Clarkson University

Prediction Analysis and System Identification of Complex Systems

A Dissertation by:

Abd AlRahman Rasheed AlMomani

A dissertation submitted in fulfillment of the requirements

for the degree of Doctor of Philosophy

in the

Electrical and Computer Engineering.

Wallace H. Coulter School Of Engineering

Clarkson University

August 2, 2019

CLARKSON UNIVERSITY

Abstract

Wallace H. Coulter School Of Engineering

Electrical and Computer Engineering

Doctor of Philosophy

Prediction Analysis and System Identification of Complex Systems

by Abd AlRahman Rasheed AlMomani

A complex system is a system composed of many components which may interact with each other. The prediction of the behavior of complex systems is important in many fields, such as weather forecasting, the motion of the planets, and temporal transitions. Philosophers and scientists have tried to formulate observational models and infer future states of such systems. Complex systems are systems whose behavior is intrinsically difficult to model due to the dependencies, competitions, relationships, or other types of interactions between their parts or between a given system and its environment. Systems that are "complex" have distinct properties that arise from these relationships, such as nonlinearity, emergence, spontaneous order, adaptation, and feedback loops, among others. Our work focuses on the prediction of future states of complex systems. Our work can be divided in to two major parts: First, an information theoretic system identification method called “Entropic Regression” that overcomes many obstacles of complex systems modeling. Second, developing a model-free method for coherent structures detection in dynamical systems called “Directed Affinity Partitioning”, where coherent structures can play a significant role in analyzing and understanding complex system dynamics.

Acknowledgements

Above all, I would like to thank my advisor Professor Erik M. Bollt for his guidance and patience, and for being one of the most intelligent, wise, kind, and active persons I ever met, who provide continuous motivation and inspiration mixed with a sense of humor. This dream could never be done without him. I am grateful for his genuine care and concern, and I sincerely appreciate every single conversation, meeting, and advice, and I will always be proud of being Professor Bollt's student.

I also want to thank my graduate committee members, professor Goodzar Ahmadi, professor William Jameson, professor Brain Helenbrook, Dr. Mahash Banavar, and Dr. Jie Sun, for serving as my committee members. I also want to thank them for the kind encouragement, brilliant comments and questions, and for making my defense a delightful moment and remarkable day.

I also want to thank my friend and academic brother, Alexander DeWitt, for his kind help in editing and proofreading my thesis, and for his helpful suggestions and feedback.

I gratefully acknowledge the funding received from the Office of Naval Research (ONR), and the Army Research Office (ARO), that I otherwise would not have been able to pursue my Ph.D.

I am deeply grateful to my father-in-law for his wise advice and genuine care, and to my mother-in-law for her warm feelings and continuous encouragement. I would like to thank them and thank all my wife's family for the help and support they provided us during my study.

My Mother; nothing in life worth more than the happiness in her eyes, and no words can thank her, or even tell her the purest feeling inside me. Thanks for every moment in this life with her. I am deeply grateful for the exceptional love she gave us, which makes all sadness and tiredness fade, and taught me that such "exceptional love" is "the secret", and it is what gave colors for life, and a purpose for the steps.

I owe a huge debt of gratitude to my family. I would like to thank all my brothers and sisters. Their belief in me was always my precious treasure and a source of motivation. It was generous luck to me to have amazing eight brothers and sisters and to be mostly homeschooled by them.

I would like to thank them all for all the love, support, and encouragement. I would like to thank them for sparking my passion for art, mathematics, engineering, physics, philosophy, history of science, literature, and poetry, from the early years of my childhood, with their diverse specialties and interests.

Special thank is reserved for my brother Ahmad, who has always been the example of the beautiful moral structure, persistence, and strong will. His belief in me is one of the main reasons for being here, pursuing my Ph.D. I could never make this without his care, support, encouragement, and patience. I am deeply thankful for him for gifting me the future.

My friend, and my brother... Khalid.
From teaching me the mechanical engineering drawing at the age of seven or less, until this moment, your hand was always very close to help me rise after every stumble. Without you, I would not be where I am today, and I would not be who I am today. I only hope that I made you proud of me, and since there are no words can thank you, I dedicate this work, and this dream, for you.

I could never achieve this dream without my love, my partner, my friend, and my wife, Heba. I owe a great debt of gratitude for your patience, sacrifices, and continuous care and support, starting from my undergraduate study, until now. I have been lucky enough to have you in my life, thank you for being always the warm heart and the deep belief.

Finally, to my kids, Matar and Jood; You have made me stronger, better, and more fulfilled than I could have ever imagined. Thank you for the amazing and beautiful life you gifted to me with your smile.

Contents

To My Brother... Khalid
and from both of us ...
To the memory of the greatest man...
My Father

Chapter 1

Introduction

> I do not know what I may appear to the world, but to myself, I seem to have been only like a boy playing on the sea-shore, and diverting myself in now and then finding a smoother pebble or a prettier shell than ordinary, while the great ocean of truth lay all undiscovered before me.
>
> *Sir Isaac Newton*
>
> *(January 4, 1643 – March 31 1727)*

Science can be defined as [201], a systematic effort that builds and organizes knowledge in the form of testable explanations and predictions about the universe, and producing more accurate natural explanations of how the natural world works, what its components are, how it comes to be what it is now, and what it could be in the future. Classically, science's primary goal has been building knowledge and understanding, regardless of its potential applications, however, scientific research is pledged with the explicit goal of solving problems, making predictions, and enabling effective technology. This goal practically the "purpose of science" according to Instrumentalism philosophical school, which is the view of scientific theories as useful tools for predicting phenomena instead of accurate or approximately accurate descriptions. Rising of Machine Learning, Deep Learning, Internet of Things, and Big Data enforced and gave the dominance for this pragmatic philosophy of John Dewey among all other philosophical views of science.

Most physical and natural systems change from a specific state to another state through time. In these systems a hidden set of rules define a relationship of these states, and a dynamical system represents these rules. When the underlying dynamics of the system are not known, it becomes compelling to use techniques that can discover these underlying dynamics

or connecting structure. This discovery process is the field of system identification, whose primary motive is prediction.

Prediction can be about the past (i.e., Evolutionary Biology), or present (i.e., Economic data). However, we focus here on the classical view: A prediction or forecast is a statement about a future event [172]. It is usually based upon experience or knowledge. And although guaranteeing accurate information about the future is in many cases not possible under the laws of uncertainty, prediction remains to be a dream for humanity throughout history. Starting from the necessary predictions to survive different kinds of catastrophes, and ending with predicting a sports game result to achieve profit.

In a non-statistical sense, the term "prediction" is usually used to refer to an informed guess or opinion [84], where such kind of prediction is based on experienced abductive reasoning, inductive reasoning, or deductive reasoning. This type of prediction can also be seen as a statistical prediction since the "data" being used—the predicting expert's cognitive experiences—forms some probability measure, which phrases the problem as a statistical prediction.

Prediction is the most significant part of statistical inference. One particular approach to such inference is known as predictive inference. Prediction can be undertaken within any of the several approaches to statistical inference.

The prediction of the behavior of complex systems is essential in many fields, such as weather forecasting, the motion of the planets, and modeling chaotic systems. Philosophers and scientists have tried to formulate observational models and infer future states of such systems. An old and straightforward example of attempts to predict the behavior of complex systems was *Old Farmer's Almanac*, which is a reference book from 1792 containing weather forecasts, planting charts, astronomical data, recipes, and articles. The book starts with making not necessarily correct predictions about the weather based on experience and accumulated personal observations through the years.

In science [201], a prediction is a rigorous, often quantitative, statement, forecasting what would happen under specific conditions, and scientific method built on testing statements that are logical consequences of scientific theories.

Constructing an underlying mathematical model which can be applied as the predictor is the basis of scientific prediction. For example, the existence of the planet Neptune was predicted

through mathematical modeling, not by observation, wherein 1821, Alexis Bouvard published astronomical tables of the orbit of Uranus, and following observations revealed deviations from the tables, which led Bouvard to a hypothesis that an unknown body was perturbing the orbit through gravitational interaction. In 1846, Urbain Le Verrier published his estimate of the planet's longitude, and in the same year, Neptune was discovered within 1° of where Le Verrier had predicted it to be.

Similar calculations on observed perturbations of the orbits of Uranus and Neptune led to a conclusion of the presence of another planet beyond the orbit of Neptune. In 1930, a new planet Pluto was discovered. Thus, the discovery of Pluto was a kind of accident.

Using statistical methods to build mathematical models of dynamical systems from measured data is known as the field of **System Identification** [120], where the mathematical model represents the description of the dynamic behavior of a system in either the time or frequency domain.

From the Instrumentalism point of view, the "goodness" of a mathematical model is determined by its prediction power, not the accuracy of the model itself. Identification for control can be considered as an example of a good model from this view. Control systems is one many possible applications of system identification, where it is the basis of modern data-driven control systems that the concepts of system identification are integrated into the controller design, and shape the foundation of controller optimality analysis.

The application to which the identified model will be applied to can play a role in examining the quality and robustness of the model. For example, in control system identification, our main purpose is to find a model that provides "enough" control for the closed-loop performance, whether or not this model is identical to the true system. This principle is know as Identification for Control (I4C). One commonly used example that describes this principle is to consider a system evolving according to the function:

$$G_{true}(s) = \frac{1}{s+1} \tag{1.1}$$

while the identified model is:

$$G(s) = \frac{1}{s}, \tag{1.2}$$

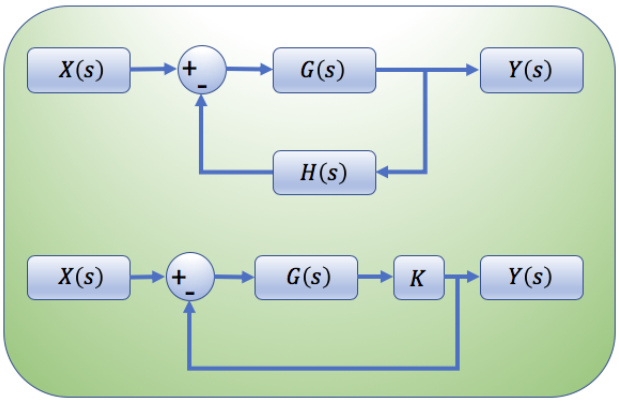

FIGURE 1.1: Standard transfer function diagram with Feedback.

where $G(s)$ is the Laplace transform of the function $g(t)$ from the frequency domain to the complex domain, and is given by:

$$G(s) = \mathcal{L}(g(t)) = \int_0^\infty g(t)e^{-st}dt. \tag{1.3}$$

We see that if we apply purely proportional negative feedback ($H(s) = 1$) control on the system as shown in Fig. (1.1), where $X(s)$ is the input and $Y(s)$ is the output, we can express the transfer function, $T(s)$, as:

$$\begin{aligned} Y(s) &= (X(s) - Y(s))\, G(s)K \\ &= X(s)G(s)K - Y(s)G(s)K \\ &= \tfrac{X(s)G(s)K}{G(s)K+1} \\ T(s) &= \tfrac{Y(s)}{X(s)} = \tfrac{G(s)K}{G(s)K+1}. \end{aligned} \tag{1.4}$$

So, we see that for the true system and model we have:

$$\begin{cases} T_{true}(s) = \frac{K}{s+K+1} \\ T_{model}(s) = \frac{K}{s+K} \end{cases} \tag{1.5}$$

and with large gain K such that $K \approx K + 1$, we have $T_{true}(s) \approx T_{model}(s)$, such that we can achieve robust real time control on the system even if the identified model is erroneous.

In other instances, estimating a model to play the role of a predictor, can be more sensitive, and any incorrect detection of features can profoundly reduce the prediction robustness, especially

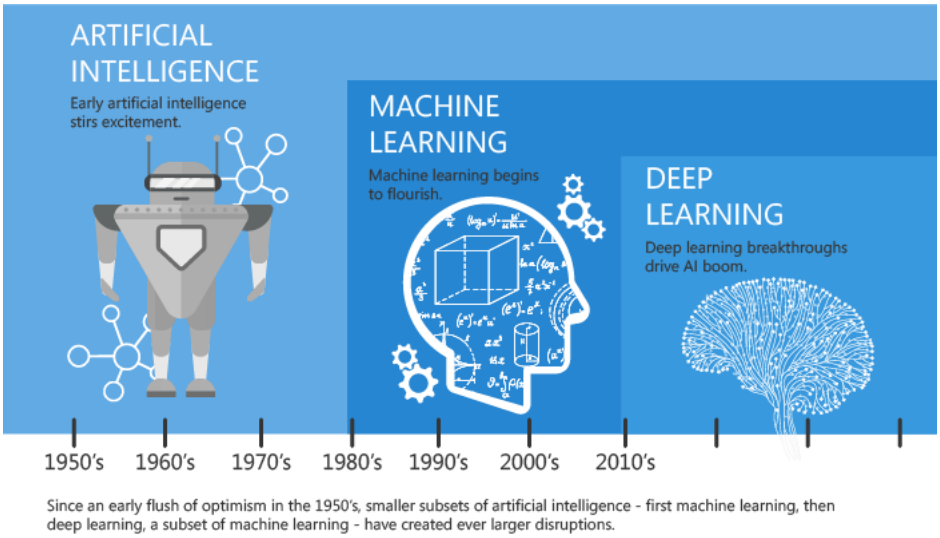

FIGURE 1.2: Time line and development of Artificial Intelligence.

for dynamical systems with chaotic behavior. Moreover, for dynamical systems with chaotic behavior, we mostly are interested in predicting the long term behavior of the system, and exploring the change in this behavior and bifurcation analysis under different assumptions, and only an accurate or nearly accurate model will be able to be the base for such long term behavior analysis.

In this work, we focus on discovering the governing dynamics of the system and concentrate our interest in constructing the optimal model that accurately simulates the actual system, which can be used as a predictor, as well as a model for control purposes.

1.1 Regression, Prediction, and Machine Learning

Deep Learning (DL), Machine Learning (ML), and Artificial Intelligence (AI) are very hot buzzwords right now, and often seem to be used interchangeably. Practically they all can be seen as Matryoshka (Russian dolls nested within each other), Deep learning is a subset of Machine Learning, and Machine Learning is a subset of AI, which is the largest space that refers to any computer program that does something smart, see Fig.1.2. People in different disciplines are trying to apply AI to make their tasks a lot easier: economists predict future market prices and crises, meteorologists use AI to predict the weather, and advertising engines try to predict buying behavior to achieve the best utilization of advertisements. The reason behind such ubiquitous use of AI is machine learning algorithms. The simplest form of machine

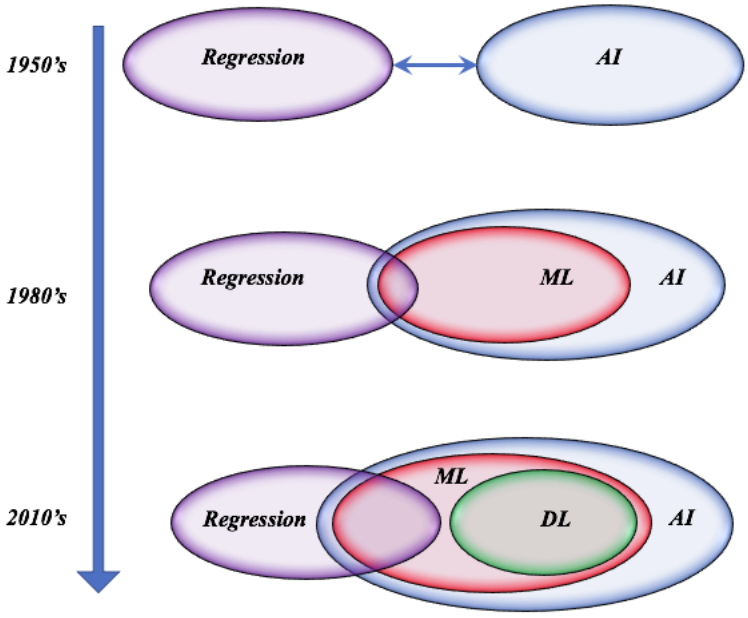

FIGURE 1.3: Regression and Machine Learning.

learning algorithms is linear regression, regardless of the time gap between them. Regression has been used for more than 200 years, while Machine learning is a very recent development. It appeared in the 1990s as steady advances in digitization, and cheap computing power enabled scientists to train computers to build finished models. The uncontrollable volume and complexity of the big data that the world is now swimming in have increased the potential of machine learning and the need for it.

The boundaries between statistical regression and machine learning are fuzzy, and finding the differences between them is controversial. However, we can say that machine learning and statistical modeling are two different fields of predictive modeling, and the difference between these two have diminished significantly over the past decade, see Fig.1.3.

Linear regression is a linear approach to modeling the relationship between a scalar response (or dependent variable, or observation, or output) and one or more explanatory variables (or independent variables, or measurements, or input, or regressors). See Fig.1.4.

Then, given the data set $\{y_i\}_{i=1}^n$ and $\{x_i^1, x_i^2, \ldots, x_i^d\}_{i=1}^n$ of n statistical units, a linear regression model assumes that the relationship between the dependent variable y and the d-dimensional vector of regressors $\boldsymbol{x}$ is linear, which means that linear regression is the algorithm that tries to find the linear function $f(\boldsymbol{x})$ that satisfies $y_i = f(\boldsymbol{x}_i)$ for all $i = 1, ..., n$.

Assume that $d = 1$, that we have only one independent variable, the function f will take the

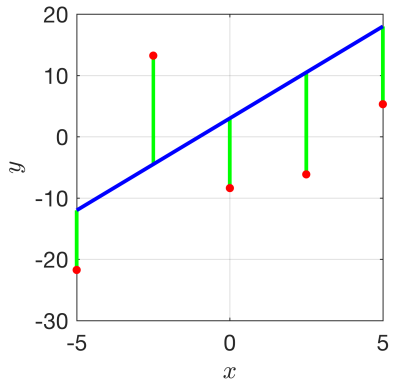

FIGURE 1.4: Random deviations (green) of the observations (red) from an governing relationship (blue) between a dependent variable y and an independent variable x.

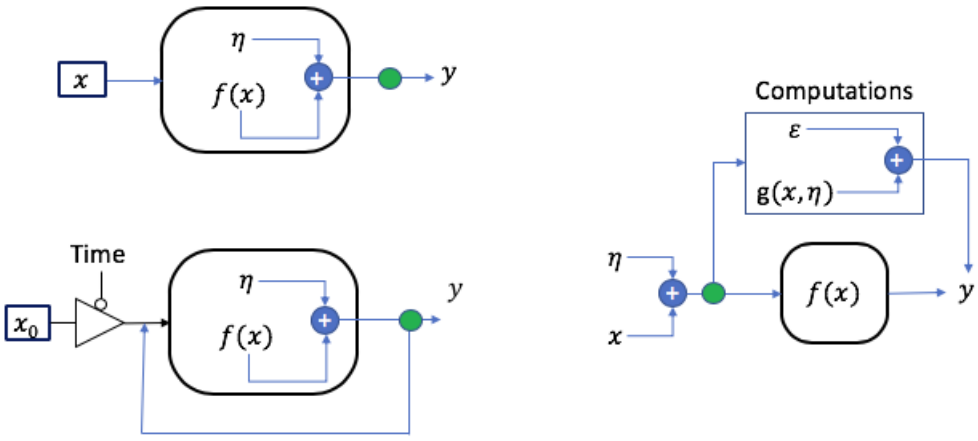

FIGURE 1.5: Noise in different sampling approaches. The green dot indicates the sampling terminal.

form $f(x) = \beta_0 + \beta_1 x$, where $\beta_0, \beta_1 \in \mathbb{R}$ are the function parameters, and then, the linear regression is the process of "parameters estimation".

In the ideal case that the measurements are exact, the problem takes its simplest closed form of solving system of linear equations, and then, it is only required $n = d$ measurements to solve the system. However, in most cases, the measurements are subject to different kind of deviations from the true underlying function, such as measurements error due to measuring tools, computations error, noise from other environment components. In the light of the physical meaning of noise, writing the function $f : X \to Y$ to include the noise effect depends on the nature of the problem, see Fig.1.5. However, for the sake of simplicity in this section, we consider a simple standard form of representation, and we discuss this issue in more details in Appendix B.

Then, considering system inputs or the independent variable $\{x_i\}_{i=1}^n$, the actual physical observations $\{y_i\}_{i=1}^n$ will have a random deviation from the true underlying governing function $f(x_i) = \beta_0 + \beta_1 x_i$, see Fig.1.4, and we can write this relation as:

$$y_i = f(x_i) + \eta_i \tag{1.6}$$

where η is random variable. (In this work we consider η to be Gaussian noise with mean $\mu = 0$, and a standard deviation σ, that $\eta \sim \mathcal{N}(0, \sigma)$, unless otherwise stated). Linear regression is the process of finding the **best** estimate for the function parameters, $\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$, that is as close as possible to the true underlying function.

The word "**best**" above, implies the need for a loss function that measures how good the estimation is. This function can be seen as an objective (cost) function and the regression process as an optimization process. One of the oldest cost functions ever known is the sum of squares given by:

$$\mathcal{C}(y, \hat{f}(x)) = \sum_{i=0}^{n} (y_i - \hat{f}(x_i))^2. \tag{1.7}$$

The use of a quadratic loss function is common, it is often more mathematically tractable than other loss functions because of the properties of variances, as well as being symmetric. In Ch.2, we discuss in more details the effect of the cost function and the details of regression methods. Finally, the goal of the regression process is find (in the ideal case) a good estimation that satisfy $\mathcal{C}(y, \hat{f}(x)) \approx 0$. In reality, $\mathcal{C}(y, \hat{f}(x)) \approx 0$ is a challenging problem due to a high levels of noise.

In the field of signal processing, noise is a general term for unwanted (and, in general, unknown) modifications that a signal may suffer during capture, storage, transmission, processing, or conversion. The nature of the noise depends on the nature of the data itself and the field of study. For example, audio noise sources can include background noise due to spurious sounds during signal capture, or audible noise due to electromagnetic vibrations in systems involving electromagnetic fields.

Noise can vary in magnitude from low values ($\sigma \approx 0$), to a significant value that may exceed the signal magnitude. Such significant noise results with measurements that largely deviate from other measurements, and these are commonly called outliers. While the noise affects the

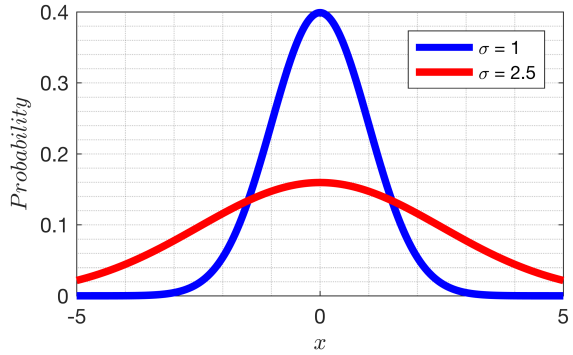

FIGURE 1.6: Normal distribution and fat tails.

accuracy of the estimated parameters proportionally, these outliers can produce a significant and radical drop in accuracy.

Outliers are one of the central points in our study, and that is because of its importance in studying and analyzing dynamical systems. Outliers can occur by chance in any distribution. However, they often indicate either measurement error or that the population has a fat-tailed distribution. In the former case, one wishes to discard them or use statistics that are robust to outliers, while in the latter case they indicate that the distribution has high skewness, see Fig.1.6, and that one should be very cautious in using tools or intuitions that assume a normal distribution.

The outlier problem appears clearly in the field of black-box modeling, where there is no prior information available, and even if the measurement "looks like" an outlier we cannot decide for sure if it is. For example, outliers due to error in data measurements can simply be discarded (i.e., when measuring temperature in Antarctica, a value of 750.5 degrees can be investigated or discarded due to instruments failure or human error), while outliers in other types of study can be the most interesting measurements, and this introduces different types of questions: are they outliers? Are they deterministic results of unknown underlying dynamic? Does this imply that the current model is incorrect? Are they the result of a the temporary external condition? What is that cause? Are they predictable?

Outliers can have harmful effects on statistical analyses. First, they generally serve to increase error variance and reduce the power of statistical tests. Second, they can increase the odds of making errors of type-I and type-II. Third, they can seriously bias or influence estimates that may be of substantive interest, [156]. Unfortunately, mathematical modeling literature

focuses on assigning low weight (such as weighted least squares), which means low importance for suspected outliers to reduce their effect on the final model.

From the other side of view, outliers can represent a significant risk. A very famous example is the Fukushima nuclear disaster. The Japanese Nuclear Commission stated in 2003 that: "The mean value of acute fatality risk by radiation exposure resultant from an accident of a nuclear installation to individuals of the public, who live in the vicinity of the site boundary of the nuclear installation, should not exceed the probability of about 10^{-6} per year (which means 1 per million years)". Fukushima Nuclear Power Plant re-study their emergency systems and the probability of natural disasters, and all calculations with considering safety margins led to a probability less than "1 in a million years" such event may occur.

After just 8 years, this "1 in a million years" event happened. The disaster initiated primarily by the "unexpected" tsunami following the "unexpected" earthquake on 11 March 2011. After the earthquake, the active reactors automatically shut down their sustained fission reactions. However, the ensuing tsunami flooded the emergency generators that were providing power to the pumps that cooled the reactors. The coolant loss led to three nuclear meltdowns, hydrogen-air explosions, and the release of radioactive material.

Interestingly, on 5 July 2012, the National Diet of Japan Fukushima Nuclear Accident Independent Investigation Commission (NAIIC) found that the causes of the accident had been **foreseeable**, and that the plant operator, Tokyo Electric Power Company (TEPCO), had failed to meet basic safety requirements such as risk assessment, preparing for containing collateral damage, and developing evacuation plans.

Such a dilemma or apparent paradox happens again and again in our daily life and scientific researches, that some event X, was unpredictable or unexpected before its occurrence, and after it occurs, we see that the signs and the primary ingredient for that event were in front of us, but we did not give attention to them. This could be true in some cases such as Fukushima disaster, where human life depends on the outcomes of probability calculations, but in many cases in scientific research, we fall in a logical fallacy when we pre-assume the conclusions starting from the given initial point, in philosophy it is called "Begging the question."

In classical rhetoric and logic, begging the question (appeared firstly in Aristotle's "Prior Analytics") is an informal fallacy that occurs when an argument's premises assume the truth

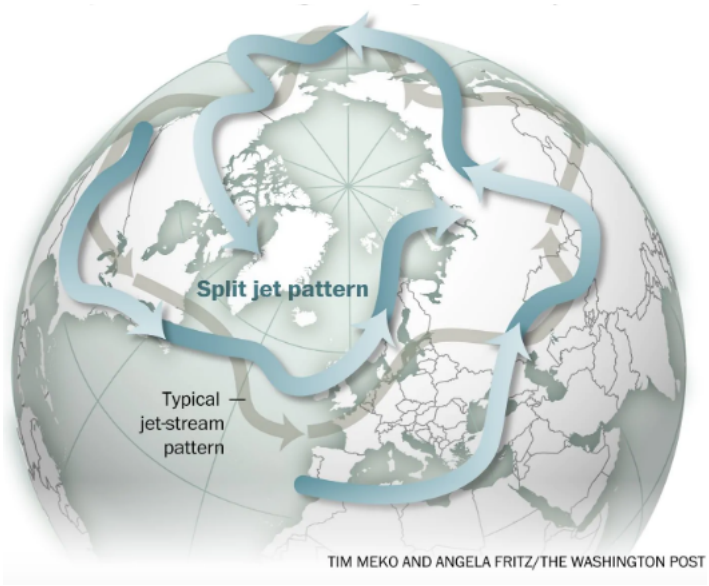

FIGURE 1.7: Atypical Jet Stream [47].

of the conclusion, instead of supporting it. It is a type of circular reasoning: an argument that requires the desired conclusion be true. This often occurs in an indirect way such that the fallacy's presence is hidden, or at least not easily apparent. The phrase begging the question originated in the 16^{th} century as a mistranslation of the Latin petitio principii, which translates to "assuming the initial point."

Rare events and their dynamics are another problem that is interpreted in terms of outliers. Rare events may have an unexpected impact on the underlying dynamic and initiate a dynamic change. One example of these phenomena on earth science is Jet Stream Break. Fig.1.7 [47], shows the change in the northern hemisphere jet stream that occurred in July 2018. As a result, the record-breaking high temperature recorded in many regions (i.e., 90 degrees in Sodankyla, Finland, which is 59 miles from the Arctic Circle. 106 degrees in Japan, highest record in history) [47], wildfires, floods, and different kind of catastrophes. Geert Jan Van Oldenborgh, a climate researcher at the Royal Netherlands Meteorological Institute, an interview with the Washington Post state that: "This kind of event was a 1-in-100-year event in 1900, now It becomes 20 times more likely" [47]. We can not state direct connections to the broken jet stream, but it is interesting to have different changes, and other rare events come after this rare event such as the fractured polar vortex in December 2018, where it split into two vortex [47].

One may consider the solution to be to make more predictions and to give more importance

to very low probability events, but this approach has many problems such as microscale probabilities that are practically immeasurable[1], and unreliable factor of safety can lead to unfeasible design or greatly increase the cost (over-engineering). The American winner of the Nobel prize in economics Paul Samuelson once said: "The stock market has predicted 9 of the past 5 recessions". Over-predicting can also lead to different kinds of crises and catastrophe, and it is a sign of poor modeling.

In the light of the previous discussion, the importance of solid modeling becomes clear, and the risk of different types of ignorance of the outliers. Fig.1.8 shows the schematic diagram of commonly used system identification methods, where different approaches try to detect bad data and reduce their effect on the model. A data pre-processing stage usually leads the modeling process, and such processing can be useful (as discussed before) in the case that we have prior information or a logical sense that allows us to ignore some measurements due to the corruption. However, If we don't have such deterministic reasons, and our only quality measure is based on the traditional statistics, then such ignorance can be harmful and can lead to a model that precisely ignore and miss-predict the most interesting events.

This importance of outliers motivates us to focus, in parallel with our main objectives, on development of regression method that takes each outliers point of measurements on the account, to construct our mathematical model without discarding, weighting, or pre-processing of any kind for our measured data.

1.2 Objectives

The field of system identification uses statistical methods to build mathematical models of dynamical systems from measured data. Our first objective in this work is:

1 Developing a robust and reliable information-theoretic based method for system identification of complex systems.

[1]Examples: 1-The probability of selecting a member from a countably infinite set of equally weighted discrete elements is immeasurable. 2- Some side effects that are written on drugs never occurred or happened to the test sample or the patients, but they are expected, without precise probability assigned to them. See also "The Howland trial".

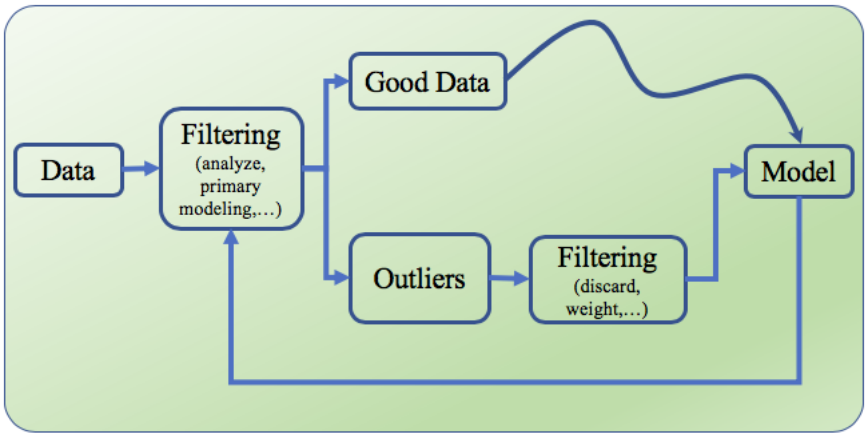

FIGURE 1.8: Schematic diagram of commonly used approaches for system identification.

Our method aims to achieve robust and reliable results on system identification and to overcome the common limitation for most of system identification methods, which is the outliers problem.

Moreover, the number of measurements required for system identification is proportional to the problem's dimension, and as the dimension increases, more measurements are needed for system identification purposes, Which makes it a challenging problem to identify the complex systems of high dimensionality. Also, the limited scalability of system identification methods makes it unreliable to cast the complex networks structure recovery as a system identification problem. This present the following objectives for our work:

2 Robust System Identification of high dimensional complex systems: Our goal is to develop a SID method that adopt the black box modeling approach for modeling high dimensional systems under high expansion orders, with robustness to noise, outliers, and ill-conditioning problems.

3 Introduce System-Identification approach for recovering network structures in complex networks: Under the consideration of the high dimensional complex systems, such as large coupled networks of high dimensional dynamical oscillators, our goal is to develop a SID method that is robust to noise and outliers for recovering coupling structures.

Prediction of complex systems behavior is commonly made based on a model, and the mathematical modeling of the image observed complex systems is a challenging problem according to the availability of observations and the difficulty of extracting the vector fields of observed

dynamic. Our final objective in this work is to introduce a system identification and prediction analysis in image observed complex systems, and then:

4 Introduce a new principle for equation-free, parameters-free, vector-field free predictions in image-observed complex systems: Our goal is to adopt the principle of coherent structures in spatiotemporal systems to introduce a method for detecting early warning signs of critical transitions.

