = \frac{\|a_5 x_3[n-1] + a_6 x_3[n-2]\|_2^2}{\|a_5 x_3[n-1] + a_6 x_3[n-2]\|_2^2 + \|a_1 x_1[n-1] + a_2 x_1[n-2]\|_2^2 + \|a_3 x_2[n-1] + a_4 x_2[n-2]\|_2^2 + \|\eta_p[n]\|_2^2}. \tag{4.9}$$

It is not difficult to verify that, with the exception of defining the new set of regressors (*i.e.*, $x_3[n] = x_1[n]x_2[n]$), this definition of NC is equivalent to the seminal definition of NC. Moreover, for linear estimated models, this definition of NC reduces to the seminal definition.

The λ^1 formulation is advantageous as it does not require defining weights for individual regressor functions. This allows it to be used with any set of candidate regressor functions. However, it creates additional regressors, which reduces the interpretability of the NC values.

4.5.2 Form 2: λ^2 - weight regressor functions equally across regressor signals

In order to avoid the creation of virtual regressors, the nonlinear contributions must be divided across the different regressor signals. Harnessing the knowledge of the arguments of each regressor functions, λ^2 splits the contributions equally between the regressors. Thus, λ^2 is defined as

$$\lambda^2_{pq}(\varphi_k) = \begin{cases} \dfrac{1}{R} & \text{if } \varphi_k \text{ is a function of } R \text{ regressor signals, including } x_q\big|_{n-M}^{n-1}, \\ 0 & \text{if } \varphi_k \text{ does not depend on } x_q\big|_{n-M}^{n-1}. \end{cases} \tag{4.10}$$

Since this approach does not require the creation of new virtual regressors, the final NC values can be easily mapped into the original signal set. For the observation model in Eq. (4.8), the NC

[2] To cover all NARMAX possibilities, permutations including $\eta_1[n]$ must also be included in the set

value can be calculated (assuming perfect model structure and parameter estimation) as

$$\mathrm{NC}_{x_3 \to x_1} = \frac{\left\| a_0 x_1[n-1] + a_1 x_1[n-2] + \frac{a_4 x_3[n-1] + a_5 x_3[n-2]}{2} \right\|_2^2}{\left\| a_0 x_1[n-1] + a_1 x_1[n-2] + \frac{a_4 x_3[n-1] + a_5 x_3[n-2]}{2} \right\|_2^2 + \left\| a_2 x_2[n-1] + a_3 x_2[n-2] + \frac{a_4 x_3[n-1] + a_5 x_3[n-2]}{2} \right\|_2^2 + \left\| \eta_p[n] \right\|_2^2}. \tag{4.11}$$

As long as the predictive model can mimic the dynamics of the observation model using a combination of regressor functions, both λ^1 and λ^2 will produce results similar to the observation model. That is, suppose an observation model can be decomposed as

$$x_1[n] = F_1^*(x_1|_{n-M}^{n-1}) + F_{12}^*(x_1|_{n-M}^{n-1}, x_2|_{n-M}^{n-1}) + F_2^*(x_2|_{n-M}^{n-1}) + \eta_1^*[n], \tag{4.12}$$

and the predictive model takes the form

$$\begin{aligned} x_{1p}[n] = &\sum_{k_1=0}^{K_1-1} \alpha_{1,k_1} \varphi_{1,k_1}(x_1|_{n-M}^{n-1}) \\ &+ \sum_{k_{12}=0}^{K_{12}-1} \alpha_{12,k_{12}} \varphi_{12,k_{12}}(x_1|_{n-M}^{n-1}, x_2|_{n-M}^{n-1}) \\ &+ \sum_{k_2=0}^{K_2-1} \alpha_{2,k_2} \varphi_{2,k_2}(x_2|_{n-M}^{n-1}), \end{aligned} \tag{4.13}$$

then, as long as

$$\begin{aligned} F_1^*(x_1|_{n-M}^{n-1}) &\approx \sum_{k_1=0}^{K_1-1} \alpha_{1,k_1} \varphi_{1,k_1}(x_1|_{n-M}^{n-1}) \\ F_{12}^*(x_1|_{n-M}^{n-1}, x_2|_{n-M}^{n-1}) &\approx \sum_{k_{12}=0}^{K_{12}-1} \alpha_{12,k_{12}} \varphi_{12,k_{12}}(x_1|_{n-M}^{n-1}, x_2|_{n-M}^{n-1}) \\ F_2^*(x_2|_{n-M}^{n-1}) &\approx \sum_{k_2=0}^{K_2-1} \alpha_{2,k_2} \varphi_{2,k_2}(x_2|_{n-M}^{n-1}) \\ \eta_1^*[n] &\approx x_1[n] - x_{1p}[n], \end{aligned} \tag{4.14}$$

the NC value for the estimated model will approximate the NC value for the observation model for both λ^1 and λ^2 forms.

The "equal splits" used in λ^2 simplify the analysis by avoiding the creation of new regressor signals. However, λ^2 does not take the characteristics of the regressor functions and distributional characteristics of the regressors into account. Particularly, for regressor functions that depend much more strongly on one regressor rather than another, it might be beneficial to implement a different weighting function.

4.5.3 Form 3: λ^3 - weight regressor functions across regressor signals according to an application (model) dependent criterion

Since no canonical form of distributing the contributions in nonlinear regressor functions exists, λ^1 and λ^2 are practical heuristic methods of approximately estimating causality strength (just as "true causality" is difficulty to define and measure [103]). Therefore, there is no need to limit NNC to pre-defined λs. The only requirement is that the function λ fit the definition of a probability mass function over the set of regressors and that it satisfy Eq. (4.6).

Certain regressor functions that depend on multiple regressors may not be equally affected by each of the regressors. The inhomogeneity of the influence may be due to the regressor function or the distributional characteristics of the regressors. For example, suppose that x_1 and x_2 are independent discrete random variables taken from Bernoulli distributions with parameters p_1 and p_2 respectively. Suppose also that $\varphi_k(x_1, x_2) = \mathtt{AND}(x_1, x_2)$, where $\mathtt{AND}()$ is the binary "and" operator. Note that this regressor function is perfectly symmetrical, as $\mathtt{AND}(x_1, x_2) = \mathtt{AND}(x_2, x_1)$, however, for $p_1 \gg p_2$, the value of x_2 contains more information on the output of φ_k than x_1. As another example, suppose that x_1 and x_2 are independent uniformly distributed random variables with support $x_1, x_2 \in [0, 1]$. Suppose also that $\varphi_k(x_1, x_2) = x_1 \sin(2\pi x_2)$. In this case, the variables are similarly distributed, but their effects upon the regressor function are not.

In such cases, it is desirable to split the contribution of the regressor function unequally across the regressors. One possible approach would be to weight the contributions by the predictive power of ${x_1}_{n-M}^{n-1}$ and ${x_1}_{n-M}^{n-1}$ to φ_k. One such method would be to use GC or TE to weight the contributions

$$\lambda_{pq}^{\mathrm{GC}}(\varphi_k) = \begin{cases} 1, & \text{if } \varphi_k \text{depends only on } x_q \\ \dfrac{\mathrm{GC}_{x_q \to \varphi_k}}{\sum_{h=1}^{N_s} \mathrm{GC}_{x_h \to \varphi_k}}, & \text{otherwise,} \end{cases} \tag{4.15}$$

where the first case is necessary as GC tends to infinity for deterministic expressions. Note that the second case tends to unity when $\mathrm{GC}_{x_q \to \varphi_k}$ approaches infinity.[3] Note also that Eq. (4.15) satisfies

[3] Although it is possible for the second case not to converge to unity, this is only true if for some $h \neq q$ there is at least one $\mathrm{GC}_{x_h \to \varphi_k}$ that also tends to infinity. This would only happen if φ_k is deterministic for more than one regressor signal, a degenerate case.

Eq. (4.6). However, there is an increased onus of estimating the GC values.

This example shows one way that contributions from regressor functions could be weighted across different regressors. Besides this example, there is an infinite set other possible variations that satisfy Eq. (4.6), but that might produce vastly different NC values. The choice and design of new λ weighting functions requires careful consideration and problem specific knowledge.

4.5.4 Spectral nonlinear new causality

A spectral expansion to NNC follows the same logic shown in Sec. 2.5.6, where the DTFT of the numerator is taken before the norm calculation, which yields

$$\mathrm{SNC}_{x_q \to x_p}(f) = \frac{\left\| \mathcal{F}\left\{ \sum_{k=1}^{K} a_{pk}\varphi_k \lambda_{pq}(\varphi_k) \right\}(f) \right\|_2^2}{\sum_{h=1}^{N_\lambda} \left\| \sum_{k=1}^{K} a_{pk}\varphi_k \lambda_{ph}(\varphi_k) \right\|_2^2 + \left\| \eta_p[n] + \sum_{k=1}^{K} a_{pk}\varphi_k \lambda_{p\eta}(\varphi_k) \right\|_2^2}. \tag{4.16}$$

Similarly to Eq. (2.49), this equation decomposes the contributions into their spectral components. However, an important distinction between Eq. (2.49) and Eq. (4.16) is that nonlinear models allow for cross-frequency couplings to be shown. These cross-frequency effects have been observed between planetary waves and tides [107] and EEG signals under various conditions [75, 83, 144, 178].

4.6 Discussion and analysis through example models

Although the choice of weighting function λ adds a additional complexity and uncertainty to the estimation of NC values, it is important to point out that the choice of function λ belongs more closely to the process of model selection and data pre-processing than causality estimation *per se*. That is, causality analysis tools are used to estimate characteristics or gain insight about systems whose internal properties are unknown.

For example, evoked potentials (EPs) are measured electrical potentials from the scalp immediately following a particular stimulus (*e.g.*, visual, auditory, tactile, etc.). EPs can be used in noninvasive tests of sensory pathway abstandardizties, language and speech disorders, among

other uses. However, due to anatomy and tissue impedance, electric potential measurements contain a significant amount of interchannel crosstalk, which may obscure the anatomical and temporal properties of the recorded EPs [191]. Since their characteristics are of the utmost importance to causality analysis, EP signals are commonly preprocessed using Current Source Density (CSD) or other spatio-temporal sharpening methods. However, the spatial component of these methods alters the recorded EPs, which in turn alter the NC values (arguably in a way that enhances the analysis).

Just as the seminal definition of NC is not transformation invariant (with the notable exception of uniform scaling and time-shifts), the nonlinear extension is not invariant to changes in the set of regressor functions. Similarly, the choice of λ weighting functions or regressor standardization falls within which assumptions better fit the current analysis. Thus, it is important to employ *a priori* knowledge about the systems being studied to obtain the most useful NC values possible. The following example shows how under typical conditions, linear transformations done as data preprocessing may affect NC values.

Example C: Effects of transformations on NC values

In many engineering applications, the desired signals cannot be directly obtained (*e.g.*, mixture ratios inside rocket engines due to the extremely high temperatures [142], or brain electric activity due to health risks and costs associated with intrusive implants [1, 207]), but instead, the signals are measured indirectly and estimated using different modeling techniques, such as Kalman filters [20, 142]. For EEG signals, the choice of reference to the unipolar measurements has also shown to affect the outcomes of the analysis [38]. Indirect measurement not only reduces the signal to noise ratio, but also limits the spatio-temporal resolution available. This example aims at demonstrating that modeling and *a priori* knowledge is critical to NC estimation. Here, a simple spatial transformation is applied to a simple three-signal model and the effect of the transformation to NC measurements will be shown.

The second-order jointly regressive model in Eq. (4.17) possesses simple relationships among its signals, *i.e.*, x_1 and x_2 "cause" x_1, but x_3 does not "cause" x_1. Similarly, x_2 and x_3 "cause" x_2,

whereas x_1 does not. Finally x_3 and x_1 "cause" x_3, but x_2 does not:

$$\begin{aligned} x_1[n] &= 0.8x_1[n-1] + 0.15x_2[n-2] + \eta_1[n], \\ x_2[n] &= 0.8x_2[n-1] + 0.1x_3[n-2] + \eta_2[n], \\ x_3[n] &= 0.5x_3[n-1] + 0.40x_1[n-2] + \eta_3[n]. \end{aligned} \tag{4.17}$$

Assuming that η_k, $k \in 1, 2, 3$ are samples taken from independent i.i.d. normally distributed processes with zero means and unity variances, the $\mathrm{NC}_{x_j \to x_k}$ values (rounded to two significant figures) are computed and displayed in table 4.3. Although extremely important to NC value estimation, we are not concerned with model topology or model parameter estimation in this example; instead, the observation model topology and parameters will be assumed to be known (or "perfectly" estimated). In general, model estimation adds an additional challenge to NC value estimation and models with highly correlated signals lead to higher variance in the parameter estimates and, in turn, higher variances in the NC value estimates [148].

Table 4.3: $\mathrm{NC}_{x_j \to x_k}$ values for the model in Eq. (4.17)

		x_j		
		x_1	x_2	x_3
x_k	x_1	0.70	0.021	0
	x_2	0	0.68	0.011
	x_3	0.26	0	0.35

The NC values in table 4.3 provide intuition about the model (*e.g.*, the current value of x_1 depends on past values of x_1 and x_2, with x_1 having a greater influence and x_3 having no direct influence on x_1).

Now suppose that the same signals cannot be measured directly, but must be estimated using surface sensors, such that the signals contain interference from surrounding sources. In this model involving three signals, the interference is assumed to be uniform and controlled by the parameter δ, as in Eq. (4.18).

$$\begin{bmatrix} y_1[n] \\ y_2[n] \\ y_3[n] \end{bmatrix} = \begin{bmatrix} x_1[n] + \delta\,(x_2[n] + x_3[n]) \\ x_2[n] + \delta\,(x_1[n] + x_3[n]) \\ x_3[n] + \delta\,(x_1[n] + x_2[n]) \end{bmatrix} = \begin{bmatrix} 1 & \delta & \delta \\ \delta & 1 & \delta \\ \delta & \delta & 1 \end{bmatrix} \begin{bmatrix} x_1[n] \\ x_2[n] \\ x_3[n] \end{bmatrix}. \tag{4.18}$$

For $\delta = 0.15$, the $\mathrm{NC}_{y_j \to y_k}$ values are shown in table 4.4. Note how the $\mathrm{NC}_{y_j \to y_3}$ row differs from $\mathrm{NC}_{x_j \to x_3}$ of table 4.3. In particular, this analysis implies that past values of y_1 have greater influence on the current value of y_3 than past values of y_3, a property which is not shared with x_1 and x_3.

Table 4.4: $\mathrm{NC}_{y_j \to y_k}$ values for the model of Eq. (4.18)

		y_j		
		y_1	y_2	y_3
y_k	y_1	0.72	0.020	0.05
	y_2	0.0033	0.68	0.0044
	y_3	0.32	0.00068	0.29

This example shows that, even for simple linear models, careful consideration is necessary not only for model topology and model parameter estimation, but also for assumptions used when preprocessing data. The preprocessing of data using *a priori* knowledge about the studied system is necessary for more "useful" NC estimates. This observation will be helpful when discussing the increased complexity that the class of nonlinear models adds to this work.

△ *End of Example C*

In the same way as data preprocessing can be used to enhance linear NC values, proper preprocessing is essential to nonlinear NC value estimation. One occasion where this is particularly apparent when regressors do not possess zero mean. For example, let $\varphi_k[n-1] = x_1[n-1]x_2[n-1]$. Intuitively, this regressor function depends equally on $x_1[n-1]$ and $x_2[n-1]$. However, suppose that $x_1[n-1]$ and $x_2[n-1]$ are distributed as multivariate normal random variables with means $\boldsymbol{\mu} = \left[\begin{smallmatrix}\mu_1\\0\end{smallmatrix}\right]$ and covariance matrix $\Sigma = \begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}$. If

$$\begin{aligned}\|\mu_1\| &\gg \sigma_1\\ \sigma_2 &\gg \sigma_1\end{aligned} \tag{4.19}$$

then the value of $x_1[n-1]$ is likely to be close to μ_1. Therefore most of the variation seen in φ_k comes from variations in $x_2[n-1]$, not $x_1[n-1]$. In other words, regressor functions $\varphi_{k_1} = x_1[n-1]x_2[n-1]$ and $\varphi_{k_2} = x_2[n-1]$ would produce very different results than $\varphi_{k_3} = (x_1[n-1] - \mu_1)x_2[n-1]$ and $\varphi_{k_4} = x_2[n-1]$. For φ_{k_1} and φ_{k_2}, the NC value for x_1 would be larger than the NC values computed

using φ_{k_3} and φ_{k_4} and, likewise, the NC value for x_2 would be smaller than the NC values computed using φ_{k_3} and φ_{k_4}.

Standardization is a common technique for data preprocessing. Standardization involves removing the means and dividing by standard deviation. While scaling of the regressors does not affect NC values, many regression methods benefit from standardization in the form of faster convergence or improved numerical stability.

For nonlinear models, standardization can have a drastic effect on NC values. The choice of removing the means of regressor signals prior to computing the regressor functions or not standardizing depends mainly on assumptions on the models and the causality information desired. Is the information contained within the signals an absolute or relative measure? Many phenomena depend linearly on absolute quantities (*e.g.*, the average sound speed on a fluid depends on the mean absolute pressure, final volume in an isobaric process depends on the absolute temperature, etc.) On the other hand, sound is a measure of relative pressure fluctuations measured at a microphone (or hydrophone for underwater measurements). The sound pressure fluctuations are several orders of magnitude smaller than the mean absolute pressure, therefore standardization is desirable. Nonetheless, in some cases, even choosing a reference value can be challenging for processes that are not wide-sense stationary and for measurements that do not have a clear reference point [218] (such as ERP and EEG signals).

When φ_k is not an odd function, even a linear regressor signal symmetrically distributed with zero mean might produce an output with nonzero mean. For example, suppose that two independent signals, x_1 and x_2 were uniformly distributed with support $[-1, 1]$. Then $|x_1[n]|$ has mean 0.5, but the regressor function $\varphi_k[n] = \left|x_1[n]\right| \cdot x_2[n]$ has zero mean. While, φ_k has zero mean, it is important to consider whether a combination of $\varphi_{pl}[n] = x_2[n]$ and $\varphi_{pm}[n] = (|x_1[n]| - 0.5)x_2[n]$ (both also having zero mean) better represent the dynamics of interest in the system being studied.

Ultimately, the differences observed between NC computed with regressors with means removed or not is a modeling issue more than a limitation of the method. Time series data must be analyzed prior to model specification [78] in order to remove undesired artifacts. Any type

of preprocessing will modify the outcomes of the analysis, but whether it will be beneficial to a particular analysis depends on the particular characteristics of the system. One must evaluate the assumptions when choosing preprocessing data as to produce "useful" models. As shown in [147], the reliability of NC value estimation is closely related to the models used, so a careful selection off preprocessing and model estimation is doubly important for NC analysis.

To demonstrate the nonlinear extension of NC, two models used in [81] are tested to demonstrate the performance of the nonlinear extension of NC. The first example model given in [81] is noise-free as shown in Eq. (4.20):

Example D: First model from [81]

$$x_1[n] = 0.5x_1[n-1] + 0.8x_2[n-2] + x_2^2[n-1] - 0.05x_1^2[n-2] + 0.5, \tag{4.20}$$

where x_2 is assumed to be sampled from an i.i.d. uniform distribution process bounded by $[-1, 1]$. x_1 has 1.42 mean and 0.4 variance, whereas x_2 has zero mean and variance ⅓. Since the equation for x_1 is noise-free, the sum of all $\mathrm{NC}_{x_j \to x_1}$ values is expected to be unity, which is confirmed by table 4.5, whereas the sum of all $\mathrm{NC}_{x_j \to x_2}$, with x_2 being i.i.d., is zero. Note that, in this instance, standardizing the regressors and regressand yield no difference, as there are no nonlinear cross-terms. The absence of nonlinear cross-terms also means that any weighting function λ following Eq. (4.6) produces identical results. An example where nonlinear cross-terms are present and the standardization affects the NC estimates and further elaboration on this effect are given in the next example.

Table 4.5: $\mathrm{NC}_{x_j \to x_k}$ values for the model of Eq. (4.20)

		x_j	
		x_1	x_2
x_k	x_1	0.25	0.75
	x_2	0	0

△ *End of Example D*

Example E: Second model from [81]

The second model example used in [81] is shown in Eq. (4.21) below. In [81], the model is used to evaluate how the robust model structure selection (RMSS) method proposed in [81] behaves when the nonlinear regressor function in the observation model is not included the candidate nonlinear regressor functions, but instead a Volterra expansion with two time lags and up to order 3 is applied to x_1 and x_2,

$$x_1[n] = -x_2[n-1]\sqrt{|x_1[n-1]|} + 0.4x_2^2[n-1] + 0.8x_2[n-1]x_2[n-2] + \eta_1[n], \tag{4.21}$$

where x_2 is assumed to be uniformly distributed on $[-1, 1]$ and $\eta_1[n]$ is white noise with zero mean and finite variation. The variance of η_1 is adjusted to produce different SNR values (*i.e.*, 0dB, 10dB, 15dB, 50dB and noise-free in the paper). Eq. (4.21) poses a particular problem for NC value estimation using Volterra expansions as the term $x_2[n-1]\sqrt{|x_1[n-1]|}$ cannot be easily expanded using polynomials since $\sqrt{|x|}$ is not differentiable at $x = 0$. Further complicating NC estimation is that a polynomial expansion of $x_2[n-1]\sqrt{|x_1[n-1]|}$, takes the form

$$x_2[n-1]\sqrt{|x_1[n-1]|} \approx x_2[n-1]\left(\alpha_0 + \alpha_1 x_1[n-1] + \alpha_2 x_1^2[n-1] + \cdots\right). \tag{4.22}$$

Note how most the terms in the right-hand side would have the same λ^1 and λ^2 value (*i.e.*, 0.5 for both x_1 and x_2 in the case of λ^1 and a separate category that depends on x_1 and x_2 for λ^1), but the term $x_2[n-1]\alpha_0$ only depends on x_2 and therefore would be counted entirely towards $\mathrm{NC}_{x_2 \to x_1}$, rather than sharing the contributions.

To observe the effect of standardization, tests were conducted at 10dB and 50dB SNR using the original and standardized regressors. The tests included NC values for $\sqrt{|x_1[n-1]|}$ as one of the candidate functions and Volterra expansions of third and fifth order. The results for 10dB and 50dB SNR are shown in tables 4.6 and 4.7, respectively.

Although the value for NC using the non-standardized $\sqrt{|x|}$ candidate regressor function differs significantly from the others values, the NC values computed with the standardized $\sqrt{|x|}$ are very similar to those computed with fifth-order polynomials. The discrepancy between NC values computed with the non-standardized $\sqrt{|x|}$ and the standardized $\sqrt{|x|}$ is a consequence of the

Table 4.6: $NC_{x_j \to x_1}$ values for the model of Eq. (4.21) with 10dB SNR

Poly. Order	Volterra				With $\sqrt{\lvert x \rvert}$			
	Not standardized		Standardized		Not standardized		Standardized	
	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
3	0.028	0.83	0.023	0.84	0.17	0.70	0.028	0.87
5	0.044	0.83	0.035	0.84				

Table 4.7: $NC_{x_j \to x_1}$ values for the model of Eq. (4.21) with 50dB SNR

Poly. Order	Volterra				With $\sqrt{\lvert x \rvert}$			
	Not standardized		Standardized		Not standardized		Standardized	
	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
3	0.038	0.92	0.029	0.93	0.18	0.82	0.040	0.96
5	0.058	0.92	0.040	0.94				

$\alpha_0 x_2[n-1]$ term from Eq. (4.22), which is assigned to solely to $NC_{x_2 \to x_1}$, whereas, being a function of both x_1 and x_2, the contributions of $x_2[n-1]\sqrt{|x_1[n-1]|}$ depends on both x_1 and x_2. If λ were set to split the contribution of $x_2[n-1]$ equally across x_1 and x_2, all the NC values would be in close agreement.

Because $\sqrt{|x|}$ cannot be well modeled with polynomials, modeling Eq. (4.21) with Volterra filters limits the accuracy of the predictive model. To observe how a similarly complex, but differentiable model behaves, the $\sqrt{|x|}$ is be replaced with a $\tanh(x)$ term, a sigmoid function. Functions that exhibit saturation, like sigmoids, are poorly approximated with polynomials at the extremes, but can produce reasonable approximations if the polynomial order is high enough and/or the input has small variance. The resulting difference equation of replacing $\sqrt{|x|}$ with $\tanh(x_1[n-1])$ in Eq. (4.21) is shown in Eq. (4.23),

$$x_1[n] = -2x_2[n-1]\tanh(x_1[n-1]) + 0.5x_2^2[n-1] + 0.5x_2[n-1]x_2[n-2] + \eta_1[n]. \tag{4.23}$$

For this modified observation model, the same tests were conducted for 10dB and 50dB SNR. Again, the Volterra expansion was applied with two time lags and polynomial orders of three and five, and a prediction model was created with $\tanh(x)$ as one of the candidate regressor functions. The results are found in tables tables 4.8 and 4.9. Note how in this case, the results between the standardized and non-standardized cases are in closer agreement as the mean of x_1 is closer to

zero [since the $x_2[n-1]\tanh(x_1[n-1])$ term does not introduce bias, only the $x_2[n-1]^2$ term does]. The Volterra results are limited by the term containing $\tanh(x)$ function being approximated only by finite order polynomials. Nevertheless, in both cases, the Volterra and $\tanh(x)$ results show good agreement.

Table 4.8: $\mathrm{NC}_{x_j \to x_1}$ values for the model of Eq. (4.23) with 10dB SNR

Poly. Order	Volterra				With $\tanh(x)$			
	Not standardized		Standardized		Not standardized		Standardized	
	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
3	0.26	0.55	0.24	0.58	0.29	0.55	0.26	0.58
5	0.28	0.55	0.26	0.59				

Table 4.9: $\mathrm{NC}_{x_j \to x_1}$ values for the model of Eq. (4.23) with 50dB SNR

Poly. Order	Volterra				With $\tanh(x)$			
	Not standardized		Standardized		Not standardized		Standardized	
	x_1	x_2	x_1	x_2	x_1	x_2	x_1	x_2
3	0.28	0.69	0.25	0.72	0.31	0.69	0.26	0.73
5	0.31	0.69	0.26	0.74				

Due to the difficulty in properly estimating parameter and topology for nonlinear models, it is not advisable to blindly increase the order of the polynomial expansions [11]. In addition to overfitting, NC value quality requires the estimated model structure and parameters to represent the observation model. Nonlinearity can often create complex relationships among regressors, such that high order regressor models might have good fitness and even generalize well, but might misrepresent the underlying model structure.

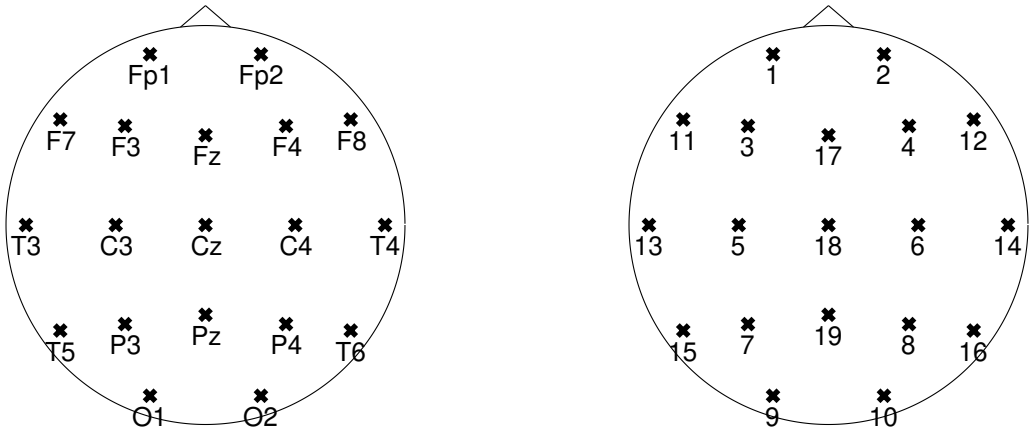
Due to the complex interaction among regressors and noise, instead of representing $\tanh(x)$ as a Taylor series, the regression algorithm will likely find a more compact set of regressor functions which produce lower prediction error. This compact set does not necessarily preserve the same relationship between x_1 and x_2, so indiscriminately increasing the model order leads to results tending towards $1/N_s$. This is similar to the behavior shown in Sec. 3.4, where the several scenarios are discussed where NC estimates exhibit bias under least squares estimation.

$\triangle$ *End of Example E*

4.7 Application: EEG data

The EEG dataset used in [150] is used to compare NNC to the performance of GC and NC. Although most of the power of EEG signals can be predicted well using simple MVAR models, EEG signals contain nonlinear components that contain important information [140, 162, 182, 186]. Since linear predictive estimation models are able to reasonably represent the gross features of EEG signals, the improvement in NNC application is expected to be modest. Experiments using digital filters are used to highlight the nonlinear components which will be compared to the unfiltered results.

The data were made publicly available by Nolte *et al.* [150], but obtained from Tom Brismar of the Karolinska Institute in Stockholm. The dataset contains EEG measurements for 10 subjects, sampled at 256Hz using the International 10–20 system, with 19 channels available using linked mastoid reference for the unipolar measurements. The measurements were made while subjects kept their eyes closed. The subjects were asked to open their eyes for 5 seconds every minute. The records contain about 200 segments of 4 seconds, which were recorded while subjects had their eyes closed. The location of the electrodes are shown in Fig. 4.3a, with the channel indices used in the dataset shown in Fig. 4.3b.



(a) Electrode labels (b) Electrode index in dataset

Figure 4.3: 10-20 International System Electrode Location Diagram

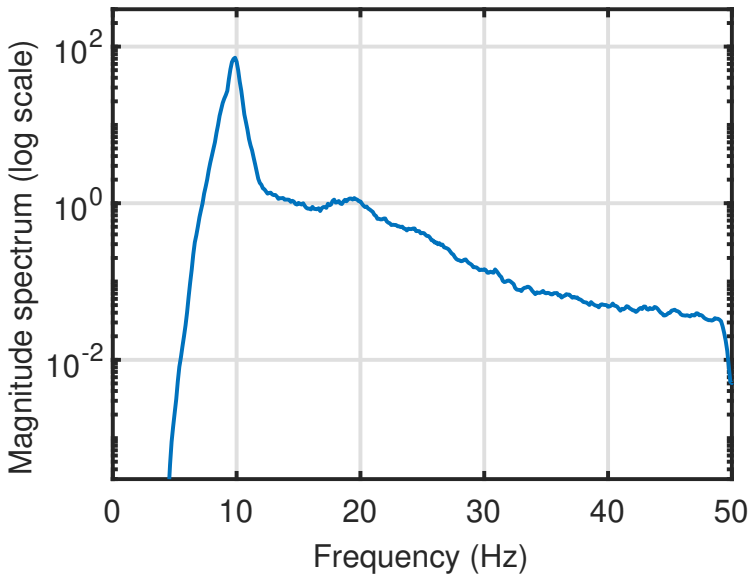


Figure 4.4: Spectrum of the Fp1 channel of the EEG recording

The signals contain a α rhythm component (8-13 Hz band) at approximately 10Hz. All apparent artifacts have been removed from the data by Nolte *et al.* prior to the publication of the data. The 10 recordings were selected out of a pool of 88 recordings based on estimated signal to noise ratio. The database contains no subject identifiable information.

While no ground truth is possible for these data, it is well established in literature that information flow for α and β waves follow a posterior-to-anterior (front to back) pattern [89, 150] during resting states. For these experiment, the flow between the left pre-frontal cortex (Fp1) channel and the right occipital (O2) and right parietal (P4) channels were considered. The θ waves flow in an anterior-to-posterior pattern [89] under similar conditions.

The time-series were further processed using a notch filter to remove 50Hz line noise and a high-pass Butterworth filter of order 10 with cutoff frequency at 7.5Hz to remove low frequency signal drifts and θ waves. The recordings were split into 202 segments of 4 seconds each. The spectrum of the entire signal and for the first segment for the Fp1 channel are shown in Fig. 4.4.

The models used to evaluate GC and NC were 3^{rd} order AR/ARX and ARX models respectively. The models used to evaluate NNC and SNNC were 3^{rd} order polynomial expansions of the regressors used to evaluate NC. The model parameters were evaluated using LASSO using four-fold

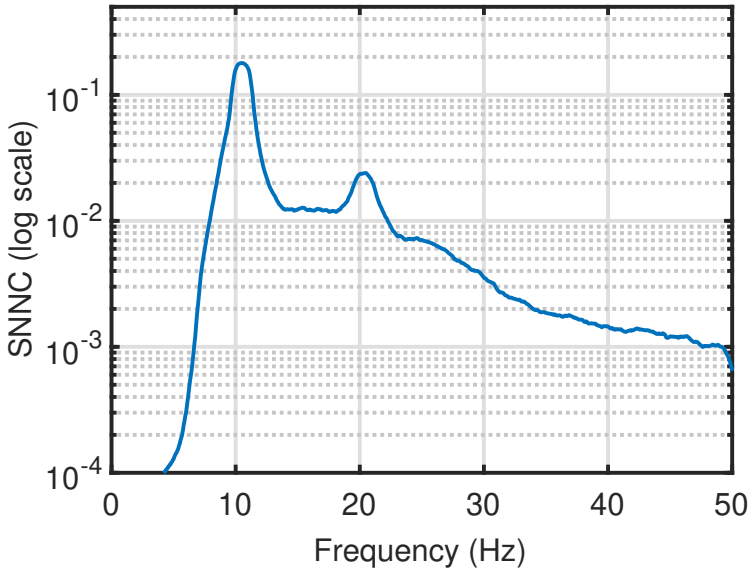


Figure 4.5: Average of $\text{SNNC}_{\text{Fp1}\rightarrow\text{O2}}$ values of subject 1

cross-validation. The average of the SNNC values for the x_{Fp1} into x_{O2} test is shown in Fig. 4.5.

The significance numbers were obtained using trial-shuffling [37, 196]. For each j^{th} segment output time-series, the GC, NC and NNC values were calculated using the input time-series of all k^{th} segments, the GC, NC and NNC values evaluated for $j \neq k$ were used to estimate the distribution of GC, NC and NNC under the non-causal assumption. The distributions were evaluated using kernel estimation technique [88]. Since the pre-frontal cortex is reasonably distant from parietal and occipital regions, no spatial sharpening procedure is applied. Additionally, since no activity is being executed by subjects and (particularly) the 4 second segments are not related to the (non) activity of the subjects, no data alignment procedure is done and trials are assumed independent. The GC, NC and NNC evaluated for $j = k$ were evaluated against that distribution to evaluate the p-value of that trial. The trials were considered significant using a Neyman-Pearson test with maximum of 1% false positives.

To highlight the nonlinear relationships in the EEG signal, the tests were repeated three times: first as described above, second by filtering the α rhythm frequencies and lower and third by filtering the β rhythm (13-35Hz) frequencies and lower. The signals were filtered using Chebyshev type II high-pass filters of order 10 at cut off frequencies 13.5Hz and 35Hz respectively. In the

models used to evaluate NNC, the filters are applied after the polynomial expansions, to preserve the contribution of the α waves into β and higher bands due to the harmonic distortion. For the tests using the 13.5Hz high-pass filters, the SNNC was also evaluated between 18Hz and 28Hz, which roughly correspond to twice the frequency of the α waves.

During the first test, all of the measures identified a strong relationship between the Fp1 and O2, but were unable to differentiate direction of flow between Fp1 and O2, having both high levels of significance in both directions, with only SNNC having significantly higher rejection in the O2 to Fp1 direction. Applying the filter with a cutoff frequency of 13.5Hz reveals the directivity and also more differences between the measures. When filtering both α and β bands, the measures fail to indicate the strong connectivity between Fp1 and O2, partially due to lower SNR and electromyographic interference [143]. The results are shown in table 4.10 and table 4.11, where the best two results[4] are in bold.

Table 4.10: GC, NC, NNC, and SNNC results on whether to accept x_{Fp1} causes x_{O2}

	Unfiltered	Filtered at 13.5Hz	Filtered at 35Hz
GC	**0.851**	0.535	**0.228**
NC	**0.851**	**0.614**	**0.267**
NNC	0.772	0.525	0.168
SNNC	0.812	**0.674**	0.891

Table 4.11: GC, NC, NNC, and SNNC results on whether to reject x_{O2} causes x_{Fp1}

	Unfiltered	Filtered at 13.5Hz	Filtered at 35Hz
GC	0.139	0.604	**0.861**
NC	0.139	0.545	**0.861**
NNC	**0.158**	**0.723**	**0.861**
SNNC	**0.386**	**0.743**	**0.99**

In the tests with the high-pass filter with cut-off frequency at 13.5Hz, SNNC performed the best at both accepting x_{Fp1} causing x_{O2} and rejecting x_{O2} causing x_{Fp1}. The NNC result was also able to reject x_{O2} causing x_{Fp1} at a comparable rate to NNC and were about 20% higher relative to GC. The NC results seem to indicate that bias towards significance as it consistently assigned

[4]When multiple measures perform equally, more than two entries may boldened.

highest significance to tests out of all measures. The GC results show no similar bias, but show lower selectivity than SNNC.

The receiver operating characteristic curves for the unfiltered tests and the tests filtered at 13.5Hz regarding Fp1 and O2 are shown in Figs. 4.6 and 4.7, where Fp1 causing O2 is assumed true positives and O2 causing Fp1 is assumed as a false positive. In Fig. 4.6, the improvement of SNNC over the other measures can be seen more clearly, where only the higher rejection of O2 causing Fp1 is seen in table 4.11. In Fig. 4.7, both NNC and SNNC perform better than the other measures, but quite similarly to each other.

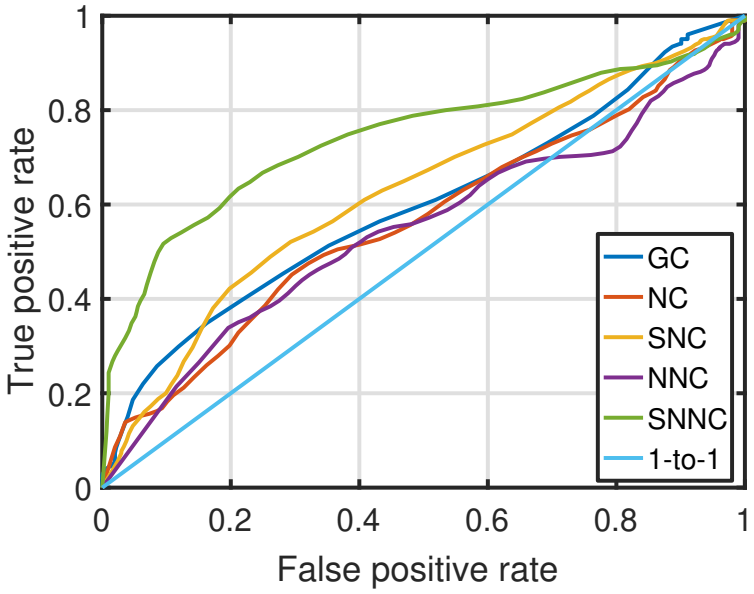


Figure 4.6: Receiver operating characteristic curves for the unfiltered tests

The tests were repeated computing the causality measures between the Fp1 and P4 channels. The results are shown in table 4.12 and table 4.13. For the unfiltered signals, all tested methods were better able to show the directionality of information flow than the tests with Fp1 and O2. Nevertheless, the rate of significant results for x_{Fp1} causing x_{P4} are also smaller. The rate of significant results for the signals filtered at 13.5Hz are higher than the unfiltered ones for x_{Fp1} causing x_{P4} and are comparable to the unfiltered results found in table 4.10. The rejection rates for x_{O2} causing x_{P4} for signals filtered at 13.5Hz are similar to table 4.11, where NNC and SNNC are both significantly superior to GC and NC (here by 27% and 42% respectively).

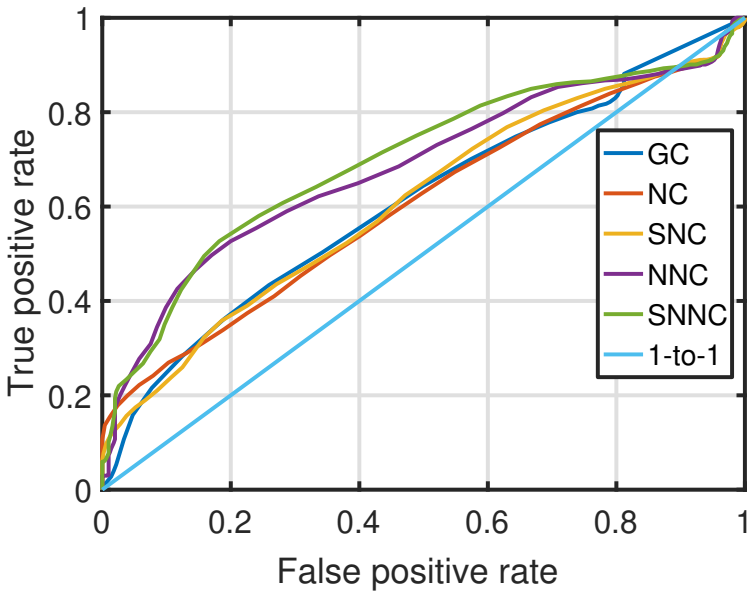


Figure 4.7: Receiver operating characteristic curves for 13.5Hz

Table 4.12: GC, NC, NNC, and SNNC results on whether to accept x_{Fp1} causes x_{P4}

	Unfiltered	Filtered at 13.5Hz	Filtered at 35Hz
GC	**0.653**	**0.891**	**0.851**
NC	**0.634**	**0.891**	**0.851**
NNC	0.593	0.842	0.743
SNNC	0.624	0.772	0.168

Table 4.13: GC, NC, NNC, and SNNC results on whether to reject x_{P4} causes x_{Fp1}

	Unfiltered	Filtered at 13.5Hz	Filtered at 35Hz
GC	0.545	0.535	0.465
NC	**0.634**	0.416	0.347
NNC	0.564	**0.683**	**0.594**
SNNC	**0.574**	**0.762**	**0.881**

The receiver operating characteristic curves for the unfiltered tests and the tests filtered at 13.5Hz regarding Fp1 and P4 are shown in Figs. 4.6 and 4.7, where Fp1 causing P4 is assumed true positives and P4 causing Fp1 is assumed as a false positive. In Fig. 4.6, the NNC and SNNC results are worse than NC, although NNC achieves similar results in the small false positive rate region and SNNC achieves similar results to NC for large false positive rates. In Fig. 4.7, the advantage of NNC and SNNC over NC is visible in the small false positive rate region, with the advantage diminishing as the false positive rate increases.

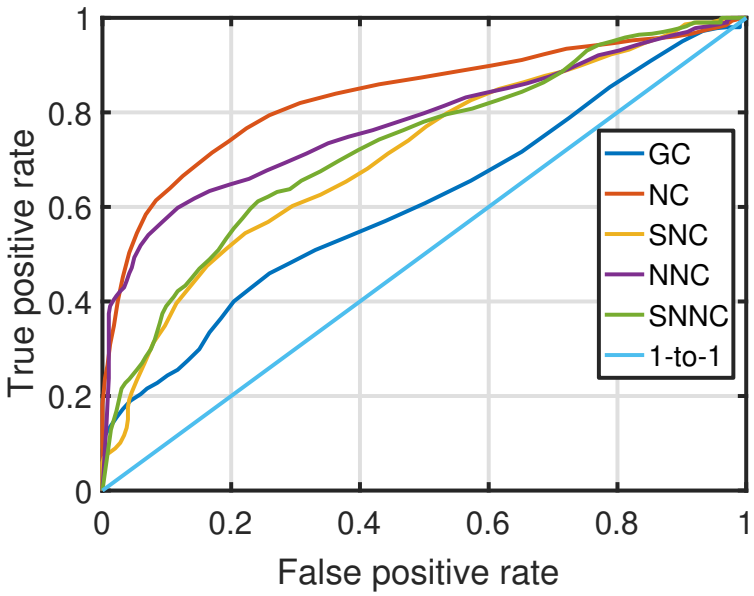


Figure 4.8: Receiver operating characteristic curves for the unfiltered tests

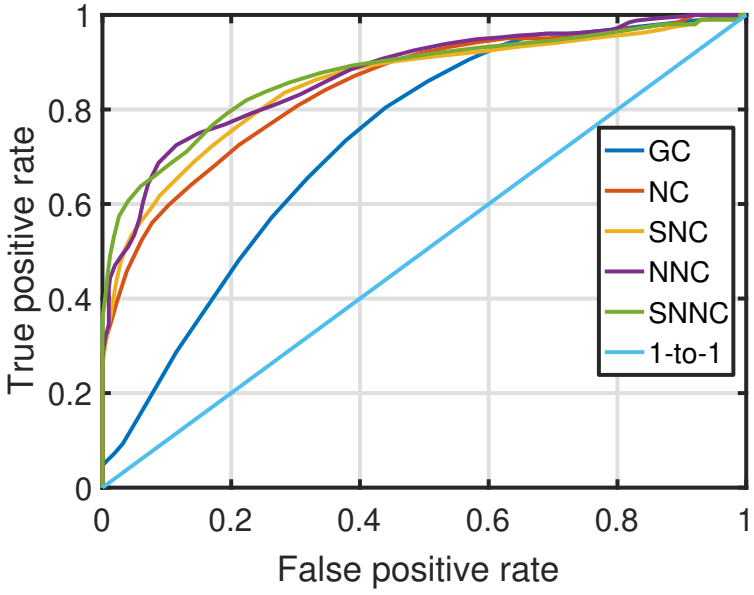


Figure 4.9: Receiver operating characteristic curves for 13.5Hz

4.8 Discussion of λ functions and preprocessing

The properties of the weighting function λ qualifies it as a probability mass function. In fact, the weighting function λ operates similarly to a probability mass function in Eq. (4.4). Since $\lambda_{pq}(\varphi_k)$ defines how much of the contribution of $a_{pk}\varphi_k$ should be attributed to x_p, this would be

equivalent of evaluating the expected value of the contribution attributed to assuming it has probability $\lambda_{pq}(\varphi_k)$ of being $a_{pk}\varphi_k$ and $(1 - \lambda_{pq}(\varphi_k))$ probability of being 0. Under the same rationale, λ^2 defines the indicator function of greatest entropy, which makes no *a priori* assumptions about the regressor functions.