1.3 Document Scope

This dissertation divided into six chapters and two appendices.

In Chapter 2, we review the basic principles of Mathematical modeling and system identification in complex systems as a leading source of analysis and prediction in complex systems. We also discuss the theoretical and applied foundations for linearizing nonlinear dynamics, and basis functions library construction. We also, review the method of least squares and related topics that are connected to the core of our research such as the law of large number, random numbers generators, L_0 minimization, and regularization techniques.

In Chapter 3, we review the basis of the information theory, the principle of causation entropy, the asymptotic equipartition property, and its implications, which is a central idea to our Entropic Regression approach.

In Chapter 4, we introduce our Entropic Regression (ER) method for sparse system identification. Our approach overcomes the competing challenges of potential overfitting, limited data, noise, and in particular outliers in observations. We demonstrate the robustness of our method in several examples varying from low dimensional systems, large dimensional systems, complex PDE, and coupled complex networks of chaotic oscillators. The work presented in this chapter primarily follows our submitted manuscript: Abd AlRahman R. AlMomani, Jie Sun, Erik Bollt, "*How Entropic Regression Beats the Outliers Problem in Nonlinear System Identification*".

In Chapter 5, we introduce our Directed Partitioning method, for detecting coherent structure and predicting critical transitions in complex fluid flows. We demonstrate the efficiency of our approach on the problem of detecting coherent structures in clouds of Jupiter, and

for predicting the critical transitions in Antarctica ice shelves. The work presented in this chapter primarily follows our published paper: Abd AlRahman R. AlMomani and Erik Bollt. "*Go With the Flow, on Jupiter and Snow. Coherence from Model-Free Video Data Without Trajectories*". In: Journal of Nonlinear Science (2018). ISSN: 14321467.

In Chapter 6, we discuss some of the extensions and future directions of our work, such as introducing an efficient approach for recovering the coupling structure of large complex networks.

In Appendix A, we provide a brief review for the K-means and spectral clustering methods, and in Appendix B we provide a summary and comments on fundamental misconceptions in sparse system identification.

1.4 Contributions

This thesis has made a number of significant contributions to the fields of systems identification, complex networks, and computer vision. We summarize the main contributions of this thesis by the following:

1. **The Entropic Regression Method**: The entropic regression method is sparse system identification method that has numerous advantages, compared to the current state of the art methods. The main advantages of this method:
    - Robustness to noise and outliers under small data regime, which was the first objective of this thesis.
    - Reliable SID under the black-box modeling, with no prior information about the order of the system, and without excluding and filtering the data based statistical inference or weighting functions. This was our second objective in this thesis.
    - Reliable sparse SID for high dimensional, and high order systems, which made it possible to tackle the causal inference in complex networks problem in terms of sparse regression. This was our third objective in this thesis.
    - Robust and reliable prediction of the long time behavior of chaotic systems, under noise, outliers, low measurements, and black box modeling.

2. **The Directed Partitioning Method**: The directed partitioning method is a spatiotemporal segmentation method that has numerous advantages.The main advantages of this method:

    - Model-free, equation-free, and vector field free detection of coherent structures in movie frames, which provides unsupervised object detection and tracking ability.
    - As a result of the previous point, our method detect the spatiotemporal changes which make it able to indicate early warning sign for spatiotemporal changes such as fractures, and merging coherent structures with different properties.
    - This was our forth and last objective in this thesis, and we applied the method for predicting critical transitions on Antarctica ice shelves.

Our work result with the following publications:

1. Abd AlRahman R. AlMomani and Erik Bollt. "*Go With the Flow, on Jupiter and Snow. Coherence from Model-Free Video Data Without Trajectories*". In: Journal of Nonlinear Science (2018). ISSN: 14321467.
2. Abd AlRahman R. AlMomani, Jie Sun, Erik Bollt, "*How Entropic Regression Beats the Outliers Problem in Nonlinear System Identification*". Submitted.

In addition to the direct contributions mentioned above, we achieved significant results in different topics, some of them are discussed in the extensions part in Chapter 6, and some them represent a straight forward applications to our methods. In Fig.1.9, we present a schematic diagram for our main contributions, and some of their applications.

Based on significant results and findings, and as extension of our work presented in this thesis, we list the following ongoing research papers in the writing stage:

1. Early Warning Sign Tool For Critical Transition in Antarctic Ice Sheet.
2. Causality inference on large complex networks of non-identical oscillators.
3. Dynamic Dictionary Learning: Unsupervised and Efficient basis construction.
4. Leadership structures in complex networks, and information accelerated synchronization.

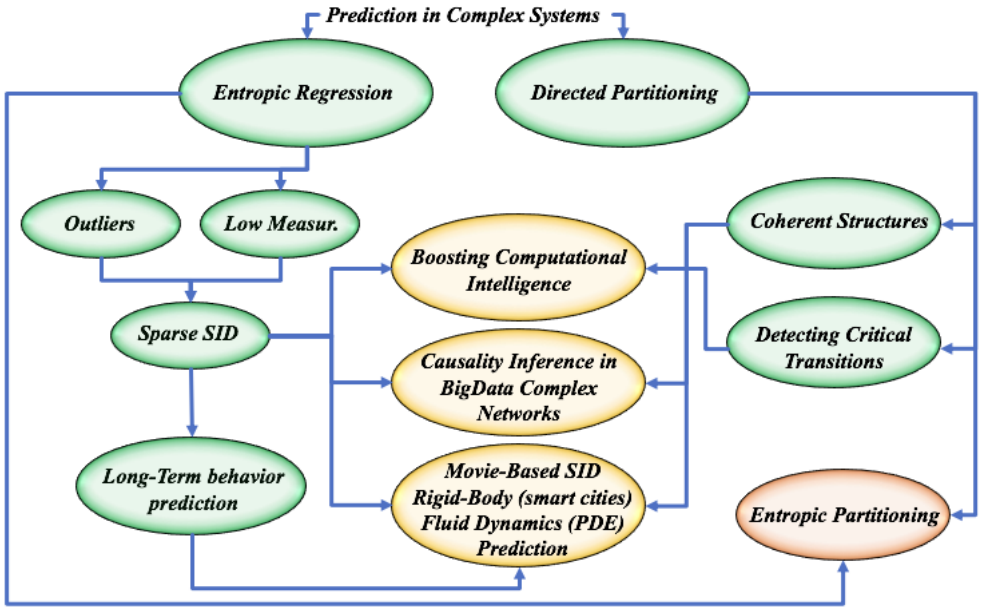

FIGURE 1.9: Thesis Direct and indirect contributions. In green we see our contributions on this thesis, and in yellow, we see some of the direct applications of our work in different fields.

5. Collective behavior modeling

Chapter 2

Identification of Complex Systems

> All the mathematical sciences are founded on relations between physical laws and laws of numbers so that the aim of exact science is to reduce the problems of nature to the determination of quantities by operations with numbers.
>
> *James Clerk Maxwell*
>
> *(13 June 1831 – 5 November 1879)*

An open system is a system that has external interactions, that can take the form of exchange of energy, matter, or information between the system and the surroundings, see Fig.2.1, depending on the discipline which defines the concept [145, 146]. A mathematical model is a description of a system using mathematical expressions, and the quality of scientific research depends on how well the mathematical models agree with results of repeatable experiments.

One could build a model based on first principles, for example, a model for a physical process from the Newton equations, but in many cases, such models are overly complicated and possibly even impossible to obtain in reasonable time due to the complex nature of many systems and processes.

A powerful alternative is, therefore, to start from measurements of the behavior of the system and the external influences (inputs to the system) and try to determine a mathematical relation between them without going into the details of what is happening inside the system. This approach is called system identification. Two types of models are common in the field of system identification:

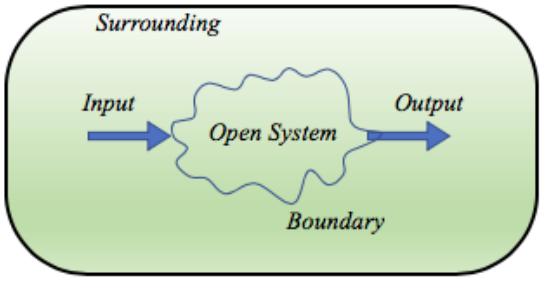

FIGURE 2.1: Open System illustration.

- Grey box model [145, 203]: Although the precise operations that are occurring inside the system are not entirely known, some information based on theoretical assumptions, experimental data, or mathematical sense could be available. However, this model still has some unknown parameters which can be estimated using system identification. One example [203], is the Monod saturation model for microbial growth. The model contains a hyperbolic relationship between substrate concentration and growth rate, but this can be justified by molecules binding to a substrate without going into detail on the types of molecules or types of binding.
- Black box model [145]: No prior model is available. Most system identification algorithms are (or supposed to be) of this type. One must therefore assume a very general model form with many parameters.

Parameter estimation in terms of gray box modeling is comparatively easy if the model form is known, but this is seldom to be the case. In the field of dynamical systems, we are mostly concerned with black box modeling, since the underlying dynamics of a highly complex system is unknown.

2.1 Dynamical Systems

In mathematics, a dynamical system is a system in which a function provides an analytical description of the evolution of a system for a long time. It is one of the primary tools employed in science to model physical systems such as electrical, mechanical, financial systems.

Dynamical systems theory is an area of mathematics used to describe the behavior of the complex dynamical systems, usually by employing differential equations (continuous system) or difference equations(discrete system). With the continuous system in mind, we can represent

a dynamical system by a set of nonlinear differential equations:

$$\dot{\mathbf{z}} = f(t, \mathbf{z}, \mathbf{u}), \tag{2.1}$$

where $t \in \mathbb{R}^+$ is the time, $\mathbf{z} \in \mathbb{R}^d$ is a d-dimensional state variable, $\mathbf{u} \in \mathbb{R}^n$ is a n-dimensional input variables, and $f : \mathbb{R}^+ \times \mathbb{R}^d \times \mathbb{R}^n \mapsto \mathbb{R}^d$. The input variables $\mathbf{u}$ can be seen as external control variables that can modify the evolution of the state variables $\mathbf{z}$ through time, which then can be called a forced dynamical system. Without the external control variable $\mathbf{u}$, we have the unforced state equation:

$$\dot{\mathbf{z}} = f(t, \mathbf{z}), \tag{2.2}$$

which is the commonly used representation of non-autonomous dynamical system, we should note that dropping the inputs term $\mathbf{u}$ in Eq.2.2 does not imply that $\mathbf{u} = 0$ since $\mathbf{u}$ can be defined as a function of the state variables and time, $\mathbf{u} = g(t, \mathbf{z})$, then the control variable can be eliminated yielding the unforced state equation.

An autonomous (or time-invariant) dynamical system is a system whose stat evolution does not depend on the time, and therefore the current state of the system is a function on the previous state only:

$$\dot{\mathbf{z}} = f(\mathbf{z}), \tag{2.3}$$

and it takes the following form in the case of discrete dynamical systems:

$$\mathbf{z}_{n+1} = f(\mathbf{z}_n), \tag{2.4}$$

where $\mathbf{z}_n$ represent the current state at iteration n, and $\mathbf{z}_{n+1}$ represent the next state of the system. The concept of a dynamical system has its origins in Newtonian mechanics, where the evolution rule of dynamical systems is a deterministic relation that gives the state of the system for only a short time into the future, and determining all future states requires iterating the relation many times, each advancing time a small step. We refer to the iteration procedure as solving the system or integrating the system.

Given an initial point, if the system is solvable, it is possible to determine all its future states which are a collection of points known as a trajectory or orbit. However, this is not always the case. In Newtonian mechanics, the n-Body problem created a broad discussion about

solvability and stability of dynamical systems.

As discussed in the introduction of this chapter, we start from measurements, observe the state's evolution of the dynamical system for a finite time, and then we try to build our mathematical model, which means discovering the function f in Eqs.(2.1-2.4). This process is known as the inverse problem, which has the goal of constructing a mathematical model with an accurate estimation of the parameters, such that the model can reproduce the dynamic.

2.2 Modeling and Parameters Estimation

Estimation theory [153], is a branch of statistics that deals with estimating the values of parameters based on measured empirical data, and an estimator aims to estimate the unknown parameters that describe the underlying physical setting using the measurements. When the data consists of multiple variables, and one is estimating the relationship between them, estimation is known as regression analysis.

Regression analysis is a set of statistical processes that aims to construct an interaction model view that describes the relationships among variables which play the rule of "Predictors". The Regression process usually includes dependent and independent predictors. Regression analysis helps one understand how the typical value of the dependent predictors changes when any one of the independent predictors is varied, while the other independent predictors are held fixed.

To describe the basic framework for regression, let us consider some dependent variable y whose governing dynamics we would like to understand through some independent variables $x_1, x_2, \ldots, x_{K-1}$ which are sometimes called features, explanatory variables, or predictors. We call the number of independent variables $K-1$ the number of features or the dimension of feature space. The main task for regression is predicting y using the features $x_i, i = 1, \ldots, K-1$ by analyzing a set of training data that consists of experimental measurements, or sampled values of the independent features together with the independent observations y.

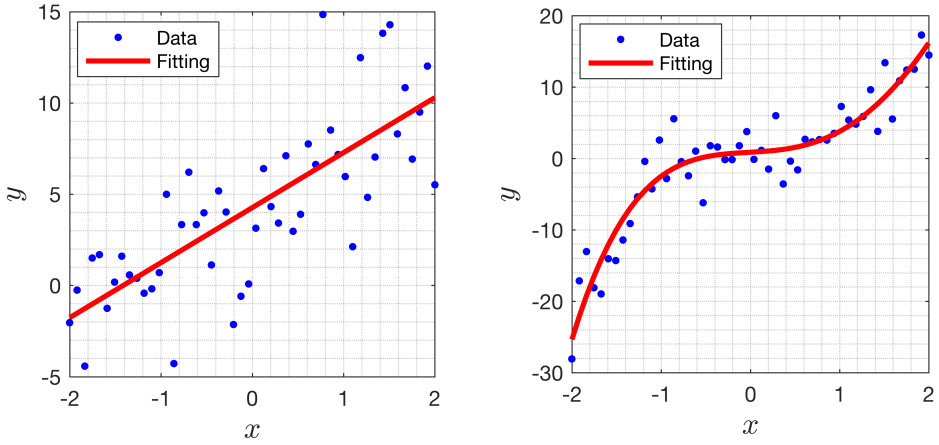

FIGURE 2.2: Linear Regression. (Left) For the observed data, in blue, we assume the relation between x and y is linear, then we assume the form $y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x$, and apply some linear regression method to find the parameters $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1$ to make Eq. (2.5) as accurate as possible. (Right) For the observed data, we assume the relation between x and y is cubic, then we assume the form $y = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 x + \boldsymbol{\beta}_2 x^2 + \boldsymbol{\beta}_3 x^3$, and apply some "linear regression" method to find the parameters $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3$. Note that we still apply linear regression on nonlinear function, because Linear regression assumes y as a linear combination from other functions, that are not necessarily linear. The example shown here assumes the gray box modeling.

The basic assumption of linear regression is that the dependent variable y is a linear combination, or linear function of the independent variables x, then y can be written as,

$$y = \boldsymbol{\beta}_0 + \sum_{i=1}^{K-1} \boldsymbol{\beta}_i x_i \tag{2.5}$$

where $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{K-1}$ are the constants, or the parameters that describe the interaction process between the features and the observations. Linear regression methods aims to to find the best choices of values for the parameters $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{K-1}$ to make Eq. (2.5) as accurate as possible. See Fig.2.2.

Practically, the disparity between the definitions of the terms "best" and "accurate" in the above sentence, is the core difference, or the main generator of the 10's of regression methods we see in literature today and we discuss this difference throughout this thesis.

Recall that the regression method aims to find the best parameters by analyzing the sampled features, the number of the sampled features can play a main rule in regression feasibility and accuracy. The number of sampled measurements N represents the number of available

independent equations available, and from linear algebra, we know that it requires $N = K$ equations to solve the system of K parameters ($K - 1$ features). When the number of measurements N, is larger than the number of unknown parameters K, then the excess of information contained in $N - K$ measurements is used to make statistical predictions about the unknown parameters. We refer to this excess of information as the degrees of freedom in the regression [192]. We can call the system:

- Underdetermined System: when it has a negative degree of freedom ($N - K < 0$), and such a system has infinitely many solutions or no solution at all.
- Determined System: when it has no degree of freedom ($N - K = 0$), and such a system has a unique solution.
- Overdetermined System: when it has a positive degree of freedom ($N - K > 0$). In general, the overdetermined system has no solution, unless the excess information is nearly dependent, and it refers to a description of the determined system.

The accurate estimation of the "best" (in some sense) parameters $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ gives more "accurate" prediction for the observations value y under new varying features through the model given in (2.5).

2.3 Linearization of Non-Linear Dynamics

Given a time-series from a chaotic dynamical system, local methods to construct the model of systems have been developed by many authors, including, [25, 28, 67, 68, 109]. These methods are called local because a model is fitted for each neighborhood of the phase space, making a grid of local models through the phase space. These local models are data-driven or weak models.

Global methods, adopted by many researchers to model continuous time systems with ODEs [27, 28, 75, 82] by ***treating each data point as an initial value problem for the ODE***. Then, the regression problem in Eq. (2.5) can be re-written as:

$$\dot{z} = F(z, \boldsymbol{\beta}) \tag{2.6}$$

where z is the observed measurements, $\dot{z}$ is the vector field, and $\boldsymbol{\beta}$ is the set of parameters. In more general form, The construction of the vector field aims to build mathematical models from observed data. In analogy to Eq. (2.5), assume that we observe the dynamic $\boldsymbol{f} \in \mathbb{R}^{N\times d}$, such that $\boldsymbol{f}_{ij}$ is the j^{th}-dimension observation of the system at time t_i, $i = 1, \ldots, N$, and our features here represent time series $z_{:j} \in \mathbb{R}^N$, $j = 1, \ldots, d$, and $\boldsymbol{z} \in \mathbb{R}^{N\times d}$. Then, one can consider the dynamical system modeled by a series expansion in some basis as:

$$\boldsymbol{f} = [\dot{\boldsymbol{z}}_{:1}, \ldots, \dot{\boldsymbol{z}}_{:d}] = [F_1(\boldsymbol{z}, \boldsymbol{\beta}), \ldots, F_d(\boldsymbol{z}, \boldsymbol{\beta})]\,, \tag{2.7}$$

where j=1,...,d, $\boldsymbol{\beta} \in \mathbb{R}^d$ denotes a set of system parameters, and $F_j : \mathbb{R}^{N\times d} \to \mathbb{R}^N$ is governing dynamics of the observed system generated from a nonlinear, high-dimensional dynamical system. Note that given our "measured" time series $\boldsymbol{z}(t)$, one can estimate the derivatives by any of the standard Newton-Cotes methods, explicit Euler's method of course being the simplest, giving $\dot{\boldsymbol{z}}_j(t_k) = \frac{z_j(t_{k+1}) - z_j(t_k)}{\tau_k} + \mathcal{O}(\tau_k)$ with $\tau_k = t_{k+1} - t_k$. The first step is to recast the nonlinear SID problem into a computational inverse problem, by considering an appropriate set of basis functions that span the space of functions including the system of interest. There are few requirements on the set of basis functions $\{\phi_k(\boldsymbol{z})\}$, and a common choice is the *polynomial basis*, represented by

$$\boldsymbol{\phi} = [\phi_0(\boldsymbol{z}), \phi_1(\boldsymbol{z}), \phi_2(\boldsymbol{z}), \ldots] = [1, z_1, z_2, \ldots, z_d, z_1^2, z_1 z_2, \ldots, z_{d-1} z_d, \ldots]. \tag{2.8}$$

Then, the individual component functions of F in (2.7) can be generally written as:

$$\boldsymbol{f}_j = F_j(\boldsymbol{z}, \boldsymbol{\beta}) = \sum_{k=0}^{\infty} a_{jk} \phi_k(\boldsymbol{z}), \tag{2.9}$$

for a linear combination of basis functions $\{\phi_k\}_{k=0}^{\infty}$. Note here that the basis functions do not need to be mutually orthogonal, which makes this approach simple with a lot of flexibility for the mathematical intuition to play a significant role in choosing the basis functions or the expansion order. However, the simplicity of this approach comes with the cost of risking the numerical stability of the solution.

In order to judge how good the model is, some researchers, including Brown, Rissanen [28, 159] consider the question, how much does truncating the model reduce accuracy? They assume

that the vector field of the dynamical system can be modeled by a set of differential equations with the functional basis of polynomials composed of all combination of state variables up to a proper order.

2.4 Function Approximation and Basis Functions

In mathematics, approximation theory is concerned with how to represent a function in the form of more simple functions with a reasonable degree of accuracy. The "more simple" term can be defined differently according to the application, but in common sense, a "more simple" function means it is more natural to be analyzed, or it requires fewer computations complexity. The aim is to make the approximation as close as possible to the actual function. In this section, we provide the very basic principles of the classical framework of function approximation.

We have a metric space $\mathcal{X}$, and we need to approximate a given element $x \in \mathcal{X}$ by an element κ of some subset $\mathscr{K} \subset \mathcal{X}$. The distance between x and κ needs to be as small as possible according to some metric d. A metric space $\mathcal{X}$ is a set $\mathcal{X}$ together with a metric d. The metric d is defined as a function $d : \mathcal{X}^2 \mapsto \mathbb{R}$ which satisfies the three properties:

1. Positive definite: $d(x, y) \geq 0$ for all $x, y \in \mathcal{X}$, with equality hold when $x = y$.
2. Symmetric: $d(x, y) = d(y, x)$ for all $x, y \in \mathcal{X}$.
3. Satisfies the triangle inequality: $d(x, z) \leq d(x, y) + d(y, z)$ for all $x, y, z \in \mathcal{X}$.

We can then define the distance between x and y as $d(x, y) = \|x - y\|^\alpha$ for any $\alpha \in (0, 1]$.

To any element $x \in \mathcal{X}$ and any subset $\mathscr{K} \subset \mathcal{X}$ we associate the distance $d(x, \mathscr{K})$ from x to $\mathscr{K}$, which by definition is:

$$d(x, \mathscr{K}) = \inf_{\kappa \in \mathscr{K}} d(x, \kappa), \qquad x \in \mathcal{X}, \mathscr{K} \subset \mathcal{X}. \tag{2.10}$$

A very good approximation is obtained when $d(x, \mathscr{K}) = 0$. $d(x, \mathscr{K}) > 0$ represents a certain error, and it may happen that:

$$d(x, \kappa) > d(x, \mathscr{K}), \forall \kappa \in \mathscr{K}. \tag{2.11}$$

which is a deviation from the best approximation that satisfies $d(x, \kappa) = d(x, \mathscr{K})$. In the case of Eq.2.11, we are interested in constructing a sequence (κ_j of elements of $\mathscr{K}$ such that $d(x, \kappa_j) \to d(x, \mathscr{K})$ as $j \to \infty$. Such sequence is called an approximating sequence.

So, the best approximation is the set:

$$\{\kappa \in \mathscr{K}; d(x, \kappa) = d(x, \mathscr{K})\}, \tag{2.12}$$

and this set can be empty, have exactly one element, or may have more than one element. A simple unique approximation can appear in the case of Hilbert space H, where if $\mathscr{K} \subset H$ is a nonempty, closed, and convex set, then there exists a unique closest point $\pi(x) \in \mathscr{K}$.

Regardless of the complexity of the function f and unavailability of a prior model (gray box modeling), polynomials expansion of the state variables was from the earliest employed techniques for function approximation and constructing mathematical model. In 1885, Karl Weierstrass introduced his approximation theorem [198], which has both practical and theoretical relevance, because polynomials are among the most straightforward functions (simple to analyze), and because (in a modern point of view) computers can efficiently evaluate polynomials.

Theorem 2.4.1. ***(Weierstrass approximation theorem)****[152, 179]: Suppose f is a continuous real-valued function defined on the real interval $[a, b]$. For every $\epsilon > 0$, there exists a polynomial p such that for all $x \in [a, b]$, we have $|f(x) - p(x)| < \epsilon$, or equivalently, $\|f - p\|_\infty < \epsilon$.*

Weierstrass approximation theorem has its significant value, and it summarizes 100 years of efforts starting from Fourier, Legendre, Chebyshev, and Gudermann. And while these efforts focused mainly on the function approximation for simplifying forward models (i.e., simplifying solving PDE), Weierstrass theorem forms the basis of the efforts toward solving the inverse problem and employing different kinds of polynomials for this purpose.

2.5 Power Series and Carleman's Linearization

For the importance of (Linearization of Non-Linear Dynamics), and for the considerable confusion and misconceptions in the literature regarding the topic, we discuss in Appendix B the

history of Carleman's Linearization, which is our adopted approach in this work. The basic idea (that we adopt in this work) of linearization of nonlinear dynamics starts with a remark given by Poincare in 1908 in the International Congress of Mathematicians in Rome, "One should be able to apply the theory of linear integral equations to the study ordinary non-linear differential equations" [90]. Torsten Carleman worked on an approach to embed a system of non-linear differential equations into an infinite set of linear equations, and in 1932, he introduced a theoretical technique to globally linearize systems of nonlinear polynomial equations, [40]. And in the recent decades, it became a favorable technique for many researchers, [30, 38, 62, 195, 205].

The Carleman matrix of an infinitely differentiable function $f(x)$ is defined as

$$M_{jk} = \frac{1}{k\,!}\left[\frac{d^k}{dx^k}(f(x))^j\right] \tag{2.13}$$

and it satisfy the Taylor series equation:

$$(f(x))^j = \sum_{k=0}^{\infty} M_{jk} x^k. \tag{2.14}$$

This original technique considers the powers of the function, j, and the order of power polynomials, k, in the approximations process, and it combines the ODEs order, model order, and the set of observation using a systematic Kronecker product technique, to model high order ODEs. Extended forms and generalization of Carleman embedding technique have been developed [110, 113, 176]. The form we adopt in our work have been discussed extensively by Erdos and Jabotinsky [64], Kowalski and Steeb [110], which takes the form of approximating the function by power polynomials in the form:

$$f(x) = \sum_{k=0}^{\infty} \boldsymbol{\beta}_k x^k \tag{2.15}$$

Additionally, when extended in the multivariate case it includes all interaction terms between variables.

The problem of nonlinear system identification is to reconstruct the functional form as well as parameters of the underlying system, that is, to infer the nonlinear function $\boldsymbol{F}$. Under the basis representation (2.9), the identification of $\boldsymbol{F}$ becomes equivalent as estimating all the

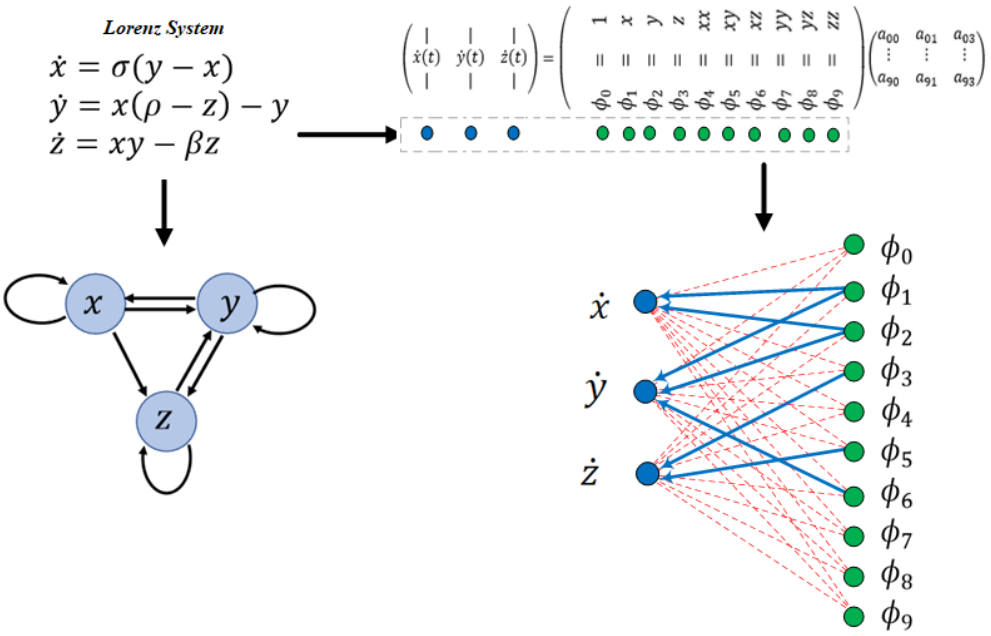

FIGURE 2.3: (Left) Lorenz system with the standard graph representation of dynamical systems. See [165]. (Right) Linear combination of nonlinear basis functions, with coupling coefficients $\boldsymbol{\beta}_{ij}$ can be described by an influence matrix $\boldsymbol{\beta}$, or equivalently a coupling graph (right-bottom) which when sparse, there can be great savings in fitting just the nonzero coefficients. Therefore here the edges describe the influence of the data between observations through the basis functions as shown. Sparsity structure of $\boldsymbol{\beta}$ corresponds to missing edges in the graph, which we discover by our proposed entropic regression as an information flow problem.

parameters $\{\beta_{jk}\}$. In practice, the infinite series is truncated after a finite number of terms; such truncation together with observational noise defines a forward model of the inverse problem:

$$F_j(\boldsymbol{z}(t), \boldsymbol{\beta}) = \sum_{k=0}^{K-1} \beta_{jk} \phi_k(\boldsymbol{z}(t)), . \tag{2.16}$$

Where $t = t_0, \ldots, t_N$, $i = 1, \ldots, d$, and K is the number of chosen basis functions. The same can be written using modern matrix methods for data analysis [30, 58, 188, 195, 205]:

$$\begin{pmatrix} | & | & | \\ \dot{z}_1 & \ldots & \dot{z}_d \\ | & | & | \end{pmatrix} \approx \begin{pmatrix} | & | & | & | \\ \phi_0(z) & \phi_1(z) & \ldots & \phi_K(z) \\ | & | & | & | \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_{00} & \ldots & \boldsymbol{\beta}_{0d} \\ \vdots & \ddots & \vdots \\ \boldsymbol{\beta}_{K0} & \ldots & \boldsymbol{\beta}_{Kd} \end{pmatrix}. \tag{2.17}$$

Then, our linear system can be written in a matrix form as:

$$\boldsymbol{f} = \Phi\boldsymbol{\beta}, \tag{2.18}$$

where $\boldsymbol{f} \in \mathbb{R}^{N\times d}$ and $\Phi \in \mathbb{R}^{N\times K}$ are given, with the goal to estimate $\boldsymbol{\beta} \in \mathbb{R}^{K\times d}$. Figure 2.3 shows the structure of the Lorenz system under standard polynomial basis up to quadratic terms.

This was the theme in [205], as well as some of the compressed sensing literature [12, 35, 36, 38, 39, 57, 187, 188, 194]. There are conflicting interests with regard to how many terms to include in the truncated series representation. While a large set of basis functions allows for a rich class of behaviors, too large a set causes problems with numerics, and convergence of fitting accuracy and overfitting, which also requires ever more data to fit an exponential explosion of terms. Recent breakthroughs in nonlinear SID has found a way to overcome this apparent paradox, by allowing a large set of basis functions and meanwhile imposing sparsity in the model, thus mitigating the issue of overfitting [55]. The success of such sparsity-based SID has been demonstrated in several recent works and applications [44, 58, 93, 103, 189].

2.6 Basis Expansion

In this section, we discuss different kinds of polynomials as basis functions. For each polynomial type, we show a numerical example for function approximation using polynomials. We choose the following two functions for this purpose:

$$\begin{aligned} f(x) &= x - x^2 + cos(3.2x) \\ g(x) &= sin(e^x) \end{aligned}$$

Note that $g(x)$ is known as an example for chaotic functions.

2.6.1 Fourier Series

Fourier series is a periodic function composed of harmonically related sinusoids, combined by a weighted summation. The Fourier series in terms of sine-cosine can be written as:

$$s_N(x) = \frac{a_0}{2} + \sum_{n=1}^{N} \left[a_n \cos\left(\frac{2\pi n x}{T}\right) + b_n \sin\left(\frac{2\pi n x}{T}\right) \right] \tag{2.19}$$

A Fourier series is an expansion of a periodic function $f(x)$ in terms of an infinite sum of sines and cosines, and it utilizes the orthogonality of the sine and cosine functions. The study of Fourier series is known as harmonic analysis and is it is useful to represent arbitrary periodic function into a set of simple terms that can be solved individually, and then recombined to obtain the solution of the original problem or an approximation to it to proper accuracy. Fourier introduced the series to solve the heat equation in a metal plate, publishing his initial results in 1807, and although the original urge was to solve the heat equation, it became apparent that the same techniques could be applied to a wide array of mathematical and physical problems. Fig.2.4 shows function approximation using Fourier series.

2.6.2 Chebyshev polynomials

Chebyshev polynomials, named after Pafnuty Chebyshev, are a sequence of orthogonal polynomials that can be defined recursively. There are two kinds: Chebyshev polynomials of the first kind (T_n), and Chebyshev polynomials of the second kind (U_n). The Chebyshev polynomials (the first and second kind), are polynomials of degree n and the sequence of Chebyshev polynomials composes a polynomial sequence.

The Chebyshev polynomials of the first kind are given by:

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_{n+1}(x) &= 2xT_n(x) - T_{n-1}(x) \end{aligned} \tag{2.20}$$

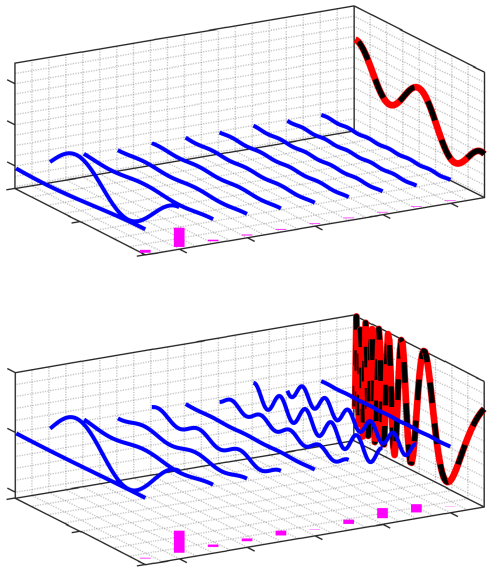

FIGURE 2.4: Function approximation with Fourier series. (Top) In red we see the function $f(x) = x - x^2 + cos(3.2x)$ to be approximated, in dashed-black we see the approximation of the function using Fourier series. The spikes in magenta color represent the magnitude of Fourier coefficients, and in the same plan of each coefficient, we see, in blue, the first few terms of Fourier series multiplied by the coefficient value. Sum of the blue signals is the dashed-black signal. (Bottom) With the same colors as above, we see the fitting result of the chaotic function $g(x) = sin(e^x)$. We see that for the simple function above, Fourier coefficients have only a few of them have significant magnitude, and all others are minimal, and even when the complexity increase in $g(x)$, some of the coefficients are negligible in terms of magnitude.

and the second kind are given by:

$$\begin{aligned} U_0(x) &= 1 \\ U_1(x) &= 2x \\ U_{n+1}(x) &= 2xU_n(x) - U_{n-1}(x) \end{aligned} \tag{2.21}$$

Fig.2.5 shows the first few terms of Chebyshev polynomials, and Fig.2.6 shows different functions approximation using Chebyshev polynomials.

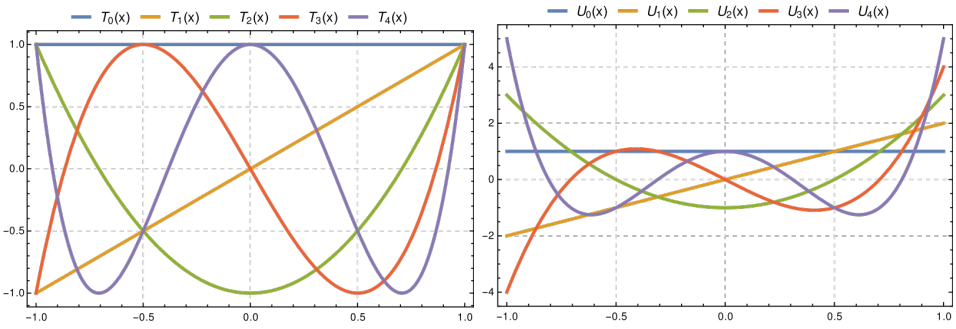

FIGURE 2.5: The first few terms of Chebyshev polynomials: (Left) First Kind Chebyshev polynomials. (Right) Second Kind Chebyshev polynomials. (source: [148]).

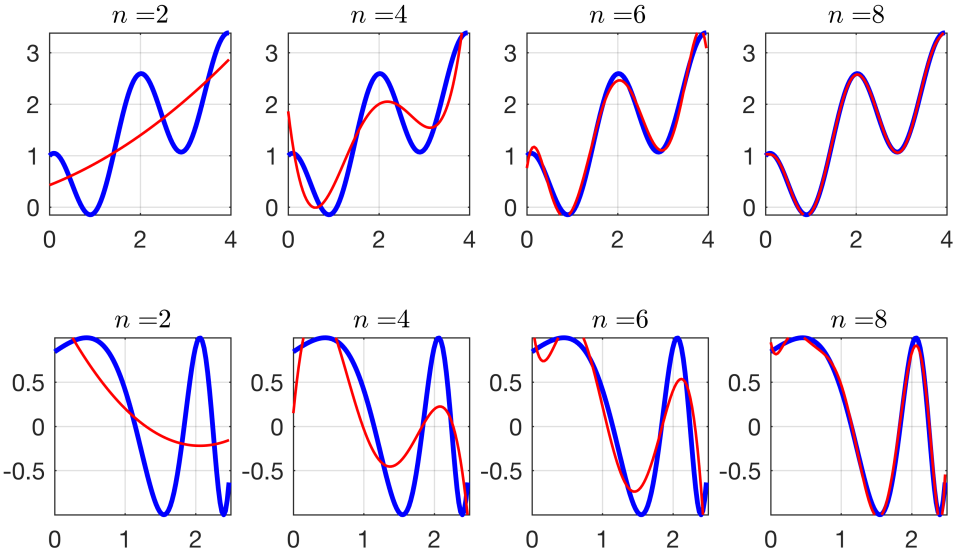

FIGURE 2.6: Function approximation with Chebyshev polynomials with different degree. In blue we see the underlying, or true function and in red we see the approximation using Chebyshev polynomials. (Top) Approximation of the function $f(x) = x - x^2 + cos(3, 2x)$. (Bottom) Approximation of the function $g(x) = sin(e^x)$.

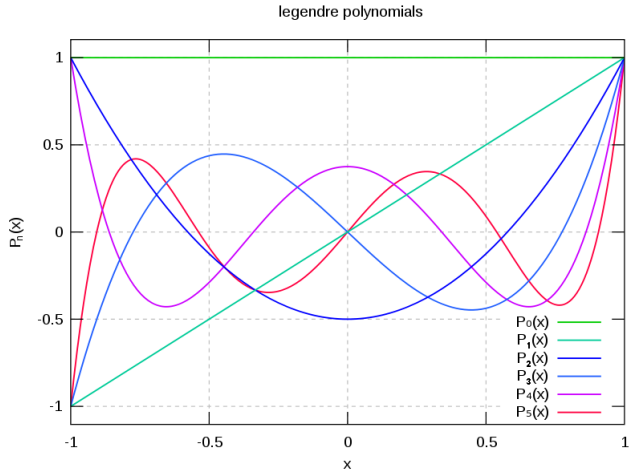

FIGURE 2.7: The first few terms of Legendre polynomial. (source: [149]).

2.6.3 Legendre polynomials

Legendre polynomials, named after Adrien-Marie Legendre, is one of the oldest orthogonal polynomials (1782), and they are a system of complete and orthogonal polynomials, and since their discovery, they have been used in numerous applications in engineering and science. They can be defined in many ways, and the various definitions highlight different aspects as well as propose generalizations and connections to different mathematical structures and physical and numerical applications.

One of the most straightforward representations for Legendre polynomials is by Bonnet's recursion formula, and it is given by:

$$\begin{aligned} P_0(x) &= 1 \\ P_1(x) &= x \\ P_{n+1}(x) &= \frac{1}{n+1}\left((2n+1)xP_n(x) - nP_{n-1}(x)\right) \end{aligned} \tag{2.22}$$

Fig.2.7 shows the first few terms of Legendre polynomials, and Fig.2.8 shows different functions approximated using Legendre polynomials.

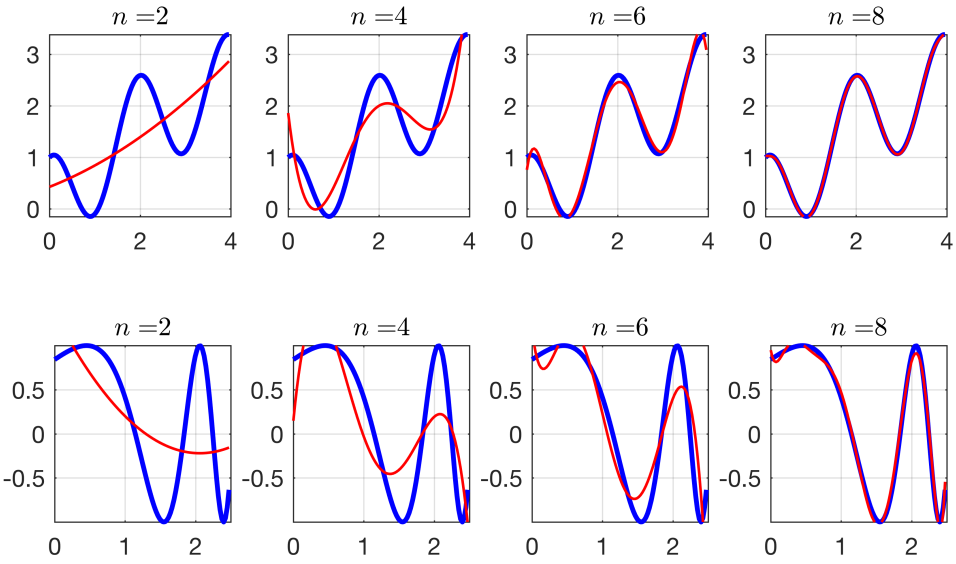

FIGURE 2.8: Function approximation with Legendre polynomials with different degree. In blue we see the underlying, or true function and in red we see the approximation using Legendre polynomials. (Top) Approximation of the function $f(x) = x - x^2 + cos(3, 2x)$. (Bottom) Approximation of the function $g(x) = sin(e^x)$. We can see that the approximation identical to the approximation shown in Fig.2.6

2.7 Limitations, Phenomena, and Pathological Functions

One may expect from Weierstrass' theorem that the accuracy of approximation will increase as the number of samples or the order of polynomial increase. However, polynomial functions are not guaranteed to have the property of uniform convergence, and they may diverge away from the target function as the order increases. This phenomenon is known as to Runge's phenomenon.

Runge's phenomenon is a problem of oscillation at the edges of an interval that occurs when using polynomial interpolation with polynomials of a high degree over a set of equispaced interpolation points, see Fig.2.10. It was discovered by Carl David Tolme Runge (1901) [52, 163], when exploring the behavior of errors when using polynomial interpolation to approximate certain functions. Divergence of approximation as polynomial degree increase, naturally, also appears in our adopted approach of power series expansion. Moreover, because of nonorthogonality of the basis, more complications regarding the magnitude of parameters arise. We will discuss this issue in the following sections since it is strongly connected with the properties of the method of least squares itself.

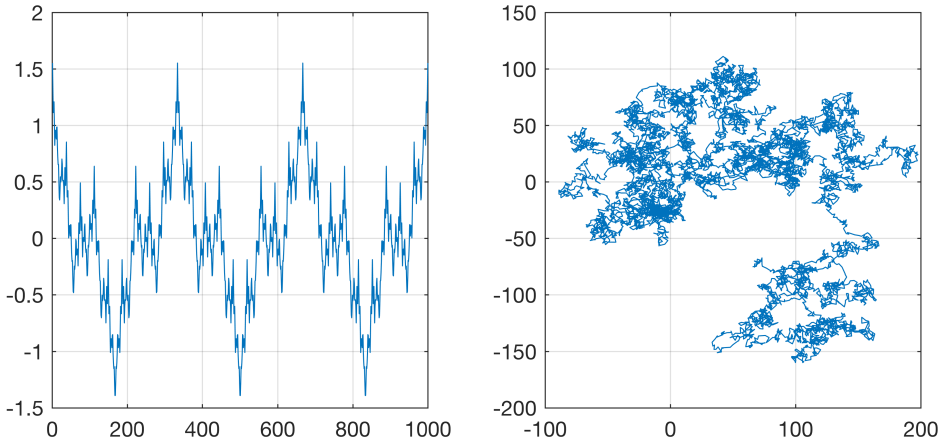

FIGURE 2.9: Pathological Functions: (Left) Weierstrass function. (Right) Wiener process.

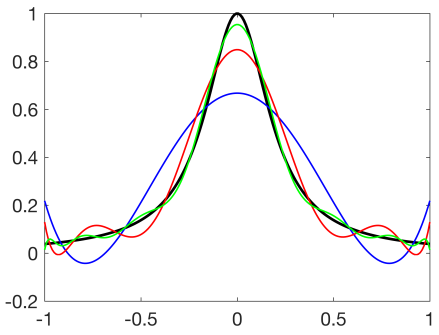

FIGURE 2.10: Runge Phenomenon. In black we see Runge function given by $f(x) = \frac{1}{1+25x}$, with equispaced sample of 401 point between [-1,1]. In blue we see the function approximation with the 5^{th} degree Legendre polynomial. In red and green we see the function approximation with the 9^{th} and 15^{th} degree Legendre polynomial respectively. At some degree of polynomial, we reach a point where the function at the sampled points close to zero, while between the sampled points, especially close to the endpoints 1 and -1, the error between the function and the approximation gets higher.

Similarly, Wilbraham-Gibbs phenomenon, discovered by Henry Wilbraham (1848) and re-discovered by J. Willard Gibbs (1899) [91, 199], is the strange manner in which the Fourier series of a piecewise continuously differentiable periodic function overshoots at discontinuities. Moreover, in the multivariate setting, there is a negative result due to Mairhuber-Curtis Theorem (1956), based on a work of Mairhuber [127] and Curtis [51], and discussed in [26], which implies that it is not possible to perform unique interpolation with multivariate polynomials of degree N to data given at arbitrary locations in $\mathbb{R}^2$.

Finally, while the polynomials approximation can be efficient in well-behaved and smooth functions, the real world data are not smooth. The pathological phenomenon is one whose properties are considered atypically defective or counterintuitive; the opposite is well-behaved, and many (if not most) of real-world data are pathological.

Example of a pathological structure is Weierstrass function, introduced by Karl Weierstrass [197], which is continuous everywhere but differentiable nowhere and Wiener process which is a continuous-time stochastic process (often called standard Brownian motion) and it is a crucial process in terms of which more complicated stochastic processes can be described. Fig.2.9 shows some pathological structures.

2.8 The Method of Least Squares

In 1805, Legendre published the earliest form of regression which was the method of **Least Squares (LS)**, and it was introduced again by Gauss in 1809 [178]. Gauss claimed that he had been using it since 1794, and this case is known as *priority dispute over the discovery of the method of least squares* [169]. Impartially, Gauss went beyond Legendre and succeeded in connecting the method of least squares with the principles of probability and in his attempts in his long and detailed paper, he introduced the Gaussian elimination and provided several examples of estimation with complete and incomplete measurements with different sample sizes. In this process he invented the normal distribution. Gauss published a further development of the theory of least squares in 1821, including a version of the Gauss–Markov theorem [41, 169].

Recall Eq.(2.18), that the objective of the regressors is to find the set of parameters that "best" fit the sampled data, and the LS defines the "best" fit to be the minimum squared residual, meaning that, the LS tries to find the optimal parameters by solving the optimization problem:

$$arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} (\boldsymbol{f_i} - \Phi_{i:}\boldsymbol{\beta})^2 \tag{2.23}$$

note that $\Phi_{i:}$ is the K-dimensional measurement point at time t_i. The sum in Eq. (2.23) can be written as the L^2-norm, where

$$\|\boldsymbol{f} - \Phi\boldsymbol{\beta}\|_2^2 = \sum_{i=1}^{N} (\boldsymbol{f_i} - \Phi_{i:}\boldsymbol{\beta})^2 \tag{2.24}$$

The differentiability of the $\|\cdot\|_2$ objective function of LS, in addition to other factors that will be discussed below, is one of the main reasons for the popularity of the LS and its wide use in many applications. LS is simple to analyze and for some applications, the optimal parameters can be found in closed form through the gradient analysis of the objective function.

The Least Squares Problem problem has a solution, and it is a unique solution, if and only if the columns of Φ are linearly independent, i.e., $rank(\Phi) = K$, where Φ. Otherwise, the solution is not unique [56, 79].

Many numerical methods have been adopted to solve the least squares problem, such as Cholesky Factorization, and QR factorization. One of the most popular methods to find the LS solution is by the Singular Value Decomposition (SVD) of the measurements matrix Φ. The SVD method is robust in solving rank-deficient problems, and it is computationally efficient. The SVD for the matrix $\Phi \in \mathbb{R}^{N\times K}$ is given by:

$$\Phi = U\Sigma V^T \tag{2.25}$$

where $U \in \mathbb{R}^{N\times N}$, $V \in \mathbb{R}^{K\times K}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{N\times K}$ is a diagonal matrix with the singular values of the matrix Φ, $\sigma_1 \leq \sigma_2 \leq \cdots \leq \sigma_p \leq 0$ on its diagonal with $p = \min\{N, K\}$. Moreover, the SVD gives efficient and accurate computation for the pseudoinverse of a matrix, which is a generalization of the inverse of a matrix. The pseudoinverse is defined and is unique for any matrix, and for a matrix that has linearly independent columns, it is

given by:

$$\begin{aligned} \Phi^{+} &= (\Phi^{T}\Phi)^{-1}\Phi^{T} \\ &= V\Sigma^{-1}U^{T} \end{aligned} \tag{2.26}$$

where U,V, and Σ given in Eq.2.25, and Φ^{+} is this equation is called the left inverse of a matrix and it satisfies $\Phi^{+}\Phi = I$, with I is the identity matrix.

The Least squares problem can simply be solved as $\boldsymbol{\beta} = \Phi^{+}\boldsymbol{f}$, and for more details, Eq. (2.25) can be written in the form of outer product as:

$$\Phi = \sum_{i=1}^{K} \sigma_i u_i v_i^{T} \tag{2.27}$$

For a full rank matrix Φ with $K \leq N$:

$$\begin{aligned} \|\Phi\boldsymbol{\beta} - \boldsymbol{f}\|_2^2 &= \|U\Sigma V^{T}\boldsymbol{\beta} - \boldsymbol{f}\|_2^2 \\ &= \|UU^{T}\Phi VV^{T}\boldsymbol{\beta} - \boldsymbol{f}\|_2^2 \\ &= \|(U^{T}\Phi V)(V^{T}\boldsymbol{\beta}) - U^{T}\boldsymbol{f}\|_2^2 \\ &= \|\Sigma(V^{T}\boldsymbol{\beta}) - U^{T}\boldsymbol{f}\|_2^2 \\ &= \sum_{i=1}^{K} \left(\sigma_i v_i^{T}\boldsymbol{\beta} - u_i^{T}\boldsymbol{f}\right)^2 + \sum_{i=K+1}^{N} \left(u_i^{T}\boldsymbol{f}\right)^2 \end{aligned} \tag{2.28}$$

and[1] it follows that this summation is minimized by setting $\sigma_i v_i^{T}\boldsymbol{\beta} = u_i^{T}\boldsymbol{f}$, which gives the LS solution by:

$$\boldsymbol{\beta}_{LS} = \sum_{i=1}^{K} \frac{u_i^{T}\boldsymbol{f}}{\sigma_i} v_i. \tag{2.29}$$

We see that $\boldsymbol{\beta}_{LS}$ make the first summation in (2.28) equal to zeros, therefore the minimum residual size is given by:

$$r_s = \sum_{i=K+1}^{N} \left(u_i^{T}\boldsymbol{f}\right)^2. \tag{2.30}$$

We carried out the SVD form of LS solution as shown above to present one major limitation of the LS. Even with a full rank matrix Φ and regardless of the noise level in our system, we see that for nearly singular system, LS solution $\boldsymbol{\beta}_{LS}$ will be sensitive to the small singular

[1] Note that in the third step of Eq.2.28, for unitary matrix U, we have $\|Ux\|_2^2 = \|x\|_2^2$.

values since:

$$\lim_{\sigma_i \to 0} \left(\sum_{i=1}^{K} \frac{u_i^T \boldsymbol{f}}{\sigma_i} v_i \right) \to \infty. \tag{2.31}$$

Note that we can not infer any indications about this sensitivity from the minimum residual size r_s since it does not depend on σ_i. This means that with a nearly singular measurement matrix Φ, we may have a bad solution while we still have a "good looking" fitting with low residual.

Parameters sensitivity for singular values did not receive significant attention in the literature, because it was the common theme in the literature to use orthonormal basis when solving the least squares problem, in such case we have $\sigma_i \approx 1$, $i = 1, 2, ..., K$. Since in last decade, it becomes favorable to use a non-orthogonal basis, and it is the theme we adopt in this work, we will discuss further this sensitivity.

To obtain a bound for the error in estimated parameters, recall that:

$$\boldsymbol{f} = \Phi \boldsymbol{\beta}. \tag{2.32}$$

Let $\delta = \hat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ be the error in estimated parameters, and $r = \hat{\boldsymbol{f}} - \boldsymbol{f}$ be the residual (error in recovered function). Then we can write:

$$\begin{aligned} r &= \hat{\boldsymbol{f}} - \boldsymbol{f} \\ &= \Phi \hat{\boldsymbol{\beta}} - \Phi \boldsymbol{\beta} \\ &= \Phi(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &= \Phi \delta. \end{aligned} \tag{2.33}$$

The Eq.2.33 has a subtle beauty in its simplicity; it shows how the projection matrix Φ works as a "transfer operator" for the error. In the forward modeling, any change (or error) in the parameters will reflect a difference in the residual, and in the backward modeling, this effect will be amplified because of the stability of the transfer operator as we will show next, and we will back to this equation when we discuss some limitations of the least squares method.

In order to get a bound for δ, follow [78], we take the norm for the inverse of Eq.2.33:

$$\begin{aligned} \|\delta\|_2 &= \|\Phi^+ r\|_2 \\ &\leq \|\Phi^+\|_2 \|r\|_2, \end{aligned} \tag{2.34}$$

multiply the right hand side with $1 = \frac{\|\Phi\boldsymbol{\beta}\|_2}{\|\boldsymbol{f}\|_2}$, we get:

$$\begin{aligned} \|\delta\|_2 &\leq \|\Phi^+\|_2 \|r\|_2 \frac{\|\Phi\boldsymbol{\beta}\|_2}{\|\boldsymbol{f}\|_2} \\ &\leq \|\Phi^+\|_2 \|r\|_2 \frac{\|\Phi\|_2 \|\boldsymbol{\beta}\|_2}{\|\boldsymbol{f}\|_2} \\ &\leq \|\Phi^+\|_2 \|\Phi\|_2 \|\boldsymbol{\beta}\|_2 \frac{\|r\|_2}{\|\boldsymbol{f}\|_2} \end{aligned} \tag{2.35}$$

noting that $\|\Phi^+\|_2 \|\Phi\|_2 = \kappa$ is the condition number of the matrix Φ, then we can write:

$$\frac{\|\delta\|_2}{\|\boldsymbol{\beta}\|_2} \leq \kappa \frac{\|r\|_2}{\|\boldsymbol{f}\|_2}. \tag{2.36}$$

Now, we develop:

$$\begin{aligned} \alpha &= \log_{10}\left(\frac{\|\delta\|_2}{\|\boldsymbol{\beta}\|_2}\right), \\ \gamma &= \log_{10}\left(\frac{\|r\|_2}{\|\boldsymbol{f}\|_2}\right), \\ k &= log_{10}\left(\kappa\right), \end{aligned} \tag{2.37}$$

where α is the order of the relative error in the parameters, γ is the order of the relative error in the recovered signal, and k is the order of magnitude of the condition number κ. Then we have:

$$\alpha \leq k + \gamma \tag{2.38}$$

In the same principle we can derive the lower bound from Eq.2.33 as the following:

$$\begin{aligned} r &= \Phi\delta \\ \|r\|_2 &= \|\Phi\delta\|_2 \\ &= \|\Phi\delta\|_2 \\ &\leq \|\Phi\|_2\|\delta\|_2 \\ \|r\|_2\|\boldsymbol{\beta}\|_2 &\leq \|\Phi\|_2\|\delta\|_2\|\boldsymbol{\beta}\|_2 \\ &\leq \|\Phi\|_2\|\Phi^+\|_2\|\delta\|_2\|\boldsymbol{f}\|_2 \\ &\leq \kappa\|\delta\|_2\|\boldsymbol{f}\|_2, \end{aligned} \tag{2.39}$$

which gives, in analogy to driving Eq.2.38:

$$\gamma - k \leq \alpha, \tag{2.40}$$

and by combining Eq.2.38 and Eq.2.40 we have:

$$\gamma - k \leq \alpha \leq \gamma + k \tag{2.41}$$

noting that $\kappa = \frac{\sigma_{max}}{\sigma_{min}}$, with σ_{max} and σ_{min} are the maximum and minimum singular values of the matrix Φ respectively, we can see that $\kappa \geq 1 \implies k \geq 0$ (k is a positive number).

The commonly used assumption is that the matrix Φ is a well-conditioned matrix, $\kappa \approx 1$, which implies that $k = \log_{10}(\kappa) \approx 0$. As a result, we will have a very narrow bound for α, and the better fitting we have will imply more accurate parameters. Then the question is: **by generating the basis matrix Φ from state measurements with power series expansion, is it possible to get a well-conditioned basis matrix?**

Unfortunately, the answer is no. We will show that with increasing expansion order, the singularity of the basis matrix become deterministic. To show that we will first give the following remark.

Remark. *The set of real numbers can be written as* $\mathbb{R} = (-\infty, -1) \cup (-1, 1) \cup (1, \infty) \cup \{-1, 1\}$, *and any real number* x_i *will belong to one of these subsets. Then, any random variable* $x \in \mathbb{R}^n$,

$x \neq 0$, can be classified to one of the following three cases:

case 1: $x_i \in \{-1, 1\}$ *for all* $i = 1, .., n$.

case 2: $|x_i| \leq 1$ *for all* $i = 1, .., n$.

case 3: At least, there exist one entry in x *such that* $|x_i| > 1$.

Recall that Weierstrass approximation theorem state that any continuous real-valued function can be approximated with a polynomial of some degree, however, there is still some phenomena that limit the practical applicability of the theorem such as Gibbs and Runge phenomena. Moreover, especially for Carleman linearization approach in 1932, there were a few discussions and use of the method in literature until the last two decade, see Appendix B, when it comes again to life as a new idea in an old book because of its simplicity and the power polynomials model that is more simple for analysis and study purposes. However, the nonorthogonal basis is known to have weak numerical stability and may be described as "bad" choice, but we can not find a description in literature of "how bad?!" is the power polynomial basis for function approximation.

With this in mind, we have developed the following theorem, which is relevant to data analysis, and it states the following:

Theorem 2.8.1. *For any vector of observation* $x \in \mathbb{R}^n$, *and power polynomial expansion* $\Phi = \{\mathbf{1}_n, x, x^2, \ldots, x^p\}$, *there exist positive integer* p *such that the matrix* Φ *is singular.*

Proof. From the above remark, we can divide our proof into the only three possible cases:

case 1: In this case, when $x_i \in \{-1, 1\}$ for all $i = 1, .., n$, it is clear that for all even powers $p = 2j, j = 1, ..., \infty$, we will have $x_i^p = 1$ for all $i = 1, ..., n$, which implies that Φ will have repeated columns, and then, it will be singular. Specifically, Φ will be definitely singular at $p = 2$.

case 2: The i^{th} row of the basis matrix Φ can be written as a series $s_j = x_i^j, j = 1, ..., \infty$, and since we have $|x_i| < 1$, then we have $\lim_{j\to\infty} x_i^j = 0$ for all $i = 1, ..., n$. So, the expansion will end up with a columns with all enties equal to zero, which approaches singularity.

case 3: Assume that we have one entry in x such that $|x_i| > 1$, then the series in the case 2 above will have the limit $\lim_{j\to\infty} x_i^j = \infty$, which means we will have at least one entry

in the matrix Φ such that $\phi_i^j = \infty$. From the inequality of the spectrum norm [78, 81],

$$\|A\|_2 \geq \max_{i,j} |a_{ij}|, \tag{2.42}$$

and since the spectrum norm $\|\Phi\|_2 = \sigma_{max}$, it become clear that the maximum singular value will be $\geq$ the largest entry in the matrix Φ. Meaning that $\lim_{j\to\infty} x_i^j \to \infty \implies \sigma_{max} \to \infty \implies \kappa \to \infty \implies$ the matrix Φ become singular.

So, for all possible entries of the vector $x \in \mathbb{R}^n$, there exist a positve integer p such that the matrix $\Phi = \{\mathbf{1}_n, x, x^2, \ldots, x^p\}$ become singular. □

It will be a subject of our future work to extend this discussion, to include quantifying the error amplification for each parameter independently, and to prove and estimate a new lower and upper bounds for the relative error in the parameters δ, combined with a cost-efficient estimation of the condition number κ.

Recall Eq.2.33 that, the condition number of the transfer operator measures the error amplification factor of function, and how the output changes under a small changes in the input. As a result, it is an essential indicator of how accurate we may estimate the parameters. For example, the condition number $\kappa = 10^k$ represents an additional cost of losing k digits of accuracy on the addition to the digits lost by numerical methods or significant figures computations.

Now, we see that the least squares solution is sensitive to the numerical stability of the system, which is a function of the state measures itself and the selected expansion order. While lower expansion orders can give better stability, it is particularly risky because, in the sense of "black box modeling", we don't have prior information about the system. The more natural assumption is to consider higher orders prior model, since singularity of the basis matrix does not imply that there are no higher order functions that influence the dynamic.

To illustrate the above ideas, numerically, we designed the following example. Consider the 1-D map

$$x_{n+1} = r x_n (10 - x_n^9) \tag{2.43}$$

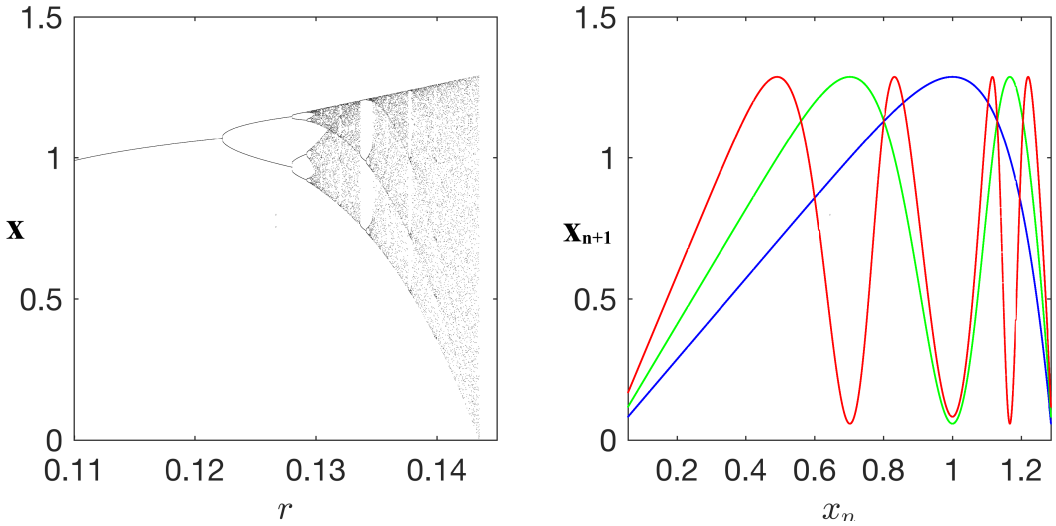

FIGURE 2.11: Bifurcation diagram for the chaotic map Eq.2.43

which is slightly modified from the standard logistic map. Observe that it shows chaotic behavior for $r = 0.143$, as shown in the bifurcation diagram in Fig.(2.11), and it has fixed points at $0, 1.322...$ and 8 fixed points $x_{fixed} \in \mathbb{C}$.

This map is designed to serve as example for the sensitivity of least squares to the numerical stability of the measurements (basis) matrix. Considering $r = 0.143$, we can write Eq.2.43 in the form:

$$x_{n+1} = 1.43x_n - 0.143x_n^{10} \tag{2.44}$$

Fig.2.11 shows the bifurcation of the map which looks similar to the well known logistic map, but in fact it is different in terms of the state probability. In order to sample data for system identification we consider carrying 1000 iteration of our map to produce our data X. Then we add a low noise by setting $X = X + \eta$, where $\eta \sim \mathcal{N}(0, (10^{-3})^2)$, which is in terms of signal to noise ratio $SNR = 53$ dB.

Fig.(2.12) shows the original signal and the sampled noisy measurements.