One of the remaining challenges for the development of a unified nonlinear extension of NC is the choice of the "correct" function λ. This begs the question of what "true causality" and the purpose of causality analysis are. As both GC and NC are based on causality as defined by Hume [103], it is helpful to point out that Hume was concerned mostly with the epistemological aspect of causality, rather than an ontological one. Similarly, it would be naive to assert that signal x_q "causes" x_p as a matter of fact, without careful consideration of *a priori* knowledge. Similarly, the appropriate choice of λ relies on understanding what is the most useful manner to assign contributions given a particular set of regressor functions and the system being observed.

The simulations concerning the model from Eq. (4.21) show how standardizing the regressors changes the NC estimates. Additionally, due to the characteristics of the nonlinear model from Eq. (4.21), the Volterra filter had limited success at estimating the contribution of past values of x_1 to the current value of x_1, as x_1 did not have zero mean and, therefore, some of the contribution of x_1 was misattributed to x_2. Analogously to the choice of candidate regressor functions, the choice of λ function relies on careful consideration of the system being modeled.

Additionally, since the NC value is derived from models, it is important to distinguish the systems from which the data are gathered from the models used to represent them. For example, one could develop very accurate models to predict sunrise and sunset times without ever considering whether the sun still exists. For such models, GC and NC would suggest that the existence of the sun has no impact on sunrise and sunset times, an absurd conclusion. Instead, an epistemological interpretation of causality analysis yields more useful interpretations, the knowledge of the effects of the sun's inexistence does not increase the knowledge of sunset and sunrise times. This argument is similar to Box's commentary on the wrongness of all models [31]. Ultimately, the goal of causality analysis is to gain knowledge on systems given limited

information available about them. Therefore, the concern should not lie on which the choice of function λ is "right" or "wrong," but rather which ones lead to most "useful" conclusions about causal relationships.

4.9 Conclusions

New Causality is a promising method for assessing causality links between two or more signals. In the seminal definition [95] NC is defined only for LTI models. This limits the use of NC to systems that can be modeled well with LTI models. In this work, a novel extension of NC to NARMAX models is presented. Three methods for choosing the λ weighting function are shown, where the first two are formally defined and a suggestion is made for the implementation of a third, while allowing for alternate implementations. All three methods produce identical results to seminal definition of NC for ARMAX models.

Results show that this extension is suitable for systems that can be modeled well by NARMAX models, producing good results in the tested models. Particularly λ^2 has shown to produce adequate results even the nonlinear functions of the observation model are not part of the set of candidate regressor functions. In tests with EEG signals, SNNC was shown to outperform NC and GC in showing the linkage between α waves in Fp1 to β waves in O2.

Just as the seminal definition of NC, the nonlinear extension depends heavily on the estimated model. Thus, it is important to highlight that careful selection of model topology and model parameter estimation is essential to obtain useful NNC estimates.

The function λ has been shown to be a probability mass function. For each suggested λ, the weights are governed by different assumptions about the distribution of "causal strength." Although λ^2 has shown promise in this work, models with non-antisymmetrical properties or regressors with non-zero means can induce shifts in the NC values. However, the seminal definition of NC is also sensitive to data preprocessing, as it pertains to modeling more than causality analysis.

This extension of NC to NARMAX models adds flexibility to NC to assess causality strength to any signals that can be well modeled with LTIiP models. The extension inherits the strengths of

NC, while also having the same requirement of accurate model topology and parameter estimation in order to produce "useful" NC values. The choice of λ function requires careful consideration, but is not unlike the choice of candidate regressor functions, in which *a priori* information about the system being modeled is used to guide the choice.

CHAPTER 5

IMPROVEMENTS TO THE EvolOBE METHOD FOR NONLINEAR CAUSALITY ANALYSIS

5.1 Overview

With the need for accurate modeling for NC analysis made clear,the focus of this chapter now shifts to a method of estimating nonlinear model structures and parameters. The current work is centered on a biologically-motivated method for both the selection of the effective regressors and the estimation of the parameters of modified NARMAX models. The approach integrates set-based parameter estimation and genetic algorithms for optimization over fitness measures derived from a set of solutions [213]. A brief sketch of the overall approach appears in Sec. 2.4. This chapter is focused on innovations in the evolutionary process by which the model regressor set is selected.

As in any nonlinear identification solution, the evolutionary–set-theoretic framework described above is computationally-intensive, as the number of regressors increases factorially with the order of the nonlinear expansion. In a general sense, this chapter addresses the need to find more efficient data-processing algorithms for brain modeling. A more efficient solution is based in the expected sparsity of the connectivity models in terms of the relatively low number of regressors that would be necessary to effectively characterize nonlinear relationships in time-series records. This assumption has significant implications for the evolutionary search over the space of regressor combinations.

In particular, modified crossover and mutation operators are incorporated in the NSGA-II [51] framework to expedite feature (regressor) selection. By adjusting the mutation and crossover operators to account for sparsity and pairwise relationships in the population, the number of generations needed to arrive at the solution is greatly reduced.

Further technical details of the operation of the model are found in previous papers [213, 214,

216]. Some portions of this chapter are quoted directly from [149] with a few modifications for improved flow and clarity.

5.2 Model form

The goal of the identification strategy in this work is to obtain a model whose internal mechanism mimics the system being studied. Note that unless *a priori* information is available, the similarity between the internal mechanism and the system cannot be measured, but instead, predictive power is often used as a surrogate measure of similarity.

The internal processing of the system is based on a subset of a ***candidate set*** of nonlinear regressor functions, $\Xi_\varphi = \left\{ \varphi_q \right\}$, of size $\left| \Xi_\varphi \right|$. Each regressor is a mapping $\varphi_q : \mathbb{R}^{\mathcal{M}N_s} \longrightarrow \mathbb{R}$. The identification strategy starts by positing that, given the appropriate candidate set, there exists a LTIiP observation model, $\mathbb{O}_{\boldsymbol{a}^*,\boldsymbol{\varphi}^*}$, of the form in Eq. (2.23) for $n \in \mathbb{Z}$, given by

$$\begin{aligned} \mathbb{O}_{\boldsymbol{a}^*,\boldsymbol{\varphi}^*} : x_p[n] &= \sum_{k=1}^{K^*} a_{pk}^* \varphi_{pk}^* \left(x_1\big|_{n-\mathcal{M}}^{n-1}, x_2\big|_{n-\mathcal{M}}^{n-1}, \ldots, x_{N_s}\big|_{n-\mathcal{M}}^{n-1} \right) + \sum_{k=1}^{K_\epsilon} b_{pk}^* \phi_{pk}^* \left(\epsilon_p^*\big|_{n-\mathcal{M}}^{n-1} \right) + \epsilon_p^*[n] \\ &\doteq \boldsymbol{a}_p^{*T} \boldsymbol{\varphi}_p^*[n] + \epsilon_p^{**}[n] \end{aligned} \tag{5.1}$$

where

$$\epsilon_p^{**}[n] = \sum_{k=1}^{K_\epsilon} b_{pk}^* \phi_{pk}^* \left(\epsilon_p^*\big|_{n-\mathcal{M}}^{n-1} \right) + \epsilon_p^*[n] \tag{5.2}$$

with $\boldsymbol{a}^* \in \mathbb{R}^{K^*}$, and ϵ^{**} an error sequence representing uncertainties in the model. The "$*$" subscript indicates a "true," but unknown, quantity associated with the observation model. [1] The arguments, $x_{-\infty}^{n}$ and $y_{-\infty}^{n-1}$, of the regressor signals φ_q (or vector $\boldsymbol{\varphi}$) indicate that a finite number of elements is selected from the subsequences $\{x_1[n-1], x_1[n-2], \ldots, x_1[n-\mathcal{M}], x_2[n-1], \ldots, x_2[n-\mathcal{M}], \ldots, x_{N_s}[n-\mathcal{M}]\}$ by each φ_q for processing at time n. For conservation of space, we define the vectors of $\mathcal{M}N_s$ signal samples used at time n by $\boldsymbol{u}_{q*}[n]$, and the matrix $\boldsymbol{U}_*[n] = \left[\begin{array}{cccc} \boldsymbol{u}_{1*}[n] & \boldsymbol{u}_{2*}[n] & \cdots & \boldsymbol{u}_{K^*}[n] \end{array} \right]$. Given observations of x and y sufficient to compute outputs on time interval $n = 1, 2, \ldots, N$, we pose an ***estimation model*** as a function of the parameters

[1] To avoid cumbersome notation, it is to be understood that φ_{q*} is the q^{th} element ***selected from*** Ξ_φ, rather than element q of Ξ_φ.

and regressor signals,

$$\mathbb{M}_{\boldsymbol{a}_p,\boldsymbol{\varphi}} : \hat{x}_p\left(n, \boldsymbol{a}_p, \boldsymbol{\varphi}\right) = \sum_{k=1}^{K} a_{pk}\varphi_{pk}\Big(\boldsymbol{u}_q[n]\Big) \doteq \boldsymbol{a}_p^T\boldsymbol{\varphi}_p\Big(\boldsymbol{U}[n]\Big), \tag{5.3}$$

in which each φ_q is drawn from the set Ξ_φ (see footnote 1), $\boldsymbol{a} \in \mathbb{R}^K$, and the $\boldsymbol{u}_q[n]$ and $\boldsymbol{U}[n]$ are defined similarly to $\boldsymbol{u}_{q*}[n]$ and $\boldsymbol{U}_*[n]$. The circumflex in $\hat{x}$ connotes "prediction" , as this estimation model corresponds to the classical prediction-error method (e.g., [128]). This is true even though the regressor functions can be highly-nonlinear functions of the observations, because (when assumed fixed in the model) they appear in a model that is linear-time-invariant-in-parameters (LTIiP). Thus, the identification of the parameters using least square errors or (theoretically) mean-squared-error techniques is a well-known problem. Our approach, however, involves a distinctly different identification method which produces parameter solution ***sets*** rather than point estimates (e.g., [54, 55]). It is the properties of these sets that couple the model creation and parameter identification problems.

5.3 Identification strategy

The EvolOBE method combines the strengths of evolutionary computing and more traditional set-theoretic parameter estimation methods to robustly obtain a family of models with different tradeoffs between accuracy and model complexity. The evolutionary algorithm is responsible for finding the subsets of regressor functions $\boldsymbol{\varphi}_p$ out of Ξ_φ, whereas the set-theoretic parameter estimation method uses of the selected $\boldsymbol{\varphi}_p$ to obtain $\boldsymbol{a}_p$. This framework simultaneously addresses selection of the model structure and the parameter estimation. Moreover, a very significant advantage of the algorithm is the lack of need for assumptions about stationarity or distributional characteristics of the noise. The specifics are outlined in the following paragraphs.

Candidate models are encoded as binary chromosomes, where each possible phenotype represents a model with different regressor functions. The chromosome is a binary sequence in which the q^{th} gene represents the presence or absence of the q^{th} regressor function. The information encoded in the chromosomes is used to generate the regressor functions which are fed to the OBE algorithm, which obtains a feasibility set according to data. The set properties are then used to

assign fitness values to each chromosome, and the fitness value is used in the genetic algorithm selection process to evolve the population toward better solutions (e.g. [167]). This fitness measure can be in the form of a single objective function that provides a summary of the quality of the model, such as FPE and AIC, or in the form of multiple objective functions, covering predictive accuracy, model complexity, and other information about the candidate model (such as the volume or sum of the semi-axes of the ellipsoid). The assigned fitness measure regulates the chance of survival of each particular model in a generation.

The algorithm starts with a random population of chromosomes. At each step, the population is evaluated, then a subset of the population is selected to generate children through mutation and crossover operations. Mutation operators work by randomly selecting genes and altering them, whereas crossover operators combine portions of the chromosomes of two or more parents to produce a new offspring, which are added into the population. The population is sorted and individuals with lower fitness are discarded. The specific mechanisms for mutation and crossover operations, as well as the selection of parents, sorting of the population, and survival criterion are often tailored for a particular application.

To reduce the computational complexity of this process, the search space of regressor models must be controlled, and the candidate and final models must use the fewest regressors that are consistent with an objective of prediction-error minimization. Since minimizing the prediction error and minimizing the number of regressors are conflicting objectives, a multi-objective optimization approach is desired. For this work, the NSGA-II [51] approach is adopted, since it generates a set of solutions (ideally the Pareto-front), providing the best solution for a given number of regressors and allowing the model with the best trade-off to be chosen.

5.3.1 NSGA-II

NSGA-II is a standard algorithm for solving multiobjective optimization problems. It requires a small number of parameters and is able to obtain solution sets spread along the pareto-front. It is especially appropriate for problems with only two objectives. The basic NSGA-II algorithm is

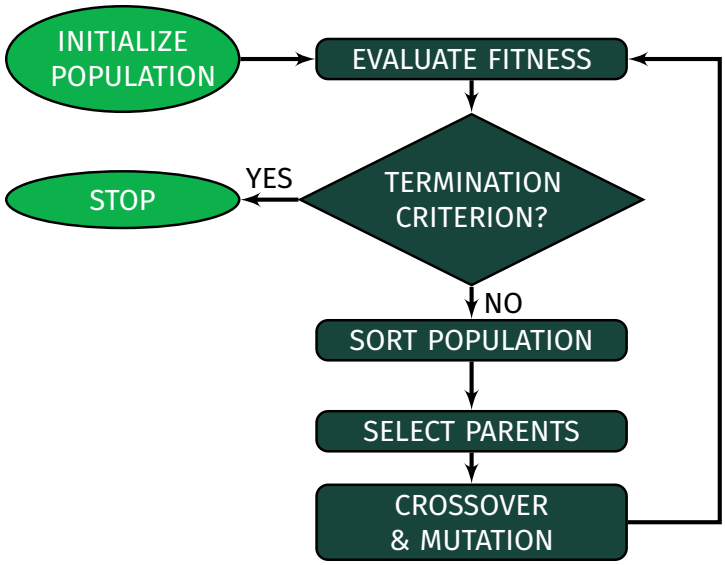


Figure 5.1: NSGA-II algorithm summary

shown in Fig. 5.1. In the original NSGA-II paper [51], Deb *et al.* use binary tournament selection, bit-wise mutation and single-point crossover with probability of $p_c = 0.9$, and mutation probability $\mu = 1/\ell$ (where ℓ is the length of the chromosome). In this work, these operators and parameters are used as a baseline for comparison, with the exception of the single-point crossover operator, which is replaced by a two-point crossover operator.

5.3.2 Asymmetric mutation operator

For sparse solutions, the mutation operator can be tuned to guide the population toward sparsity. Although judicious selection alone can effect sparse solutions, a properly tuned mutation operator can increase the convergence rate significantly. Here, an asymmetric mutation (AM) operator is developed. Classic mutation operators use a fixed probability to flip each chromosome regardless of its previous value. This is effective for blind exploration, but imposes pressure toward solutions with 50% active genes.

For a given μ probability of mutation, the expected number of active (N_1) and inactive genes

(N_0) at step $n + 1$ is given by

$$\begin{bmatrix} N_1^{n+1} \\ N_0^{n+1} \end{bmatrix} = \begin{bmatrix} (1-\mu) & \mu \\ \mu & (1-\mu) \end{bmatrix} \begin{bmatrix} N_1^{n} \\ N_0^{n} \end{bmatrix} \tag{5.4}$$

This matrix has eigenvalues 1 and (1-2μ). The eigenvector for 1 is $[\ 1 \quad 1\ ]^T$, which means that, in the absence of selection operators, the number of active and inactive genes tends to equality at a rate depending on μ .

An asymmetric mutation operator can be used to achieve any desired rate of activation. Two distinct mutation operators are introduced to implement this effect: μ_{10} the probability of deactivating an active gene, and μ_{01} the probability of activating an inactive gene. The matrix system (5.4) becomes

$$\begin{bmatrix} N_1^{n+1} \\ N_0^{n+1} \end{bmatrix} = \begin{bmatrix} (1-\mu_{10}) & \mu_{01} \\ \mu_{10} & (1-\mu_{01}) \end{bmatrix} \begin{bmatrix} N_1^{n} \\ N_0^{n} \end{bmatrix} \tag{5.5}$$

The eigenvalues of this system are 1 and $1 - \mu_{10} - \mu_{01}$ with corresponding eigenvectors $[\ \mu_{01} \quad \mu_{10}\ ]^T$ and $[\ 1 \quad -1\ ]^T$. The desired ratio of active to inactive genes is given by

$$r_d = \frac{\mu_{01}}{\mu_{10} + \mu_{01}} \tag{5.6}$$

For this scheme, the mutation rate is defined as

$$\mu = r_c\ \mu_{10} + (1 - r_c)\mu_{01} \tag{5.7}$$

where r_c is the ratio of active to total genes (i.e. $N_1/(N_1 + N_0)$). By combining Eqs. (5.6) and (5.7), the following expressions for the mutation probabilities are obtained

$$\begin{aligned} \mu_{01} &= \frac{\mu\ r_d}{r_c + r_d - 2r_c r_d} \\ \mu_{10} &= \frac{\mu(1 - r_d)}{r_c + r_d - 2r_c r_d} \end{aligned} \tag{5.8}$$

This extended solution reduces to that for the traditional mutation operator when $r_d = 0.5$. Decoupling the mutation probabilities yields a more flexible mutation operator with which pressure can be applied toward a desired sparsity level.

5.3.3 Reduced surrogate crossover

Evolution is improved by a crossover operator that generates novel individuals. A method to achieve novelty is to use reduced surrogate crossover (RSX) [30]. With RSX, only non-matching alleles are crossed between individuals. This is especially important as the genetic diversity decreases with evolution. Thus the likelihood of generating a novel individual from two similar parents becomes smaller in traditional two-point crossover operations.

A varying-minimum Hamming distance between chromosomes is suggested in [134]. In the present work, a fixed unity Hamming distance yielded small improvements in convergence speed. The fixed distance avoids the shortcomings of the minimum Hamming approach, but results in less efficient sampling of the search space.

5.3.4 Linkage tree crossover

One of the tenets for the convergence of genetic algorithms is that the population will shift from the initial randomly generated solutions into a population that increasingly has characteristics found in the pareto-optimal solution set. Under this assumption, the statistical characteristics of the population at a generation can be used to estimate what operations are more likely to produce helpful results.[2] Linkage tree crossover (LTX), introduced by Thierens in [192], crosses solutions over at positions that are more likely to generate fit offspring.

First, LTX collects information of the statistical characteristics of the population and clusters the genes into a binary tree that summarizes how clusters are linked together. Each cluster is initialized with a single gene, and clusters are then progressively linked together until all genes are included in a single cluster. The clustering uses a distance metric based on mutual information and entropy [114]. For clusters C_1 and C_2, the mutual information is computed as

$$I(C_1; C_2) \doteq \sum_{c_1 \in \mathfrak{C}_1} \sum_{c_2 \in \mathfrak{C}_2} p_{C_1,C_2}(c_1, c_2) \log \left(\frac{p_{C_1,C_2}(c_1, c_2)}{p_{C_1}(c_1) p_{C_2}(c_2)} \right), \tag{5.9}$$

[2] However, care must be taken not to heavy-handedly influence the evolution, as a stronger emphasis on exploitation is likely to diminish the ability of the GA for exploration.

where $\mathfrak{C}_1$ and $\mathfrak{C}_2$ are the sets of all possible values for C_1 and C_2, $p_{C_1,C_2}(c_1, c_2)$ is the joint probability of c_1 and c_2, $p_{C_1}(c_1)$ is the probability of c_1 and $p_{C_2}(c_2)$ is the probability of c_2. Alternatively, the mutual information may be computed using the entropies. The entropy for a cluster $C \in \mathfrak{C}$ is defined as

$$H(C) \doteq -\sum_{c \in \mathfrak{C}} p_C(c) \log\left(p_C(c)\right), \tag{5.10}$$

and using the following identity

$$I(C_1; C_2) \doteq H(C_1) + H(C_2) - H(C_1; C_2). \tag{5.11}$$

The distance metric is then defined as

$$D(C_1, C_2) \doteq 2 - \frac{H(C_1) + H(C_2)}{H(C_1, C_2)} = \frac{H(C_1, C_2) - I(C_1, C_2)}{H(C_1, C_2)}. \tag{5.12}$$

The general procedure for generating the linkage tree is found in Alg. B.4 in Appendix B. An example of a linkage tree is shown in Fig. 5.2. In the example, the order of the crossover operations would be combining φ_1, φ_2, and φ_4 from one parent and φ_3 and φ_4 from the other parent, then φ_1 and φ_4 with φ_2, then φ_3, and φ_5, and finally combining φ_1 with φ_4.