From the noisy sample X, we create our data such that:

$$x = X(1, ..., n-1)$$

$$y = X(2, ..., n)$$

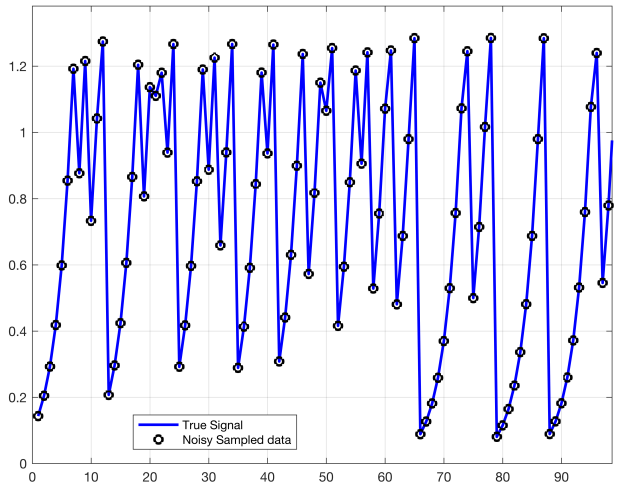

FIGURE 2.12: (Blue) the original iteration of the map. (Black) The sampled data with noise. For clear view, this figure shows only 100 points out of the 1000 sample points, and it shows clearly the low level of noise added to the signal.

which makes the problem is to find a function $f(x)$ such that $y = f(x)$. Our 10^{th} order polynomial basis matrix:

$$\Phi = \left[1, x, x^2, x^3, ..., x^{10}\right]$$

Note that we consider the 10^{th} order expansion "assuming" that the true expansion order is known. Now, the parameters estimation problem takes the form

$$y = \Phi\boldsymbol{\beta} \tag{2.45}$$

The least square solution is given in Table (2.1), which shows very poor parameters, although that the mean squared error of the residual was of order -3.

On the other hand, the least squares solution is known to be the *Best Linear Unbiased Estimator* (**BLUE**) according to the well-known Gauss-Markov theorem [129]. We remark that this does not contradict with our previous discussion, because, in Gauss-Markov theorem, the least squares solution for the problem $y = Ax + \eta$, is BLUE if and only if the following assumptions concern the set of error random variables (η_i) hold:

1. They have mean zero: $E[\eta_i] = 0$, where E is the expected value.

2. They have a constant variance: $Var(\eta_i) = \sigma^2 < \infty$.

1.4381	0
-0.1479	1.43
1.0514	0
-4.1005	0
10.1945	0
-17.0364	0
18.9159	0
-13.2051	0
5.1903	0
-1.0132	-0.143

TABLE 2.1: (Left) The least squares solution given by $\boldsymbol{\beta} = \Phi^{+}y$. (Right) The true solution. We see that, according to the sensitivity to the condition number (which is of order 5 at the assumed expansion order), the least squares solution is poor and far from the true solution. The important point here, we see that the smallest magnitude parameter in the least square solution (red) and the next smallest (blue) are the only two true parameters, meaning that **whatever** the choice of the threshold parameter in any method depends on the hard threshold of the least squares solution, the true parameters will be the first to be threshold.

3. Distinct error terms are uncorrelated: $Cov(\eta_i, \eta_j) = 0, \forall i \neq j$.

This theorem, is one part of a long chain of arguments, starting with Laplace 1774 [53, 202] (before Gauss), continuing through Chebyshev (1821-1891) [34], his students Markov (1856-1922) [129, 167], and Lyapunov (1857-1918) [123, 174], before finally receiving significant advancement by A. Kolmogorov (1903-1987) [206]. As a summary, the physical feasibility of the least squares to be the best estimator depends highly on the distribution and properties of the noise vector η, and the Central Limit Theorem, specially a Lyapunov variant of the theorem [19, 69]:

Theorem 2.8.2. ***Lyapunov Central Limit Theorem:*** *Suppose* $\{x_1, x_2, \dots\}$ *is a sequence of independent random variables, each with finite expected value* μ_i*, variance* σ_i^2*, and absolute moment* $E[|x_i - \mu_i|^{2+\delta}]$ *and let* $s_n = \sum_{i=1}^{n} \sigma_i^2$*. Then, if for some* $\delta > 0$*,*

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} E|x_i - \mu_i|^{2+\delta}}{s_n^{1+\frac{\delta}{2}}} = 0 \tag{2.46}$$

satisfied, then the probability of the inequality:

$$\alpha_1 < \frac{\sum_{i=1}^{n}(x_i - \mu_i)}{\sqrt{s_n}} < \alpha_2 \tag{2.47}$$

tends to the limit:

$$\frac{1}{\sqrt{2\pi}} \int_{\alpha_1}^{\alpha_2} e^{-x^2/2} dx \tag{2.48}$$

as $n \to \infty$*, uniformly with respect to all values of* α_1 *and* α_2*.*

Eq.2.46 is called Lyapunov condition, which limits the rate of growth of the absolute moment of the random variable. Lyapunov theorem implies that an identical random variable, and not necessarily identically distributed, that satisfy Lyapunov condition, will converge as standardized sum to a normal distribution:

$$\sum_{i=1}^{n} \frac{x_i - \mu_i}{s_n} \to \mathcal{N}(0, 1). \tag{2.49}$$

All of the above leads us to the Law of Large Numbers (which we are discussing in the following sections) and its main effect on the success of the method of least squares.

Through at least the last 200 years, the LS method has been a favorite method in the inverse problems literature, where we can find thousands of papers that use the LS method in different applications, and we can find 10's (or more) of modified LS methods. However, LS suffers from large number of limitations as we will discuss later in this sections, and its popularity does not come from its robustness, but from other different reasons that (in the core) are not related to the technique itself or to its robustness against the different experiments conditions to make accurate predictions. Some of the reasons of LS popularity are:

- It is the best linear unbiased estimator (BLUE), provided it exists. But according to the previous discussion, it worth to say that the BLUE property should be treated carefully with considerations for the feasibility of its conditions, since as we showed before that the "best" estimator based on the residual as a measure of quality does not necessarily reflect a satisfactory solution. So, while the LS is BLUE in the residual sense, it could be PURPLE (Poor and UnReliable Parameters with Large Error) in the parameters sense.

- The LS method, in its basic formulations, is easy to understand and follow even in the non-mathematical fields.

- The solution of LS method was simple enough to do by hand 10's of years ago, and it is one of the most cost efficient methods in nowadays computers.
- In its time, it created a revolution in the inverse problem solution in many different fields. For many decades scientists used the method of LS because it was the most robust and accurate method available.

On the other hand, many problems may appear when using LS, some of them are not related to the method itself, but the numerical computations in general such as the curse of dimensionality and using non-relevant features to describe the dynamics. More problems also can prevent the LS from obtaining a reasonable estimate to the parameters are:

1. Outliers, as discussed in Chapter 1, and will be discussed in more details in the following sections.
2. Non-Linearities: When the relation between independent and dependent variables is non-linear, meaning that the parameters are not constants.
3. Dependent Variables: When there exists some basis function that can be accurately (within some tolerance) be represented as a linear combination of the other basis functions.
4. Heteroskedasticity: When the variance of the noise (error) is not constant.
5. Too many parameters: A system with too many parameters, is subject to the curse of dimensionality, and some parameters may work opposite to other parameters.
6. Different Scale: When the basis functions and/or the parameters are on different scale order.
7. Sloppy Parameters [85, 196]: When a small change in the output, requires a large change in the magnitude of the parameter. We strongly believe that parameters sloppiness is a determining property for the above two problems.

Notes on Outliers.

We have discussed in our introduction the outliers problem and its dangerous implications, and it is our goal in this work to develop parameter estimation method that does not try to apply any filtering or preprocessing of data to drop the outliers or assign a low weight for them.

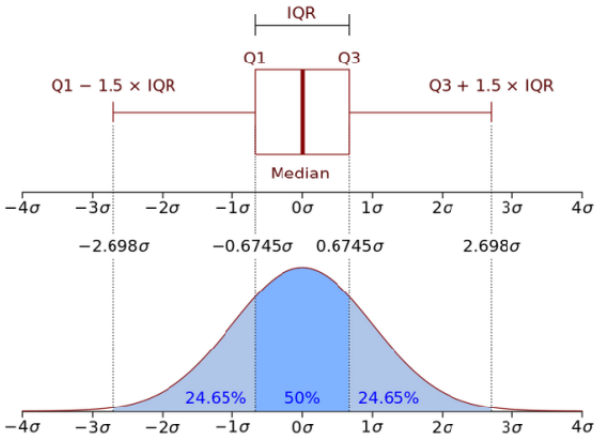

FIGURE 2.13: Boxplot of Gaussian Distribution.

However, we will discuss here the general approaches in the literature to detect outliers, since if some prior information is available and we can know that error in measurements techniques or measuring tools may occur, then such tools of detecting outliers can be handy.

The first method for detecting outliers is the Box plot construction. The box plot is a graphical representation of the data to describe the behavior of the data in the center and the terminal ends of the distributions.

If we divide the data into four quartiles, the median will be the center that divides between the second and third quartiles, and the separation edges can be marked as: $[Q_1, Q_2 = \text{median}, Q_3, Q_4]$, where there is no data below the 0 mark or above the Q_4 mark. The data between Q_1 and Q_3 marks represent 50% of the population, and we call it the Interquartile Range (IQR). See Fig.2.13. Then, the main idea is to construct fences to judge the data points inside and outside these fences. One suggestion of these fences boundaries is as the following:

- Lower fence: $Q_1 - 1.5 \times IQR$
- Upper fence: $Q_3 + 3 \times IQR$

Then, a point below the Lower fence, or beyond the Upper fence, can be classified as outlier point.

The second method depends on the standard deviation of the data. For a Gaussian distribution:

- 68.0% of the data lie within 1 standard deviation from the mean.

- 95.0% of the data lie within 2 standard deviation from the mean.
- 99.7% of the data lie within 3 standard deviation from the mean.
- 99.9% of the data lie within 4 standard deviation from the mean.

The standard deviation of 4 is commonly used as a cut point, and any point with distance more than 4 standard deviations from the mean is considered to be an outlier point, that is for a random variable $x \in \mathbb{R}^n$, if the inequality:

$$\frac{|x_i - \mu|}{\sigma} \geq 4, \tag{2.50}$$

holds, then the point x_i is said to be an outlier point, where μ is the sample mean and σ is the sample standard deviation.

2.9 Law of Large Numbers, and Gambler's Fallacy

In solving the least squares problem, the assumption that the noise signal has zero mean and constant variance stands mostly (at least in many cases) on the law of large number, that is why it plays a primary rule in understanding the outcomes of the LS. The term "law of large numbers" coined by Siméon Denis Poisson, and the original theorem is for his academic advisor Jacob Bernoulli. The very first version was published in Bernoulli book "The Art of Conjecturing" in 1713 [17, 87, 190], eight years after his death.

Bernoulli theorem state that: in a sequence of independent trials, in each of which the probability of occurrence of a specific event x has the same value $p, 0 < p < 1$, then:

$$\lim_{n\to\infty} P\left(|\frac{c_n}{n} - p| > \varepsilon\right) = 0, \forall \varepsilon > 0, \tag{2.51}$$

where c_n is the number of occurrence of the event in the first n trials. Poisson extended Bernoulli theorem to the case of non-identical probability, that the probability of the event A changes with the number of trials. For the sake of summary and clarity, We can write in more straightforward form:

$$\lim_{n\to\infty} P\left(|E[x_n] - \mu| > \varepsilon\right) = 0, \forall \varepsilon > 0, \tag{2.52}$$

where E is the expected value and μ is sequence mean. Then, the law of large numbers says that for a sequence of random numbers, the mean of the sample will converge to the expected mean as the sample size increase. The law of large number had a great interest because of its implications to the probability theory. Rigorous proofs and extensions of the theorem were introduced in 1846 by Chebyshev.

Misunderstanding law of large numbers, naturally, can lead to falling in the Gamblers Fallacy, which also known as Monte Carlo Fallacy. In 1913, in a game of roulette at the Monte Carlo Casino, Monaco, the ball fell in black 26 times. The probability for such sequence to occur is less than 1 in 66 million, many gamblers lost millions betting against black, because of the incorrect reasoning that since the mean of probability should be 50-50 between the red and black, then repeated black implies higher probability for the red. Randomness has no memory. Even if we get 9 heads in a row in a coin toss, the probability in the 10^{th} toss will be 0.5.

Another type is called retrospective gambler's fallacy, which is when the one expects that a rare event with low probability needs a very long sequence to occur, or that the rare event comes from very long sequences. We discuss an example of this fallacy in our introduction which is Fukushima nuclear disaster, where an event with probability ("1 in a million years") event occurred within 8 years.

So, we have two sides of view, from one side the law of large numbers, and from the other side, randomness has no memory in the real world. Far from arguing which one is "more" accurate regarding the real world problems, both of them are accurate in some sense, and the one may say that "in real-world problems, randomness has no detailed short memory, but it has accumulative knowledge, and it obeys the law of large numbers".

The major problem is that simulating this behavior in our modern computers can not (yet) reflect this reality. In the International Encyclopedia of Statistical Science (2010) [118, 119], we can see the following alarm: "The list of widely used generators that should be discarded is long". This last recommendation has been made over and over again over the past 50 years. Perhaps amazingly, it remains as relevant today as it was 50 years ago.

A pseudorandom number generator (PRNG) [119], also known as a deterministic random bit generator (DRBG), is an algorithm for generating a sequence of numbers whose properties

approximate the properties of sequences of random numbers. The PRNG-generated sequences are not really random, because it is a deterministic sequence initiated with a value that is closer to be called random since it is hardware-based value [143], and the key point here is nonetheless that:

- We don't know the initial condition, so, randomness is that "one row" choice.
- The system is ergodic with absolutely continuous invariant measure. However, it is repeatable "random".

One interesting example is IBM's RANDU [118], a linear congruential pseudorandom number generator (LCG) that has been used since the 1960s until 1999. It is found that RANDU generates "random numbers" in 15 parallel two dimensional planes.

The recurrence relations, creating random numbers by recurrence functions and using the gambler's fallacy (from the gambler side of view) to generate sequences with predefined properties, all of that, is entirely the opposite of what are random numbers in the real world. However, what can we do?!!

Figs.2.14-2.16 shows the mean and variance of random numbers generated by Matlab on Mac machine, which can be considered good PRNG. We see the law of large numbers and how the converged the predefined mean and variance with high accuracy at a sample size of order 4, which is relatively small.

Our goal here is to state that: **In generating data sets to investigate the effect of noise and the robustness of algorithms against noise and outliers, the smaller the data set the more close the data from real-world behavior of error and noise.**

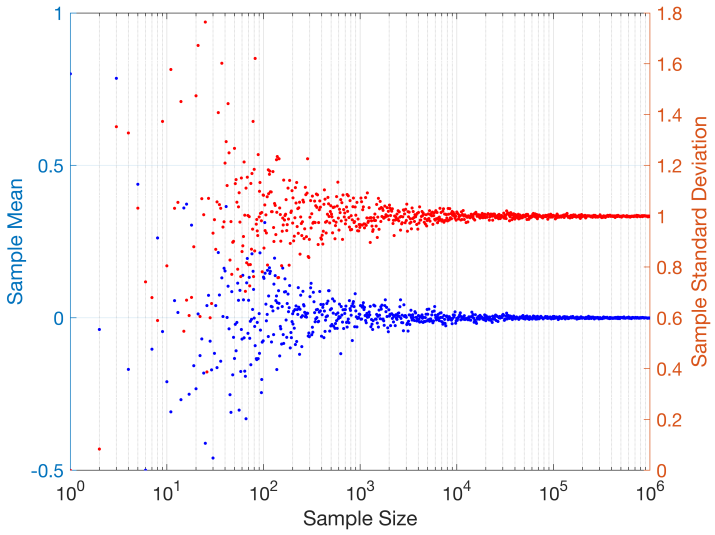

FIGURE 2.14: Random sample: Mean of the sample on the left axis and blue dots, and variance of the sample on the right axis and red dots. The generated signal was with zeros mean and 1 standard deviation. At each sample size, we create a new sequence with a new PRNG seed. We see that as the sample size increase, the more deterministic the signal will be regarding the mean and variance.

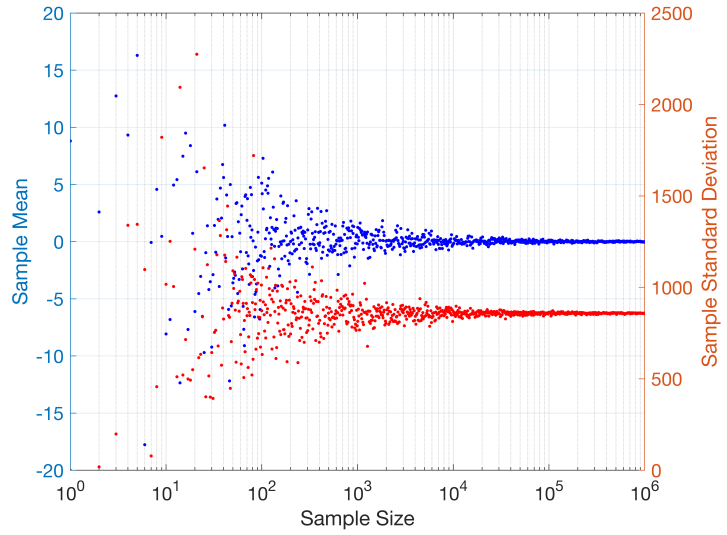

FIGURE 2.15: Random sample with random mean and random variance: For the same setting as in Fig.2.14, we create the data such that at each sample size, we create a new sample with random variance uniformly chosen between [1 100], and random Gaussian mean $\mu \sim \mathcal{N}(0, 1)$. Again, we see the weak convergence at approximately the same sample size as in the case of fixed mean and variance.

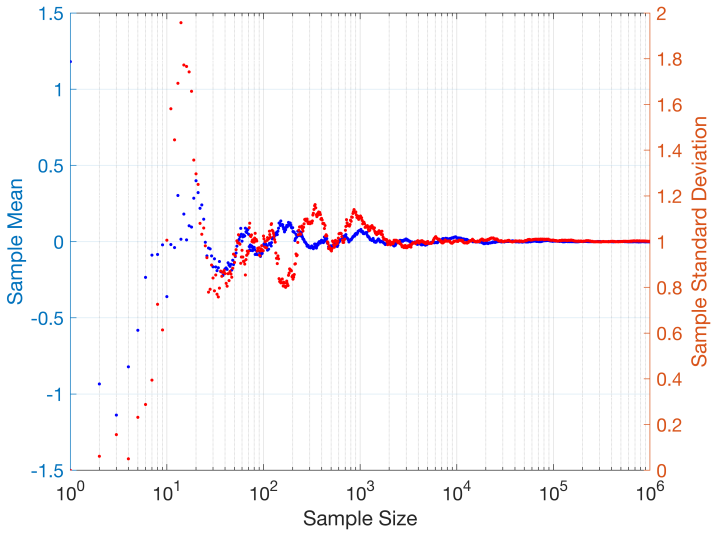

FIGURE 2.16: LLN and Gambler's Fallacy: For random number $x \sim \mathcal{N}(0,1)$, we create the sample in an accumulative manner, meaning that we create one random number $x_i \sim \mathcal{N}(0,1)$ and concatenate it to the previously chosen sequence. We see that in small sample size, some regions have continuously increasing/decreasing values for the mean and variance, which means for example that we have for long iterations generated number that is always above the accumulated mean. However, finally, we converge again at a large sample size, where the effect of new samples on the mean and variance become negligible.

2.10 L0 Minimization

Recall the least squares problem:

$$\min_{\boldsymbol{\beta}} \left(\|\Phi\boldsymbol{\beta} - \boldsymbol{f}\|_2 \right) \tag{2.53}$$

The problem is said to be well-posed, according to Jacques Hadamard definition [183], if it satisfies three main properties

1. A solution exists.
2. The solution is unique.
3. The solution's behavior changes continuously with the initial conditions.

Problems that are not well-posed are termed Ill-Posed problems. As discussed in many sections in this chapter, most of inverse problems are ill-posed. In the ill-posed problem, where the LS method deals with positive (or negative) degree of freedom ill-conditioned matrices with error amplification in the inverse problem, it is desired to give some preference to a particular

solution over all other solutions based on some quality measure. This can be achieved by adding a regularization term to the minimization objective function such as:

$$\min_{\boldsymbol{\beta}} \left(\|\Phi\boldsymbol{\beta} - \boldsymbol{f}\|_2 + \mathcal{R}(\boldsymbol{\beta}) \right) \tag{2.54}$$

where $\mathcal{R}(\boldsymbol{\beta})$ is the regularization term, or penalty term. In general, this equation is called Tikhonove regularization where the first term called fidelity and the function $\mathcal{R}$ is called regulatory. In fact, we can interpret $\mathcal{R}$ in many ways related to the optimization literature [10, 175], we can see it as a constraint that bound a feasible region for the solution, or the Lagrange form of the constrained optimization. Choosing the function $\mathcal{R}(\boldsymbol{\beta})$ is the core subject for too many research papers and books [89, 182, 183], and it has a great interest during the last 60 years. Different approaches use different judging criteria on the parameters, and such criteria usually depend on the nature of the application or the computation complexity [183].

For many dynamical systems, it is usually observed that the governing equations for a very complex dynamics usually have just a few parameters, and there is a famous quote for John von Neumann says: *With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.* Finally, it becomes a favorable approach in solving the inverse problem with the assumption that the dynamic is sparse in some basis, and just a few parameters are governing the dynamic.

A sparse model is a model that have only a small number of nonzero parameters. It can be easier to estimate and interpret than a dense model, and in the sense of big data, the number of features (which implies the number of basis) measured can be huge, and much larger than the number of measurements for each feature. The sparsity assumption allows us to efficiently construct models that consider only the most significant features, which helps in understanding and interpreting such complex systems.

Such an approach is known as (in its general form and topic-independent) the **optimal subset**. Which is optimally finding a small subset from a large set of features according to some objective function. In the inverse problem, it is known as L_0 minimization, meaning that optimally finding the parameters vector $\boldsymbol{\beta}$ such that $\|\boldsymbol{\beta}\|_0$ is minimized, where $\|\boldsymbol{\beta}\|_0$ is number of non-zero entries in the vector $\boldsymbol{\beta}$.

To illustrate the L_0 minimization numerically (and to introduce for important misconception

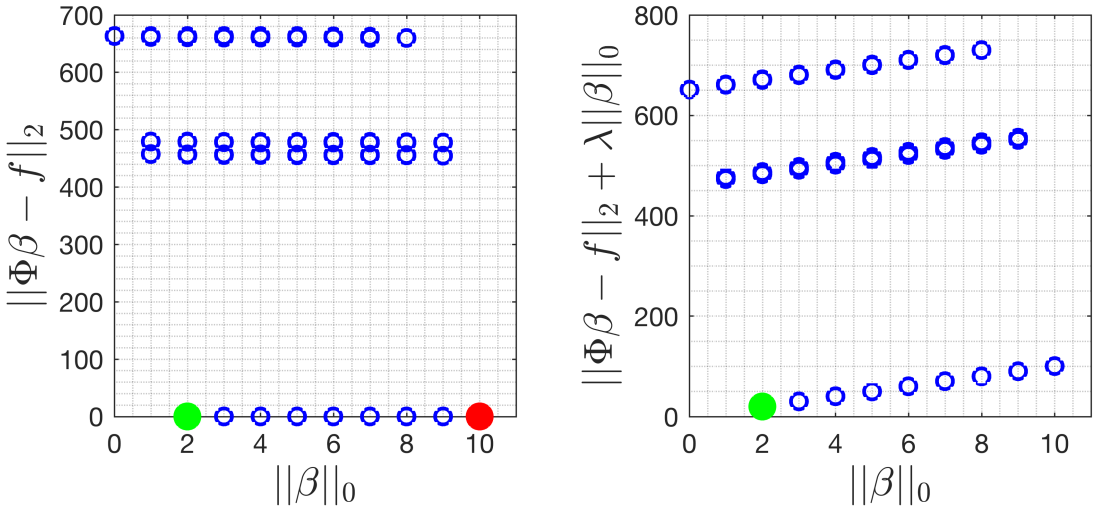

FIGURE 2.17: Exhaustive search and the effect of penalty in L_0 minimization. (Left) for 10 dimensions linear system $\boldsymbol{f} = \Phi\beta$, the blue markers shows $\|\boldsymbol{f} - \Phi\boldsymbol{\beta}\|_2$ for all possible combination of basis, the green marker shows the true solution, and the red marker shows the solution with minimum value. (Right) The value of objective function after adding a penalty term $\|\boldsymbol{f} - \Phi\boldsymbol{\beta}\|_2 + \lambda\|\boldsymbol{\beta}\|_0$. It is clear how the solution with minimum objective function value found to be the true solution after adding the penalty term.

in sparse regression litrature), let $\mathcal{R}(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|_0$, where $\lambda > 0$ is the regularization (penalty) parameter. Now, let $\Phi \in \mathbb{R}^{1000\times 10} \sim \mathcal{N}(0,1)$, $\boldsymbol{\beta} \in \mathbb{R}^{10\times 1} = 0$ except for $\boldsymbol{\beta}_2 = 15, \boldsymbol{\beta}_3 = -15$, and we find $\boldsymbol{f} = \Phi\boldsymbol{\beta} + \eta$, where $\eta \in \mathbb{R}^{1000\times 1} \sim \mathcal{N}(0,2)$ is a large noise. We see that $\boldsymbol{\beta}$ is sparse vector with only two entries are non-zero.

Now, we assume $\boldsymbol{\beta}$ is unknown, and we try to find the best estimation for it. Since we have 10 basis function, that means we have $2^{10} = 1024$ possible combinations of basis that we can use. Fig.2.17-(Left) shows the value of least squares fit $\|\Phi\Phi^+\boldsymbol{f} - \boldsymbol{f}\|_2$ for all possible combination of basis. It is clear that there are many solutions lies within the very small neighborhood (vertically) of the best solution, and the best solution was the dense solution that used all the basis.

Now, we try to find it through the minimizing optimization problem [38]:

$$\min_{\boldsymbol{\beta}} \left(\|\Phi\boldsymbol{\beta} - \boldsymbol{f}\|_2 + \lambda\|\boldsymbol{\beta}\|_0\right) \tag{2.55}$$

where the parameters found by $\boldsymbol{\beta} = \Phi\Phi^+\boldsymbol{f}$ for the chosen basis, and we choose $\lambda = 10$.

Fig.2.17-(Right) shows the value of the objective function for all possible combinations, and we see clearly how the true solution is simply the one with minimum objective function value.

This technique is very efficient, and even with hard problems with a rank deficient matrix with condition number of a high order, it becomes easy to classify a small set of solutions and choose the optimal one.

The only, and unfortunately a major, problem in exhaustive search is the computation complexity, which makes the L_0 minimization an NP-Hard problem. It is a well-known fact, but we showed the robustness of the method in low dimension to clarify a significant misconception about the regression methods.

Almost every sparse regression paper contains the sentence: "Unfortunately, L_0 minimization is NP-Hard", but also in most of the previous work, a low dimensional chaotic system considered as test problem, such as Lorenz and Rossler systems (3-dimensions). Moreover, using low expansion order for the low dimensional system will result with very few basis (it is usually 10 basis under the 2^{nd} expansion order of Lorenz system). So, **we firmly believe that the efficiency and robustness of any sparse regression method should be evaluated under high dimensional systems**.

In the following sections, we discuss the major regularization methods and sparse regression methods.

2.10.1 Tikhonov regularization

Tikhonov regularization is named for Andrey Tikhonov, known for his important contributions to topology, and ill-posed problems, and he worked on the regularization techniques in the early 1960s. Tikhonov regularization is used in many fields aside from linear regression, such as classification with logistic regression.

Although that his original work discuss many forms and general formulation of the regularization function, see our discussion on Eq.2.54, but the commonly used form is the use of L_2 norm in the regularization term such that $\mathcal{R}(\boldsymbol{\beta}) = \|\Gamma\boldsymbol{\beta}\|_2^2$, where Γ is Tikhonov matrix, and it is usually chosen to be $\Gamma = \lambda I$, with λ is the regularization parameter and I is the identity

matrix [80, 183]. Then, the optimization problem takes the form:

$$\min_{\boldsymbol{\beta}} \left(\|\Phi\boldsymbol{\beta} - \boldsymbol{f}\|_2^2 + \|\Gamma\boldsymbol{\beta}\|_2^2 \right) \tag{2.56}$$

Since for $\boldsymbol{\beta} \in \mathbb{R}^K$, we have [78]:

$$\|\boldsymbol{\beta}\|_\infty \leq \|\boldsymbol{\beta}\|_2 \leq \|\boldsymbol{\beta}\|_1 \leq \sqrt{K}\|\boldsymbol{\beta}\|_2 \leq K\|\boldsymbol{\beta}\|_\infty \tag{2.57}$$

Tikhonov regularization re-invented and formulated in many other forms, with different strategies of choosing Γ.

2.10.2 LASSO

A relaxation for L_0 minimization is the regularized optimization problem:

$$\min_{\boldsymbol{\beta}} \left(\alpha\|\Phi\boldsymbol{\beta} - \boldsymbol{f}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \right), \tag{2.58}$$

which is the Lagrangian form of constrained optimization problem that have $\|\boldsymbol{\beta}\|_1 < t$ is the feasible search space for some predefined constant t, and the factor α, which have been used in literature as $\alpha = 1/2N, \alpha = 1/2$, and $\alpha = 1$ corresponds to a different parametrization of λ, to make λ values comparable for different sample sizes, which is useful when using cross-validation technique. Similar to [89], we choose $\alpha = 1$. The parameter $\lambda \geq 0$ controls the extent to which sparsity is desired: as $\lambda \to \infty$ the second term dominates and the only solution is a vector of all zeros, whereas at the other extreme $\lambda = 0$ and the problem becomes identical to a least squares problem which generally yields a full (non-sparse) solution. Values of $0 < \lambda < \infty$ balances the "model fit" quantified by the 2-norm and the sparsity of the solution characterized by the 1-norm. For a given problem, the parameter λ needs to be tuned to specify a particular solution. A common way to select λ is via cross validation [89].

2.10.3 SINDy

Judging the parameters relevancy based on their magnitude, and iteratively applying a threshold to discard the low magnitude parameters, can be useful and computationally efficient when

having stable basis matrix, with low condition number, and unified scale of the basis. Hard thresholding has been extensively studied through last two decades [21, 22, 23, 49, 71, 96].

In 2015, the authors in [30], introduced SINDy (Sparse Identification of Nonlinear Dynamics) as a way to perform nonlinear system identification in similar way by applying iterative hard thresholding on least squares solution. Different versions of SINDy have been introduced in [31, 102, 128, 154], but all of them commonly share the same basic form of hard thresholding which can be described as the following:

Let:

$$\mathcal{T}(\boldsymbol{\beta}, \lambda) = \begin{cases} \boldsymbol{\beta}_i, & \forall |\boldsymbol{\beta}_i| \geq \lambda, i =, 1, ..., K \\ \\ 0\ , & \text{otherwise.} \end{cases} \tag{2.59}$$

be the threshold operator that thresholds all parameters with a magnitude less than λ in the parameters vector $\boldsymbol{\beta}$, and let $x = L_s(\Phi, \boldsymbol{f}, \boldsymbol{\beta})$ be the least squares solution using the columns of Φ with index j such that $\boldsymbol{\beta}_j \neq 0$, and $x_j = 0$ for all $\boldsymbol{\beta}_j = 0$. Then, starting from $\boldsymbol{\beta}_0 = \Phi^+ \boldsymbol{f}$, SINDy solution obtained iteratively by:

$$\boldsymbol{\beta}_k = L_s(\Phi, \boldsymbol{f}, \mathcal{T}(\boldsymbol{\beta}_{k-1}, \lambda)). \tag{2.60}$$

Given that Lasso can be computationally costly, SINdy proposed to use sequential least squares with (hard) thresholding as an alternative. For a (pre-chosen) threshold λ, the method starts from the least squares solution and abandons all basis functions whose corresponding parameter in the solution has absolute value smaller than λ; then the same is repeated for the data matrix associated with the remaining basis functions, and so on, until no more basis function (and the corresponding parameter) are removed.

2.10.4 Extended SINDy

In their recent paper [186], the authors considered the SID problem where a certain fraction of data points are corrupted and proposed a method to simultaneously identify these corrupted data and reconstruct the system assuming that the corrupted data occurs in sparse and isolated time intervals. In addition to an initial guess of the solution and corresponding

residual, which can be assigned using standard least squares, TW approach requires a pre-determination of three additional parameters: A tolerance value to set the stopping criterion, threshold value λ used in each iteration to set those parameters whose absolute values are below λ to be zero, and another parameter μ to control the extent to which data points that do not (approximately) satisfy the prescribed model are to be considered as "corrupted data" and removed.