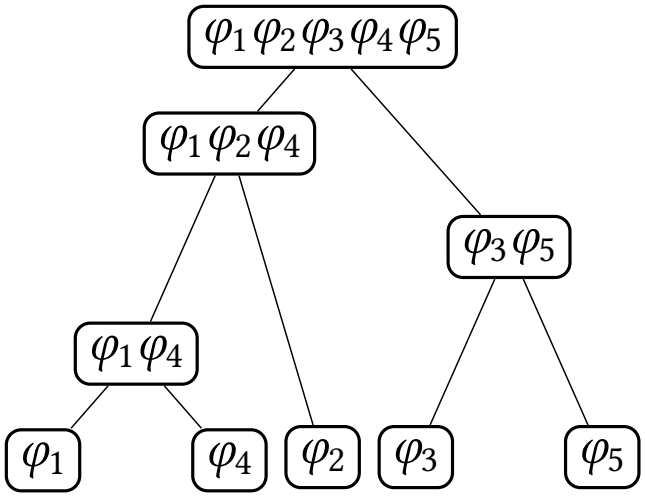


Figure 5.2: Linkage tree example

Once the linkage tree is generated, the algorithm traverses the tree executing crossovers exchanging the clustered genes. In the seminal algorithm, if at least one offspring is superior to

both parents, the parents are replaced by the children. When the tree is fully traversed, the best individuals are copied into the next generation. The detailed LTX procedure is shown in Alg. B.5 in Appendix B.

In this work, a special consideration is necessary, as LTX was not envisioned for multi-objective problems. For single-objective problems, a solution may either be superior, inferior or equivalent to a second solution, whereas for multi-objective problems, solutions may also be neither superior (dominate) nor inferior (dominated), but simply offer a different trade-off (*i.e.* superior in at least one objective function, but inferior in at least one solution). When the offspring and parents neither dominate nor are dominated by each other, there is the choice to keep or replace the parents or a stochastic combination of both. Preliminary tests showed no clear advantage of either choice, but further investigation on this topic is planned as future work. The detailed procedure for the use of LTX in multi-objective problems is shown in Alg. B.6 in Appendix B.

One possible downside of the use of LTX is the substantial computational cost of evaluating the large number of entropy calculations needed to construct the linkage tree [155]. For problems with fitness functions that are computationally costly, LTX is more advantageous. The overhead of computing the linkage tree is becomes more significant as population sizes increase, but small population sizes can produce poor estimates of entropy [199].

5.4 Results of AM and RSX

A randomly generated NARMAX model with five regressors was used to evaluate the modified operators. Three delayed outputs and two delayed inputs to the system were extracted as linear regressors and expanded to a 3$^{\text{rd}}$ order Volterra series, obtaining a total of 55 nonlinear regressors.

The estimated Pareto front is shown in Fig. 5.3. The ordinate shows the RMSE of the prediction error (dB scale) and the abscissa shows the number of regressors in each model. As expected, there is a knee located at five regressors, corresponding to the number in the generative model. There is some improvement in the RMSE for models with more regressors, but only due to overfitting.

To assess the improvement relative to the unmodified NSGA-II method, simulations were used

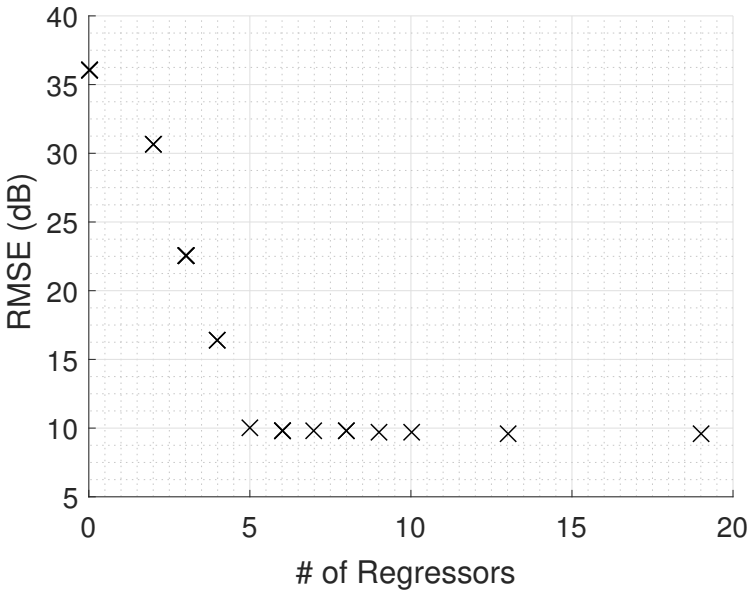


Figure 5.3: Estimated pareto front

to estimate the number of generations required for each genetic algorithm (GA) to converge to the generative model. The number of generations follows a probability distribution with parameters depending on the GA and its internal parametrization.

Clearwater *et al.* [45] have shown that the number of generations required by a GA to find a solution asymptotically approaches a log-normal distribution. Due to the long-tailed nature, the mean and variance of this distribution are both significant. A lower-variance estimator can provide a more meaningful measure of number of generations to convergence, even at the expense of mild estimator bias.

To examine the how well the number of generations required to find the solution fits a log-normal distribution, 16384 runs of our algorithm were evaluated using NSGA-II with the same parameters for each run. The number of generations required by each run was recorded and a log-normal distribution fitted to the data. Fig. 5.4 shows the histogram with "×" markers and the fitted distribution with the solid line. The fitted distribution tracks the histogram remarkably well, especially in the long tail of the distribution. For clarity, the histogram is omitted from further figures.

The NSGA-II algorithm was implemented using the symmetric bit-wise mutation operator

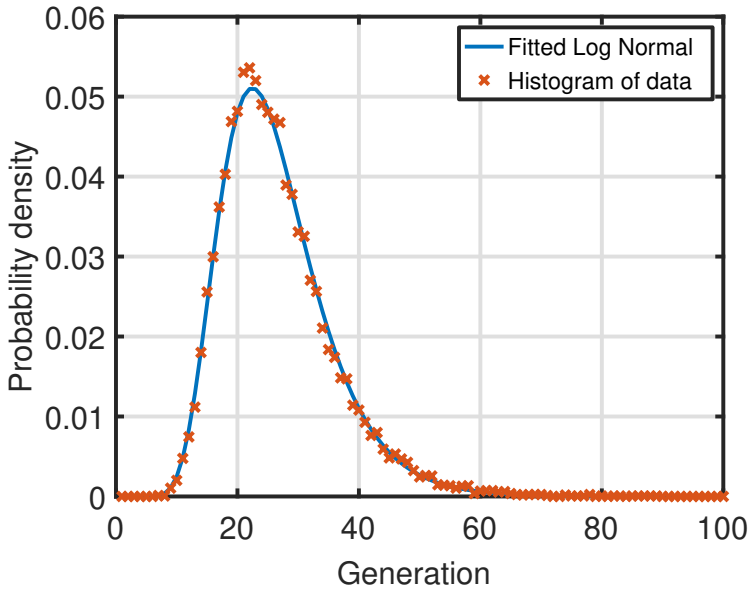


Figure 5.4: Histogram vs. Fitted distribution

(equivalent to $r_d = 0.5$) and two-point crossover. The asymmetric binary mutation with $r = 0.1$ and was applied to all remaining simulations. The modified domination criterion (unique sorting) was added to the third simulation onwards. The fourth simulation incorporated the RSX operator and the fifth added a minimum Hamming distance (HD) of 1 to the mutation operator. The results can be seen in Fig. 5.5. The parameters for the fitted models are compiled in Table 5.1. Since in the

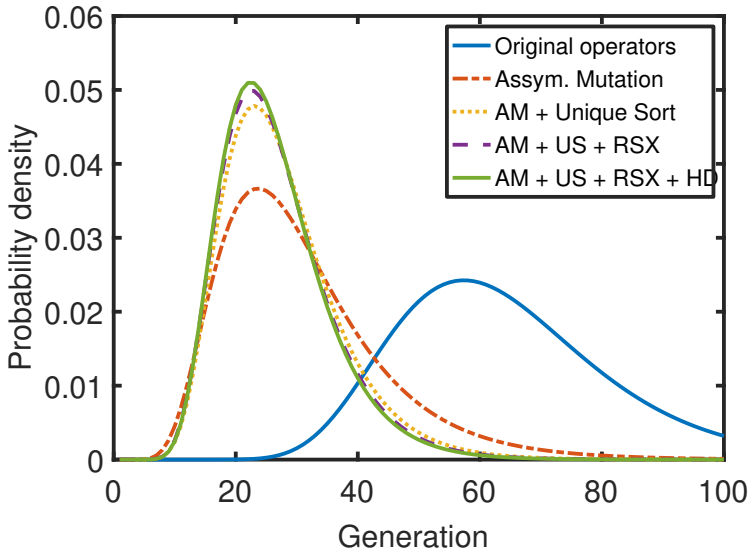


Figure 5.5: Generations to arrive at the desired model

log-normal distribution, μ and σ do not correspond to the mean and standard deviations, these

values are also calculated and shown in separate columns.

Table 5.1: Fitted parameters for different methods

Method	μ	σ	Mean	Std. Dev.
Original operators	4.13	0.28	64.3	18.1
Assymetrical Mutation	3.34	0.421	30.9	13.6
AM + Unique Sort	3.25	0.341	27.5	9.64
AM + US + RSX	3.23	0.335	26.7	9.19
AM + US + RSX + HD	3.22	0.329	26.4	8.95

The asymmetric mutation operator causes a drastic change in the simulation results, reducing both the mean and standard deviation significantly. It reduces the number of generations to reach 99% confidence from 118 to 76 generations (reduction of 35%) and 99.9% from 145 to 104 generations (a reduction of 28%).

While the modified domination criterion caused a smaller reduction in μ, the σ is reduced more significantly. As seen in the graph, the modes are largely unchanged (ranging from 22 to 23 generations), but the reduction in σ greatly reduces the number of generations to reach high confidence. The old sorting algorithm passes 99% confidence at 76 generations and the 99.9% confidence region at 104 generations, while the new sorting algorithm passes 99% confidence in 58 (a further reduction of 23%) and 99.9% in 75 generations (a further reduction of 27%).

The remaining improvements improve convergence, albeit in a smaller scale, with the reduced surrogate without minimum Hamming distance and the reduced surrogate with minimum distance of 1 needing 55 and 54 generations to reach 99% confidence respectively and needing 71 and 70 generations to reach 99.9% confidence.

5.5 Results of LTX

To test the improvement given by LTX, a test with the nonlinear observation model of Eq. (4.23) was run under various conditions. The advantage of the tested model over the previous model is that the tanh(·) term in the difference equation means polynomial estimation models will provide better prediction with increasing model orders, but no finite set of polynomial regressor functions will manage to perfectly represent the tanh(·) term. The comparison tests were done against the

results in the previous section under similar conditions as the previous test.

One challenge shown in the literature is that regression algorithms will adapt to the absence of certain regressors functions by using other correlated regressor functions, regardless of their presence in the observation models [23, 157]. This choice is further complicated by the presence of noise in the measurements. When the variance of the contribution is comparable with the variance of the noise, then discerning the optimum parameter values or even the optimum regressor sets for large numbers of regressors is not always possible.

Fig. 5.6 shows the estimated set of best regressors for models with seven or fewer regressor functions, where dark squares indicate the presence of a particular regressor function in the model. The RMSE of the prediction error is shown in Fig. 5.7. The regressor sets were obtained using the EvolOBE method and a realization of Eq. (4.23) of 1024 consecutive epochs with 15dB SNR, with two delay taps (exact value) and the linear regressors were expanded to a polynomial order of ten, which results in a total of 1000 candidate regressor functions. Note how $u^6[n-6]$ is present for the model with two regressors, even though it is not present in the observation model. In fact, in this realization, replacing $u^6[n-6]$ with either $u^2[n-1]$ or $u[n-1]u[n-2]$ yields slightly worse RMSE than $u^6[n-6]$, where the model containing $u^6[n-6]$ has -11.89dB and the ones containing $u^2[n-1]$ or $u[n-1]u[n-2]$ have -11.88dB and -11.38dB, respectively. This can be interpreted as $u^6[n-6]$ being able to better fit the missing terms than either $u^2[n-1]$ or $u[n-1]u[n-2]$ individually, given the distributional characteristics of $u[n-1]^2$ and $u[n-1]u[n-2]$ and the particular realization being used. However, $u^2[n-1]$ and $u[n-1]u[n-2]$ are synergistic in the sense that together they provide better fit to the data than the combination of $u^6[n-1]$ and any other regressor function, as shown by the presence of both regressor functions in all subsequent models.

Also in Fig. 5.6, note that other terms of the McLaurin series of $\tanh(y[n-1])u[n-1]$ are present, such as $y^3[n-1]u[n-1]$ and $y^5[n-1]u[n-1]$, but other terms provide so little improvement in the prediction that even though they might be present (*e.g.*, $y^7[n-1]u[n-1]$) in a later model, they appear in conjunction with spurious regressor functions (*e.g.*, $y[n-1]y^8[n-2]u[n-2]$). The same is true for $y^9[n-1]u[n-1]$, which appears in the estimated best model with eight regressors (not

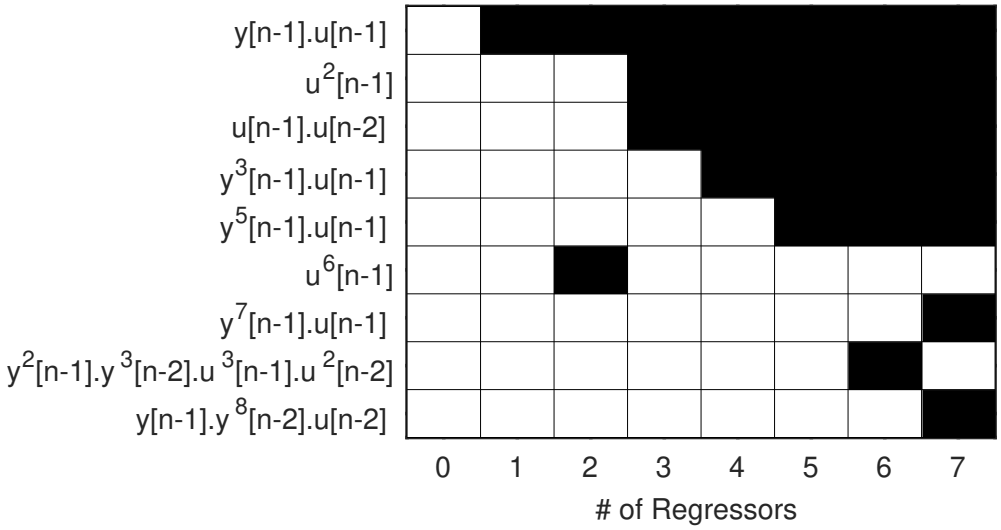


Figure 5.6: Estimated regressor functions present in best models

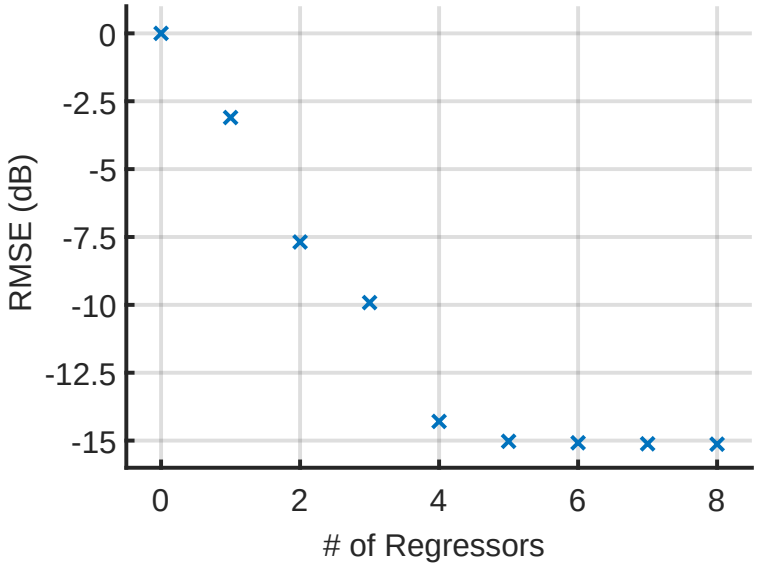


Figure 5.7: Estimated pareto-front for 15dB SNR

shown in Fig. 5.6 for clarity), which also includes the spurious term $y^2[n-1]y^2[n-2]u^3[n-1]u[n-2]$.

While the challenge of appropriate choice of regressor functions for nonlinear estimation models is unique to modeling nonlinear observation models, a similar challenge exists when there are missing input signals in any (linear or not) regression problem. An intuitive example is given in Sec. 4.2, where the increase of a temporally correlated quadratic term of the past value of x_2 resulted in the decrease of the linear NC measure of x_2 into x_1, since this increase caused x_1 to be

more temporally correlated with itself, whereas the quadratic term in x_2 is uncorrelated with x_2.

The first test used a noise free realization of Eq. (4.23) of 1024 consecutive epochs, with two delay taps (exact value) and the linear regressors were expanded to a polynomial order of eight, which results in a total of 494 candidate regressor functions. The histogram of the number of evaluations needed to find the best models with eight or fewer regressor functions is shown in Fig. 5.8 and the fitted log-normal probability density distribution is shown in Fig. 5.9. The number of evaluations needed was reduced by 73% for 99% confidence and 79% for 99.9% confidence.

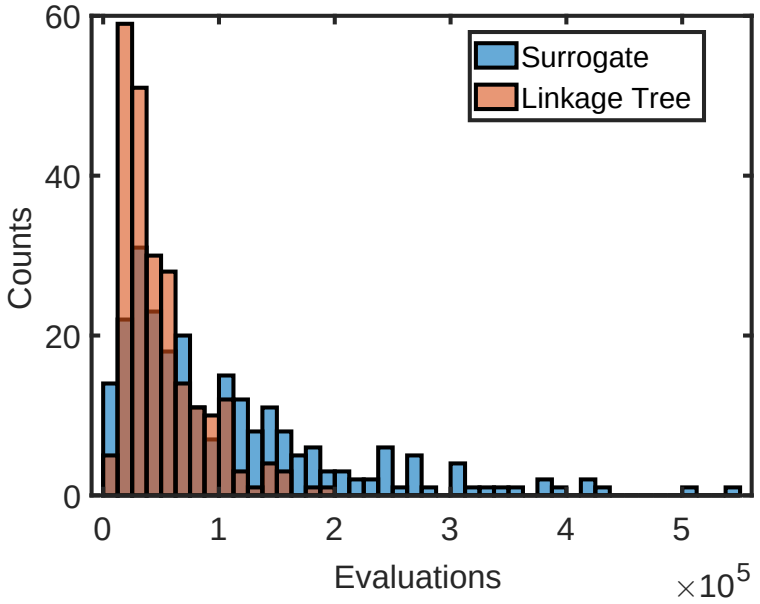


Figure 5.8: Histogram of required evaluations for RSX and LTX

It is important to note, that there are times where LTX performs worse than RSX in terms of required number of evaluations. Figs. 5.10 and 5.11 show the estimated CDF of the number of required evaluations. Fig. 5.10 demonstrates that LTX clearly outperforms RSX under most circumstances. However, looking closely in the region of fewer than 15,000 evaluations (shown in Fig. 5.11), RSX outperforms LTX in about 3% of cases. This behavior can be traced back to the assumptions of LTX, that the current population contains characteristics of the desired solutions. In the first few evaluations, this assumption is less valid and leads to some runs requiring more evaluations to find the desired solutions. There are mitigation measures to avoid this increase in evaluations and to hasten the search overall, like waiting for a few generations prior to switching

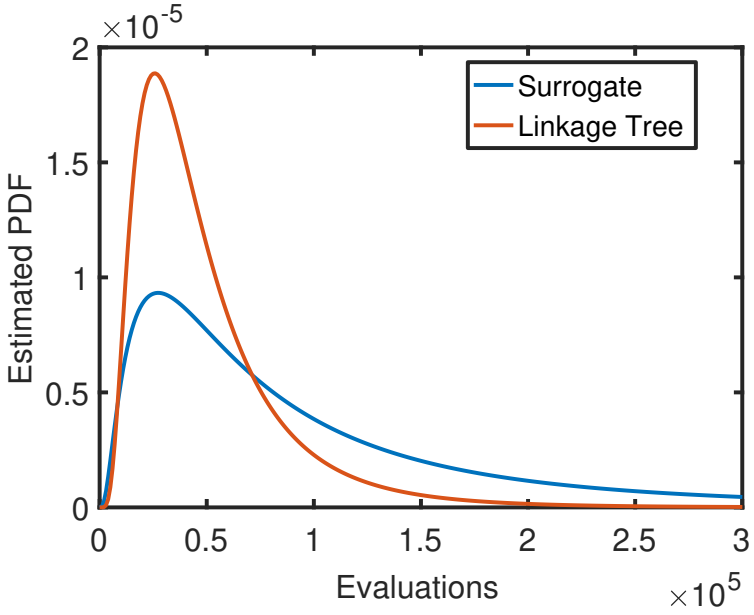


Figure 5.9: Fitted PDF for the required evaluations

to LTX or using a greedy or suboptimal algorithm to find the initial candidate solution set that is fed into the LTX operator, such as the bitwise hillclimber algorithm employed in the seminal LTGA paper [192]. A review of LTGA variants is given by Goldman and Tauritz in [73].