TW algorithm can be summarized as the following:

$$\begin{cases} \text{Given: } \varepsilon_0, b_0, tol, \lambda, \mu \\ \text{while } \|\varepsilon_k - \varepsilon_{k-1}\|_\infty > tol \\ \boldsymbol{\beta}_{k+1} = \mathcal{T}(\Phi^+(\boldsymbol{f} - \varepsilon_k - b_k), \lambda) \\ \varepsilon_{k+1} = W(\boldsymbol{f} - b_k - \Phi\boldsymbol{\beta}_{k+1}, \mu) \\ b_{k+1} = b_k + \Phi\boldsymbol{\beta}_{k+1} + \varepsilon_{k+1} - \boldsymbol{f} \end{cases} \tag{2.61}$$

where $\varepsilon_0 = \mathbf{0}_N$ is column of zeros, $b_0 \in \mathbb{R}^N$ is random variable such that $b_0 \sim \mathcal{N}(0,1)$, $\mathcal{T}(\cdot, \lambda)$ given in Eq.2.59, and $W(x, \mu)$ is the weight function:

$$W(x_j, \mu) = \max\left(1 - \frac{1}{\mu\|x_j\|}, 0)\right) x_j \tag{2.62}$$

with x_j indicate the rows of a matrix x.

2.11 Compressed Sensing and L1 Magic

Compressed sensing is a paradigm developed in recent years to reconstruct sparse signals using only limited data [12, 35, 36, 38, 39, 57, 194]. Mathematically, the problem of compressed sensing is to reconstruct the vector $\boldsymbol{\beta} \in \mathbb{R}^K$ in Eq. (2.18) from linear measurements, and it mainly focuses on the under-determined systems where that measurements matrix $\Phi \in \mathbb{R}^{N \times K}$ has negative degree of freedom with $N < K$, which also called "short-fat" matrix, and it provides a new sampling scheme given that the signal of interest is sparse in a certain basis. Moreover, compressive Sensing assumes the non-sparsity structure to be just a few terms [36,

38] that are lower than the basis functions by introducing the inequality:

$$s < N << K \tag{2.63}$$

Where s is the number of nonsparse entries in the vector $\boldsymbol{\beta}$, N is the number of measurements and K is the number of candidate functions. Then, the main objective is to solve (2.18) by finding the s-sparse vector $\boldsymbol{\beta}$. So, the basic optimization problem we have is:

$$\begin{cases} arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_0, \\ \text{subject to } \|\Phi\boldsymbol{\beta} - \boldsymbol{f}\| = 0, \end{cases} \tag{2.64}$$

As a relaxation of L^0-norm minimization, as discussed before, it has been shown that a recovery of the original signal is possible through the solution of the following convex optimization problem [35, 36, 38, 39], with equality constraints optimization problem:

$$\begin{cases} arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1, \\ \text{subject to } \|\Phi\boldsymbol{\beta} - \boldsymbol{f}\| = 0, \end{cases} \tag{2.65}$$

and with quadratic constraints optimization problem:

$$\begin{cases} arg\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1, \\ \text{subject to } \|\Phi\boldsymbol{\beta} - \boldsymbol{f}\| \leq \epsilon, \end{cases} \tag{2.66}$$

where $\|\boldsymbol{\beta}\|_1$ is the L^1 norm of the vector $\boldsymbol{\beta}$, and the tolerance ϵ is a user specified parameter represent a relaxation for the equality constrained optimization problem (2.65) with considering the presence of noise.

Recall that CS is the principle of recovering an unknown sparse vector with many fewer measurements than the systems dimension (Underdetermined Systems). The idea behind this recovery process is that we required our few measurements to be informative not just random sampling. So, CS combines the important task of compression directly with the measurement task [20, 37].

The recovery process by CS is highly dependent on the basis matrix Φ, and while CS has high

performance with orthogonal basis, but it is sensitive to the chosen basis. Restricted Isometry Property (RIP), Nullspace Property (NSP), and Coherence Criteria, [3, 11, 20, 35, 36, 37, 38, 184], become the main properties to analyze and decide the robustness of the recovery process.

From Eq. (2.17) and Eq. (2.18), we see that our linearization approach depends on creating a set of candidate functions as our basis that is not necessarily orthogonal or even independent, which can cause the measurement matrix to be unstable and sensitive for the CS recovery process [20, 81], as will be shown in our numerical results in Ch. (4).

Moreover, by considering the system of interest in Eq. (2.18) as black-box, with no prior information about number of non-sparse entries or the boundaries of the signal values, CS will tend to oversparse the solution with the presence of noise. In order to construct an example that clearly shows the oversparse mechanism in CS, consider the three-dimensional linear system:

$$\begin{pmatrix} 6 & 3 & 2 \\ 2 & 1 & 1 \\ 1 & 2 & 1 \end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix} 6 \\ 2 \\ 4 \end{pmatrix}. \tag{2.67}$$

It is easy to find that the solution for the above system is $\boldsymbol{\beta} = \begin{pmatrix} 0 & 2 & 0 \end{pmatrix}^T$. Now, suppose that the third "measurement" is missing, and we have the under-determined system

$$\begin{pmatrix} 6 & 3 & 2 \\ 2 & 1 & 1 \end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix} 6 \\ 2 \end{pmatrix} \tag{2.68}$$

where infinitely many solutions lie on the line of intersection of the two planes:

$$\begin{cases} 6x + 3y + 2z &= 6 \\ 2x + y + z &= 2 \end{cases}$$

Figure 2.18 shows this simple example, where the solution for $\boldsymbol{\beta}$ lies on the intersection of the two planes shown, and we see the true solution, the LS solution and CS solution on the solution line. We see how the LS solution is far from the true solution with a high margin of

error.

CS uses a different mechanism, since within all feasible solutions, it tends to select the one with minimum $\|\boldsymbol{\beta}\|_1$, even if there is another solution with the same sparsity that has a residual $\|A\boldsymbol{\beta}-b\|_2 = 0$, and it is the case in our example where $\|A \begin{pmatrix} 0 & 2 & 0 \end{pmatrix}^T - b\|_2 = 0$, while the CS solution has the residual $\|A\boldsymbol{\beta}_{CS}-b\|_2 = 2.5\times10^{-5}$. In other words, for the system $A\boldsymbol{\beta} = b$, if there exist two solutions such that $\|\boldsymbol{a_1}\|_1 < \|\boldsymbol{a_2}\|_1$ and $\|A\boldsymbol{a_2}-b\|_2 < \|A\boldsymbol{a_1}-b\|_2 \leq \epsilon$, where ϵ is the tolerance for CS optimization, then CS will select a_1 as a solution, even it has higher residual, and regardless of the structure of the sparse or the information flow between the basis functions and the observations. Numerically, assume the system in Eq. 2.68 to be:

$$\begin{pmatrix} (6+1e^{-10}) & 3 & 2 \\ 2 & (1+1e^{-16}) & 1 \end{pmatrix} \boldsymbol{\beta} = \begin{pmatrix} 6 \\ 2 \end{pmatrix} \tag{2.69}$$

and consider a reasonable tolerance for CS solver to be $\epsilon = 1e^{-9}$, then CS will always select [1 0 0] as a solution even though it has a higher residual.

For many applications, it is acceptable to have such solution since it lies on the solution line and such residual difference will have negligible effect on the final result, But in discovering the governing equations of dynamical systems, such solution can often lead to a completely inaccurate structure of the system.

On the other hand, some applications are not highly sensitive if they are missing a few non-sparse elements, and CS shows a high performance in such applications which makes it an efficient algorithm in communication in Wireless Sensors Networks (WSN), signal processing, signal compression, and antenna characterization. The common factor between the applications where CS exhibits high performance is the availability of prior informations, while such prior information may not be available in the case of discovering the governing equations of a complex system.

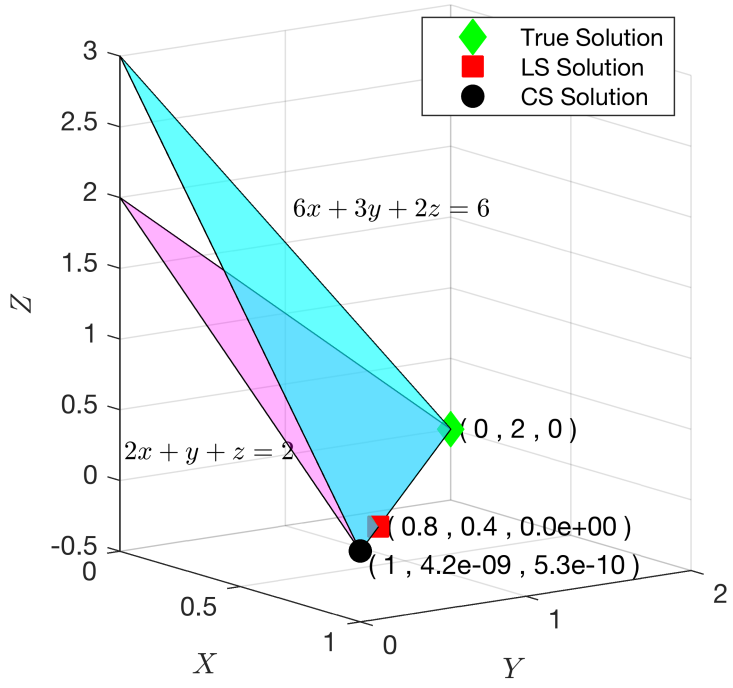

FIGURE 2.18: Oversparsity: The line of intersection of the two planes (triangles) shows the solution plane. We see that compressed sensing solution is oversparsed.

2.12 Orthogonal Least Squares

A Greedy Algorithms approach have been developed 1988 [43], which called Orthogonal Least Squares (OLS),[2] and have been re-introduced [18, 59, 108, 142, 151, 193], and it extended in [207, 208, 209].

Greedy search depends on dividing the optimization problem to iterative local ones, by iteratively search for the function (column) that is locally optimal according to the objective function. It is one of the most efficient approaches, that has better performance with fewer measurement than other regularization techniques.

In orthogonal least squares (OLS), the idea is to iteratively select the columns of Φ that minimize the (2-norm) model error, which corresponds to iterative assigning nonzero values to the components of $\boldsymbol{\beta}$. In particular, the first step is to select basis ϕ_{k_1} and compute the

[2] Not to be confused with the Ordinary Least Squares which we simply call here the Least squares method.

corresponding parameter $\boldsymbol{\beta}_{k_1}$ and residual $\boldsymbol{r}_1$ according to

$$\begin{cases} (k_1, \boldsymbol{\beta}_{k_1}) = arg\min_{k,c} \|\boldsymbol{f} - c\boldsymbol{\phi}_k\|_2, \\ \boldsymbol{r}_1 = \boldsymbol{f} - \boldsymbol{\phi}_{k_1}\boldsymbol{\beta}_{k_1}. \end{cases} \tag{2.70}$$

Then, one iteratively selects additional basis functions (until stopping cretia is met) and then computes the corresponding parameter value and residual, as

$$\begin{cases} (k_{\ell+1}, \boldsymbol{\beta}_{k_{\ell+1}}) = arg\min_{k,c} \|\boldsymbol{r}_\ell - c\boldsymbol{\phi}_k\|_2, \\ \boldsymbol{r}_{\ell+1} = \boldsymbol{r}_\ell - \boldsymbol{\phi}_{k_{\ell+1}}\boldsymbol{\beta}_{k_{\ell+1}}. \end{cases} \tag{2.71}$$

To end the process there are several choices of stopping criteria, including AIC and BIC. In this work, in the absence of knowledge of the error distribution, we adopt a commonly used criterion where the iterations terminate when the norm of the residual is below a prescribed threshold. To determine the threshold, we consider 50 log-spaced candidate values in the interval $[10^{-6}, 100]$ and select the best using 5-fold cross validation.

2.13 Conclusions

Regardless of the particular method or system, most previous works focus on observational data that are either perfectly sampled data from a known system or with some very low level of noise. In practice, since an observation process can be subject to large disturbances in unpredictable ways, the effective noise can be large and even contain "outliers" that can contaminate the otherwise excellent data. Can SID still work under the presence of significant noise and outliers? At a glance, the answer should be yes, given that several recent SID methods for nonlinear systems are readily deployable in the presence of noise. For example, compressive sensing can handle noise by relaxing the constraint set. In the Sec. (4.3), we report numerical evidence that relatively large noise and outliers in the observational data generally cause issues for standard SID methods, including those that specialize in finding sparse models. More alarmingly, while the underlying model may be quite sparse with respect to a chosen basis, the truth may well not be optimally sparse and consequently sparse optimization methods (such as compressive sensing) can be brittle in the sense that when

they fail, they may fail spectacularly, even worse, the presence of outliers makes this issue more pronounced.

This general problem is traditionally discussed in the language of inverse problems, solved by assuming various forms of noise, or alternatively in the language of optimization, by least squares, orthogonal least squares, lasso, compressed sending, to name a few, each of these being mentioned in Sec. (4.3).

We depart from the standard approaches to SID. We identify the error quantification via metric norms as a root cause for existing methods to fail under large noise and outliers because outliers tend to deviate from the rest of sample data as measured by metric distance; thus trying to "fit" these outliers will cause the model to put less weight on the "good" data points. Instead, we propose to infer the (sparsity) structure of a general model together with its parameters using a novel *information theoretic* approach that we call entropic regression because of the inclusion of both entropy optimization and regression.

Real-world data sets are invariably often "too small" or "smaller than we wish," since collecting data is expensive, and too small is related to the dimensionality of the problem, clearly methods that can succeed to the same degree in the SmallData regime (in contrast to the trending phrase BigData) must be developed. We therefore suggest performing BigData analysis using SmallData.

As we will show in Ch. (4), while standard metric-based methods emphasize the data in ways as designed by the chosen metric, the proposed entropic regression is robust with regards to the presence of noise and outliers in the data. Instead of searching for the sparsest model and thus possibly becoming brittle, entropic regression emphasizes "relevance" according to a model-free, information-theoretic criterion. Basis terms are included in the model only because they are relevant and not because they together make up a sparse model. We demonstrate the effectiveness of entropic regression in several examples. We also remark on the computational complexity and convergence in a small-data regime.

Chapter 3

Information Theory

> In order to arrive at knowledge of the motions of birds in the air, it is first necessary to acquire knowledge of the winds, which we will prove by the motions of water in itself, and this knowledge will be a step enabling us to arrive at the knowledge of beings that fly between the air and the wind.
>
> *Leonardo da Vinci*
> *(1452 - 1519)*

Information theory in it's most basic form can be explained by considering the everyday learning process of our minds. The more "information" we have about a specific topic, the less "new" information we may find in the following days, and lower the probability of finding information resources that can *influence* you with *new* information that updates your previous assumptions. In other words, if the event A has a high probability of happening in our daily life, then there is no (or less) surprise when observing that event A occurs. On the other hand, seeing that event B happens—which is a rare event with low occurrence probability—will be a "surprise".

We can think of the "surprise" term as an indicator of the uncertainty. Applying this to our learning process example, the less one is surprised about information they receive, the more "certain" they are about the topic they are learning. More surprise indicates a higher uncertainty. That leads us to the first subsection about the fundamental measure in the information theory, which is Entropy; The measure of uncertainty. Since the methodology for our entropic regression in the following chapter depends on information theory, we review some of the relevant terms of this beautiful theory here.

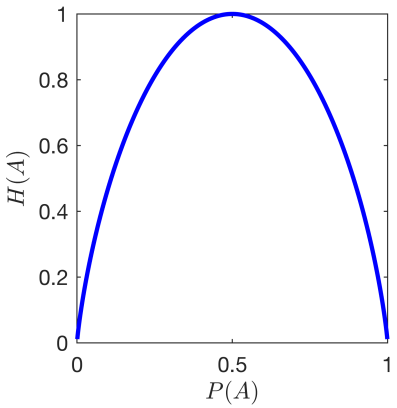

FIGURE 3.1: Entropy of the event A. Here we assume A has two mutually exclusive states, and $P(A)$ represent the probability of the state 1. This figure shows the uncertainty about the event A. In x-axis we have the probability $P(A) = p$ of state 1, then by Eq. 3.2, $H(A) = -p\log(p) - (1-p)\log(1-p)$ is the measure of uncertainty of the event A, where $(1-p)$ is the probability of state 2. Starting from $P(A) = 1$, meaning that the event A is always at state 1, then $H(A) = 0$, meaning that there is no uncertainty and we are sure of the event A state. As the probability decrease, the entropy (uncertainty) increase to reach its maximum at $P(A) = 0.5$, which is the case the state 1 and state 2 have equal probability. Continuing decreasing $P(A)$ will reduce the entropy again since we become more certain that the event A tends to have the state 2, until we become completely certain that A will only be in state 2, $H(A) = 0$ at $P(A) = 0$.

3.1 Entropy

Entropy was first known as an extensive property of a thermodynamic system [16]. The entropy of a thermodynamic system is a function of the number of possible microscopic states consistent with the macroscopic quantities that characterize the system. Assuming microstates with equal probability, the entropy is given by:

$$S = k_B \ln(W) \tag{3.1}$$

where W are the number of microscopic states and k_B is Boltzmann constant named after Ludwig Eduard Boltzmann [16]. This quantity, reffered to as Boltzmann's entropy, is a measure of statistical disorder in the system.

An analog to thermodynamic entropy is information entropy introduced by Claude Shannon in 1948 as "measures of information, choice, and uncertainty". To describe Shannon's entropy,

consider a discrete random variable X whose probability mass function is denoted by $p(x) = Prob(X = x)$. One can calculate its entropy as [50, 168],

$$H(X) = -K \sum_{x} p(x) \log p(x), \tag{3.2}$$

where K is positive constant, and $H(X)$ is a measure of the uncertainty or unpredictability of X. Note that if we assume uniform probability distribution for the states of X, then we have $p(x) = \frac{1}{N}$, where N is the number of states, and then Eq. 3.2 can be written as $H(X) = K \log(N)$ similar to Boltzmann's entropy under the same assumption of equal probability of the states. The constant K, as Shannon sates, is determined by a choice of a unit of measurement, and we will consider $K = 1$ and $\log = \log_2$ for the rest of this document for simplicity. Fig. 3.1 shows the entropy function for a random event with different probabilities.

Shannon's work provides an extended and generalized view and understanding for the entropy. One of the extended perspectives of Shannon's entropy is dealing with the continuous random variables, and it takes the form:

$$H(X) = \int_{-\infty}^{\infty} f_X(x) \log(f_X(x)) dx, \tag{3.3}$$

where $f_X(x)$ is the probability density function. The entropy shown in Eq. (3.3) is referred to the *differential entropy*.

3.2 Mutual Information

The entropy defined in Eq. (3.2) naturally extends to the case of multiple random variables. For example, the joint entropy $H(X, Y)$, and conditional entropy $H(X|Y)$ of two random variables X and Y is given, respectively, by [50, 168],

$$H(X, Y) = -\sum_{x,y} p(x, y) \log p(x, y) \tag{3.4}$$

$$\begin{aligned} H(X|Y) &= -\sum_{y} p(y) H(Y|X=x) \\ &= -\sum_{x,y} p(x,y) \log p(x|y), \end{aligned} \tag{3.5}$$

where $p(x,y)$ is the joint probability distribution, and $H(X|Y)$ (read as entropy of X given Y) is the measure of the uncertainty in X if Y is known. Some of the main properties of the entropy, joint entropy, and conditional entropy can be summarize as follows:

- The entropy of a discrete variable X is positive ($H(X) \geq 0$), while the differential entropy does not necessarily satisfy this property.
- For two independent random variables X and Y, $H(X,Y) = H(X) + H(Y)$.
- The chain rule: $H(X,Y) = H(X) + H(Y|X)$.
- One important property is that for a random variable X, the conditional entropy of X given any other variable Y will reduce the entropy of X, meaning that $H(X) \geq H(X|Y)$. The equality holds when X and Y are independent with $H(X,Y) = 0$. This property implies that the information from Y reduces the uncertainty about X. When $Y = X$ we have been given all the information about X, and we are completely certain about X, therefore $H(X|X) = 0$.

The joint and conditional entropies lead to measures that detect the statistical dependence or independence between random variables. Such a measure is called the mutual information between X and Y, and it is given by [50, 168],

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y), \end{aligned} \tag{3.6}$$

where the mutual information $I(X;Y)$ (reads as mutual information between X and Y) is a measure of the mutual dependence between the two variables. In terms of joint probability distribution, mutual information can be written as,

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right), \tag{3.7}$$

and in its continuous form,

$$I(X;Y) = \int_Y \int_X f_{X,Y}(x,y) \log \left(\frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} \right), \tag{3.8}$$

where $f_{X,Y}(x,y)$ is the joint probability density function for the two continuous random variables X and Y.

In case of independence of the two random variables, we have

$$p(x,y) = p(x)p(y), \tag{3.9}$$

and then we have

$$\log \left(\frac{p(x,y)}{p(x)p(y)} \right) = \log(1) = 0 \implies I(X;Y) = 0. \tag{3.10}$$

The same principle holds for the continuous variables in Eq. (3.8), although $I(X;Y)$ satisfies the inequality $I(X;Y) \leqslant min[H(X), H(Y)]$ only in the discrete variables case.

3.3 Transfer Entropy

Transfer entropy is a non-parametric statistic measuring the amount of directed transfer of asymmetric information between two random processes [166], it is related to Granger causality [83], and equivalent if the process is linear stochastic Gaussian [13]. For two stochastic processes X_t and Y_t, the reduction of uncertainty about X_{t+1} due to the information of the past τ_Y states of Y is represented by

$$Y^{(\tau_Y)} = (Y_t, Y_{t-1}, ..., Y_{t-\tau_Y+1}),$$

in addition to the information of the past τ_X states of X , represented by

$$X^{(\tau_X)} = (X_t, X_{t-1}, ..., X_{t-\tau_X+1}),$$

this reduction of uncertainty about X_{t+1} is measured by "Transfer Entropy" which given by [50, 168],

$$\begin{aligned} T_{Y\to X} &= H(X_{t+1}|Xt^{\tau_X}) - (X_{t+1}|Xt^{\tau_X}, Y_t^{\tau_Y}), \\ &= I(X_{t+1}; Y_t^{\tau_Y}|Xt^{\tau_X}). \end{aligned} \tag{3.11}$$

3.4 Causation Entropy

Causation, which is referred to as causality or cause and effect, is the effectiveness by which one process contributes or leads to the generation of another process. Here, the first process is the "cause" and the later the "effect." The cause being a least partially responsible for the effect, and the effect being partially dependent on the cause.

In Aristotle's Metaphysics, the word "cause" used to mean explanation or "answer to a 'why' question," and a general question in metaphysical philosophy is: what kind of entity can be a cause, and what kind can be an effect. One viewpoint is that any entity can be a cause or effect, with causation an asymmetric relation between them.

After Newtons work, and the new explanations of the world and the physical phenomenons, causation received a lot of interest, and Immanuel Kant introduced significant contributions to the field. In 1754 [32], while contemplating on Berlin Academy prize question about the problem of Earth's rotation, he claimed that the Moon's gravity would slow down Earth's spin and the gravity would eventually cause the Moon's tidal locking to coincide with the Earth's rotation. In 1755 he published his work: *"Universal Natural History and Theory of the Heavens: Attempt to Account for the Constitutional and Mechanical Origin of the Universe upon Newtonian Principles"*.

In the early 20^{th} century, different points of view start to appear for causation; Bertrand Russell argued that "In the motions of mutually gravitating bodies, there is nothing that called a cause and nothing that called an effect, there is merely a formula". In his metaphysics work, he takes the relation of causation to be a relation of determination, and there are no profound metaphysical truths about causation. In Russell view, A causes B if and only if A determines B to occur, with asymmetric relation. His ideas set the foundation of Eliminativism

philosophy.

Practically, the beginning of the 20's Century was the golden age for deterministic knowledge, where many scientists were hoping to create deterministic relationships in order to explain different phenomenon. However, at the same time, and initiated by Heisenberg, there was a deterministic existence of uncertainty, and many efforts went in the direction of quantifying the uncertainty, to have a better understanding to the nature of the relationship between processes.

Causes can be classified to [61, 141], necessary causes, sufficient causes, and contributory causes.

- Necessary causes: If x is a necessary cause of y, then the presence of y implies the prior occurrence of x. However, the presence of x does not imply the future occurrence of y.
- Sufficient causes: If x is a sufficient cause of y, then the presence of x implies the subsequent occurrence of y. However, the presence of y does not imply the prior occurrence of x, since another cause z may alternatively cause y.
- Contributory causes: For some specific effect y, a cause x_i is a contributory cause among several co-occurrent causes $x_j, j = 1, ...,$ and $j \neq i$. There is no implication that a contributory cause is necessary, though it may be so.

Here, we discuss one idea that set the ground for our approach for system identification. Recall the linear system

$$\boldsymbol{f} = \Phi\boldsymbol{\beta}, \tag{3.12}$$

the basic assumption is that among the set of processes $\phi_1, \phi_2, ..., \phi_K$, there is a small set $S \subset \{1, .., K\}$ of processes $\phi_i, i \in S$, and the cardinality $card(S) << K$, represent "together" a sufficient cause for the dynamic f. However, because of the noise, outliers, and many other parameters we discussed in Ch.2, we have an ill-conditioned system with a high degree of uncertainty. As a result, all the processes in Φ appear to be contributory causes, especially with the absence of a **reliable measure of causality**. Our approach in this thesis aims to find the minimal set S, such that the processes $\phi_i, i \in S$ represent a sufficient cause for $\boldsymbol{f}$.

The traditional approach of inferring causality between two stochastic processes is to perform the Granger causality test [83]. The main limitation of this test is that it can only provide

information about linear dependence between two processes, and therefore fails to capture intrinsic nonlinearities that are common in real-world systems. To overcome this difficulty, Schreiber developed the concept of transfer entropy between two processes [166]. Transfer entropy measures the uncertainty reduction in inferring the future state of a process by learning the (current and past) states of another process.

In [180], the authors showed by several examples that causal relationship inferred by transfer entropy is often misleading when the underlying system contains indirect connections, a dominance of neighboring dynamics, or anticipatory couplings. For example, the approaches that consider the transfer entropy in order to find the weak terms in Φ that has no influence on $\boldsymbol{f}$ to construct the sparse matrix $\boldsymbol{\beta}$, these approaches neglect the simple and clear idea that the terms of Φ have an indirect influence on $\boldsymbol{f}$ through the other terms of Φ, meaning that it neglects the information flow between the processes in Φ. To account for these effects, J. Sun, D. Taylor, and E. Bollt [180], developed a measure called *Causation Entropy* (CSE), that can (in contrast to transfer entropy) distinguish the directed and indirect influence, and they showed that its appropriate application reveals true coupling structures of the underlying dynamics.

Consider a stochastic network of N processes (nodes) denoted by:

$$X_t = \{X_t^{(1)}, X_t^{(2)}, \ldots, X_t^{(N)}\} \tag{3.13}$$

where $X_t^{(i)} \in \mathbb{R}^d$ is a random variable representing the state of process (or node) i at time t, and $i \in \mathcal{V} = \{1, 2, \ldots, N\}$, and let I, J, and K be a subsets of $\mathcal{V}$, then we can define the causation entropy as the following:

Definition 1 [180]: The causation entropy from the set of processes J to the set of processes I conditioning on the set of processes K is defined as

$$C_{J \to I|K} = H(X_{t+1}^{(I)} | X_t^{(K)}) - H(X_{t+1}^{(I)} | X_t^{(K)}, X_t^{(J)}). \tag{3.14}$$

The Causation entropy is a natural generalization of transfer entropy from measuring pairwise causal relationships to network relationships of many variables. In particular, we can list the

main properties for the causation entropy, noting that if $J = \{j\}$ and $I = \{i\}$, we simplify the notation as $C_{j\to i|K}$:

- If $j \in K$, then the causation entropy $C_{j\to i|K} = 0$, as j does not carry extra information (compared to that of K).
- If $K = \{i\}$, then the causation entropy recovers the transfer entropy $C_{j\to i|i} = T_{j\to i}$ which is given by $T_{j\to i} = H(X_{t+1}^{(i)}|X_t^{(i)}) - H(X_{t+1}^{(i)}|X_t^{(i)}, X_t^{(j)})$..

In [180], the authors introduced the principle of optimal Causation Entropy (oCSE) in a network of N processes to find the minimum subset that maximizes the causation entropy. We can see this minimal subset as the dominant subset of a network of N processes, and they rule the underlying dynamic of the network.

3.5 Mutual Information Estimators

We relay on a non-parametric entropy estimator based on the principle of k-nearest neighbors (Knn) developed by Kozachenko et al. (1987) [111], and in 2004, Knn mutual information estimator introduced by Kraskov et al. [112], which is known as KSG estimator.

Let $x, y \in \mathbb{R}^n$ be two random variables, then, for each data index $i \in \{1, ..., n\}$, KSG finds the radii R_i of the spherical volumes in the joint space $J = (x, y)$ between J_i and its k^{th} smallest element in the set $S = \{\|J_i - J_j\| : j \in \{1, ..., n\}, j \neq i\}$. Then, by centering the volume of the sphere at the i^{th} data point, we find $N_{i,x}$, which is the number of neighbors in x that are inside the sphere with respect to x_i, and similarly we find $N_{i,y}$, which is the number of neighbors in y that are inside the sphere with respect to y_i. Then, the mutual information between x and y is given by:

$$I(x; y) = \psi(k) + \log(n) - \frac{1}{n}\sum_{i=1}^{n}(\psi(N_{i,x} + 1) + \psi(N_{i,y} + 1)) \tag{3.15}$$

where ψ is the Digamma function defined as the logarithmic derivative of the Gamma function $\psi(x) = \frac{d}{dx}\Gamma(x)$. Different approaches can be considered to choose distance function to find R_i in the joint space, the above equation, which we adopt in our work, is for using the max-norm sphere ($\|\cdot\|_\infty$).

Similar approach introduced by Vejmelka et al. [191], to estimate the conditional mutual information $I(x;y|z)$, $x,y,z \in \mathbb{R}^n$. Where R_i found in the joint space $J=(x,y,z)$, between J_i and its k^{th} smallest element in the set $S=\{\|J_i-J_j\|_\infty : j \in \{1,...,n\}, j \neq i\}$. Then the volume centered at the i^{th} data point on the subspace $J^x=(x,z), J^y=(y,z)$ and $J^z=z$ to find $N_{i,xz}, N_{i,yz}$ and $N_{i,z}$ respectively. The conditional mutual information is then given by:

$$I(x;y|z) = \psi(k) - \frac{1}{n}\sum_{i=1}^{n}\left(\psi(N_{i,xz}+1)+\psi(N_{i,yz}+1)-\psi(N_{i,y}+1)\right) \tag{3.16}$$

KSG's computational complexity comes from distance computations for the nearest neighbor, but the using Kd trees search algorithm, the complexity can be reduced from $\mathcal{O}(n^2)$ to $\mathcal{O}(n\log n)$.

3.6 Mutual Independence Test

Another important issue in practice is the determination of mutual independence: in theory, mutual information $I(x;y|z)$ is always non-negative and equals zero if and only if x and x are statistically independent given z. However, in practice, due to finite sampling and estimation inaccuracies, the estimated mutual information does not always equal to zero even when x and y are independent and can be negative. Thus, one needs a way to decide whether x and y should be deemed independent given the estimated value of $I(x;y|z)$.

In [180], the authors introduced a standard shuffle test, with a "confidence" parameter $\alpha \in [0,1]$ for tolerance estimation. The shuffle test requires randomly shuffling of one of the variables n_s times, to build a test statistic. In particular, for the i-th random shuffle, a random permutation $\pi^{(i)} : [T] \to [T]$ is generated to shuffle one of the variables, say y, which produces a new variable $(\tilde{y}^{(i)})$ where $\tilde{y}^{(i)} = y_{\pi^{(i)}}$; x and z are kept the same. Then, we estimate the mutual information $I(x;\tilde{y}^{(i)}|z)$ using the (partially) permuted variable $(x,\tilde{y}^{(i)},z)$, for each $i=1,\ldots,n_s$. For given α, we then compute a threshold value $I_\alpha(x;y|z)$ as the α-percentile from the values of $I(x;\tilde{y}^{(i)}|z)$. If $I(x;y|z) > I_\alpha(x;y|z)$, we determine z and z as dependent; otherwise independent. Algorithm 1 shows the shuffle test algorithm.

In [180], the authors showed the robustness of shuffling test for optimal causation entropy calculations, especially in complex dynamics, although it is computationally expensive. For

Algorithm 1 Shuffle Test

procedure SHUFFLE TEST($\mathbf{f}, \phi, \mathbf{\Phi_K}, \alpha, \mathbf{n_s}$)
  $i = 1$, $I = \emptyset$
  while $i \leq n_s$ **do**
    $I \leftarrow C_{\phi \rightarrow \mathbf{f}_{\pi \mathbf{i}} | \mathbf{\Phi_K}(\mathbf{\Phi_K^+ f})}$,
    $i := i + 1$,
  return I
  $\mathcal{I} \leftarrow I$ s.t. $\mathcal{I}_j \leq \mathcal{I}_{j+1}$, $j = 1, \ldots, n_s - 1$
  $tol = \mathcal{I}_k$, where $k = \lceil \alpha n_s \rceil$.
return *tol*

more efficient computations complexity, we considered a simplified version of the shuffle test that will be discussed in Ch.4.