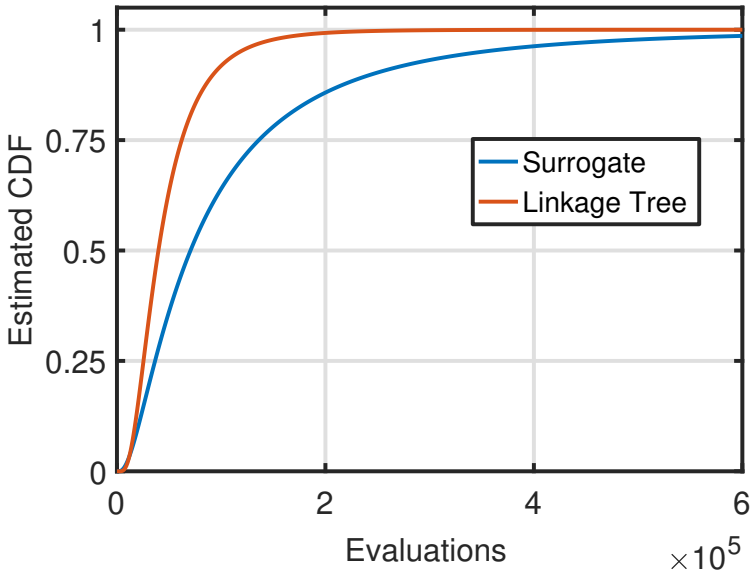


Figure 5.10: CDF for the required number of evaluations to find the desired solution

For simulations with lower SNR, the improvements are less pronounced. Fig. 5.12 shows the estimated CDF for the required number of evaluations for RSX and LTX. The CDFs intersect at

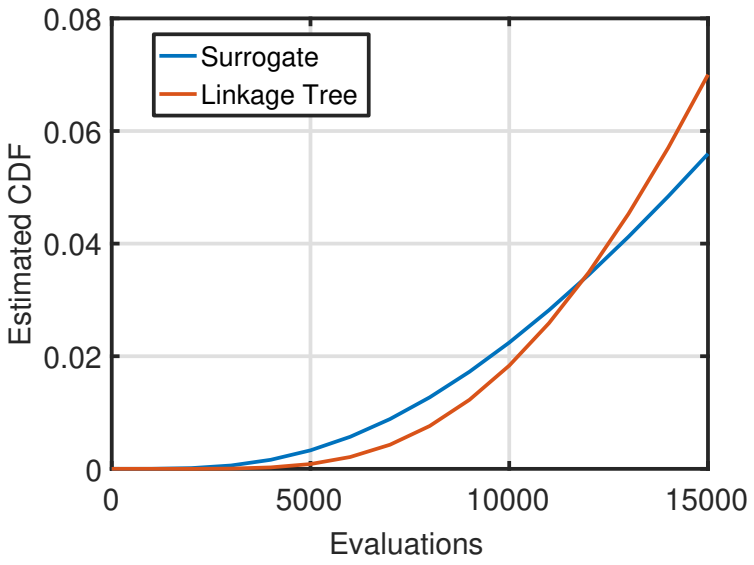


Figure 5.11: Close-up for fewer than 15000 evaluations

94.7%, before which RSX requires fewer evaluations. At 99% and 99.9% confidence levels, LTX still outperforms RSX, with 7.5% and 15% respective reductions in the required numbers of evaluations. Nevertheless, this small reduction in number of required evaluations is negated by the additional computational cost of computing the linkage tree. The overhead added by LTX is not negligible and increases with population size and number of regressor functions. This drawback is less evident when the cost of evaluating the fitness function is large in comparison to the computation of the linkage tree.

With lower SNR, nonlinear terms with smaller contributions are obfuscated by the noise and thus the final set of candidate solutions cannot include these terms with certainty. Finding this smaller set of solutions requires less exploration and gives LTX less chance to improve the search. Fig. 5.13 shows the comparison between the estimated Pareto fronts for the dataset expanded to polynomial order 10. Note how the noise free case continues to improve significantly until seven regressors are added, while at 15dB SNR, the improvements are greatly diminished beyond four regressors. The prediction error is not reduced beyond seven regressors, as the search space was limited to tenth-order polynomials, with the following term ($y^{11}[n-1]u[n-1]$) requiring an expansion to order 12, increasing the chromosome sizes to 1819 genes and the search space to over 10^{547} possible solutions.

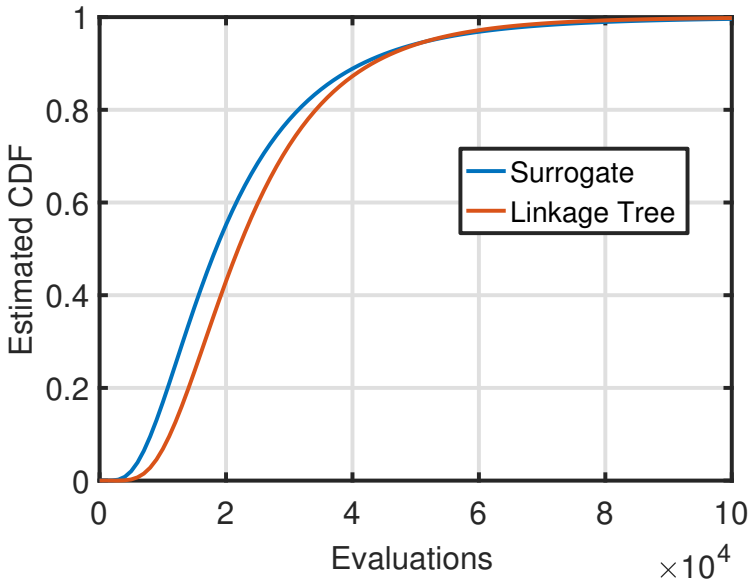


Figure 5.12: CDF for the required number of evaluations for 15dB SNR

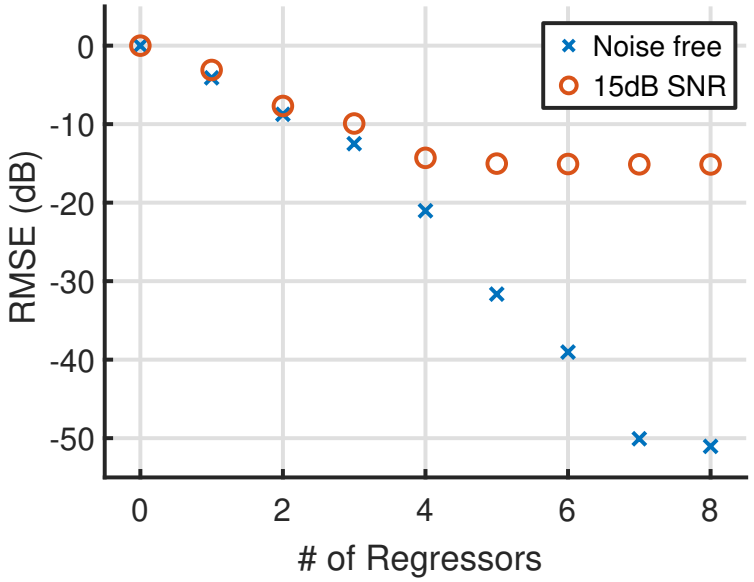


Figure 5.13: Comparison between estimated pareto fronts for different SNR values

5.6 Application to NNC analysis

Once the final set of models is obtained, these models can be used for NNC analysis. One of the advantages of the biobjective optimization approach is that at the end of the optimization process, the algorithm provides the set of best models for different levels of tradeoff between complexity

and predictive power.

In Fig. 5.14, the NNC values are given for the observation model of Eq. (4.23) for 10dB SNR. The NNC values using the observation model (found in Table 4.8) are 0.29 and 0.55. Note that the values converge quickly, with the model with four regressors being very close to the expected values and negligible changes with five or more regressors.

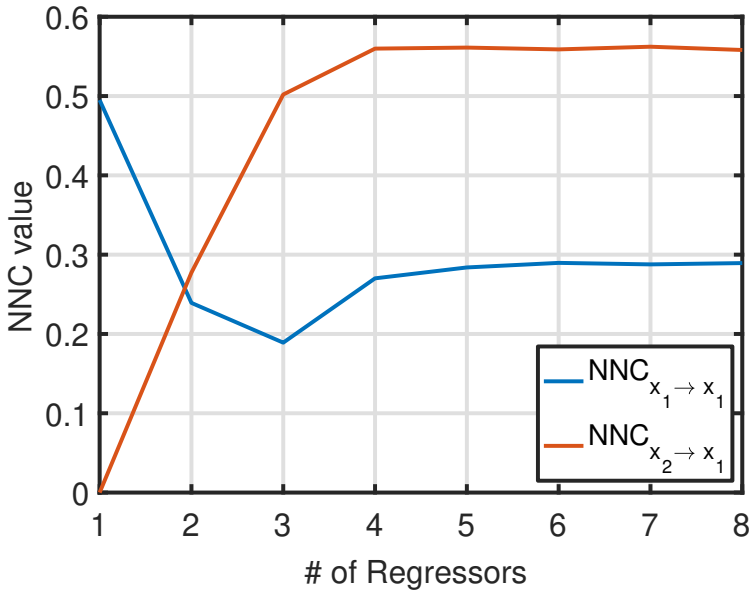


Figure 5.14: NNC values for the final candidate model set for 10dB SNR

In Fig. 5.15, the NNC values are given for the same model and 50dB SNR. The NNC values using the observation model (found in Table 4.9) are 0.31 and 0.69. Again, the values converge quickly and do not diverge in the observed range, as the contributions from the higher order terms are small compared to the first four chosen regressors.

Note that in situations where the observational model possesses large SNR, the GC value would increase with the increase of the polynomial order expansion, even when the contribution of the new terms is small, finally converging when the variance of residual is comparable to that of the noise. On the other hand, a small residual causes the sum of NNC values to approach unity, but Figs. 5.14 and 5.15 shows that the NNC values do not vary much when the residuals become smaller.

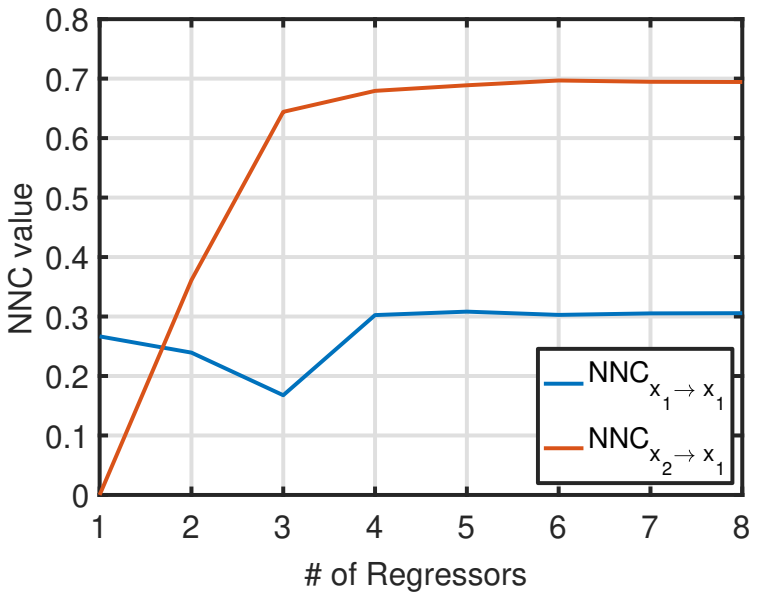


Figure 5.15: NNC values for the final candidate model set for 50dB SNR

In the studied cases, NNC performs well with simpler estimation models, provided the estimated parameters well represent the internal mechanism of the observation models. This echoes Ljung's advice to "try simple things first" [128], even though those models are "wrong[3]."

5.7 Discussion and conclusions

In this chapter, modified crossover and mutation operators were presented for use in NARMAX model estimation as part of ongoing improvements to EvolOBE. These modifications yield significant performance improvements over the pure NSGA-II algorithm for this application. The operators take advantage of posited characteristics of the population and final solution sets, such as sparsity and pairwise relationships between genes.

When a modeling problem provides little guidance on the selection of an effective model form, a GA must search a wide space of candidate features but determine a reliable and consistent solution in a limited number of generations. The 99% and 99.9% confidence metrics resulting in the modified search methods provide a stronger measure of performance than the estimated mean and variance, even though the confidence information is theoretically inherent in the two statistics.

[3]but "useful" [31].

The asymmetric mutation operator guides the mutation towards arbitrarily sparse solutions for any desired mutation rate. Tests have shown the asymmetric mutation operator to be an effective way to reduce the number of required evaluations. The change in the mutation operator also does not prevent exploration, as it simply increases the probability of the mutation to produce an offspring with the set sparsity, but does not prevent the mutation from generating offspring with other sparsity in the genes.

The modified crossover operators increase the search speed by finding valid crossover locations (RSX) or finding more crossover masks that are more likely to produce fit offspring (LTX). The LTX operator estimates pairwise proximity between genes in the current population to define crossover masks. In the simulations presented here, LTX required fewer evalutions to find the desired set of solutions for high confidence levels, but a small percentage of simulations completed faster when using RSX exclusively. Since LTX requires that the population to provide useful information on linkage between genes, LTX requires a minimum number of evaluations until the population can provide such information. More complex crossover operators that make use of information beyond just pairwise linkage, such as *covariance matrix adaptation* [85], are very powerful, but require even more evaluations until the crossover operator can perform appropriately.

At lower SNR values, the improvements given by LTX are less pronounced. Additionally, a larger set of regressor functions and the noise also adversely affect the parameter estimation, which further indicate the need for parsimony in the final candidate model set.

The resulting models were used to compute the NNC values of the models and compare them to NNC values obtained with the observation model parameters. Since NC and NNC are susceptible to error in the model estimation, it is important to carefully consider the estimation models used to compute these measures. In the tests, the EvolOBE method produced NNC estimates with very good agreement with the theoretical values. As the EvolOBE method produces the set of most accurate models for any number of regressors, NNC can be estimated for the entire set of fittest models for comparison and analysis of the models.

At this point, some characteristics of the algorithm have not yet been explored. For example,

the sparsity parameter for the asymmetric mutation operator is currently fixed at the beginning of the operation, but could potentially be set dynamically by observing the population and/or evolution. The algorithm also does not regard the relationships among regressors (e.g. $y[n-1]$ and $y^3[n-1]$), which could potentially provide useful information that will likely result in further improvement, especially when modeling non-polynomial regressor functions with polynomial expansions.

CHAPTER 6

CONCLUSION

6.1 Overview

Causality analysis is a very important area of study, ranging from philosophy and econometrics to physics, neurology and engineering. The topic is highly debated and somewhat controversial. Indeed, a concise universal definition of causality or causality measures has not been reached. This work focuses on statistical methods of evaluating evidence of causality, rather than the philosophy of causality. This work possesses two synergistic goals: the characterization and development of a causality measure for nonlinear parametric models, and the investigation of an evolutionary search algorithm for sets of the best nonlinear parametric models for different levels of tradeoff between complexity and predictive power. NC is shown to be sensitive to parameter estimation error and prone to bias, which is compounded when extending NC to nonlinear models, so a method of finding and comparing models complements NNC by using the optimum set of models for NNC estimation.

NC is a recent method to assess causality between signals in parametric models. In this work, a thorough critical study and nonlinear extension to NC are shown. In Summary, NC does have advantages over GC and similar causality measures in that it is more proportional to internal model parameters, it is normalized and does not require a choice on the order of the conditioning signals unlike CGC [95]. In Ch. 4, the seminal definition of NC is extended to cover all LTIiP models with a flexible weighting method that reduces to the seminal definition for LTI models.

This work also explores aspects of NC that have been overlooked in the the seminal papers. In much of the literature surrounding causality, the distinctions among systems, observation models and estimation models are often not clearly stated. Although very powerful methods for parameter estimation exist, estimated models are not the systems they represent and should not be taken as anything greater (or lesser) than that - a representation. Under the risk of repeating

a truism, "all models are wrong, but some are useful." As shown in Ch. 3, the usefulness of the NC estimates is strongly tied to the quality of the estimated models. This is arguably even more substantive for nonlinear model estimation, as nonlinear models entail increased difficulty in accurately estimating the parameters and selecting regressor sets.

Another aspect that is often overlooked is the validity of some models found in the literature. Models that are not representative of practical applications should not be used to compare causality analysis tools unless their use is justified. Sec. 3.2 contains a list of example models and a discussion on the validity of such models.

These two overlooked aspects in NC literature are very unfortunate, especially as it undercuts the argument for the unique characteristics of NC. The models shown in Sec. 3.2 could mislead a reader into thinking that NC is only superior in these impractical scenarios, which, without overlooking the merits of alternative methods, is not true in general. NC is unique in comparison to other methods in that NC values depend much more on internal model parameters and that, granted that the models represent the internal dynamics of the system well, it can better measure causal relationships for systems with quasi-periodic and slow dynamics.

In its seminal form, NC was only fully defined for ARX models. In Ch. 4, a nonlinear extension of NC was presented. For models with strong nonlinearities, the seminal form can behave counterintuitively as shown in Fig. 4.1. The extension presented in this work, NNC, produces results that are in line with intuition (shown in Fig. 4.2) and shares all the strengths of NC while allowing application to a much wider set of models. As is the case with NC (and GC), NNC can also be spectrally expanded into a frequency dependent measure. The definition of NNC also offers a flexible approach to partitioning the contribution of nonlinear regressor functions that depend on more than a single regressor signal. Tests were conducted on synthetic and real data with promising results.

With the need for a robust nonlinear model estimation framework having been demonstrated, improvements to the EvolOBE method are reported in Ch. 5. The EvolOBE method combines a genetic search algorithm for regressor selection with a set-theoretic approach for parameter

estimation. In this work, enhanced mutation and crossover are described and introduced to the EvolOBE method. The introduction of these new operators is shown to increase convergence speed, decrease the number of evaluations needed for convergence and reduce the variance of the number of evaluations needed for high confidence rates.

6.2 Contributions

The major contributions of this work are the following:

1. Shown that NC is susceptible to two sources of variation, natural variations in the specific realization (*e.g.*, difference between the sample variances and the observational model variances) and parameter estimation errors. In the sames tests, GC was shown to be significantly more robust to errors in the parameter estimation;

2. Shown that NC is prone to bias in the estimates that increase with parameter estimation errors;

3. Analytically explored four cases of the source of bias in NC estimates including regularization;

4. Provided an extension to NC to the set of all LTIiP models, which are considered interpretable and transparent [200]. This enables the use of NC to a much wider range of applications. The extension is equivalent to the seminal definition for linear models and can be spectrally expanded in the same way as the seminal definition. The extension is applied to real data (EEG signals) with encouraging results;

5. Introduced new operators into the EvolOBE method that significantly reduce the computation time and required number of evaluations to reach convergence;

6.3 Future Work

The large improvements seen in the EvolOBE method are encouraging and also indicate that further improvements are possible. Particularly, further enhancements in mutation and crossover

operators are likely to yield significant benefits to the genetic search.

In the most current variant of the EvolOBE method is currently blind to the particular relationship between regressor functions. When using Volterra expansions of the regressor signals on signals whose observational model has non-polynomial nonlinear terms, the set of optimal regressor functions are often related by the regressor signals used. Implementing a method to account for these relationships is likely to further improve convergence.

In its current form, the multi-objective adaptation to LTX treats all situations where the offspring neither dominate or are dominated by the parents by randomly selecting whether to keep the offspring or parents. Using a different heuristic to guide the choice of whether to choose offspring or parents might help speed up the genetic search.

Also, candidate models with larger number of active genes require longer computations than models with fewer active genes. Currently, the algorithm waits until all candidate models are evaluated to proceed, reducing the computational efficiency in parallel computing environments. Enhancements in the computational efficiency are possible and have not been explored.

Additionally, the set-membership parameter estimation provides other indicators of set quality, such as bounds for each parameter, or size and shape of the final ellipsoid. These indicators have not been studied yet as a complement or substitute for the currently used fitness functions.

Nonlinear NC is a new technique and its application has not yet been fully explored. The ability to describe the effect of a single regressor into the regressand in a complex function could have application areas outside of causality analysis, such as multi-criterion decision making. The normalized nature of NNC and sensitivity to changes in the model parameters make it particularly suitable as the results are more easily interpretable.

While the susceptibility of NC to bias in the estimates, a detailed statistical characterization of NC have not yet been explored. This could lead to enhanced significance tests for NC and better understanding of how it relates to other causality analysis tools.

APPENDICES

APPENDIX A

DERIVATION OF CLOSED-FORM EXPRESSIONS FOR GC AND NC FOR FIRST-ORDER BIJOINTLY REGRESSIVE OBSERVATION MODELS

A.1 Overview

Closed-form solutions for the GC and NC measures are useful in evaluating relative performance of the techniques. In [95], closed form expressions for GC and NC are derived for certain first-order ARX observation models. However, no general formula is given for NC or GC, and GC is only asymptotically evaluated for large M. Closed form expressions for GC depend on M and the process of obtaining closed form expressions laborious, but understanding the intricacies of GC and NC provide insight into what each technique measures.

In [95], it is argued that, GC does not depend on the feedback loop formed from the product of a_{21} and a_{12}, reflecting a coupling between x_1 and x_2. The argument is supported by a closed form expression given for GC in [95, Eq. 13] for a particular form of first order ARX model which does not include any term that depends on the product. However, this expression is only true if GC is allowed to compare models with unlimited order. When the model orders are finite, the expression for GC does depend on $a_{21}a_{12}$.

A large portion of this appendix is quoted directly from [147] with a few modifications for improved flow and clarity. *Long equations are placed at the end of the appendix.*

A.2 Derivations

In order to increase clarity, the time-delay index i superscript will be omitted. For all first-order models, $a_{pq}^{i} = 0$ for $i > 1$, so a_{pq} is used to mean a_{pq}^{1}. Instead, in this section, the superscript will be used to denote the exponent. The GC and NC measures can be evaluated in both directions (*i.e.*, $x_1 \longrightarrow x_2$ and $x_2 \longrightarrow x_1$). For simplicity, only the $x_2 \longrightarrow x_1$ direction will be used. The derivation for $x_1 \longrightarrow x_2$ follows the same basic steps.

First, the observation model is defined

$$\begin{aligned} x_1[n] &= a_{11}^* x_1[n-1] + a_{21}^* x_2[n-1] + \eta_1^*[n], \\ x_2[n] &= a_{12}^* x_1[n-1] + a_{22}^* x_2[n-1] + \eta_2^*[n], \end{aligned} \tag{A.1}$$

where η_1^* and η_2^* are discrete-time white noise processes with zero mean. The two estimated models that are compared for GC estimation follow an ARX model [Eq. (2.35)] under the joint case assumption, and follow an AR model under the disjoint case assumption. The ARX estimated model for the joint case is of the form

$$\begin{aligned} x_1[n] &= a_{11} x_1[n-1] + a_{21} x_2[n-1] + \eta_1[n], \\ x_2[n] &= a_{12} x_1[n-1] + a_{22} x_2[n-1] + \eta_2[n], \end{aligned} \tag{A.2}$$

where the model parameters η_1 and η_2 are discrete-time white noise processes with zero mean and a_{pq} are the estimated model parameters. The AR estimated model for the disjoint case is of the form

$$x_1[n] = \sum_{m=1}^{M} \alpha_m x_1[n-m] + \epsilon_1[n], \tag{A.3}$$

where ϵ_1 is a discrete-time white noise process with zero mean and α_m are the autoregressive model parameters. These estimated models will be used in the following derivations. The first derivation will be a generalization of the closed form expression given in [95, Eq. (12)], where the a_{11} and a_{22} are equal to zero and the estimated model order is unconstrained.