3.7 Asymptotic Equipartition Property

In information theory, the asymptotic equipartition property is the direct analog of the law of large numbers. Indeed a large fraction of the information theory can be built on the AEP [50]. Recall that the weak law of large numbers state that the sample average converges in probability towards the expected value:

$$\lim_{n \to \infty} Pr\left[|\bar{x}_n - \mu| > \varepsilon\right] = 0 \tag{3.17}$$

which implies that for any nonzero small ε, there will be a very high probability that the average of the observations will be close to the expected value with a sufficiently large sample.

The Asymptotic equipartion property states the following [50]:

Theorem 3.7.1. ***AEP Theorem [50]:*** *If $x_1, x_2, \ldots, x_n$ are independent and identically distributed random variables, then*

$$\lim_{n \to \infty} Pr\left[|-\frac{1}{n}\log p(x_1, x_2, \ldots, x_n) - H(X)| > \varepsilon\right] = 0. \tag{3.18}$$

where $p(x)$ is a probability function and $H(X)$ is the entropy of the sample.

Proof. Since x_i are i.i.d, then a function $\log p(x)$ is also i.i.d., Hence, by the weak law of large numbers:

$$\begin{aligned} -\frac{1}{n}\log p(x_1, x_2, \ldots, x_n) &= -\frac{1}{n}\sum_{i=1}^{n}\log p(x_i) \\ &\to E[-\frac{1}{n}\log p(X)], \text{(converge in probability)} \\ &= H(X). \end{aligned} \tag{3.19}$$

□

The AEP allows dividing the set of all sequences into two sets, the typical set, where the sample entropy is close to the actual entropy, and the non-typical set, which contains the other sequences. The typical set is our primary interest, so we will extend the discussion in [50], in this section to discuss the implications of AEP on our approach for system identification.

From [50] we have:

Definition 3.7.1. *A set of sequences* $(x_1, x_2, ..., x_n) \in \mathcal{X}$*, that satisfy the property:*

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, ..., x_n) \leq 2^{-n(H(X)-\epsilon)} \tag{3.20}$$

is said to be a typical set $A_\epsilon^{(n)}$ *with respect to* $p(x)$*.*

As a consequence of the AEP, the set $A_\epsilon^{(n)}$ has the following properties:

1. If $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$, then

$$H(X) - \epsilon \leq -\frac{1}{n}\log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon. \tag{3.21}$$

2. $Pr\{A_\epsilon^{(n)}\} > 1 - \epsilon$ for sufficiently large n.
3. $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$, where $|A_\epsilon^{(n)}|$ is number of elements in the set $A_\epsilon^{(n)}$.
4. $|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$, for sufficiently large n.

Note that number of elements is the number of possible sequences in the set entries. Then we summarize that:

- The typical set has a probability of nearly 1.

- All elements of the typical set are nearly equiprobable.
- The number of elements in the typical set is nearly $2^{nH(X)}$.

One example to give clear view for the above properties is to consider random binary number in the form Bernoulli probability $Ber(p)$, where for a uniform random number r and a predefined value p, $r, p \in [0, 1]$, then $Ber(p)$ is given by:

$$Ber(p) = \begin{cases} 1, r < p \\ 0, r \geq p \end{cases} . \tag{3.22}$$

Practically, $Ber(p)$ should be written as $Ber(p, r)$, however, since r is a random number, we assume the implicit generation of r and we use the reduced notation $Ber(p)$ in all this text.

With $p = 0.5$, $x = Ber(p)$ generate a sequence of 0 and 1 with equal probability for each state, and as discussed in Fig.3.1, such sequence will have a maximum entropy $H(X) = 1$. Then the typical set will have $2^{nH(X)}$ elements with requires $nH(X)$-bits to represent them, and in our case in it is required n bits to represent a sequence of Bernoulli trials with $p = 0.5$. Then, the number of elements in the typical set equal the total number of elements.

Now, consider the case where $p = 0.05$, which will give a sequence $x_n = Ber(p)$ with "most" entries equal to 0. Such sequence has entropy $H(x) = 0.1414$ with large n. Then, the typical set will have $2^{nH(X)} = 2^{0.1414n}$ elements.

This idea is mainly used in coding, to use lower bitrate to store or transmit the data, where it is known that the optimal coding will code the signal with bit rate equal to the entropy of the signal, as the signal length is large. So, for a "large" signal, there will be a subset of the signal that has most of the information, and the non-typical set will only be a small fraction of the information. On the other hand, if we have two identical signals x and y, with some entries in x have a deviation (outliers) from y, then by the principle of AEP and the typical sets in x, and y, the information $I(x; y)$ we remain the same and will not be affected by the non-typical sets, as the sample size is large.

Then, the information measure $I(x, y)$ and the L_p measures $||x - y||_p$, since they are functions of i.i.d., will converge in probability to their mean for large sample size. The only difference between them is the term "large". We will show that the information measures $I(x, y)$ reach

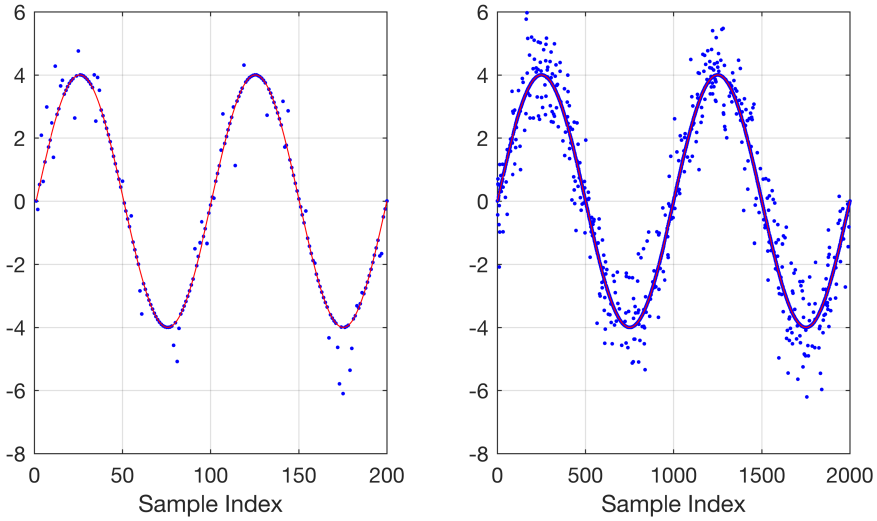

FIGURE 3.2: The clean signal x (red) and the noisy signal $\tilde{y}$ (blue) generated by Eq.3.23. (Left) 200 data points. (Right) 2000 data points.

stable and robust estimation to outliers with the number of measurements that is 1 to 2 orders less than L_2.

To investigate this difference numerically, let $t \in \mathbb{R}^n$ be equally spaced span such that $t_i \in [-4\pi, 4\pi]$, and let the signal $x_i = y_i = 4\sin(t_i), \forall i \in \{1, ..., n\}$, that the two vectors x and y are identical. We want to investigate how the L_2 measure $\|x - y\|_2$ and the information measure $I(x, y)$ perform if the signal y has a "some" of its entries deviate from the true value. Moreover, we investigate the change in performance for increasing sample size n.

So, for each sample size $n_k \in [10^2, 10^4]$, we add high noise to y such that:

$$\tilde{y}_i = y_i + \eta_i Ber(p) \tag{3.23}$$

where $\eta \sim \mathcal{N}(0, \sigma_\eta^2)$, $\sigma_\eta^2 = \sigma_x^2$, meaning that the signal to noise ratio $SNR = 1$, which is high noise (see Fig.3.2), and $Ber(p)$ is Bernoulli probability with $p = 0.25$, meaning that only 25% of the sample size is noisy. For each sample size n_k, we repeat the experiment 100 times and we record the values $std(\{\|x - \tilde{y}_j\|_2\}_{j=1}^{100})_k$ and $std(\{I(x, \tilde{y}_j)\}_{j=1}^{100})_k$. Fig.3.3 shows the standard deviation for mutual information and the L_2-norm for different sample sizes. We see that the mutual information estimated by the KSG estimator is more robust to the presence of outliers in a subset of the sample, since according to the AEP, there is a subset from the sample that have entropy (information) equal within small tolerance the total entropy of the signal.

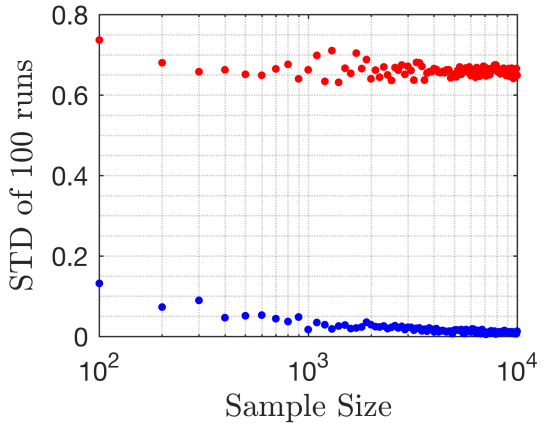

FIGURE 3.3: (Blue) STD in mutual information $I(x, \tilde{y})$. (Red) STD in $\|x-\tilde{y}\|_2$. We see that the mutual information and the L_2 converged as the sample size increased. However, mutual information has low sensitivity to outliers and has a lower magnitude and faster convergence at a smaller sample size than L_2-norm. Moreover, it shows that mutual information did not affect by adding outliers points to a fraction of the sample, and the standard deviation of mutual information converges to zero as the sample size increase.

Chapter 4

Entropic Regression

> Science is built up of facts, as a house is with stones. But a collection of facts is no more a science than a heap of stones is a house.
>
> *Henri Poincare*
>
> *(1854 - 1912)*

To overcome the competing challenges of potential overfitting, efficiency when limited data points are available, and robustness to noise and in particular outliers in observations, we propose a novel framework that combines the advantage of information-theoretic measures and iterative regression methods. The framework, which we term *Entropic Regression* (ER), is model-free[1], noise-resilient, and efficient in discovering a "minimally sufficient" model to represent data. The work presented in this chapter primarily follows our submitted manuscript: "*How Entropic Regression Beats the Outliers Problem in Nonlinear System Identification*" [6].

In entropic regression, we use (conditional) mutual information as an information-theoretic criterion and iteratively select relevant basis functions, analogous to the optimal causation entropy algorithm previously developed for causal network inference [180, 181]; between each iteration, the corresponding parameters are updated using a standard regression method, e.g., least squares. Thus, ER can be thought of as an information-theoretic extension of the orthogonal least squares regression, or as a regression version of optimal causation entropy.

[1] It is a black box modeling method.

4.1 Theoretical and Applied Foundations

Recall that in this work, we consider the linear regression problem in the following form:

Observing a state space measurements $\boldsymbol{z} \in \mathbb{R}^{N\times d}$, where N is the number of measurements in d-dimensional space, $\boldsymbol{z}$ is equally spaced measurements with time step τ between observations, and we assume that the observations $\boldsymbol{z}$ deviates from some unknown ideal (or true) states such that:

$$\boldsymbol{z} = \boldsymbol{z}^{\star} + \eta \tag{4.1}$$

where $\boldsymbol{z}^{\star} \in \mathbb{R}^{N\times d}$ is the unknown true states and $\eta \in \mathbb{R}^{N\times d}$, $\eta \sim \mathcal{N}(\mu, \sigma^2)$, is Gaussian noise with μ mean and variance σ^2. Here we assume $\mu = 0$, and the measurements contaminated with large noise and outliers ($\sigma >> 0$).

We adopt the approach that the underlying dynamic is governed by a set of high dimensional nonlinear ODE, and we obtain the vector field from the measurements such that:

$$\begin{aligned} \boldsymbol{f} &= \mathcal{E}(\boldsymbol{z}^{\star} + \eta, \tau) + \epsilon \\ &= \mathcal{E}(\boldsymbol{z}, \tau) + \epsilon \end{aligned} \tag{4.2}$$

where $\mathcal{E} : \mathbb{R}^{N\times d} \times \mathbb{R} \to \mathbb{R}^{N-n\times K}$ is a numerical derivative estimation function such as using the finite difference method, n is a number of measurements that we may lose in derivative estimation by finite difference and we will consider it negligible in the following discussion since we can consider our measurements are $N + n$ measurements, and ϵ is the rounding and numeric error. We use in this work the central difference method given by:

$$\begin{aligned} \boldsymbol{f} &= \mathcal{E}(\boldsymbol{z}, \tau) \\ &= \frac{\boldsymbol{z}_{i+1} - \boldsymbol{z}_{i-1}}{2\tau} \end{aligned} \tag{4.3}$$

where $\boldsymbol{z}_i \in \mathbb{R}^{1\times d}$ is the d-dimensional i^{th} measurement, $i = 1, ..., N$. We see that here we assumed $n = 2$, and we observed $N + 2$ measurements, we discard the first and last measurement from $\boldsymbol{z}$ to have $\boldsymbol{f}, \boldsymbol{z} \in \mathbb{R}^{N\times d}$. Note that derivative estimation is an error

amplification process[2]:

$$\begin{aligned}\mathcal{E}(\boldsymbol{z}^\star + \eta, \tau) &= \frac{\boldsymbol{z}^\star_{i+1} + \eta_{i+1} - \boldsymbol{z}^\star_{i-1} - \eta_{i-1}}{2\tau} \\ &= \frac{\boldsymbol{z}^\star_{i+1} - \boldsymbol{z}^\star_{i-1}}{2\tau} + \frac{\eta}{2\tau}\end{aligned} \tag{4.4}$$

which gives that the estimated vector field $\boldsymbol{f}$ in terms of the true unknown one is given by:

$$\boldsymbol{f} = \boldsymbol{f}^\star + \frac{\eta}{2\tau}, \tag{4.5}$$

which is results with large error in the estimated derivative, considering that we usually use small step size of order -2 or -3. This small fact created a confusion and even misconceptions in literature, that we have discussed in more details in Appendix A.

As discussed in Sec.2.5, we construct our basis matrix from the state noisy measurements such that:

$$\Phi = \mathcal{C}(\boldsymbol{z}, l) \tag{4.6}$$

where $\mathcal{C} : \mathbb{R}^{N\times d} \times \mathbb{R} \to \mathbb{R}^{N\times K}$ is the power polynomials expansion function, is the expansion order, and $K = \frac{(d+l)\,!}{d\,!l\,!}$ is the dimension (number of columns, number of functions) of the basis matrix Φ. Now, our problem can be set in linear regression matrix form[3]:

$$\boldsymbol{f} = \Phi\boldsymbol{\beta} \tag{4.7}$$

which have the least squares solution given by:

$$\begin{aligned}\mathcal{L}(\boldsymbol{f}, \Phi) &= (\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{f} \\ &= \Phi^+\boldsymbol{f}\end{aligned} \tag{4.8}$$

[2]Note that in the equation below, sum of two random numbers is again a random number of the same order.

[3]We will write few functions in detail in order to simplify the discussion in the following sections, and we try to match the function letter with its physical meaning. i.e., $\mathcal{L}$::Least squares solution, $\mathcal{V}$:: reconstructed Vector field.

with Φ^+ is the pseudoinverse of the matrix Φ. The vector field reconstructed using the least squares solution is given by:

$$\begin{aligned} \mathcal{V}(\boldsymbol{f}, \Phi) &= \Phi\Phi^+\boldsymbol{f} \\ &= \Phi\mathcal{L}(\boldsymbol{f}, \Phi) \end{aligned} \tag{4.9}$$

Sparse Regression problem now finding the minimal set of index $s \subset \{1, 2, ..., K\}$, such that $\mathcal{V}(\boldsymbol{f}, \Phi_s)$ as close as possible for $\boldsymbol{f}$ according to adopted quality measure (or objective function, cost function, loss function), where $\Phi_s \subset \Phi$ is the matrix with only the columns with index $i \in s$.

We discussed in Chapter. 2 the challenges of sparse regression; noise, outliers, dimension, numerical stability,...etc. And we discussed the current state of the art approaches for solving sparse regression problems. In Chapter 3 we discussed the information theory, and the principles of asymptotic equipartition property, optimal causation entropy, and conditional mutual information. In light of the previous discussions, we have developed the Entropic Regression approach that we discuss in the following section.

4.2 Entropic Regression Algorithm

The ER method contains two stages: Forward ER and Backward ER; in both stages, selection and elimination of basis functions are based on an entropy criterion (conditional mutual information), and parameters are updated in each iteration using a standard regression (e.g., least squares).

4.2.1 Forward Entropic Regression

In the forward stage, our objective is to select the subset $s \subset \mathcal{S} = \{1, 2, ..., K\}$, that represent sufficient causes, or in other words, a strong candidate functions with high probability to be accurate.

Starting from empty set $s_0 = \emptyset$, the forward selection stage can be written as:

$$\begin{aligned} u_k &= arg \max_{i \in \mathcal{S}, i \notin s_{k-1}} I(\boldsymbol{f}; \mathcal{V}(\boldsymbol{f}, \Phi_i) | \mathcal{V}(\boldsymbol{f}, \Phi_{s_{k-1}})), \\ s_k &= s_{k-1} + u_k \end{aligned} \tag{4.10}$$

where $k = 1, ...,$ is the iteration index, u_k is the set[4] of index with the maximum objective function value. Note that $s_0 = \emptyset \implies \mathcal{V}(\boldsymbol{f}, \Phi_{s_0}) = \emptyset$ which reduces the conditional $I(\cdot;\cdot|\cdot)$ to the mutual information $I(\cdot;\cdot)$.

The forward stage have a reward function, where at each iteration k, given the information $(\mathcal{V}(\boldsymbol{f}, \Phi_{s_{k-1}}))$ we already have from the set s_{k-1} we are looking for the function that maximally add extra information to the model.

The process terminates when either all basis functions are exhausted (with maximum number of iterations equal K), or the reward function $I(\boldsymbol{f}; \mathcal{V}(\boldsymbol{f}, \Phi_i) | \mathcal{V}(\boldsymbol{f}, \Phi_{s_{k-1}})) = 0$ indicating that none of the remaining basis functions are *relevant*, in an information-theoretic sense. In another words, the process terminates when the strongest candidate is weak, and have no extra information than what we already have.

As discussed in Sec.3.6, the value of the conditional mutual information may not reach a value of zero for the termination condition, and to estimate reliable tolerance value, one may adopt the shuffle test in analogy to oCSE approach in [180]. However, we developed a simplified, cost-efficient, and insensitive tolerance estimation method based on a modified shuffle test approach, we discuss tolerance estimation in section 4.2.3. Algorithm 2 shows the forward entropic regression algorithm.

Algorithm 2 Entropic Regression

procedure FORWARD ER:$(\mathbf{f}, \mathbf{\Phi}, \mathbf{tol})$
  $s = \emptyset, index = \emptyset, value = \infty$
  while $value > tol$ **do**
    $s \leftarrow index$ ▷ Add the index to the selection set.
    $\mathcal{I} := -\boldsymbol{\infty}_K$ ▷ Vector of length K of $-\infty$ entries.
    $\mathcal{I}_k := I(\boldsymbol{f}; \mathcal{V}(\boldsymbol{f}, \Phi_k) | \mathcal{V}(\boldsymbol{f}, \Phi_s))$, for $k = 1, ..., K$ and $k \notin s$.
    $(value, index) := \max_k(\mathcal{I}_k)$ ▷ Find the value and the index of maximum $\mathcal{I}$.
  end
return s.

[4] Note that we may have more than one index with the same maximum value, that u_k is not necessarily has one element, although it usually does.

4.2.2 Backward Entropic Regression

After the forward ER, we have the set s that has the indices of the strong candidate functions. Eventually, s may have a few non-relevant functions that are selected due to a high degree of uncertainty and the rounding error at the end of forward ER. Since we have reduced set ($card(s) << K$), it will be inexpensive to perform a validation operation to ensure the accuracy of the model, and the Backward ER represent this operation.

The backward stage is an elimination stage, where the functions indexed by s re-examined for their information-theoretic relevance and these that are redundant will be removed. In particular, we label the set s as initial set $s_0 = s$ for the backward stage, and we perform the following computations and updates,

$$\begin{aligned} u_k &= arg \min_{i \in s_{k-1}} I(\boldsymbol{f}; \mathcal{V}(\boldsymbol{f}, \Phi_i) | \mathcal{V}(\boldsymbol{f}, \Phi_{\{s_{k-1}-i\}})), \\ s_k &= s_{k-1} - u_k. \end{aligned} \tag{4.11}$$

The backward stage have a loss function, where at each iteration k, we examine that what information will be lost if we remove the index i from the set s_{k-1}, and we continue the elimination process as well as this information lose is zeros, and the process terminate when $I(\boldsymbol{f}; \mathcal{V}(\boldsymbol{f}, \Phi_i) | \mathcal{V}(\boldsymbol{f}, \Phi_{\{s_{k-1}-i\}})) > tol$. The result of the backward ER is a set of indices s, and Algorithm 3 shows the backward entropic regression algorithm.

Algorithm 3 Entropic Regression

procedure BACKWARD ER:($\mathbf{f}, \mathbf{\Phi}, \mathbf{s}, \mathbf{tol}$)
  $index = \emptyset, value = \infty$
  while $value < tol$ **do**
    $s := s - index$ ▷ Remove index from the set s.
    $\mathcal{I} := \boldsymbol{\infty}_{card(s)}$ ▷ Vector of length $card(s)$ of ∞ entries.
    $\mathcal{I}_i := I(\boldsymbol{f}; \mathcal{V}(\boldsymbol{f}, \Phi_{s_i}) | \mathcal{V}(\boldsymbol{f}, \Phi_{\{s-s_i\}}))$, for $i = 1, ..., card(s)$.
    $(value, index) := \min_i(\mathcal{I}_i)$ ▷ Find the value and the index of minimum $\mathcal{I}$.
  end
return s.

The corresponding parameters $\boldsymbol{\beta} \in \mathbb{R}^K$, can be found by updating the vector of zeros $\boldsymbol{\beta} = \mathbf{0}_K$ such that:

$$\boldsymbol{\beta}_s = \mathcal{L}(\boldsymbol{f}, \Phi_s) \tag{4.12}$$

where $\boldsymbol{\beta}_s$ is the entries of $\boldsymbol{\beta}$ indexed by the elements of s, which gives that $\|\boldsymbol{\beta}\|_0 = card(s)$.

Note that the ER focus on finding the optimal set of basis functions, and after finding this set, we find the value of the parameters by the ordinary least squares, without any attempts to apply any advanced techniques for optimizing the parameter's value.

4.2.3 Tolerance Estimation

We can interpret the tolerance as the minimum effective quantity of information. In this sense, in the forward ER we are selecting the functions as well as they add a significant quantity of information to the model, while in the backward ER, we discarding functions as well as the information added by them is below the minimum effective quantity, or, negligible.

And since the objective of our model is to maximize the mutual information with the dynamic $\boldsymbol{f}$, we considered the mutual information between $\boldsymbol{f}$ and a random shuffling of $\boldsymbol{f}_\pi$, where π is shuffling function in analogy to Sec.3.6, as the minimum accepted quantity of information. Meaning that, even if the information added by a function is practically a random permutation of the dynamic, we will accept that function.

Algorithm. 4 shows the tolerance estimation algorithm with n_s is the number of shuffle test. We see that the value of $I(\boldsymbol{f}, \boldsymbol{f}_{\pi^i})$ is independent of the loop index, which enables the parallel computation and increases the efficiency. Fig.4.1 shows the histogram of the shuffle test performed on sample function and the tolerance estimation.

Algorithm 4 Tolerance Estimation

procedure TOLERANCE SHUFFLE TEST($\mathbf{f}, \mathbf{n_s}$)
  $I_j^{sh} := I(\boldsymbol{f}, \boldsymbol{f}_{\pi^i})$, for all $i = 1, ..., n_s$.
  $\mathcal{I} \leftarrow I^{sh}$ s.t. $\mathcal{I}_j \leq \mathcal{I}_{j+1}$, $j = 1, \ldots, n_s - 1$. ▷ Sort entries.
  $tol = \mathcal{I}_k$, where $k = \lceil \alpha n_s \rceil$. ▷ Return the k^{th} entry, where $\lceil \cdot \rceil$ is the ceil function.
return *tol*

Our method considers the system in the SID process as a black-box, without assuming the availability for any prior information about the system, and the tolerance estimation process itself is within the algorithm itself, and our numerical experiments shows that the Entropic regression is insensitive to the value of α, chosen within reasonable interval such as $\alpha = [0.95, 1)$. Our numerical results show the robustness of the method under this assumption.

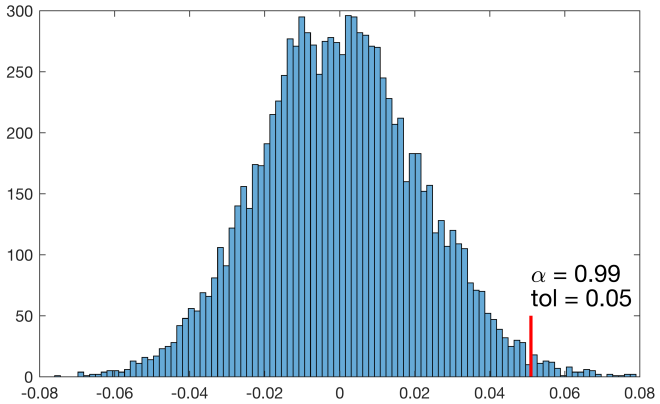

FIGURE 4.1: Tolerance Estimation with shuffle test. The figure shows the histogram of the information $I(\boldsymbol{f}, \boldsymbol{f}_{\pi^i})$, $i =, 1, ..., n_s$ with $n_s = 10000$, for the chaotic function $f(x) = \sin(e^x)$. We see that with a confidence parameter $\alpha = 0.99$, we have a tolerance estimation of $tol = 0.05$.

4.3 Numerical Results

To demonstrate the utility of ER for nonlinear system identification under noisy observations, we compare its performance against existing methods including the standard least squares (LS), orthogonal least squares (OLS), LASSO, compressed sensing (CS), SINDy, and extended SINDy algorithm which we will use TW to abbreviate it. The details of existing approaches are described in Chapter 2. The examples we cover represent different types of systems and scenarios, including both ODEs and PDEs, differential and difference equations, and network-coupled dynamics. Besides, we consider different noise models with a focus on the presence of outliers in evaluating the robustness of the respective methods.

The sampling and noise addition methodology discussed in Sec.4.1, Where we assume the observations are noisy, and the vector field $\boldsymbol{f}$ is subject to the amplification factor of order $\frac{1}{\tau}$, with τ is the sampling step size. The vector field $\boldsymbol{f}$ is estimated using the central difference method. The number of nearest neighbors of Knn estimator is set to $k = 2$, and the confidence parameter $\alpha = 0.99$, for all the numerical results, to show that performance of the method is insensitive to the parameters k and α choice.

4.3.1 Double Well Potential

The single point side of view.

In this example, we will show the performance of each method, in recovering the true parameters for exact observations from a simple 1-dimensional system, with deviation of one single point from its true value, which represent one outlier point. We consider the equation

$$f(x) = x^4 - x^2. \tag{4.13}$$

and we sample 61 equally spaced measurements for $x \in [-1.2, 1.2]$, and we construct Φ using the 10^{th} order polynomial expansion with $K = 11$ is the number of candidate functions. Then, we consider a single fixed value corrupted measurement to be $f(0.52) = 0.5$.

In this example, we see that the true solution will have a residual δ equal to outliers deviation from its true position,

$$\delta = \sqrt{(f(0.52) - 0.5)^2} = 0.6973 \tag{4.14}$$

Fig. 4.2 shows the result for LS. The LS with its BLUE property (Best Linear Unbiased Estimator), succeed to minimize the residual to have better fitting residual than the true solution, but it is clear that the residual value does not reflect reliable solution. Practically, when the true solution gives a fitting residual δ, then any other solution deviates in its residual from δ will have a reduction in the solution accuracy, no matter the direction of deviation from δ. In Fig. 4.3, we see the result of OLS. We see that the results with the best residual of OLS is almost identical to LS result. Here it worth to say a detailed review for the 1000 OLS solutions under different threshold showed us a small interval that gives solutions closer in structure to the true solution more than the minimum residual solution is shown, which is if treated with suitable trade-off strategy can give a better solution.

Fig. 4.4 shows the result for CS, where it failed to find any feasible solution for all values of $\epsilon < \delta$. Such outliers makes it hard to find a parameter vector $\boldsymbol{\beta}$ that can fit the data including the outliers point, and even with considering high resolution for ϵ span, so, CS as discussed before tends to select the solution with minimum $\|\boldsymbol{\beta}\|_1$ within the best feasible residuals. CS solution simulation for different outliers values is provided on our YouTube channel here. Fig. 4.5 shows the result for LASSO, and it shows the sparse solution with wrong structure of

LASSO. We considered the bounds of λ to be $\lambda \in \left[\|\Phi\Phi^{\dagger}\boldsymbol{f} - \boldsymbol{f}\|, \|\boldsymbol{f}\|\right]$, where $\lambda = \|\Phi\Phi^{\dagger}\boldsymbol{f} - \boldsymbol{f}\|$ is the penalty on the solution with all entries are non-sparse and $\lambda = \|\boldsymbol{f}\|$ is the penalty on the solution with all entries are sparse. For this example, different from others (see Methods section), and because of its small dimensions, we considered very large span (1000 values) of the tununing parameter value for OLS, LASSO and CS to investigate the best expected outcomes of the methods.

Fig. 4.6 and Fig.4.7 shows the results for SINDy and TW, respectively, where we see that none of them was able to detect the correct sparse structure. Since different tuning parameters affect TW performance, and there is no unsupervised method to estimate these parameters, for the fair comparison, we evaluate TW performance for different values of the tuning parameters. Fig.4.8 shows the results of the TW method for different values of tuning parameters, and we found that only a very narrow choice of tuning parameters can result with accurate estimation. Selecting these correct parameters, practically, requires prior information about the true answer to judge the tuning parameters goodness.

Fig. 4.9 shows the accurate structure found by ER. Even with a slight difference in the magnitude of the parameters, we see how ER recovers the true basis functions. The residual of the ER was 0.865, which is higher more than most other methods, but the ER focuses on the information flow between the basis and dynamic and not the residual of solution magnitudes.

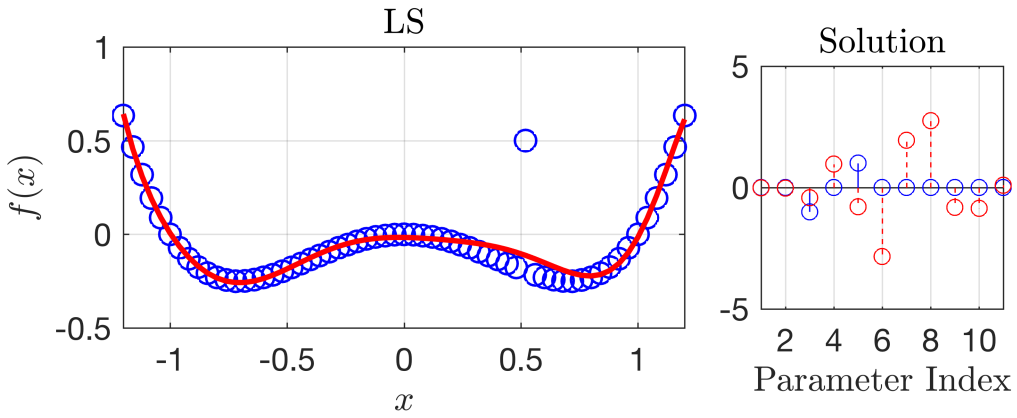

FIGURE 4.2: The LS solution for the data given by Eq. 4.13. This result shows how the LS invest in all available parameters to reach the best possible fitting. In fact, the residual of the least square solution was lower than the residual of the true solution, $0.6535 = \|\Phi\Phi^{\dagger}\boldsymbol{f} - \boldsymbol{f}\| < \|\Phi\boldsymbol{\beta}_{\boldsymbol{true}} - \boldsymbol{f}\| = 0.6973$, and in sparse regression literature, this initiate the need for developing trade off algorithms that considers different measures such as $\|\boldsymbol{\beta}\|_1$ and $\|\boldsymbol{\beta}\|_0$.

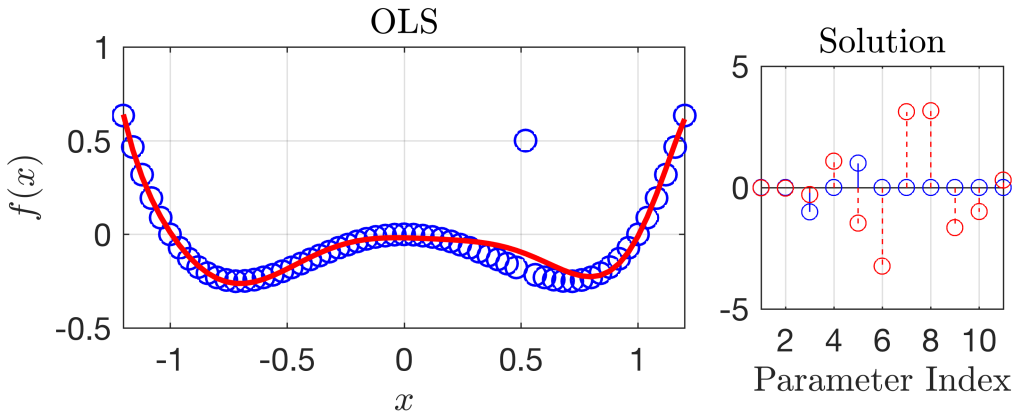

FIGURE 4.3: The OLS solution with 1000 log-spaced span for the threshold value $\epsilon \in [10^{-6}, 10^{2}]$. We see that the OLS failed to find solution better than the LS and they are almost identical.