For simplicity, the analysis assumes that enough epochs are available such that the sample variances and variances are assumed equal, and that the estimated models are the MMSE estimators, such that $a_{pq} \approx a_{pq}^*$ $(p, q \in \{1, 2\})$ for the joint model. These assumptions are not reasonable in many circumstances, but still provide insight on the "ideal" GC and NC estimates. Nevertheless, it is important to reinforce the point made in Sec. 2.2, that the observation models and estimation models must not be confused, even when the parameter estimation is assumed to be "perfect."

A.2.1 Derivation for the GC value for $M = 1$ and $M = 2$

Obtaining the GC value for different M values is tedious, but not complicated. First the expected values for the variances of x_1, x_2 and the covariance between x_1 and x_2 are calculated. Since η_1

and η_2 are white and zero mean,

$$\begin{aligned}
\mathcal{E}\{x_1[n] \cdot x_1[n]\} = \sigma_1^2 &= \frac{2a_{11}a_{12}\sigma_{12}^2 + \sigma_{\eta_1}^2}{1 - a_{11}^2}, \\
\mathcal{E}\{x_2[n] \cdot x_2[n]\} = \sigma_2^2 &= \frac{2a_{22}a_{21}\sigma_{12}^2 + \sigma_{\eta_2}^2}{1 - a_{22}^2}, \\
\mathcal{E}\{x_1[n] \cdot x_2[n]\} = \sigma_{12}^2 &= \frac{a_{11}a_{21}\sigma_1^2 + a_{12}a_{22}\sigma_2^2}{1 - a_{11}a_{22} - a_{12}a_{21}},
\end{aligned} \tag{A.4}$$

where $\mathcal{E}$ represents the expectation operator. Solving the system yields Eq. (A.5). The covariance between $x_1[n-1]$ and $x_1[n]$ can be succinctly expressed in terms of σ_1^2 and σ_{12}^2 as

$$\mathcal{E}\{x_1[n-1] \cdot x_1[n]\} = a_{11}\sigma_1^2 + a_{12}\sigma_{12}^2, \tag{A.6}$$

so that, by evaluating the conditional distribution of $x_1[n]$ given only $x_1[n-1]$ [64], the variance of ϵ_1 becomes

$$\sigma_{\epsilon_1}^2 = (1 - a_{11}^2)\sigma_1^2 - 2a_{11}a_{12}\sigma_{12}^2 - a_{12}^2\frac{\sigma_{12}^4}{\sigma_1^2}. \tag{A.7}$$

In this case, the GC value for evaluated when fitting first-order disjoint and bijointly regressive systems becomes

$$\mathrm{GC}_{2\to1} = \ln\left[\frac{(1 - a_{11}^2)\sigma_1^2 - 2a_{11}a_{12}\sigma_{12}^2 - a_{12}^2\frac{\sigma_{12}^4}{\sigma_1^2}}{\sigma_{\eta_1}^2}\right], \tag{A.8}$$

which can be expanded into Eq. (A.9). The expression demonstrates clearly that GC *does* take the $a_{12}a_{21}$ feedback loop into consideration. Although the analysis of the contribution of these terms using Eq. (A.9) is not straightforward, the terms in Eq. (A.8) that depend $a_{12}\sigma_{12}^2$ can be shown to contain $a_{12}a_{21}$ by using Eq. (A.8). A similar approach can be taken to evaluate GC using higher-order models. This is done by evaluating $\mathcal{E}\{x_1[n-\Delta n] \cdot x_1[n]\}$ for $\Delta n \in [1, \cdots, M]$ and using the conditional distributions to obtain the expected $\sigma_{\epsilon_1}^2$. One helpful identity to evaluate these covariances is

$$\mathcal{E}\left\{\begin{bmatrix} x_1[n] \\ x_2[n] \end{bmatrix} x_1[n-\Delta n]\right\} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}^{\Delta n} \begin{bmatrix} \sigma_1^2 \\ \sigma_{12}^2 \end{bmatrix}, \tag{A.10}$$

where $[\boldsymbol{A}]^{\Delta n} = \overbrace{\boldsymbol{A} \cdot \boldsymbol{A} \cdots \boldsymbol{A}}^{\Delta n \text{ times}}$. For the sake of brevity, a detailed derivation for higher orders is omitted, but the expression for $M = 2$ is found in Eq. (A.11).

Increasing M causes GC monotonically decrease, which can be intuitively explained by remembering that GC compares two models, an AR and an ARX model. For the ARX model, there should be no improvement for higher order estimated ARX models over the first-order estimated ARX model, since the observation model is a first-order ARX model.[1] Meanwhile, AR models cannot perfectly mimic the dynamics of ARX models. While a second-order AR model is also not able to perfectly represent an ARX model, it is better able to predict $x_1[n]$ than a first order AR model. Similarly, the variance of the residual of a third-order AR model is smaller (or equal) to that of the second-order model. The variance of the residual of the ARX models should be equal to $\sigma^2_{\eta_1}$ for any $M \geq 1$, but the variance of the residual of the AR model should decrease monotonically with M. Thus, the GC values for $M = 2$ will be lower than with $M = 1$.

A.2.2 Derivation for a lower bound of the GC measure for large M

The fact that the GC value decreases monotonically with the model order is well known [95] and can be argued qualitatively, as is done in Sec. A.2.1. However, it is helpful to define a lower bound for the GC value, so that the range of possible GC values for any M may be known, that is, $\mathrm{GC}_{M\to\infty} \leq \mathrm{GC}_{M\in\mathbb{Z}^+} \leq \mathrm{GC}_{2\to 1|M=1}$.

Since GC compares the sample variance of the error sequence of a joint and a disjoint model, the two variances must be obtained. For the joint model, the expected variance of the error sequence is simply the variance of η_1 (following the assumption that $a_{pq} \approx a^*_{pq}$ for $p, q \in \{1, 2\}$). To find the disjoint model of the form Eq. (A.3), one can start by expanding Eq. (A.1) into

$$x_1[n] = a_{11}x_1[n-1] + a_{12}\overbrace{(a_{21}x_1[n-2] + a_{22}x_2[n-2] + \eta_2[n-1])}^{x_2[n-1]} + \eta_1[n], \tag{A.12}$$

which, for $a_{22} \neq 0$, can be further expanded by recursively replacing the x_2 terms, yielding

$$\begin{aligned} x_1[n] =& a_{11}x_1[n-1] + a_{12}\sum_{m=2}^{M}(a_{21})^{m-1}x_1[n-m] + \\ & a_{12}\,(a_{22})^{M-1}x_2[n-M] + a_{12}\sum_{m=1}^{M}(a_{22})^{m-1}\eta_2[n-m]) + \eta_1[n]. \end{aligned} \tag{A.13}$$

[1] Nevertheless, a larger number of model parameters lead to larger variance in the parameter estimation.

For $a_{22} = 0$, the expansion reduces to

$$x_1[n] = a_{11}x_1[n-1] + a_{12}\sum_{m=2}^{M}(a_{21})^{m-1}x_1[n-m] + a_{12}\eta_2[n-1]) + \eta_1[n], \tag{A.14}$$

and the asymptotic MMSE parameter values for the disjoint model are

$$\alpha_m \approx \begin{cases} a_{11} & \text{for } m = 1, \\ a_{12}(a_{21})^{m-1} & \text{for } m > 1, , \\ 0 & \text{for } m < 0, \end{cases} \tag{A.15}$$

so that the prediction error for the MMSE of the disjoint model is

$$\epsilon_1[n] = a_{12}\eta_2[n-1] + \eta_1[n]. \tag{A.16}$$

Using the fact that η_1 and η_2 are white and uncorrelated, the variance of ϵ_1 is

$$\sigma^2_{\epsilon_1} = (a_{12})^2\sigma^2_{\eta_2} + \sigma^2_{\eta_1}, \tag{A.17}$$

and the GC value can be expressed as

$$\mathrm{GC}_{2\to 1} = \ln\left(1 + \frac{(a_{12})^2\sigma^2_{\eta_2}}{\sigma^2_{\eta_1}}\right), \tag{A.18}$$

which does not depend on a_{21} or a_{11}. This expression was used to argue that GC overlooks important parameters of the model by Hu *et al.* in [95]. However, it is important to remember that this expression is only valid for $a_{22} = 0$ and large M, nevertheless, as is shown below, the expression is still useful for the purpose of establishing a lower bound. Returning briefly to Eq. (A.12), note that for any $M > 1$, the $x_1[n-1]$ and $x_1[n-2]$ terms are available in the AR model and therefore will add no additional residual error. Note also that the $\eta_1[n]$ and $\eta_2[n-1]$ terms cannot be predicted in any way by the AR model. This is a consequence of $\eta_1[n]$ being uncorrelated with any past values of η_1 and that $\eta_2[n-1]$ is only correlated with $x_1[n]$, but not with $x_1[n-1]$ or any other past values of x_1. Thus, the only remaining question is how well can the $x_2[n-2]$ term be predicted by past values of x_1.

Supposing that $x_2[n-2]$ could be perfectly predicted by past values of x_1 produces a prediction error equivalent to Eq. (A.16). Although this is only exactly true for $a_{22} = 0$, it becomes clear that for any $M \geq 1$,

$$\sigma_{\epsilon_1}^2 \geq (a_{12})^2 \sigma_{\eta_2}^2 + \sigma_{\eta_1}^2, \tag{A.19}$$

which shows that Eq. (A.18) is indeed a lower bound for all GC values for $M \geq 1$ and is the asymptotic value for large M and $a_{22} = 0$.

A.2.3 Derivation of the NC value

In [95, Eq. (21)], a partial expansion of NC was expressed for models for which $a_{11} = a_{22} = 0$. The motivation to eliminate these parameters is to simplify the expressions, but this simplification arguably reduces the representativeness of the model [147]. Here, the general expression is shown. By expanding the difference equations as done in Eq. (A.12),

$$NC_{2\to1} = \frac{\sum_{n=3}^{N} \left(a_{12}a_{21}x_1[n-2] + a_{22}x_2[n-2] + a_{21}\eta_2[n-1]\right)^2}{\sum_{n=3}^{N} \left(a_{12}a_{21}x_1[n-2] + a_{21}\eta_2[n-1]\right)^2 + \sum_{n=3}^{N} \eta_1^2[n]}, \tag{A.20}$$

which shows the clear dependence of NC on the $a_{12}a_{21}$ term. The expression is only valid for observation models with $a_{11} = 0$. A general expression for first order bijointly variate models is

$$NC_{2\to1} = \frac{\sum_{n=2}^{N} \left(a_{12}x_2[n-1]\right)^2}{\sum_{n=2}^{N} \left(a_{12}x_2[n-1]\right)^2 + \sum_{n=2}^{N} \left(a_{11}x_1[n-1]\right)^2 + \sum_{n=2}^{N} \eta_1^2[n]}, \tag{A.21}$$

which, as $N \longrightarrow \infty$, converges to

$$NC_{2\to1} = \frac{a_{12}^2\sigma_2^2}{a_{12}^2\sigma_2^2 + a_{11}^2\sigma_1^2 + \sigma{\eta_1}^2}. \tag{A.22}$$

This expression shows the dependence of NC on the product $a_{12}a_{21}$. It is important to note that the variances σ_1^2 and σ_1^2 themselves depend on the model parameters as Eq. (A.5) shows. This means that a change on any of the model parameters will also cause a change in σ_1^2 and σ_2^2, thus the interactions between the model parameters and the NC values is also not straightforward. Combining Eqs. (A.5) and (A.22) yields Eq. (A.23).

A.3 Discussion

In this appendix closed form expressions for NC and GC are derived. The GC expressions are shown for $M = 1$, $M = 2$, and an asymptotic expression for large M. The technique can be expanded to any order, although the complexity for the closed form expressions grows with M. Although the process of obtaining these estimates is laborious, they can quickly and accurately be numerically evaluated.

These expressions show that, for finite M, the expression for GC does indeed depend the product of a_{12} and a_{21}. In fact, the relationship between the model parameters shows many intricate relationships between model parameters and GC values. The interaction between the model parameters and the GC values is not straightforward, so expressions in terms of η_1 and η_2 and in terms of x_1 and x_2 are provided.

When doing theoretical analysis on GC and NC estimation, it is helpful to be able to evaluate the analytical values for comparison. These closed form expressions are used in Ch. 3 to compare the effects of estimation errors and sample variances on GC and NC estimates.

$$
\begin{aligned}
\sigma_1^2 &= \frac{\left(1 + a_{11}a_{22}^3 - a_{11}a_{22} - a_{22}^2 - a_{12}a_{21} - a_{12}a_{21}a_{22}^2\right)\sigma_{\eta_1}^2 + \left(a_{12}^2 - a_{12}^3a_{21} + a_{11}a_{12}^2a_{22}\right)\sigma_{\eta_2}^2}{(1 + a_{12}a_{21} - a_{11}a_{22})(1 - a_{11} - a_{12}a_{21} - a_{22} + a_{11}a_{22})(1 + a_{11} + a_{22} + a_{11}a_{22} - a_{12}a_{21})}, \\
\sigma_2^2 &= \frac{\left(a_{21}^2 - a_{21}^3a_{12} + a_{22}a_{21}^2a_{11}\right)\sigma_{\eta_1}^2 + \left(1 + a_{22}a_{11}^3 - a_{22}a_{11} - a_{11}^2 - a_{12}a_{21} - a_{12}a_{21}a_{11}^2\right)\sigma_{\eta_2}^2}{(1 + a_{12}a_{21} - a_{11}a_{22})(1 - a_{11} - a_{12}a_{21} - a_{22} + a_{11}a_{22})(1 + a_{11} + a_{22} + a_{11}a_{22} - a_{12}a_{21})}, \\
\sigma_{12}^2 &= \frac{\left(a_{12}a_{22}a_{21}^2 + a_{11}a_{21} - a_{11}a_{21}a_{22}^2\right)\sigma_{\eta_1}^2 + \left(a_{21}a_{11}a_{12}^2 + a_{22}a_{12} - a_{22}a_{12}a_{11}^2\right)\sigma_{\eta_2}^2}{(1 + a_{12}a_{21} - a_{11}a_{22})(1 - a_{11} - a_{12}a_{21} - a_{22} + a_{11}a_{22})(1 + a_{11} + a_{22} + a_{11}a_{22} - a_{12}a_{21})}.
\end{aligned}
\tag{A.5}
$$

$\mathrm{GC}_{2\to1}=$

$$
\ln\left[\frac{\left\{\left[a_{22}\left(a_{11} - a_{12}a_{21} + a_{11}a_{22} - 1\right) - 1\right]\sigma_{\eta_1}^2 - a_{12}^2\sigma_{\eta_2}^2\right\}\left\{\left[1 - \left(1 + a_{11} + a_{12}a_{21}\right)a_{22} + a_{11}a_{22}^2\right]\sigma_{\eta_1}^2 + a_{12}^2\sigma_{\eta_2}^2\right\}}{\sigma_{\eta_1}^2\left(1 + a_{12}a_{21} - a_{11}a_{22}\right)\left\{\left[\left(1 + a_{22}^2\right)a_{12}a_{21} + a_{22}\left(a_{11} + a_{22}\right) - a_{11}a_{22}^3 - 1\right]\sigma_{\eta_1}^2 + a_{12}^2\left(a_{12}a_{21} - a_{11}a_{22} - 1\right)\sigma_{\eta_2}^2\right\}}\right].
\tag{A.9}
$$

$\mathrm{GC}_{2\to1}=$

$$
\ln\left[\frac{\left[\left(a_{11}a_{22} - 1\right)\sigma_{\eta_1}^2 - a_{12}^2\sigma_{\eta_2}^2\right]\left\{\left[1 - \left(1 + 2a_{12}a_{21}\right)a_{22}^2 + a_{11}a_{22}\left(a_{22}^2 - 1\right)\right]\sigma_{\eta_1}^4 + a_{12}^2\left(2 - a_{11}a_{22} + a_{22}^2\right)\sigma_{\eta_1}^2\sigma_{\eta_2}^2 + a_{12}^4\sigma_{\eta_2}^4\right\}}{\sigma_{\eta_1}^2\left\{\left[a_{22}\left(a_{11} - a_{12}a_{21} + a_{11}a_{22} - 1\right) - 1\right]\sigma_{\eta_1}^2 - a_{12}^2\sigma_{\eta_2}^2\right\}\left\{\left[1 - a_{22}\left(1 + a_{11} + a_{12}a_{21}\right) + a_{11}a_{22}^2\right]\sigma_{\eta_1}^2 + a_{12}^2\sigma_{\eta_2}^2\right\}}\right].
\tag{A.11}
$$

$$
\begin{aligned}
\mathrm{NC}_{2\to1} = \frac{a_{12}^2a_{21}^2\left(a_{12}a_{21} - a_{11}a_{22} - 1\right)\sigma_{\eta_1}^2 + a_{12}^2\left[a_{12}a_{21} + a_{11}\left(a_{11} + a_{11}a_{12}a_{21} + a_{22} - a_{11}^2a_{22}\right) - 1\right]\sigma_{\eta_2}^2}{\left\{2a_{11}a_{12}^2a_{21}^2a_{22} + a_{22}\left(a_{11} + a_{22} - a_{11}a_{22}^2\right) + a_{12}a_{21}\left[1 + a_{22}^2 - 2a_{11}^2\left(a_{22}^2 - 1\right)\right] - 1\right\}\sigma_{\eta_1}^2} \\
+ a_{12}^2\left[a_{12}\left(a_{21} + 2a_{11}^2a_{21}\right) + a_{11}a_{22} - 2a_{11}^3a_{22} - 1\right]\sigma_{\eta_2}^2
\end{aligned}.
\tag{A.23}
$$

APPENDIX B

LISTINGS FOR ALGORITHMS

B.1 Overview

This appendix contains the listings for key algorithms used in this work. Deeper discussion and more thorough description of the algorithms are found in the references.

B.2 OBE-related algorithms

The unified OBE framework is more thoroughly described and discussed in [54], whereas a summary is given here as a reference. The general algorithm follows a recursion similar to weighted recursive least squares (WRLS) [53, 54, 101], but with dynamically evaluated optimal forgetting factor calculations. The algorithm shown here assumes a MISO model, since this is the focus of this work, but UOBE defined for general MIMO models in [54].

Given a sequence of error bounds $\gamma[n]$ [as in Eq. (2.27)], output signal $x_p[n]$ and vector of regressors (or regressor functions) $\boldsymbol{\varphi}_p[n]$, the UOBE framework is given in Alg. B.1 and the recursion in Alg. B.2.

The optimum weights are selected according to different optimization criteria. With the exception of the Dasgupta-Huang OBE [50] which optimizes $\kappa[n]$, other algorithms under the UOBE umbrella choose the weights that minimize either the determinant of $\kappa[n]\boldsymbol{P}[n]$ (proportional to the square of the volume of the ellipsoid) or the trace of $\kappa[n]\boldsymbol{P}[n]$ (proportional to the sum of the squares of the semi-axes of the ellipsoid).

Defining $q[n] = \beta[n]/\alpha[n]$, the weights that minimize the volume, if they exist, are obtained by finding the unique positive root of the following equation

$$F_v(s) = a_2 s^2 + a_1 s + a_0, \tag{B.1}$$

where

$$a_2 = (K-1)\gamma[n]G^2[n], \tag{B.2}$$

$$a_1 = \left[(2K-1)\gamma[n] + \|\epsilon[n]\|^2 - \kappa[n-1]G[n]\right] G[n], \tag{B.3}$$

$$a_0 = K[\gamma[n] - \|\epsilon[n]\|^2] - \kappa[n-1]G[n], \tag{B.4}$$

Algorithm B.1: Unified Optimum Bounded Ellipsoid Algorithm

procedure UOBE
  $\boldsymbol{\theta}[1] = \mathbf{0}$ ▷ Set initial ellipsoid as a very large hyper-sphere centered at origin
  $\kappa[1] = 1$
  $\boldsymbol{P}[1] = \frac{1}{\mu}\boldsymbol{I}$
  for $n = 2$ **to** N **do**
    $\epsilon[n] = x_p[n] - \boldsymbol{\theta}^T[n-1]\boldsymbol{\varphi}_p[n]$ ▷ Calculate prediction error
    $G[n] = \boldsymbol{\varphi}_p^T[n]\boldsymbol{P}[n-1]\boldsymbol{\varphi}_p[n]$ ▷ Obs.: $G[n]$ is a scalar
    Evaluate if optimum $\alpha[n]$ and $\beta[n]$ exist ▷ $\alpha[n]$ and $\beta[n]$ are described later
    if optimum $\alpha[n]$ and $\beta[n]$ exist **then**
      do UOBE-Recursion (Alg. B.2)
      count = 0
    else ▷ Ellipsoid does not change
      $\boldsymbol{P}[n] = \boldsymbol{P}[n-1]$
      $\boldsymbol{\theta}[n] = \boldsymbol{\theta}[n-1]$
      $\kappa[n] = \kappa[n-1]$
      count = count + 1
      if count > N_{abe} **then**
        do EstimateBounds (Alg. B.2) ▷ If using ABE
        count = 0
      end if
    end if
  end for
end procedure

Algorithm B.2: UOBE Recursion

procedure UOBE-Recursion
  $\boldsymbol{P}[n] = \frac{1}{\alpha[n]}\left[\boldsymbol{P}[n-1] - \frac{\beta[n]\boldsymbol{P}[n-1]\boldsymbol{\varphi}_p[n]\boldsymbol{\varphi}_p^T[n]\boldsymbol{P}[n-1]}{\alpha[n] + \beta[n]G[n]}\right]$ ▷ Update direction & shape of ellipsoid
  $\boldsymbol{\theta}[n] = \boldsymbol{\theta}[n-1] + \beta[n]\boldsymbol{P}[n]\boldsymbol{\varphi}_p[n]\epsilon[n]$ ▷ Update centroid
  $\kappa[n] = \alpha[n]\kappa[n] + \beta[n]\gamma^2[n] - \frac{\alpha[n]\beta[n]\epsilon^2[n]}{\alpha[n] + \beta[n]G[n]}$ ▷ Update size of ellipsoid
end procedure

such that $F_v(q[n]) = 0$. When no such root exists, none of the ellipsoids that contain the intersection between the hyperstrip and the previous ellipsoid have smaller volume than the current ellipsoid. Equivalently, the positive root indicates that the value of $q[n]$ that defines the ellipsoid with smallest volume out of the set of all ellipsoids that fully contains the intersection between the previous ellipsoid and the hyperstrip. Note here that setting $\alpha[n]$ to unity and $\beta[n]$ to zero is equivalent to ignoring or discarding the current $x_p[n]$ and $\boldsymbol{\varphi}_p[n]$ and making no changes to the ellipsoid.