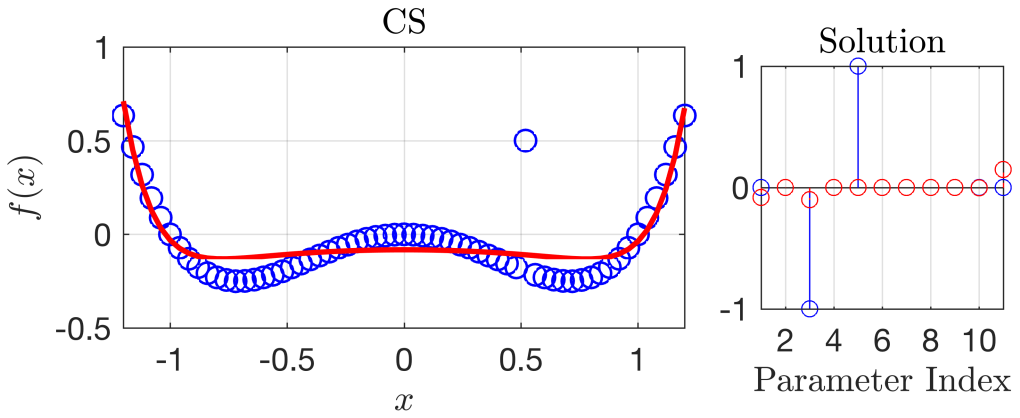

FIGURE 4.4: The CS solution, with 1000 log-spaced span for $\epsilon \in [10^{-6}, 10^{2}]$. The solution with minimum residual is shown to the right. As expected, the CVX solver failed to find any feasible solution for all values of $\epsilon < 0.69$, and that was the reason to consider 10^2 as the upper bound of epsilon although it represent a high value for tolerance.

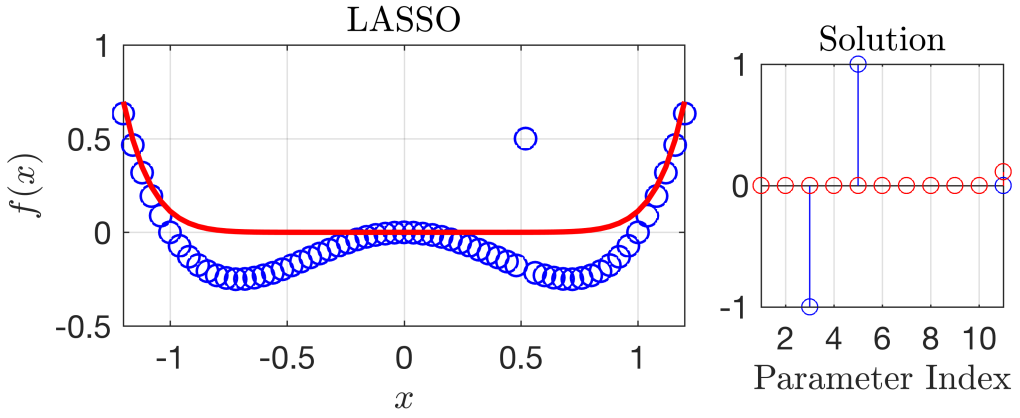

FIGURE 4.5: The LASSO solution, with 1000 equally-spaced span for $\lambda \in \left[\|\Phi\Phi^{\dagger}\boldsymbol{f} - \boldsymbol{f}\|, \|\boldsymbol{f}\|\right]$. The solution with minimum residual is shown to the right and it found at $\lambda = 0.818$.

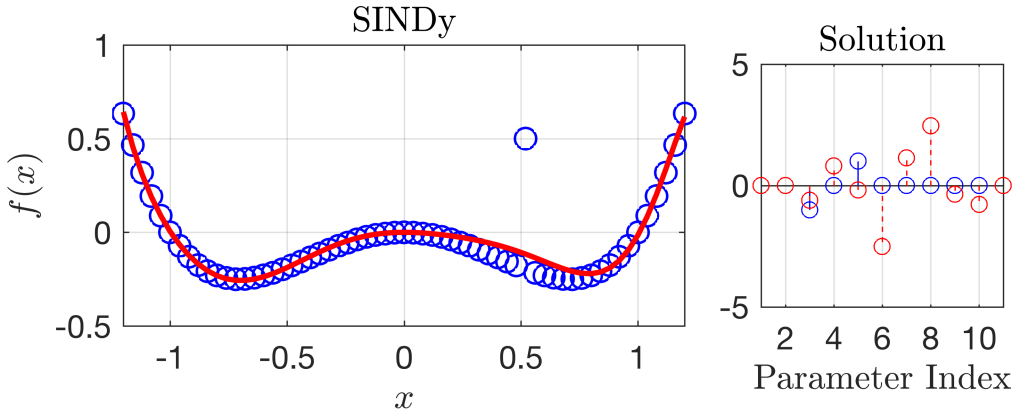

FIGURE 4.6: SINDy solution. We choose the threshold value of SINDy to be $\lambda = 0.42$, which is the optimal value (chosen manually since there is no unsupervised method for such choice) that prevent SINDy from oversparse the true parameters.

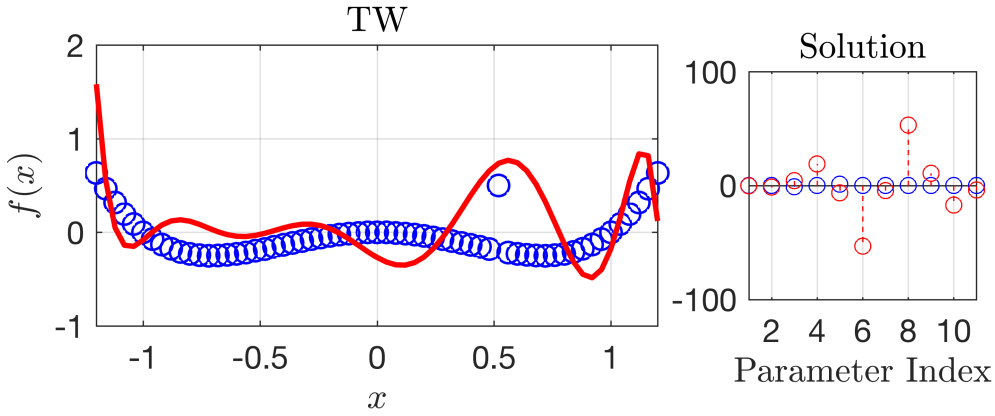

FIGURE 4.7: TW solution. Under the default values for TW method, $\mu = 0.0125$ and $\lambda = 0.1$, the results was very poor, and that was surprising since the problem setting match the exact assumptions of availability of "exact" measurement, and here we assume only one outlier point. So, in analogy to Fig. 4.11 and for the fair comparison, we explored TW results under varying tuning parameters and the results are shown in Fig. 4.8.

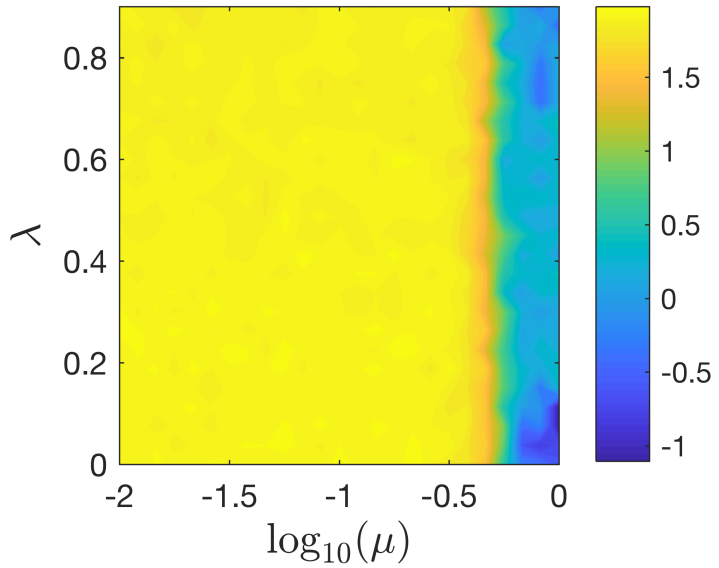

FIGURE 4.8: Double Well potential example. Error in recovered solution by TW under different values of λ and μ for the example shown in Fig.(4.7). Although the problem measurements are fixed, TW is also depended in random number generator seed, so, we averaged the results over 100 runs. We see that TW has overall failed in recovering the parameters. Although it has some degree of success in the very narrow lower-right corner with error = 0.1

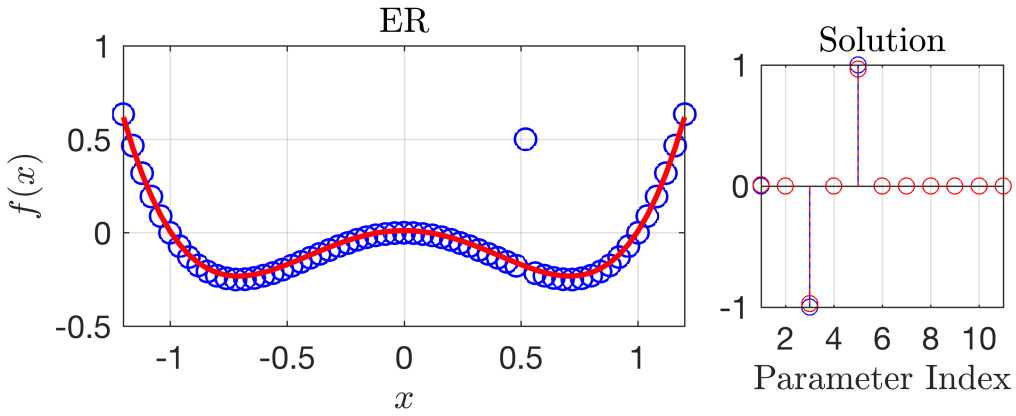

FIGURE 4.9: The ER solution. We see that ER recovered the true solution, No trade-off, No-tuning parameter and large span with expensive computations.

4.3.2 Lorenz system.

Our first detailed example data set was generated by noisy observations from a chaotic Lorenz system, which is represented by a three-dimensional ODE which is a prototype system as a minimal model for thermal convection obtained by a low-ordered modal truncation of the Saltzman PDE [164], and for many parameter combinations exhibits chaotic behavior [121]. In our standard notation, we have $\boldsymbol{z} = [z_1, z_2, z_3]^\top$ and

$$\begin{cases} \dot{z}_1 = F_1(\boldsymbol{z}) = \sigma(z_2 - z_1), \\ \dot{z}_2 = F_2(\boldsymbol{z}) = z_1(\rho - z_3) - z_2, \\ \dot{z}_3 = F_3(\boldsymbol{z}) = z_1 z_2 - \beta z_3, \end{cases}$$

with default parameter values $\sigma = 10$, $\rho = 28$ and $\beta = 8/3$ unless otherwise specified. We consider a standard polynomial basis as in Eq. (2.8).

Over the recent years, the Lorenz system has become a favorable and standard example for testing SID methods, which make it convenient to demonstrate several system identification misconceptions in recent literature that we have discussed theoretically in Chapter 2, and demonstrate numerically in Appendix B. In recent literature, the Lorenz system requires tens of thousands of measurements for accurate reconstruction, see Table B.1.

We compare several nonlinear SID methods in reconstructing the Lorenz system when the state observational noise is drawn independently from a Gaussian distribution, $\eta \sim \mathcal{N}(0, \epsilon^2)$. As we discussed before, this translates into effective noise that is not necessarily Gaussian or even independent due to the error amplification effect on the derivative.

Fig. 4.10 shows the error in the estimated parameters where,

$$error = \|\boldsymbol{\beta}^\star - \boldsymbol{\beta}\|_2 \tag{4.15}$$

where $\boldsymbol{\beta}^\star$ is the true parameters and $\boldsymbol{\beta}$ is the estimated parameters. As shown in Fig. 4.10, even with observational noise as low as $\epsilon = 10^{-4}$, ER and OLS outperform all other methods. In this low noise regime, SINDy required more measurements (around 4 times) to reach similar accuracy as ER.

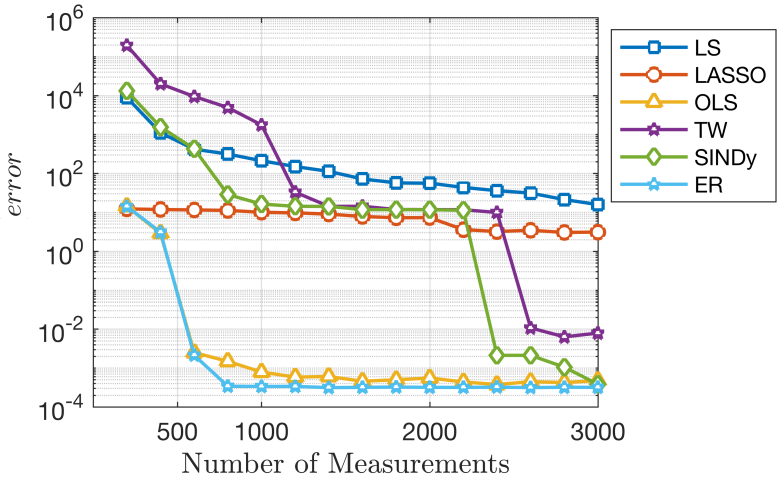

FIGURE 4.10: Lorenz system. We perform 100 runs for the comparison, no outliers, 0.0005 step size, and we considered the median result out of 100 runs. The figure shows the error in the parameter estimation for a Lorenz system but subject to noisy measurements by Gaussian noise, with $\epsilon = 10^{-4}$, and using a 5^{th}-order polynomial expansion. We see that ER and OLS has an overall superior performance compared to other standard methods. We see that SINDy and TW are less successful (under a large span of tuning parameters, see Fig.4.11) at this number of measurements even with low noise levels. Box plot for all the runs at 1500 measurements is shown in Fig.4.12

In comparison, as noted in [30, 186] and the implementation provided by the authors, for SINDy and TW methods to yield accurate reconstruction the number of measurements is at the order of 10^4. Additionally, it is worth pointing out that in the TW method there are two (free) parameters, λ and μ as discussed in Chapter 2, and their impact on the reconstruction quality is further evaluated and reported in Fig. (4.11).

Next, to explore the performance of SID methods under the presence of outliers, we conduct additional numerical experiments. The extent to which outliers present is controlled by a single parameter p: each observation is subject to an added noise $\eta = \eta^{noise} + \eta^{outliers}$, where $\eta^{noise} \sim \mathcal{N}(0, \epsilon_1^2)$ with probability 1 and $\eta^{outliers} \sim \mathcal{N}(0, \epsilon_2^2)$ with probability p. Then the measurements are given by

$$\begin{aligned} \boldsymbol{z}_i &= \boldsymbol{z}_i^\star + \eta_i^{noise} + \eta_i^{outliers} Ber(p), \\ &= \boldsymbol{z}_i^\star + \eta_i, \end{aligned} \tag{4.16}$$

for all $i = 1, \ldots, N$. Fig.4.13 shows the effect of error amplification on derivative estimation.

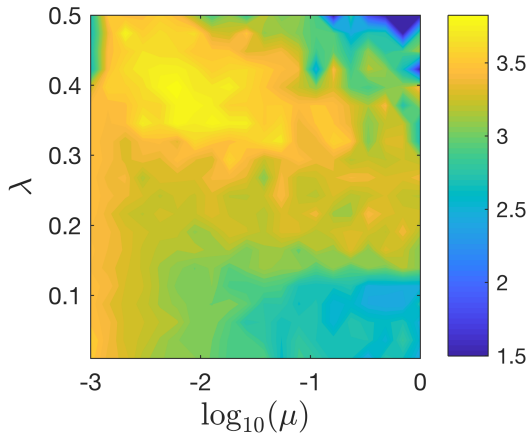

FIGURE 4.11: Contour plot of the error in the recovered solution of Lorenz system (Fig.(4.10)) by TW method for a grid of μ and λ values and using 2000 measurements, 5^{th} order polynomial expansion, low noise with $\epsilon_1 = 10^{-4}$ and no corrupted data. The color bar indicates the value of $\log_{10}(error)$ in the recovered solution, and it shows a large error at all levels of tuning parameters, meaning that no tuning parameter may improve performance of TW at this level of measurements. And this result applied to SINDy too since TW uses SINDy as the sparse recovery method.

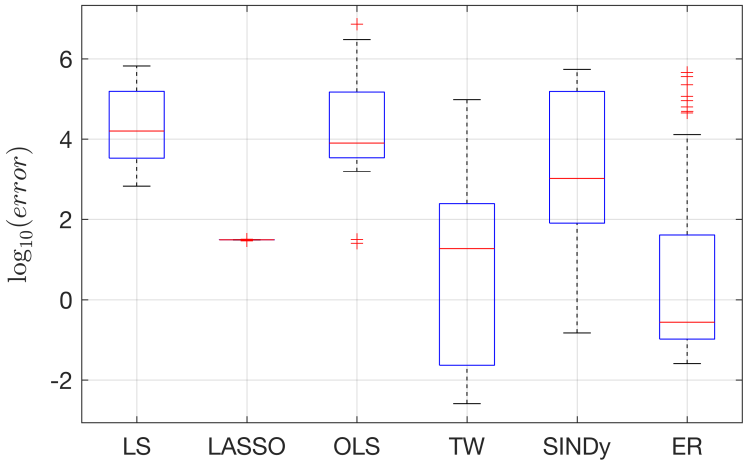

FIGURE 4.12: Boxplot for Lorenz. Referes to main text Fig.(4.14) at 1500 measurements, this figure shows the results of the all 100 runs.

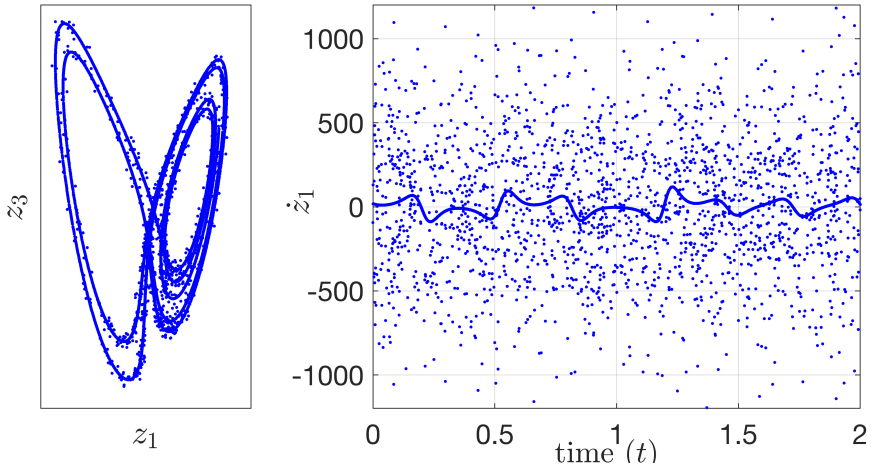

FIGURE 4.13: Error amplification and outliers. (Left) Sampled data $\boldsymbol{z}$, from Lorenz system with step size $\tau = 0.0005$, $N = 10^4$, and with the presence of noise and outliers according to Eq.4.16 with $\epsilon_1 = 10^{-5}$, $\epsilon_2 = 0.2$, and $p = 0.1$. (Right) Estimated derivative $\boldsymbol{f_1}$ (only $\dot{\boldsymbol{z}}_1$ direction for clear view) using central difference method.

For methods comparison, we use $\epsilon_1 = 10^{-5}$, $\epsilon_2 = 0.2$ and $p = 0.2$. The results of SID are shown in Fig. 4.14.

Compared to Fig.(4.10), we see that with $p > 0$ OLS performance drops due to the increasing occurrence of significant noise and outliers whereas ER remains its capacity of accurately identifying the underlying system. As an example, in each of the side panels of Fig. 4.14, we show the trajectory of the identified dynamics using the median solution of each method. It is clear that under such noisy, chaotic dynamics and at a relatively under-sampled regime, ER method successfully recovers the system dynamically. As an ample amount of data becomes available, we note that the TW method starts to produce excellent reconstruction, which is consistent with recent findings reported in Ref. [186].

Given that a major theme of modern SID is to seek for *sparse* representations, and the Lorenz system under standard polynomial basis is indeed sparse, it is worth asking: what are the respective structure identified by the different methods? In Fig. 4.16, we compare the structure of the identified model using different methods across a range of parameter values for ρ. In this case, under the presence of large noise and outliers ($p = 0.2$), none of the methods examined here, including recently proposed sparsity-promoting (CS, SINDy) and outlier-resilient (TW) methods, can identify the correct structure. The proposed ER method, however, does determine the correct structure. It is worth pointing out that, often when

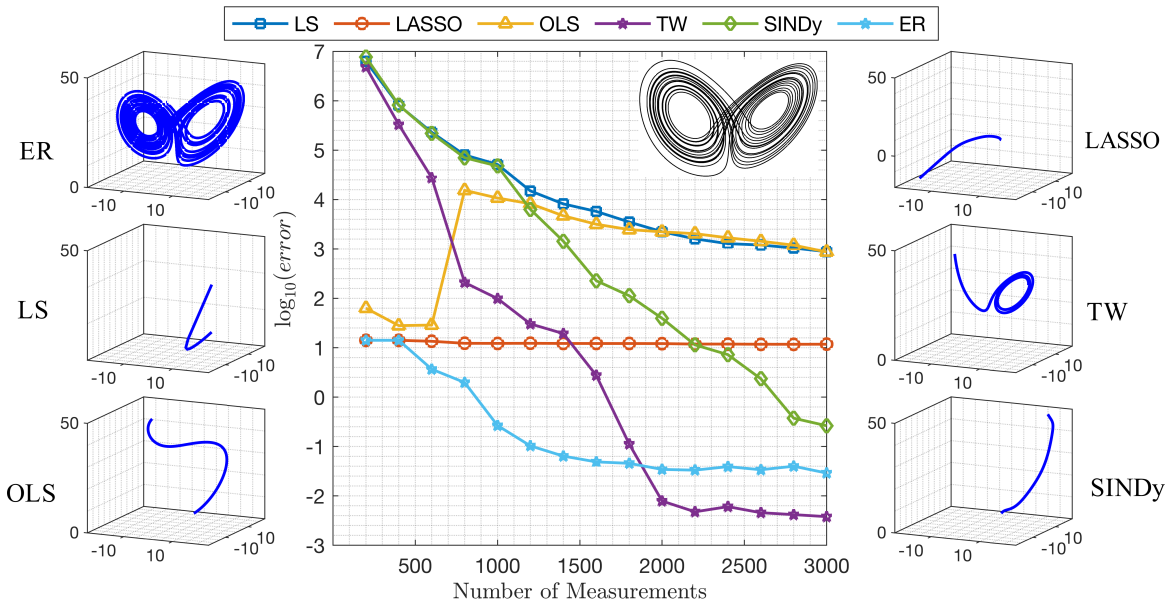

FIGURE 4.14: SID for the Lorenz system when outliers corrupt the observations. Contrast to Fig. 4.10. As before, we specify a level of persistent Gaussian observation noise following Eq.4.16. **(Middle)** Error in estimated parameters for Lorenz system given in Eq. 4.15 with noise, $\epsilon_1 = 10^{-5}$, $\epsilon_2 = 0.2$, 5^{th}-Order polynomial expansion, and $p = 0.2$. Lorenz system dynamics is shown in the upper right corner. We see that ER has fast convergence at a low number of measurements, followed by TW, which required twice number of measurements. Different from TW, in our ER method we focus in detecting the true sparse structure with the presence of outliers, without any attempts to neglect outliers based on some weight function to achieve higher accuracy which is the case in TW method. This point clearly appears in Fig.(4.15 where we see that although TW achieved higher accuracy, it has low exact recovery probability, while ER reached exact recovery probability more than 90%. A detailed statistics box-plot (quartiles, median,...,etc.) over the 100 runs with 1500 measurements is shown in Fig. (4.12). **(Side panels)** Typical trajectories generated by the reconstructed dynamical systems, where for each method we show results using the "median" solution, that is, recovered system whose corresponding parameter estimation error is at the median over a large number of independent simulation runs. In each such simulation, 1500 samples are used. Comparing with the true Lorenz attractor (upper right corner in the main panel), we see that the only reasonable reconstruction, in this case, was produced by ER.

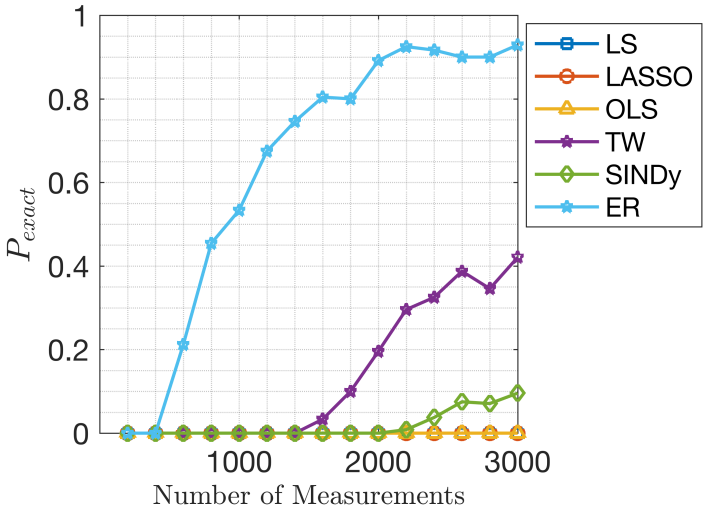

FIGURE 4.15: Probability of exact recovery for Lorenz system. For the same results shown in Fig.(4.14), *Pexact* here represent the number of runs in which a method recovered the exact sparse structure over the total number of runs. We see that although TW reached high accuracy at a high number of measurements, its exact recovery probability remains low.

expressed in the right basis, a model will appear to be sparse, the converse is not true: just because a method returns a sparse solution does not suggest (at all) that such a solution gives a reasonable approximation of the true model structure.

Moreover, one more interesting observation on the results (obtained with the presence of outliers and shown in Fig.4.14) is the ability of the solution to reproduce the dynamic, since it is the main objective of mathematical modeling.

Consider the results in Fig.4.14, at 2000 measurements, which shows that TW has a more accurate solution (in the magnitude of the parameters) while it miss-detect the correct sparse structure of the solution as shown by the probability of exact recovery in Fig.4.15. In Fig.4.17 we show the 2000 measurements training data, and starting from the same initial condition as the training data, we reproduce the dynamic iteratively using the median solution obtained by ER and TW. We see that TW miss-detect the overall dynamic, and Fig.4.18 and Fig.4.19, we show more examples for the same phenomena.

This observation will be a subject for future research, as primary comments, we can see in Fig.4.18 and Fig.4.19 a connection between the cases where the TW solution with more accurate parameters and few wrong features was not able to reproduce the dynamic. Since in

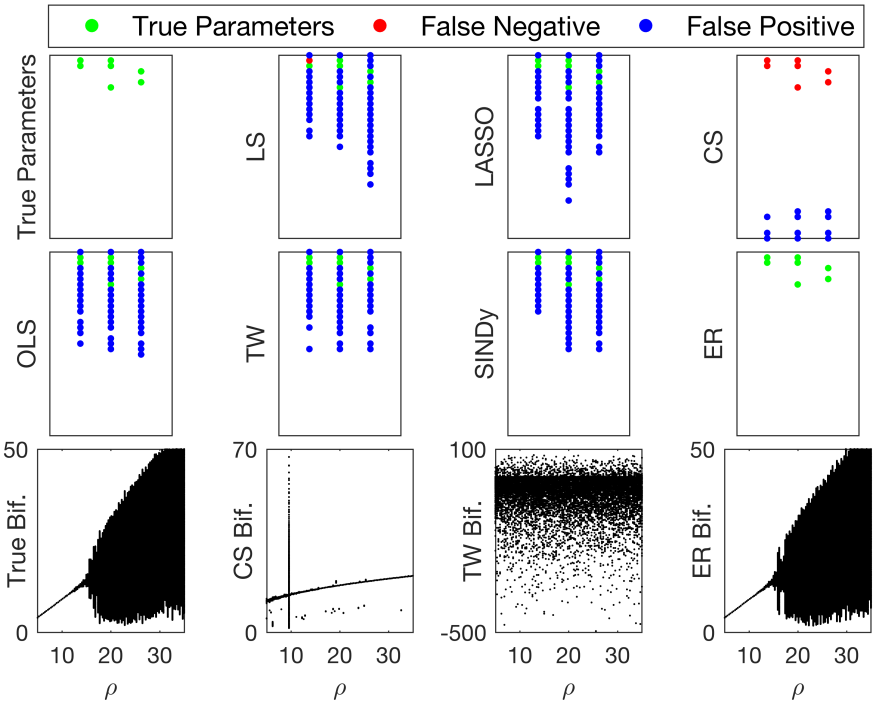

FIGURE 4.16: Sparse representation of the solution found by solvers using 1500 measurements, and $p = 0.2$ on Fig.(4.14). The upper left corner shows the true solution of Lorenz system. The bottom column shows the bifurcation diagram on z dimension of Lorenz system with $\rho \in [5, 30]$ as bifurcation parameter, created using 5000 initial conditions evolved according the recovered solution.

all these cases, we see that the training data is collected from a small fraction of the attractor, and it does not include all possible oscillations of the dynamic.

We will investigate this observation in more details in our future work with the following questions in mind:

- From bifurcation diagram, we see that the **long time** behavior of some chaotic systems such as Lorenz system is not "very" sensitive to the bifurcation parameters (i.e., $\rho = 30$ instead of $\rho = 28$ in Lorenz system will produce the same dynamic on the long run although the trajectories will be different). However, is it possible that the dynamic will be "very" sensitive to including wrong features? (i.e., with the standard Lorenz parameters, does including a term such as $0.00001z_3^3$ to any dimension of the system can produce utterly different dynamic?).
- If the answer for the previous point is yes, then how to find these great influence features and how to construct a model validation technique that avoids such undesirable long term influence.
- What is the minimum fraction of the overall the dynamic that we need to observe not to misunderstand the dynamic?.

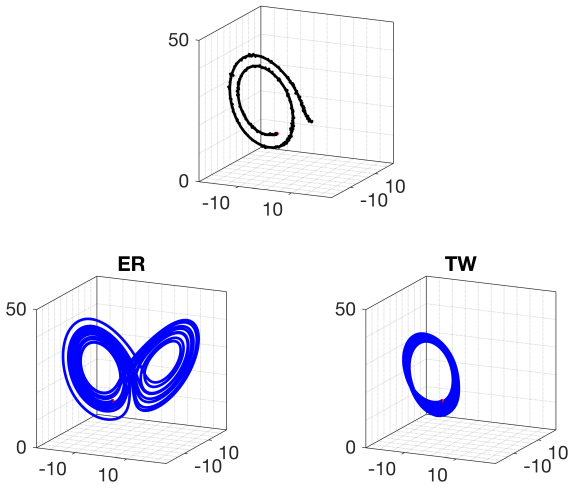

FIGURE 4.17: Re-produced dynamic of Lorenz system. (Top) Training data (2000 measurements) in black dots and the initial condition in the red dot. (Bottom) Re-produced dynamic (10000 iterations) by ER and TW starting from the same initial condition. We see that the training data collected from one side of the attractor, and TW reproduced dynamic converged to a periodic dynamic (limit cycle).

4.3.3 Network coupled logistic maps.

Our third example is a network of coupled logistic maps which is typical of either coupled map lattices [104], but also cellular automata [107] and more generally the scenario of high dimensional and complex systems that have become the thrust of recent analysis included in the synchronization literature [7, 130]. In this example, we assume that not only the governing dynamics are unknown, but so is the structure of the network that moderates the coupling between individual chaotic elements; both of these must be (simultaneously) identified from observed dynamic data alone. In Fig. 4.20, we compare the results of several system identification methods, including the proposed ER approach. We now offer here a rough description of the dramatic difference in performance. In particular, the improvement on systems with noisy data subject to outliers; a more detailed mathematical analysis will be the subject of our future work.