To minimize the square sum of the semi-axes, the optimum weights, if they exist, are obtained by finding the unique positive root of

$$F_t(s) = b^3 s^3 + b_2 s^2 + b_1 s + b_0, \tag{B.5}$$

where

$$b_3 = \gamma[n] G^2[n] \left[G[n] - I[n-1] H[n] \right], \tag{B.6}$$

$$b_2 = 3\gamma[n] G[n] \left[G[n] - I[n-1] H[n] \right], \tag{B.7}$$

$$b_1 = H[n] G[n] I[n-1] \kappa[n-1] - 2H[n] I[n-1] \left[\gamma[n] - \|\epsilon[n]\|^2 \right] \tag{B.8}$$

$$- G[n] \|\epsilon[n]\|^2 + 3\gamma[n] G[n], \tag{B.9}$$

$$b_0 = \gamma[n] - \|\epsilon[n]\|^2 - H[n] I[n-1] \kappa[n-1], \tag{B.10}$$

where $H[n] \doteq \boldsymbol{\varphi}_p^T[n] P^2[n] \boldsymbol{\varphi}_p[n]$ and $I[n] \doteq \operatorname{tr}\left\{ \boldsymbol{P}^{-1}[n] \right\}$, where $\operatorname{tr}\{\cdot\}$ is the trace operator.

A stochastic method to estimate error bounds is developed by Joachim *et al.* in [106]. The algorithm starts with an overestimated bound. If no update to the ellipsoid is made for $\mathrm{N_{ABE}}$ samples, it finds the largest error in the last $\mathrm{N_{ABE}}$ samples and reduces the bounds accordingly. This is repeated until the error bound estimate is close enough to the true bounds. The general algorithm is shown in Alg. B.3.

B.3 Linkage tree crossover

In [192], Thierens introduces the Linkage Tree Genetic Algorithm (LTGA). The algorithm initializes the population randomly, but applies a steepest ascent hill climber to each member of the population

Algorithm B.3: Automatic Bounds Estimation

procedure EstimateBounds
  $N_{max} = \underset{m \in [n-N_{ABE}+1, n]}{\arg\max} \epsilon^2[m]$ ▷ Find largest prediction error in last N_{ABE} samples
  $\Delta_\gamma = \kappa[N_{ABE} - 1]G[N_{ABE}]/K - \varepsilon(2\sqrt{\gamma[N_{ABE} - 1]} - \varepsilon)$ ▷ Find appropriate reduction in bound for $n = N_{ABE}$
  if $\Delta_\gamma > 0$ **then**
    $\gamma[n] = \gamma[n-1] - \Delta_\gamma$ ▷ If a bound reduction is possible, reduce it
  else
    $\gamma[n] = \gamma[n-1]$ ▷ If γ cannot be reduced, keep old bounds
  end if
end procedure

to increase its fitness. The resulting population undergoes crossover until the termination criterion is reached (without further mutation).

The initial hill climbing is desirable so that the population can provide useful statistical pairwise linkage information to LTX. While it is possible to achieve convergence without this step, convergence is slower, and the linkage points will less likely be at helpful locations.

The first step for LTX is generating the linkage tree. The general steps are given in Alg. B.4. The distance metric used by LTX is introduced by Kraskov *et al.* in [113], and is a normalized mutual information distance metric. Following the generation of the linkage tree, the crossover occurs. The general steps for LTX are given in Alg. B.5.

Algorithm B.4: Generate Linkage Tree

procedure GenerateLinkageTree
  Initialize each gene as one cluster
  repeat
    Compute the distance between clusters
    Merge closest clusters together
  until Only one cluster remains
  Organize clustering information into tree
end procedure

In its seminal form, LTX is defined for single-objective optimization problems, where two solutions can be superior, inferior, or equivalent to one another. In multi-objective problems, comparing two solutions is less straighforward. Although the categories of superior (dominating),

Algorithm B.5: Linkage Tree Crossover

```
1: procedure LinkageTreeCrossover
2:     Select parents
3:     Start at the largest cluster
4:     while Tree is not fully traversed do
5:         Crossover parents using the current cluster
6:         if one or more offspring are superior to both parents then
7:             Replace parents with offspring
8:         end if
9:         Move down the linkage tree and repeat
10:    end while
11: end procedure
```

inferior (dominated) and equivalent are still present, a solution might be superior to a second solution in one objective function, but inferior in a second objective function (non-dominated). To accommodate for non-domination between parents and offspring, this work introduces a new variant of LTX that includes an additional conditional statement which can be tuned to choose keep offspring, parents or randomly select one of them. The general algorithm for LTX adapted to multi-objective problems is shown in Alg. B.6.

Algorithm B.6: Linkage Tree Crossover for multi-objective problems

```
1: procedure LinkageTreeCrossover2
2:     Select parents
3:     Start at the largest cluster
4:     while Tree is not fully traversed do
5:         Crossover parents using the current cluster
6:         if one or more offspring dominate both parents then
7:             Replace parents with offspring
8:         else if neither offspring dominate both parents or dominated by both then
9:             Randomly decide which to keep[1]
10:        else if one or more offspring are dominated by both parents then
11:            Do not replace parents
12:        end if
13:        Move down the linkage tree
14:    end while
15: end procedure
```

[1] By setting the probability of each option to 1 or 0, a deterministic behavior can be set

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Abdulkader, S.N., Atia, A., & Mostafa, M.S.M. (2015). Brain computer interfacing: Applications and challenges. *Egyptian Informatics Journal*, 16(2), 213–230.

[2] Afsharnia, F., Madadi, A., & Menhaj, M.B. (2019). Iterative learning identification and control for dynamic systems described by NARMAX model. *AUT Journal of Modeling and Simulation*.

[3] Akaike, H. (1968). On the use of a linear model for the identification of feedback systems. *Annals of the Institute of Statistical Mathematics*, 20(1), 425–439.

[4] Akaike, H. (1969). Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21(1), 243–247.

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

[6] Amisigo, B.A., Van de Giesen, N., Rogers, C., Andah, W.E.I., & Friesen, J. (2008). Monthly streamflow prediction in the Volta Basin of West Africa: A SISO NARMAX polynomial modelling. *Physics and Chemistry of the Earth, Parts A/B/C*, 33(1-2), 141–150.

[7] Ancona, N., Marinazzo, D., & Stramaglia, S. (2004). Radial basis function approach to nonlinear Granger causality of time series. *Physical Review E*, 70(5), 056221.

[8] Anderson, S.R., Lepora, N.F., Porrill, J., & Dean, P. (2010). Nonlinear dynamic modeling of isometric force production in primate eye muscle. *IEEE Transactions on Biomedical Engineering*, 57(7), 1554–1567.

[9] Aviyente, S., Bernat, E.M., Evans, W.S., & Sponheim, S.R. (2011). *A phase synchrony measure for quantifying dynamic functional integration in the brain.* Technical report, Wiley Online Library.

[10] Aviyente, S., Evans, W.S., Bernat, E.M., & Sponheim, S. (2007). A time-varying phase coherence measure for quantifying functional integration in the brain. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 4 (pp. IV–1169–IV–1172).

[11] Aysal, T.C. & Barner, K.E. (2006). Hybrid polynomial filters for Gaussian and non-Gaussian noise environments. *IEEE Transactions on Signal Processing*, 54(12), 4644–4661.

[12] Baccalá, L.A. & Sameshima, K. (2001). Partial directed coherence: A new concept in neural structure determination. *Biological Cybernetics*, 84(6), 463–474.

[13] Baek, E. & Brock, W. (1992). *A general test for nonlinear Granger causality: Bivariate model*. Technical report, Iowa State University and University of Wisconsin at Madison.

[14] Barnett, L., Barrett, A.B., & Seth, A.K. (2009). Granger causality and transfer entropy are equivalent for Gaussian variables. *Physical Review Letters*, 103(23), 238701.

[15] Barnett, L. & Seth, A.K. (2011). Behaviour of Granger causality under filtering: Theoretical invariance and practical application. *Journal of Neuroscience Methods*, 201(2), 404–419.

[16] Barnett, L. & Seth, A.K. (2014). The MVGC multivariate Granger causality toolbox: A new approach to Granger-causal inference. *Journal of Meuroscience Methods*, 223, 50–68.

[17] Barnett, L. & Seth, A.K. (2015). Granger causality for state-space models. *Physical Review E*, 91(4), 040101.

[18] Barnett, L. & Seth, A.K. (2017). Detectability of Granger causality for subsampled continuous-time neurophysiological processes. *Journal of Neuroscience Methods*, 275, 93–121.

[19] Barrett, A.B. & Barnett, L. (2013). Granger causality is designed to measure effect, not mechanism. *Frontiers in Neuroinformatics*, 7, 6.

[20] Barton, M.J., Robinson, P.A., Kumar, S., Galka, A., Durrant-Whyte, H.F., Guivant, J., & Ozaki, T. (2009). Evaluating the performance of Kalman-filter-based EEG source localization. *IEEE Transactions on Biomedical Engineering*, 56(1), 122–136.

[21] Beleites, C., Baumgartner, R., Bowman, C., Somorjai, R., Steiner, G., Salzer, R., & Sowa, M.G. (2005). Variance reduction in estimating classification error using sparse datasets. *Chemometrics and Intelligent Laboratory Systems*, 79(1), 91–100.

[22] Billings, S.A. (2013). *Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains*. John Wiley & Sons.

[23] Billings, S.A. & Aguirre, L.A. (1995). Effects of the sampling time on the dynamics and identification of nonlinear models. *International Journal of Bifurcation and Chaos*, 5(06), 1541–1556.

[24] Billings, S.A. & Chen, S. (1989). Extended model set, global data and threshold model identification of severely non-linear systems. *International Journal of Control*, 50(5), 1897–1923.

[25] Billings, S.A., Korenberg, M.J., & Chen, S. (1988). Identification of non-linear output-affine systems using an orthogonal least-squares algorithm. *International Journal of Systems Science*, 19(8), 1559–1568.

[26] Billings, S.A. & Wei, H.L. (2005). The wavelet-NARMAX representation: A hybrid model structure combining polynomial models with multiresolution wavelet decompositions.

International Journal of Systems Science, 36(3), 137–152.

[27] Billings, S.A. & Zhu, Q. (1994). A structure detection algorithm for nonlinear dynamic rational models. *International Journal of Control*, 59(6), 1439–1463.

[28] Billingsley, P. (1995). *Measure and Probability*. John Wiley & Sons.

[29] Blomgren, P. & Chan, T.F. (1998). Color TV: Total variation methods for restoration of vector-valued images. *IEEE Transactions on Image Processing*, 7(3), 304–309.

[30] Booker, L. (1987). Improving search in genetic algorithms. *Genetic Algorithms and Simulated Annealing*, (pp. 61–73).

[31] Box, G.E.P. (1979). Robustness in the strategy of scientific model building. In *Robustness in statistics* (pp. 201–236). Elsevier.

[32] Box, G.E.P., Jenkins, G.M., Reinsel, G.C., & Ljung, G.M. (1970). *Time Series Analysis: Forecasting and Control*. John Wiley & Sons.

[33] Bressler, S.L. & Seth, A.K. (2011). Wiener–Granger causality: A well established methodology. *Neuroimage*, 58(2), 323–329.

[34] Brito, A.G., Leite Filho, W.C., & Hemerly, E.M. (2013). Identification of a Hammerstein model for an aerospace electrohydraulic servovalve. *IFAC Proceedings Volumes*, 46(19), 459–463.

[35] Brovelli, A., Ding, M., Ledberg, A., Chen, Y., Nakamura, R., & Bressler, S.L. (2004). Beta oscillations in a large-scale sensorimotor cortical network: Directional influences revealed by Granger causality. *Proceedings of the National Academy of Sciences*, 101(26), 9849–9854.

[36] Cartwright, N. (1989). *Nature's Capacities and Their Measurement.* Oxford University Press.

[37] Chávez, M., Martinerie, J., & Le Van Quyen, M. (2003). Statistical assessment of nonlinear causality: application to epileptic EEG signals. *Journal of Neuroscience Methods*, 124(2), 113–128.

[38] Chella, F., D'Andrea, A., Basti, A., Pizzella, V., & Marzetti, L. (2017). Non-linear analysis of scalp EEG by using bispectra: The effect of the reference choice. *Frontiers in Neuroscience*, 11, 262.

[39] Chen, S. & Billings, S.A. (1989). Representations of non-linear systems: The NARMAX model. *International Journal of Control*, 49(3), 1013–1032.

[40] Chen, S., Billings, S.A., Cowan, C.F.N., & Grant, P.M. (1990). Practical identification of NARMAX models using radial basis functions. *International Journal of Control*, 52(6), 1327–1350.

[41] Chen, S., Cowan, C., & Grant, P. (1991). OLS learning algorithm for RBF networks. *IEEE Transactions on Neural Networks*, 2(2), 302–309.

[42] Chen, S. & Donoho, D. (1994). Basis pursuit. In *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, volume 1 (pp. 41–44).: IEEE.

[43] Chen, Y., Bressler, S.L., & Ding, M. (2006a). Frequency decomposition of conditional Granger causality and application to multivariate neural field potential data. *Journal of Neuroscience Methods*, 150(2), 228–237.

[44] Chen, Y., Bressler, S.L., & Ding, M. (2006b). Frequency decomposition of conditional Granger causality and application to multivariate neural field potential data. *Journal of Neuroscience Methods*, 150(2), 228 – 237.

[45] Clearwater, S.H., Hogg, T., & Huberman, B.A. (1992). Cooperative problem solving. In *Computation: The Micro and the Macro View* (pp. 33–70).: World Scientific.

[46] Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate behavioral research*, 18(1), 115–126.

[47] Connor, J.T., Martin, R.D., & Atlas, L.E. (1994). Recurrent neural networks and robust time series prediction. *IEEE Transactions on Neural Networks*, 5(2), 240–254.

[48] Csáji, B.C. (2001). Approximation with artificial neural networks. *Faculty of Sciences, Etvs Lornd University, Hungary*, 24(48), 7.

[49] Cybenko, G. (1989). Approximation by superposition of sigmoidal functions. *Mathematics of Control, Signals and Systems*, 2(4), 303–314.

[50] Dasgupta, S. & Huang, Y.F. (1987). Asymptotically convergent modified recursive least-squares with data-dependent updating and forgetting factor for systems with bounded noise. *IEEE Transactions on Information Theory*, 33(3), 383–392.

[51] Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6(2), 182–197.

[52] Deller Jr., J.R. (1989). Set membership identification in digital signal processing. *IEEE ASSP Magazine*, 6(4), 4–20.

[53] Deller Jr., J.R. & Huang, Y.F. (2002). Set-membership identification and filtering for signal processing applications. *Circuits, Systems, and Signal Proc.*, 21, 69–82.

[54] Deller Jr., J.R., Nayeri, M., & Liu, M. (1994). Unifying the landmark developments in OBE identification. *International Journal of Adaptive Control and Signal Processing*, 8, 43–60.

[55] Deller Jr., J.R., Nayeri, M., & Odeh, S.F. (1993). Least-square identification with error bounds for real-time signal processing and control. *Proceedings of the IEEE*, 81(6), 815–849.

[56] Deller Jr., J.R. & Odeh, S.F. (1992). SM-WRLS algorithms with an efficient test for innovation. *IFAC Proceedings Volumes*, 25(15), 267 – 272. 9th IFAC/IFORS Symposium on Identification and System Parameter Estimation 1991 , Budapest, Hungary, 8-12 July 1991.

[57] Deshpande, G., LaConte, S., Peltier, S., & Hu, X. (2006). Directed transfer function analysis of fMRI data to investigate network dynamics. In *2006 International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 671–674).

[58] Dewdney, A.K. (1997). *Nonlinear system identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains*. John Wiley & Sons.

[59] Dhamala, M., Rangarajan, G., & Ding, M. (2008). Estimating Granger causality from Fourier and wavelet transforms of time series data. *Physical Review Letters*, 100(1), 018701.

[60] Diks, C. & Panchenko, V. (2006). A new statistic and practical guidelines for nonparametric Granger causality testing. *Journal of Economic Dynamics and Control*, 30(9-10), 1647–1669.

[61] Dimitriadis, S., Laskaris, N., & Tzelepi, A. (2013). On the quantization of time-varying phase synchrony patterns into distinct functional connectivity microstates (FCμstates) in a multi-trial visual ERP paradigm. *Brain Topography*, 26(3), 397–409.

[62] Ding, J., Tarokh, V., & Yang, Y. (2017). Bridging AIC and BIC: a new criterion for autoregression. *IEEE Transactions on Information Theory*.

[63] Durbin, J. (1960). The fitting of time-series models. *Revue de l'Institut International de Statistique*, (pp. 233–244).

[64] Eaton, M.L. (1983). *Multivariate Statistics: A Vector Space Approach*. Wiley series in probability and mathematical statistics. Probability and mathematical statistics. New York: John Wiley & Sons.

[65] Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, 32(2), 407–499.

[66] Faes, L., Nollo, G., & Porta, A. (2011). Information-based detection of nonlinear Granger causality in multivariate processes via a nonuniform embedding technique. *Physical Review E*, 83(5), 051112.

[67] Frank, I.E. & Friedman, J.H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2), 109–135.

[68] Friston, K.J., Harrison, L., & Penny, W. (2003). Dynamic causal modelling. *NeuroImage*, 19(4),

1273 – 1302.

[69] Friston, K.J., Moran, R., & Seth, A.K. (2013). Analysing connectivity with Granger causality and dynamic causal modelling. *Current Opinion in Neurobiology*, 23(2), 172–178.

[70] Fu, W. & Knight, K. (2000). Asymptotics for LASSO-type estimators. *The Annals of Statistics*, 28(5), 1356 – 1378.

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

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

[73] Goldman, B.W. & Tauritz, D.R. (2012). Linkage tree genetic algorithms: variants and analysis. In *Proceedings of the 14th Annual Conference on Genetic and Evolutionary Computation* (pp. 625–632).

[74] Golub, G.H. & Van Loan, C.F. (2012). *Matrix Computations*, volume 3. Johns-Hopkins University Press.

[75] Goshvarpour, A., Goshvarpour, A., Rahati, S., & Saadatian, V. (2012). Bispectrum estimation of electroencephalogram signals during meditation. *Iranian Journal of Psychiatry and Behavioral Sciences*, 6(2), 48.

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

[77] Granger, C.W.J. (1980). Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, 2, 329–352.

[78] Granger, C.W.J. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics*, 16(1), 121 – 130.

[79] Grassmann, G. (2020). New considerations on the validity of the Wiener-Granger causality test. *Heliyon*, 6(10), e05208.

[80] Greenblatt, R.E., Pflieger, M.E., & Ossadtchi, A.E. (2012). Connectivity measures applied to human brain electrophysiological data. *Journal of Neuroscience Methods*, 207(1), 1–16.

[81] Gu, Y. & Wei, H.L. (2018). A robust model structure selection method for small sample size and multiple datasets problems. *Information Sciences*, 451, 195–209.

[82] Guo, Y., Guo, L.Z., Billings, S.A., & Wei, H.L. (2015). Identification of nonlinear systems with non-persistent excitation using an iterative forward orthogonal least squares regression

algorithm. *International Journal of Modelling, Identification and Control*, 23(1), 1–7.

[83] Hagihira, S., Takashina, M., Mori, T., & Mashimo, T. (2004). Bispectral analysis gives us more information than power spectral-based analysis. *British Journal of Anaesthesia*, 92(5), 772–773.

[84] Hand, M.L. (1978). *Aspects of linear regression estimation under the criterion of minimizing the maximum absolute residual.* PhD thesis, Iowa State University.

[85] Hansen, N. & Ostermeier, A. (2001). Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2), 159–195.

[86] Haufe, S., Nikulin, V.V., & Nolte, G. (2011). Identifying brain effective connectivity patterns from EEG: performance of Granger causality, DTF, PDC and PSI on simulated data. *BMC Neuroscience*, 12(S1), P141.

[87] Hiemstra, C. & Jones, J.D. (1994). Testing for linear and nonlinear Granger causality in the stock price-volume relation. *The Journal of Finance*, 49(5), 1639–1664.

[88] Hill, P.D. (1985). Kernel estimation of a distribution function. *Communications in Statistics-Theory and Methods*, 14(3), 605–620.

[89] Hillebrand, A., Tewarie, P., Van Dellen, E., Yu, M., Carbo, E.W.S., Douw, L., Gouw, A.A., Van Straaten, E.C.W., & Stam, C.J. (2016). Direction of information flow in large-scale resting-state networks is frequency-dependent. *Proceedings of the National Academy of Sciences*, 113(14), 3867–3872.

[90] Hoerl, A.E. & Kennard, R.W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67.

[91] Hoke, M., Lehnertz, K., Pantev, C., & Lütkenhöner, B. (1989). Spatiotemporal aspects of synergetic processes in the auditory cortex as revealed by the magnetoencephalogram. In E. Başar & T. H. Bullock (Eds.), *Brain Dynamics* (pp. 84–105). Berlin, Heidelberg: Springer Berlin Heidelberg.

[92] Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396), 945–960.

[93] Hosoya, Y. (1991). The decomposition and measurement of the interdependency between second-order stationary processes. *Probability Theory and Related Fields*, 88(4), 429–444.

[94] Hu, S., Cao, Y., Zhang, J., Kong, W., Yang, K., Li, X., & Zhang, Y. (2011a). Evidence for existence of real causality in the case of zero Granger causality. In *International Conference on Information Science and Technology* (pp. 1385–1389).

[95] Hu, S., Dai, G., Worrell, G.A., Dai, Q., & Liang, H. (2011b). Causality analysis of neural

connectivity: Critical examination of existing methods and advances of new methods. *IEEE Transactions on Neural Networks*, 22(6), 829–844.

[96] Hu, S., Jia, X., Zhang, J., Kong, W., & Cao, Y. (2016a). Shortcomings/limitations of blockwise Granger causality and advances of blockwise New causality. *IEEE Transactions on Neural Networks and Learning Systems*, 27(12), 2588–2601.

[97] Hu, S. & Liang, H. (2012). Causality analysis of neural connectivity: New tool and limitations of spectral Granger causality. *Neurocomputing*, 76(1), 44–47.

[98] Hu, S., Wang, H., Zhang, J., Kong, W., & Cao, Y. (2014). Causality from Cz to C3/C4 or between C3 and C4 revealed by Granger causality and New causality during motor imagery. In *2014 International Joint Conference on Neural Networks (IJCNN)* (pp. 3178–3185).

[99] Hu, S., Wang, H., Zhang, J., Kong, W., Cao, Y., & Kozma, R. (2016b). Comparison analysis: Granger causality and New causality and their applications to motor imagery. *IEEE Transactions on Neural Networks and Learning Systems*, 27(7), 1429–1444.

[100] Hu, X., Hu, S., Zhang, J., Kong, W., & Cao, Y. (2016c). A fatal drawback of the widely used Granger causality in neuroscience. In *2016 Sixth International Conference on Information Science and Technology (ICIST)* (pp. 61–65).

[101] Huang, Y.F. (1986). A recursive estimation algorithm using selective updating for spectral analysis and adaptive signal processing. *IEEE transactions on acoustics, speech, and signal processing*, 34(5), 1331–1334.

[102] Hume, D. (1904). *Enquiry Concerning Human Understanding*. Clarendon Press.

[103] Hume, D. (1978). A Treatise of Human Nature [1739]. *British Moralists*, (pp. 1650–1800).

[104] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*, volume 112. Springer.

[105] Jia, X., Hu, S., Zhang, J., & Kong, W. (2015). Blockwise Granger causality and blockwise new causality. In *2015 Seventh International Conference on Advanced Computational Intelligence (ICACI)* (pp. 421–425).

[106] Joachim, D., Deller Jr., J.R., & Nayeri, M. (1997). Practical considerations in the use of a new OBE algorithm that blindly estimates error bounds. In *Proceedings of the 40th Midwest Symposium on Circuits and Systems, 1997*, volume 2 (pp. 762–765).

[107] Kamalabadi, F., Forbes, J., Makarov, N., & Portnyagin, Y.I. (1997). Evidence for nonlinear coupling of planetary waves and tides in the antarctic mesopause. *Journal of Geophysical Research: Atmospheres*, 102(D4), 4437–4446.

[108] Kamiński, M., Ding, M., Truccolo, W.A., & Bressler, S.L. (2001). Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biol Cybern*, 85(2), 145–157.

[109] Kharchenko, V.S. (2019). Internet of things for industry and human application. In *Modelling and Development*, volume 1 (pp. 547). Ministry of Education and Science of Ukraine, National Aerospace University - Kharkiv Aviation Institute.

[110] Kim, J.H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, 53(11), 3735–3745.

[111] Kimeldorf, G.S. & Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2), 495–502.

[112] Kozma, R., Hu, S., Sokolov, Y., Wanger, T., Schulz, A.L., Woldeit, M.L., Gonçalves, A.I., Ruszinkó, M., & Ohl, F.W. (2021). State transitions during discrimination learning in the gerbil auditory cortex analyzed by network causality metrics. *Frontiers in systems neuroscience*, 15.

[113] Kraskov, A., Stögbauer, H., Andrzejak, R.G., & Grassberger, P. (2005). Hierarchical clustering using mutual information. *Europhysics Letters (EPL)*, 70(2), 278–284.

[114] Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, 69(6), 066138.

[115] Kubiny, H. (1994). Variable selection in QSAR studies. I. An evolutionary algorithm. *Quantitative Structure-Activity Relationships*, 13(3), 285–294.

[116] Kukreja, S.L., Galiana, H.L., & Kearney, R.E. (2003). NARMAX representation and identification of ankle dynamics. *IEEE transactions on biomedical engineering*, 50(1), 70–81.

[117] Kukreja, S.L., Galiana, H.L., & Kearney, R.E. (2004). A bootstrap method for structure detection of NARMAX models. *International Journal of Control*, 77(2), 132–143.

[118] Kukreja, S.L., Löfberg, J., & Brenner, M.J. (2006). A least absolute shrinkage and selection operator (LASSO) for nonlinear system identification. *IFAC Proceedings Volumes*, 39(1), 814–819.

[119] Lachaux, J.P., Lutz, A., Rudrauf, D., Cosmelli, D., Le Van Quyen, M., Martinerie, J., & Varela, F. (2002). Estimating the time-course of coherence between single-trial brain signals: An introduction to wavelet coherence. *Neurophysiologie Clinique/Clinical Neurophysiology*, 32(3), 157–174.

[120] Lachaux, J.P., Rodriguez, E., Martinerie, J., & Varela, F.J. (1999). Measuring phase synchrony in brain signals. *Human Brain Mapping*, 8(4), 194–208.

[121] Leardi, R., Boggia, R., & Terrile, M. (1992). Genetic algorithms as a strategy for feature selection. *Journal of Chemometrics*, 6(5), 267–281.

[122] Leontaritis, I.J. & Billings, S.A. (1985). Input-output parametric models for non-linear systems Part I: Deterministic non-linear systems. *International Journal of Control*, 41(2), 303–328.

[123] Liao, S.H. & Wen, C.H. (2007). Artificial neural networks classification and clustering of methodologies and applications–literature analysis from 1995 to 2005. *Expert Systems with Applications*, 32(1), 1–11.

[124] Libal, U. (2011). Feature selection for pattern recognition by lasso and thresholding methods - a comparison. In *2011 16th International Conference on Methods Models in Automation Robotics* (pp. 168–173).

[125] Lin, T., Nayeri, M., & Deller Jr., J.R. (1998). A consistently convergent OBE algorithm with automatic estimation of error bounds. *International Journal of Adaptive Control and Signal Processing*, 12(4), 305–324.

[126] Liu, Y. & Aviyente, S. (2009). Directed information measure for quantifying the information flow in the brain. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 2188–2191).

[127] Lizier, J.T. & Prokopenko, M. (2010). Differentiating information transfer and causal effect. *The European Physical Journal B*, 73(4), 605–615.

[128] Ljung, L. (1987). *System Identification: Theory for the User.* Prentice-Hall, Inc.

[129] Lu, Z., Pu, H., Wang, F., Hu, Z., & Wang, L. (2017). The expressive power of neural networks: A view from the width. In *Advances in Neural Information Processing Systems* (pp. 6231–6239).

[130] Lundberg, S. & Lee, S. (2017). A unified approach to interpreting model predictions. *Computing Research Repository*, abs/1705.07874.

[131] Mallat, S.G. & Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12), 3397–3415.

[132] Marinazzo, D., Pellicoro, M., & Stramaglia, S. (2008). Kernel method for nonlinear Granger causality. *Physical Review Letters*, 100(14), 144103.

[133] Markovsky, I. & Van Huffel, S. (2007). Overview of total least-squares methods. *Signal Processing*, 87(10), 2283–2302.

[134] Mauldin, M.L. (1984). Maintaining diversity in genetic search. In *Proceedings of the Fourth*

AAAI Conference on Artificial Intelligence, AAAI'84 (pp. 247–250).: AAAI Press.

[135] Maziarz, M. (2015). A review of the Granger-causality fallacy. *The Journal of Philosophical Economics: Reflections on Economic and Social Issues*, 8(2), 86–105.

[136] McCrorie, J.R. & Chambers, M.J. (2006). Granger causality and the sampling of economic processes. *Journal of Econometrics*, 132(2), 311–336.

[137] Mellin, W.D. (1957). Work with new electronic 'brains' opens field for army math experts. *The Hammond Times*, 10, 66.

[138] Miller, A. (2002). *Subset Selection in Regression*. Chapman and Hall/CRC.

[139] Morse, G. & Stanley, K.O. (2016). Simple evolutionary optimization can rival stochastic gradient descent in neural networks. In *Proceedings of the Genetic and Evolutionary Computation Conference 2016* (pp. 477–484).

[140] Munia, T.T.K. & Aviyente, S. (2019). Time-frequency based phase-amplitude coupling measure for neuronal oscillations. *Scientific Reports*, 9(1), 1–15.

[141] Munia, T.T.K. & Aviyente, S. (2021). Granger causality based directional phase-amplitude coupling measure. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1070–1074).

[142] Musgrave, J.L. (1992). Linear quadratic servo control of a reusable rocket engine. *Journal of Guidance, Control, and Dynamics*, 15(5), 1149–1154.

[143] Muthukumaraswamy, S. (2013). High-frequency brain activity and muscle artifacts in MEG/EEG: a review and recommendations. *Frontiers in Human Neuroscience*, 7, 138.

[144] Myles, P.S., Leslie, K., McNeil, J., Forbes, A., & Chan, M.T.V. (2004). Bispectral index monitoring to prevent awareness during anaesthesia: the B-Aware randomised controlled trial. *Lancet*, 363(9423), 1757–1763.

[145] Narendra, K. & Gallman, P. (1966). An iterative method for the identification of nonlinear systems using a Hammerstein model. *IEEE Transactions on Automatic control*, 11(3), 546–550.

[146] Nariyoshi, P. & Deller Jr., J.R. (2021). Nonlinear extensions of new causality. *Neuroscience Informatics*. To appear.

[147] Nariyoshi, P., Deller Jr., J.R., & Goodman, E.D. (2021a). On models for assessing causality strength. *Heliyon*. To appear.

[148] Nariyoshi, P., Deller Jr., J.R., & Goodman, E.D. (2021b). On the robustness of Granger and new causality measures to model order and parameter uncertainty in multivariate regressive

models. In review.

[149] Nariyoshi, P., Deller Jr., J.R., & Yan, J. (2017). Modified genetic crossover and mutation operators for sparse regressor selection in NARMAX brain connectivity modeling. In *2017 8th International IEEE/EMBS Conference on Neural Engineering (NER)* (pp. 660–663).

[150] Nolte, G., Ziehe, A., Nikulin, V.V., Schlögl, A., Krämer, N., Brismar, T., & Müller, K.R. (2008). Robustly estimating the flow direction of information in complex physical systems. *Physical Review Letters*, 100(23), 234101.

[151] Ogata, K. (2001). *Modern Control Engineering*. USA: Prentice-Hall, Inc., 4th edition.

[152] Oppenheim, A.V., Willsky, A.S., & Nawab, S.H. (1996). *Signals and Systems (2nd Ed.)*. USA: Prentice-Hall, Inc.

[153] Papoulis, A. & Pillai, S.U. (2002). *Probability, Random Variables, and Stochastic Processes*. Tata McGraw-Hill Education.

[154] Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition.

[155] Pelikan, M., Hauschild, M.W., & Thierens, D. (2011). Pairwise and problem-specific distance metrics in the linkage tree genetic algorithm. *Genetic & Evolutionary Computation Conference*, (pp. 1005 – 1012).

[156] Pereda, E., Quiroga, R.Q., & Bhattacharya, J. (2005). Nonlinear multivariate analysis of neurophysiological signals. *Progress in Neurobiology*, 77(1-2), 1–37.

[157] Piroddi, L. & Spinelli, W. (2003). A pruning method for the identification of polynomial NARMAX models. *IFAC Proceedings Volumes*, 36(16), 1071–1076.

[158] Plackett, R.L. (1950). Some theorems in least squares. *Biometrika*, 37(1/2), 149–157.

[159] Politis, D.N. & Romano, J.P. (1992). *A circular block-resampling procedure for stationary data*. Technical report, Stanford University.

[160] Politis, D.N. & Romano, J.P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89(428), 1303–1313.

[161] Poor, H.V. (1994). *An Introduction to Signal Detection and Estimation (2nd Ed.)*. Berlin, Heidelberg: Springer-Verlag.

[162] Pradhan, C., Jena, S.K., Nadar, S.R., & Pradhan, N. (2012). Higher-order spectrum in understanding nonlinearity in EEG rhythms. *Computational and Mathematical Methods in Medicine*, 2012.

[163] Punch III, W.F., Goodman, E.D., Pei, M., Chia-Shun, L., Hovland, P.D., & Enbody, R.J. (1993). Further research on feature selection and classification using genetic algorithms. In *ICGA* (pp. 557–564).

[164] Rabiner, L.R. & Gold, B. (1975). *Theory and Application of Digital Signal Processing*. Prentice-Hall, Inc.

[165] Rahim, H.A., Ibrahim, F., & Taib, M.N. (2007). A novel prediction system in dengue fever using NARMAX model. In *2007 International Conference on Control, Automation and Systems* (pp. 305–309).

[166] Rasmussen, C.E. (2004). Gaussian processes in machine learning. In *Advanced Lectures on Machine Learning* (pp. 63–71). Springer.

[167] Reeves, C.R. & Rowe, J.E. (2003). *Genetic Algorithms: Principles and Perspectives: A Guide to GA Theory*, volume 20. Springer.

[168] Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, 14(5), 465–471.

[169] Rodrigues, P.L.C. & Baccalá, L.A. (2016). Statistically significant time-varying neural connectivity estimation using generalized partial directed coherence. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (pp. 5493–5496).

[170] Rosenblum, M.G. & Pikovsky, A.S. (2001). Detecting direction of coupling in interacting oscillators. *Physical Review E*, 64(4), 045202.

[171] Rudin, L.I., Osher, S., & Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4), 259–268.

[172] Sayed, A.H. (2003). *Fundamentals of Adaptive Filtering*. John Wiley & Sons.

[173] Schelter, B., Winterhalder, M., Eichler, M., Peifer, M., Hellwig, B., Guschlbauer, B., Lücking, C.H., Dahlhaus, R., & Timmer, J. (2006). Testing for directed influences among neural signals using partial directed coherence. *Journal of Neuroscience Methods*, 152(1-2), 210–219.

[174] Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2), 461 – 464.

[175] Seth, A.K., Barrett, A.B., & Barnett, L. (2015). Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience*, 35(8), 3293–3297.

[176] Shahabi, H., Moghimi, S., & Moghimi, A. (2013). Investigating the effective brain networks related to working memory using a modified directed transfer function. In *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)* (pp. 1398–1401).

[177] Sheikhattar, A., Miran, S., Liu, J., Fritz, J.B., Shamma, S.A., Kanold, P.O., & Babadi, B. (2018). Extracting neuronal functional network dynamics via adaptive Granger causality analysis. *Proceedings of the National Academy of Sciences*, 115(17), E3869–E3878.

[178] Shen Minfen, Sun Lisha, & Beadle, P.J. (2000). Parametric bispectral estimation of EEG signals in different functional states of brain. In *2000 First International Conference Advances in Medical Signal and Information Processing (IEE Conf. Publ. No. 476)* (pp. 66–72).

[179] Shimmura, T., Ohashi, S., & Yoshimura, T. (2015). The highest-ranking rooster has priority to announce the break of dawn. *Scientific Reports*, 5, 11683.

[180] Shimmura, T. & Yoshimura, T. (2013). Circadian clock determines the timing of rooster crowing. *Current Biology*, 23(6), R231 – R233.

[181] Shrikumar, A., Greenside, P., & Kundaje, A. (2017). Learning important features through propagating activation differences. *Computing Research Repository*, abs/1704.02685.

[182] Sigl, J.C. & Chamoun, N.G. (1994). An introduction to bispectral analysis for the electroencephalogram. *Journal of clinical monitoring*, 10(6), 392–404.

[183] Silver, D., Huang, A., Maddison, C.J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, s., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 484–489.

[184] Sims, C.A. (1972). Money, income, and causality. *The American Economic Review*, (pp. 540–552).

[185] Sposito, V.A. (1976). Minimizing the maximum absolute deviation. *SIGMAP Bulletin*, 1(20), 51–53.

[186] Stam, C.J. (2005). Nonlinear dynamical analysis of EEG and MEG: Review of an emerging field. *Clinical neurophysiology*, 116(10), 2266–2301.

[187] Stoean, R., Stoean, C., & Sandita, A. (2017). Evolutionary regressor selection in ARIMA model for stock price time series forecasting. In *International Conference on Intelligent Decision Technologies* (pp. 117–126).: Springer.

[188] Stokes, P.A. & Purdon, P.L. (2017). A study of problems encountered in Granger causality analysis from a neuroscience perspective. *Proceedings of the Mational Academy of Sciences*, 114(34), E7063–E7072.

[189] Sun, R. (1994). A neural network model of causality. *IEEE Transactions on Neural Networks*, 5(4), 604–611.

[190] Taylor, S.J. (1994). Modeling stochastic volatility: A review and comparative study. *Mathematical Finance*, 4(2), 183–204.

[191] Tenke, C.E. & Kayser, J. (2012). Generator localization by current source density (CSD): Implications of volume conduction and field closure at intracranial and scalp resolutions. *Clinical Neurophysiology*, 123(12), 2328–2345.

[192] Thierens, D. (2010). The linkage tree genetic algorithm. In *International Conference on Parallel Problem Solving from Nature* (pp. 264–273).: Springer.

[193] Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 267–288).

[194] Tikhonov, A.N. (1943). On the stability of inverse problems. In *Dokl. Akad. Nauk SSSR*, volume 39 (pp. 195–198).

[195] Ulrich, T. & Thiele, L. (2011). Maximizing population diversity in single-objective optimization. In *Proceedings of the 13th Annual Conference on Genetic and Evolutionary Computation* (pp. 641–648).

[196] Vicente, R., Wibral, M., Lindner, M., & Pipa, G. (2011). Transfer entropy—a model-free measure of effective connectivity for the neurosciences. *Journal of Computational Neuroscience*, 30(1), 45–67.

[197] Vinyals, O., Babuschkin, I., Czarnecki, W.M., Mathieu, M., Dudzik, A., Chung, J., Choi, D.H., Powell, R., Ewalds, T., Georgiev, P., *et al.* (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782), 350–354.

[198] Volterra, V. (1887). *Sopra le Funzioni che Dipendono da Altre Funzioni (About Functions That Depend on Other Functions.* Tip. della R. Accademia dei Lincei.

[199] Wassenaar, G. (2020). Empirical performance evaluation of the linkage tree genetic algorithm.

[200] Wei, H.L. (2019). Sparse, interpretable and transparent predictive model identification for healthcare data analysis. In *International Work-Conference on Artificial Neural Networks* (pp. 103–114).: Springer.

[201] Wei, H.L. & Billings, S.A. (2008). Model structure selection using an integrated forward orthogonal search algorithm assisted by squared correlation and mutual information. *International Journal of Modelling, Identification and Control*, 3(4), 341–356.

[202] Wei, H.L. & Billings, S.A. (2019). NARMAX model as a sparse, interpretable and transparent machine learning approach for big medical and healthcare data analysis. In *Proceedings of the 5th IEEE International Conference on Data Science and Systems*: IEEE.

[203] Westwick, D. & Kearney, R. (2003). *Identification of Nonlinear Physiological Systems*, volume 7. John Wiley & Sons.

[204] Whittle, P. (1963). On the fitting of multivariate autoregressions, and the approximate canonical factorization of a spectral density matrix. *Biometrika*, 50(1-2), 129–134.

[205] Widrow, B. & Hoff, M.E. (1960). *Adaptive switching circuits*. Technical report, Stanford Electronics Laboratories - Stanford University.

[206] Wiener, N. (1956). The theory of prediction. In E. F. Beckenbach & R. Weller (Eds.), *Modern Mathematics for the Engineer* chapter 8, (pp. 165–190). New York: McGraw-Hill.

[207] Wolpaw, J. & Wolpaw, E.W. (2012). *Brain-Computer Interfaces: Principles and Practice*. OUP Usa.

[208] Yamashita, O., Sadato, N., Okada, T., & Ozaki, T. (2005). Evaluating frequency-wise directed connectivity of BOLD signals applying relative power contribution with the linear multivariate time-series models. *Neuroimage*, 25(2), 478–490.

[209] Yan, J. (2018). *Two Studies in Nonlinear Biological System Modeling and Identification*. PhD thesis, Michigan State University.

[210] Yan, J. & Deller Jr., J.R. (2014). Biologically-motivated system identification: Application to microbial growth modeling. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 322–325).

[211] Yan, J. & Deller Jr., J.R. (2015a). Set-theoretic measures as evolutionary fitness criteria in nonlinear system identification. In *Proceedings of the 17th IFAC Symposium on System Identification (SYSID)* Beijing. Published on CD-ROM and at IFAC-PapersOnLine.net.

[212] Yan, J. & Deller Jr., J.R. (2015b). Set-theoretic measures as evolutionary fitness criteria in nonlinear system identification. *17th IFAC Symposium on System Identification SYSID*, 48(28), 178–183.

[213] Yan, J. & Deller Jr., J.R. (2016). NARMAX model identification using a set-theoretic evolutionary approach. *Signal Processing*, 123(5), 30–41.

[214] Yan, J., Deller Jr., J.R., Fleet, B., Goodman, E.D., & Yao, M. (2013). Evolutionary identification of nonlinear parametric models with a set-theoretic fitness criterion. In *2013 IEEE China Summit and International Conference on Signal and Information Processing* (pp. 44–48).

[215] Yan, J., Deller Jr., J.R., Yao, M., & Goodman, E.D. (2014a). Biologically-motivated system identification: Application to microbial growth modeling. In *Proc. 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC),*.

[216] Yan, J., Deller Jr., J.R., Yao, M., & Goodman, E.D. (2014b). Evolutionary model selection for identification of nonlinear parametric systems. In *Proceedings of the 2014 IEEE China Summit and International Conference of Signal and Information Processing* (pp. 693–697).

[217] Yan, J., Nariyoshi, P., & Deller Jr., J.R. (2021). Sparse nonlinear model structure selection and parameter estimation using bi-objective optimization. In review.

[218] Yao, D., Qin, Y., Hu, S., Dong, L., Bringas Vega, M.L., & Sosa, P.A.V. (2019). Which reference should we use for EEG and ERP practice? *Brain topography*, 32(4), 530–549.

[219] Yassin, I.M., Zabidi, A., Amin Megat Ali, M.S., Md Tahir, N., Zainol Abidin, H., & Rizman, Z.I. (2016). Binary particle swarm optimization structure selection of nonlinear autoregressive moving average with exogenous inputs (NARMAX) model of a flexible robot arm. *International Journal on Advanced Science, Engineering and Information Technology*, 6(5), 630–637.

[220] Zhuo, H., Hu, S., Myers, M.H., Zhang, J., Kong, W., Cao, Y., & Kozma, R. (2016). Causality analysis during shared intentionality. In *2016 12th World Congress on Intelligent Control and Automation (WCICA)* (pp. 2215–2219).