Consider that each of the other methods we reviewed involves minimizing a functional $J(\mathbf{a})$ of the data a, and that when $\mathbf{a}$ is subject to noise, that the functionals are each continuous

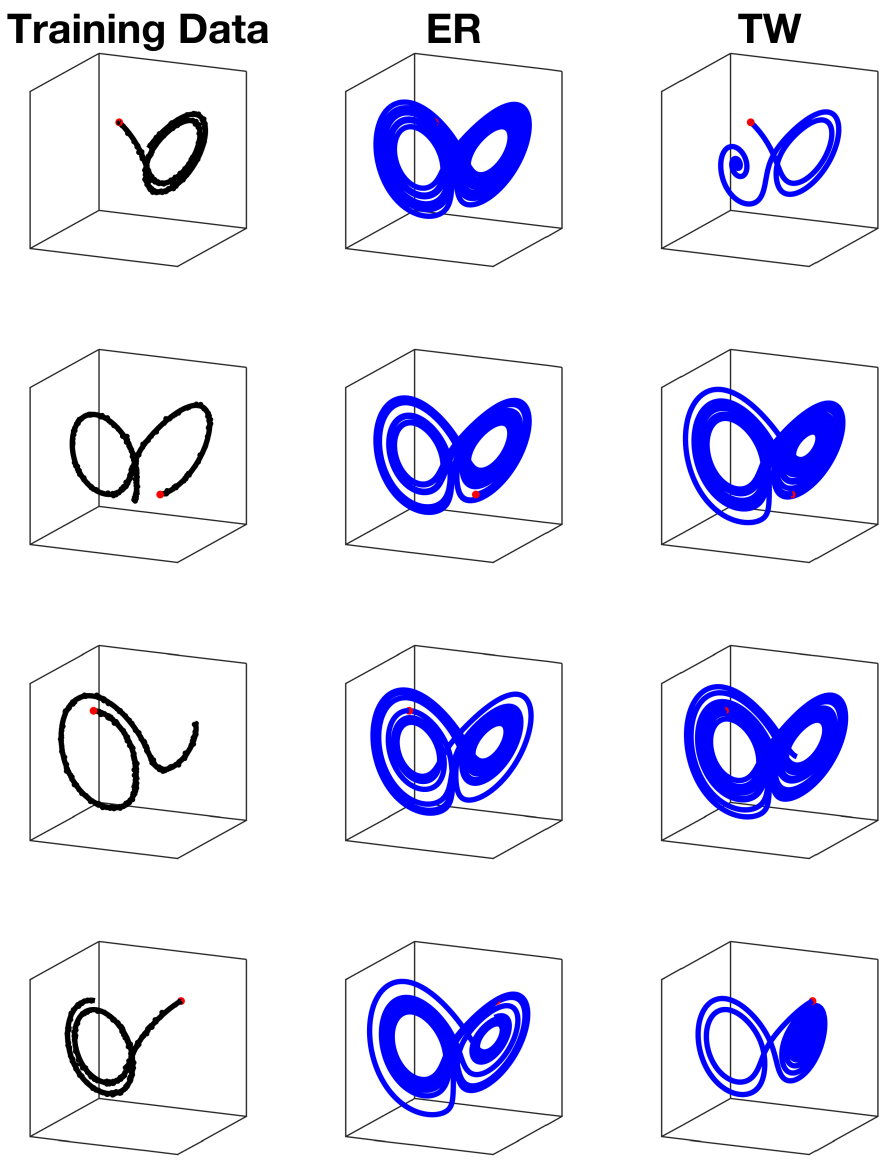

FIGURE 4.18: Re-produce Lorenz system dynamics. In analogy to Fig.4.17, this figure shows the re-produced dynamic for selected runs. In the top and bottom rows, TW solution converged to a fixed point. In second and third rows, where the training data spans both sides of the attractor, TW re-produced the dynamic accurately.

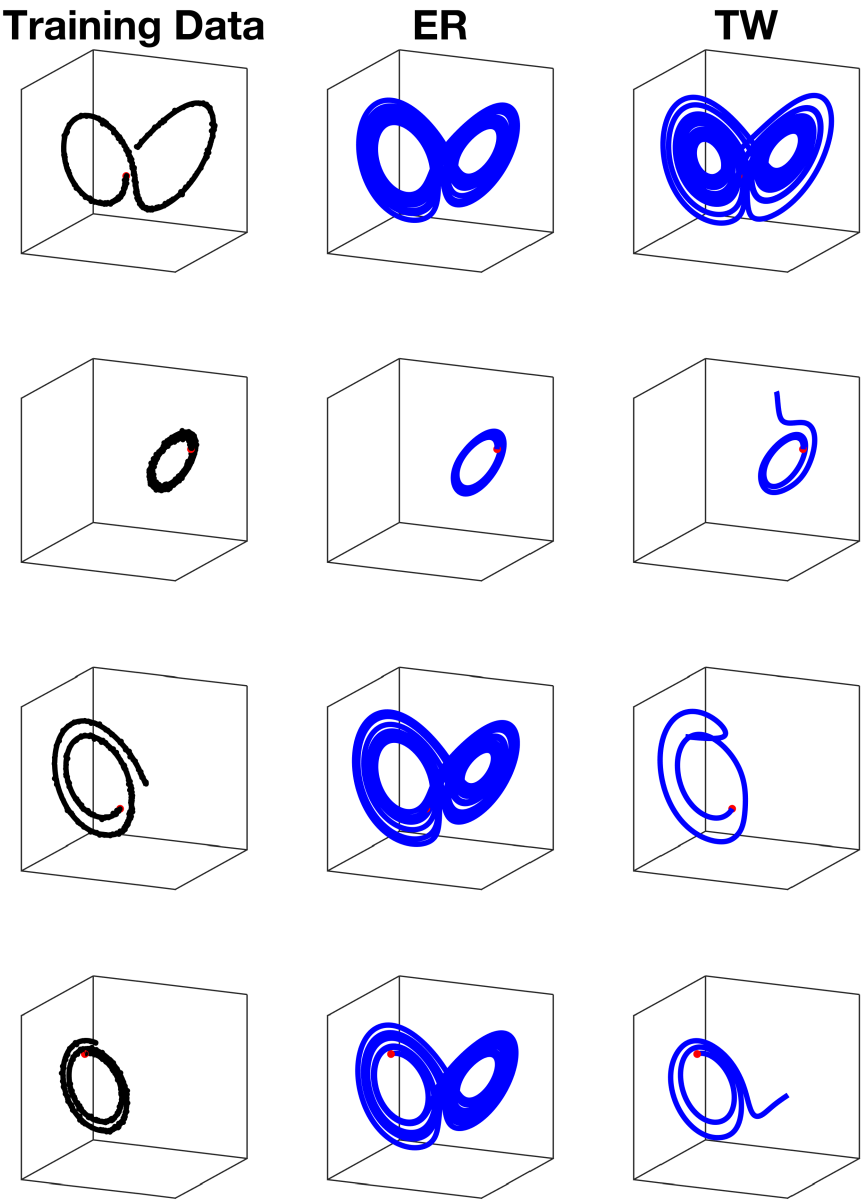

FIGURE 4.19: Re-produce Lorenz system dynamics. In analogy to Fig.4.17, this figure shows the re-produced dynamic for selected runs. In the second row, where the training data span very small fraction of the attractor, ER converged to periodic dynamic, and TW diverge completely.

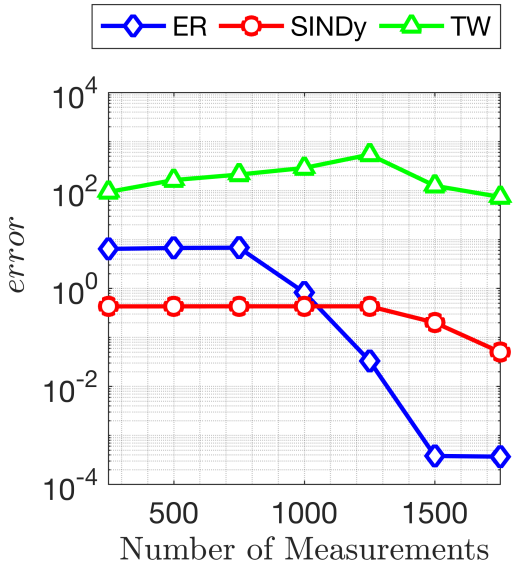

FIGURE 4.20: Coupled Logistic map example. The error in recovered parameters with noise $\epsilon = 10^{-3}$, second-order expansion. As discussed in the Methods section in our main text, we see that TW could not conserve SINDy error level and it diverge to higher error levels until SINDy starts to slightly converge (but still with high error) to the solution with 1500 measurements. While we see that 1500 measurements were enough for the ER to recover the exact sparse structure with high accuracy.

with respect to their argument. We assume that the underlying system is,

$$f(x) = ax(1-x), \tag{4.17}$$

describing the individual elements as Logistic maps, the coupled network of N such oscillators is of the form,

$$F(x_i) = f(x_i) + \sum_{j=1}^{N} A_{ij} W_{ij} \left(f(x_j) - f(x_i) \right) \tag{4.18}$$

where $i, j = 1, ..., N$, A is the adjacency matrix of the coupled network, W_{ij} is the coupling strength between the nodes i and j, and $f(x_i)$ is the image of the point x_i under the logistic map given in Eq.4.17.

To present a specific example, let $N = 50$, we construct the adjacency matrix A to have simple coupling such that:

$$1 < D_{ii} \leq 4 \tag{4.19}$$

Where D is the degree matrix of A, and the coupling adjacency matrix A constructed randomly

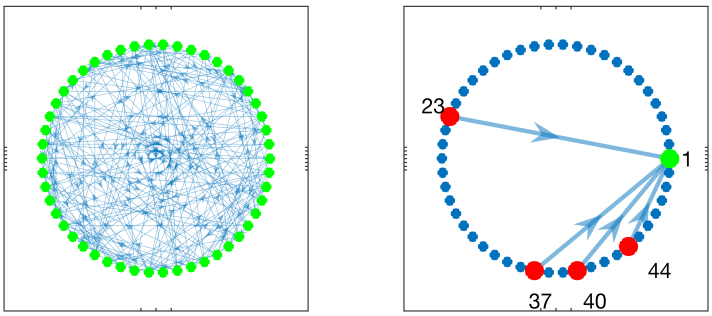

FIGURE 4.21: Graph representation of the coupled network of logistic map example. (Left) The 50-nodes network in the directed graph representation. (Right) For a selected node, we see that it is basically influenced by few other nodes.

such that the above inequality holds. Fig. (4.21) show the graph of the coupled network. Then if we consider only the second order expansion (where the basis matrix Φ is the second-order expansion of the 50 time-series of all nodes), we will have 1326 terms in our expansion matrix. We focus on this example on solving the underdetermined system by considering 2000 measurements as maximum available measurements. So, exclude OLS which only solve overdetermined systems and cannot be investigated at a number of measurements less than 1326, and we exclude LASSO and CS for their high computation complexity. Fig. 4.20 shows the error in recovered parameters for this example. For simplicity and the computation complexity, we performed the experiment to find the parameters for one single dimension, and results are averaged over 50 runs.

This example shows the robustness of the ER in recovering the coupling structure in complex coupled networks. The computations complexity in such problem can be highly reduced by considering basic and trivial assumptions. For example, we can consider each node N_i as a default influence source for itself, and then instead of starting the forward step in ER from the empty set, we may initialize the index set with the terms that purely includes N_i.

Fig.(4.22) shows the sparse representation of the Logistic map discussed with number of nodes $N = 20$.

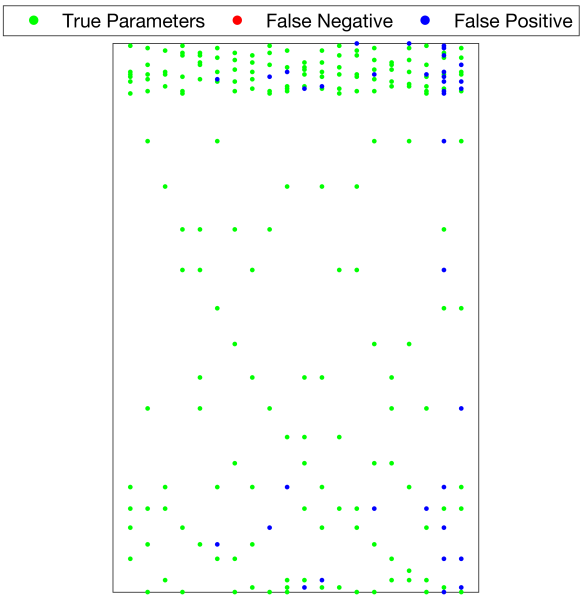

FIGURE 4.22: ER solution sparse representation for the coupled Logistic map created by Eqs. (4.17 - 4.19), with $\epsilon_1 = 0.001$, $\epsilon_2 = 0$, and using 2000 measurements and number of nodes $N = 20$. The true solution contained 192 non-zero entries (out of 4620, the total number of parameters) and all of them detected accurately (green dots) with Zero false negative rate, and we see that there was few false positives in ER solution which have 226 total non-zero entries.

4.3.4 Kuramoto-Sivashinsky equations.

To further demonstrate the power of ER, we consider a nonlinear PDE, namely the Kuramoto-Sivashinsky (KS) equation [94, 114, 115, 117, 173], which arises as a description of flame front flutter of gas burning in a cylindrically symmetric burner. It has become a popular example of a PDE that exhibits chaotic behavior, in particular spatiotemporal chaos [45, 92]. We will consider Kuramoto-Sivashinsky system in the following form,

$$u_t = -\nu u_{xxxx} - u_{xx} + 2uu_x, \qquad (t, x) \in [0, \infty) \times (0, L) \tag{4.20}$$

in periodic domain, $u(t, x) = u(t, x + L)$, and we restrict our solution to the subspace of odd solutions $u(t, -x) = -u(t, x)$. The viscosity parameter ν controls the suppression of solutions with fast spatial variations, and is set to $\nu = 0.029910$ under which the system exhibit chaotic behavior [45].

Since a PDE corresponds to an infinite-dimensional dynamical system, in practice, we focus on an approximate finite-dimensional representation of the system, for example, by Galerkin-projection onto basis functions as infinitely many ODE's in the corresponding Banach space.

To develop the Galerkin projection, we follow the procedure as presented in [8], to expand a periodic solution $u(x, t)$ using a discrete spatial Fourier series,

$$u(x, t) = \sum_{-\infty}^{\infty} b_k(t) e^{ikqx}, \qquad \text{where } q = \frac{2\pi}{L}. \tag{4.21}$$

Notice that we have written this Fourier series of basis elements e^{ikqx} in terms of time varying combinations of basis elements. For simplicity, consider $L = 2\pi$, then $q = 1$ for the following analysis. This is typical [160] with the representation of a PDE as infinitely many ODE's in the Banach space, where orbits of these coefficients therefore become time varying patterns by Eq. (4.21). Substituting Eq. (4.21) into Eq. (4.20), we produce the infinitely many evolution equations for the Fourier coefficients,

$$\dot{b}_k = (k^2 - \nu k^4) b_k + ik \sum_{m=-\infty}^{\infty} b_m b_{k-m} \tag{4.22}$$

In general, the coefficients b_k are complex functions of time t. However, by symmetry, we

can reduce to a subspace by considering the special symmetry case that b_k is pure imaginary, $b_k = ia_k$ and $a_k \in \mathbb{R}$. Then,

$$\dot{a}_k = (k^2 - \nu k^4)a_k - k \sum_{m=-\infty}^{\infty} a_m a_{k-m}. \tag{4.23}$$

where $k = 1, .., N_m$. However, the assumption that there is a slow manifold (slow modes as an inertial manifold [99, 100, 155, 160]) suggests the practical matter that a finite truncation of the series Eq. (4.21), and correspondingly a reduction to finitely many ODEs will suffice. Therefore we choose a sufficiently large number of modes N_m. Then we solve the resulting N_m-dimensional ODE (4.23) to produce the estimated solution of $u(x,t)$ by (4.21), and use such data for the purpose of SID, have meaning to estimate the structure and parameters of the ODE model (4.23).

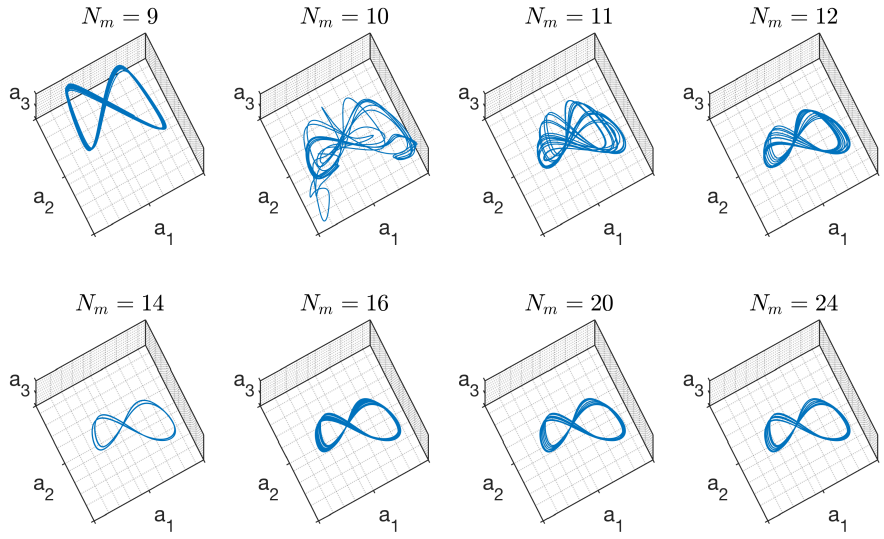

FIGURE 4.23: The first three modes of the ODE Eq.(4.23) solution. We show the modes a_1, a_2 and a_3 for selected number of modes. For clear view, we fixed the axis limits to be $a_1 \in [-1.21, 1.06], a_2 \in [-0.75, 0.98]$ and $a_3 \in [-1.1, 1.12]$ for all plots. We found that there was no significant addition to the dynamic with $16 < N_m$. (meaning that $N_m = 16$ was enough to describe the system).

Fig. 4.23 shows the first three dimensions plot under different number of modes. We see that using just a few numbers of modes ($N_m = 8, ..., 11$) is insufficient to capture the true dynamical behavior of the system whereas too large a number of modes ($N_m = 20, 24$) may be unnecessary. In this example, an adequate but not excessive number of modes seems to be

around $N_m = 16$, as no significant information is gained by increasing N_m.

Fig. 4.24 shows the sparse structure of the recovered solution by different methods. Here we mention that the true non-zero parameters of KSE using $N_m = 16$ are 200 parameters that vary in the magnitude from 0.9701 to 1705. With the second order expansion, our basis matrix will have 153 candidate functions, and it will be nearly singular with condition number 4×10^7. Likely due to such high condition number, neither TW nor SINDy gives reasonable reconstruction. In particular, we note that the solution of SINDy is already optimized by selecting the threshold value λ that is slightly above λ_* where here $\lambda_* \approx 0.1731$ is the smallest magnitude of the true nonzero parameter of the full least squares solution. A larger value of λ only worsens the reconstruction, as we found numerically.

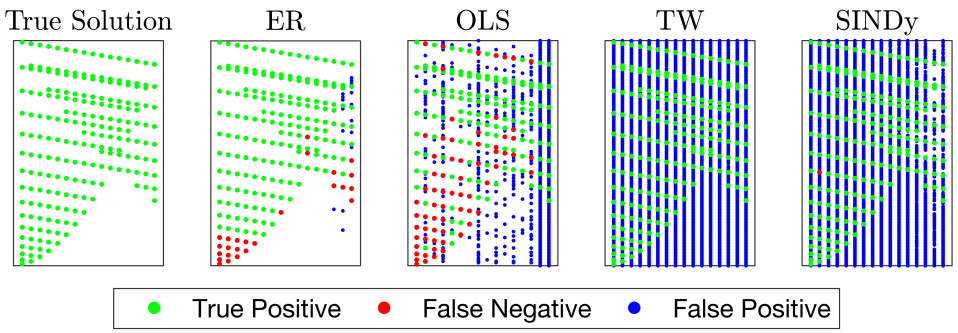

FIGURE 4.24: In analogy to Fig. 4.16, sparse representation of KSE solution by different methods. CS, LASSO have been excluded for their high computation complexity.

The OLS method overcomes the disadvantage of LS by iteratively finding the most relevant "feature" variables, where relevance is measured in terms of (squared) model error; but it comes at a price: similar to LS, the OLS is sensitive to outliers in the data, and such sensitivity seems to be even more amplified due to the smaller number of terms typically included in OLS as compared to LS, which cause the high false negative rate in the OLS solution. Although ER solution has few false negatives, but was completely able to recover the overall dynamic of the system as shown in Fig.(4.25), while all other solutions diverge and failed to recover $u(x, t)$.

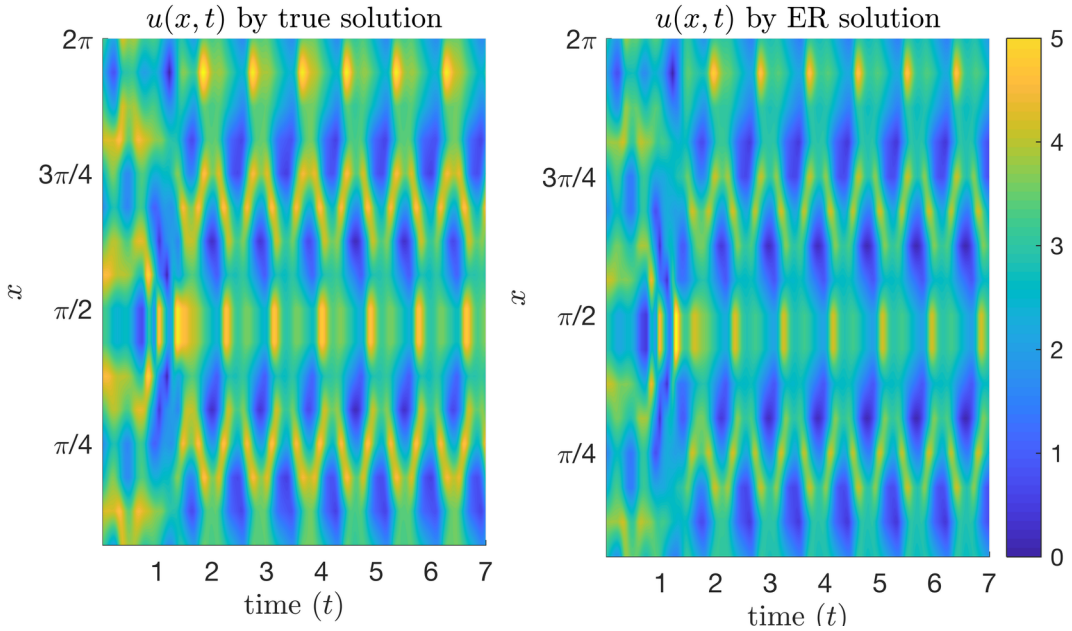

FIGURE 4.25: $u(x,t)$ constructed by the true solution (left) and the ER solution (right) using Eq.(4.21). OLS and TW were not able to reproduce the dynamic, and they diverge after a few iterations. We see that the reconstructed dynamic using ER solution is identical to the true solution with a minor difference in the transient time, although there was a false negative in the ER solution. ER detected the stiff parameters that dominate the overall dynamic. The sloppiness of some KSE parameters make there influence practically negligible to the overall dynamic.

4.4 Limitations

Our Entropic Regression Method shows superior performance over the current state of the art approaches, and we showed its performance for systems that range from the simple one-dimensional equation to a large network of coupled oscillators. However, there is still a challenging problem for ER, as well as all system identification method:

1. High sloppiness parameters: we showed in KSE example that ER was able to detect all the parameters with considerable influence to the dynamic, including a large part of the sloppy parameters. However, the parameters with a high degree of sloppiness (large change in parameters cause a micro scale change in the dynamic) were not detected.

2. Big Data: Thinking of a system with 10^6 observations and 10^4 candidate functions, such matrix requires around 80GB of storage space, and it becomes very hard and costly to deal with such data. Practically, if we consider a coupled network of 10^4 nodes of three-dimensional oscillators, the basis matrix can exceed the dimensions above. It is our focus in the future stage of our work to design algorithms that can combine the ER with the efficient Big Data computations.

4.5 Advantages and Future Directions

It is clear from the examples discussed on our numerical results that the ER has the following main advantages:

1. ER selects the basis based on their relevance in information theoretic sense, which makes ER robust to noise and outliers in the observed data.

2. ER investigates this relevance locally using a greedy algorithm, which makes ER robust against the numerical stability of the basis matrix.

3. In Algorithm 2-Line 6, and Algorithm 3-Line 6, which are the main loops to compute the conditional mutual information, we see that they are independent iteration loops, which means they are completely parallelizable loops. Moreover, the value of $\boldsymbol{f}$ and Φ are constant, and the only change is the index k, which means we only need to pass a vector of index pointers to the program function, which makes the algorithm

consumes no extra memory during computation. Parallel computations and efficient memory allocation can be a great advantage when dealing with large systems.

In Chapter 6, we discuss our future direction, which is based on promising primary results achieved during our research. Some of our future directions, but not limited to:

1. Large Networks of Non-Identical Oscillators: After the success of ER in detecting the sparse structure in coupled networks (logistic map example), and complex high dimensional dynamic such as KSE, we concern on the problem of detecting the coupling structure (adjacency matrix) of large non-identical oscillators.

2. Learning PDEs: A spatiotemporal PDE model $u(t, \boldsymbol{x})$ is a linear combination of partial derivatives. In this sense, we have ongoing research in learning PDEs directly from data.

3. Efficient Basis Construction: By using the principle of causation entropy [180], it is possible to detect the relevant features and the optimal expansion order before constructing the basis matrix, which highly reduces the computations complexity and the uncertainty in regression process.

4. Dynamic Dictionary Learning (DDL): We also focus on developing an efficient method to construct a basis matrix that can handle different type of functions such as $x\sin(xy)$, $x^2ye^{-t^2}$, $xy\cos(x^2)$, and the most important, the fractional functions.

Chapter 5

Image-Observed Complex Systems

> The profound study of nature is the most fertile source of mathematical discovery.
>
> *Joseph Fourier*
> *(1768 - 1830)*

In the last decades, imaging science has had a significant rule and impact on many fields of science such as medical, biological, computer science, and algorithms.

The work presented in this chapter primarily follows our published paper: "*Go With the Flow, on Jupiter and Snow. Coherence from Model-Free Video Data Without Trajectories*" [5].

Acquisition of information about an object or phenomenon without making physical contact with the object is known as **Remote Sensing** [33], which is in contrast to on-site observation. Remote sensing is a significant source of information for many systems including geography, land surveying, and most Earth Science disciplines (i.e., hydrology, ecology, oceanography, glaciology, and geology) [33, 98], particularly from "cameras" on artificial satellites on space or otherwise on aircraft. It also has military, intelligence, commercial, economic, planning, and humanitarian applications [76]. Remote sensing can be even the *only* source of observations for other complex system fields such as astronomy. Fig. (5.1) shows the Cassini space probe, which was the fourth space probe to visit Saturn and the first to enter its orbit. It was named after astronomer Giovanni Cassini. The Cassini Spacecraft was equipped with many Optical Remote Sensing instruments such as the Composite Infrared Spectrometer (CIRS), Imaging Science Subsystem (ISS), Ultraviolet Imaging Spectrograph (UVIS), and Visible and Infrared

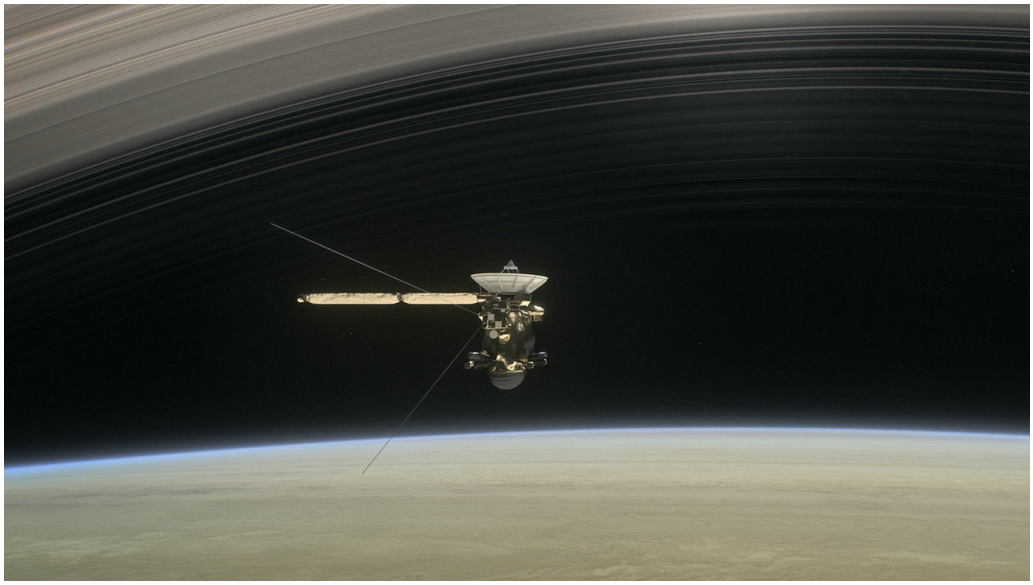

FIGURE 5.1: Cassini mission, 1997-2017. In the im from the short film Cassini's Grand Finale, the spacecraft is shown diving between Saturn and the planet's innermost ring. Credits: NASA/JPL-Caltech.

Mapping Spectrometer (VIMS), in addition to several particles and wave analysis instruments and microwave remote sensing instruments.

The images and data provided by Cassini played a significant role in understanding and analyzing planetary dynamics. For example, Cassini made its closest approach to Jupiter on December 30, 2000, and collected many scientific measurements. About 26,000 images of Jupiter, its faint rings, and its moons were taken during the six-month flyby. It produced the most detailed global color portrait of the planet yet, in which the smallest visible features are approximately 60 km across.

In this chapter, we discuss the basic image processing techniques used in dynamical system analysis and the reasons for their weakness in analyzing fluid dynamics.

5.1 Object Detection and Tracking

Object detection [171], is a computer technology related to image processing and computer vision that deals with detecting instances of objects of a particular class in digital images and videos. Researched domains of object detection include pedestrian detection and face detection.

Video tracking which is also known as **Object tracking** [60, 147], is the process of locating a moving object over time using a camera, and it has many applications such as human-computer interaction, security and surveillance, video communication and compression, augmented reality, traffic control, medical imaging, and video editing.

Before going further in defining image processing terms, here we present the question: ***What is an Object?!***.

Image processing science provides robust methods for detecting and tracking *rigid bodies*. One of the most popular principles for object detection and tracking is by applying features detection in images $I(t)$[1] at time t, and in the next image $I(t+\tau)$ at time $t+\tau$, and matching the equivalent features according to the selected measure of quality.

One of the most robust and frequently used methods of extracting features from images is the corner points method. Corner detection is an approach used within computer vision systems to extract certain kinds of features and infer the contents of an image and is commonly used in motion detection, image registration, video tracking. Fig. (5.2) the example of corner points features matching between two images.

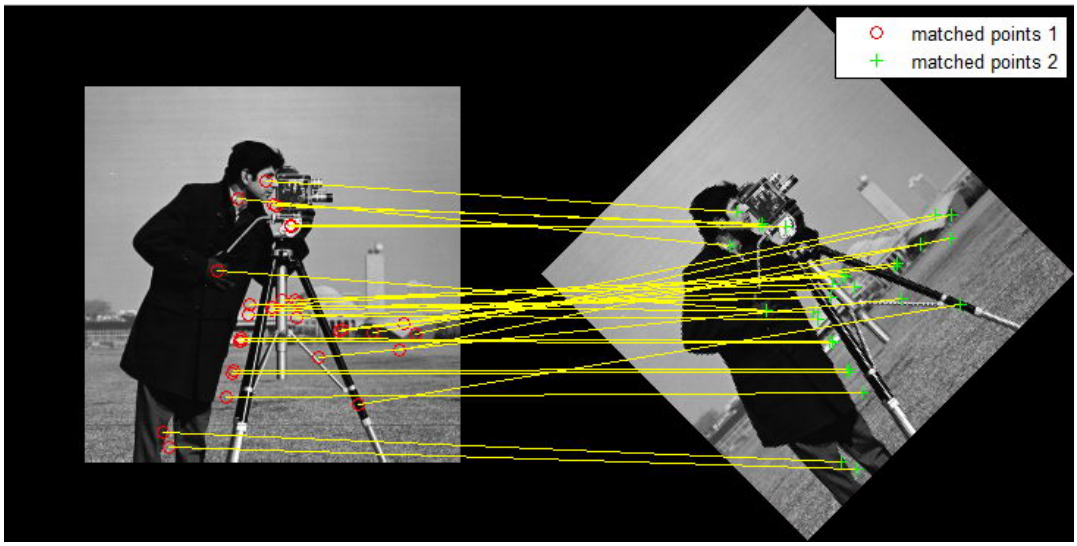

FIGURE 5.2: Cameraman image, with corner points matching. Source: Matlab documentation.

A corner can be defined as the intersection of two edges, and it also can be defined as a point for which there are two dominant and different edge directions in a local neighborhood of the point. Within the image, there will be some pixels that have a well-defined position and

[1]Note that in this chapter we use I notation to indicate an image.

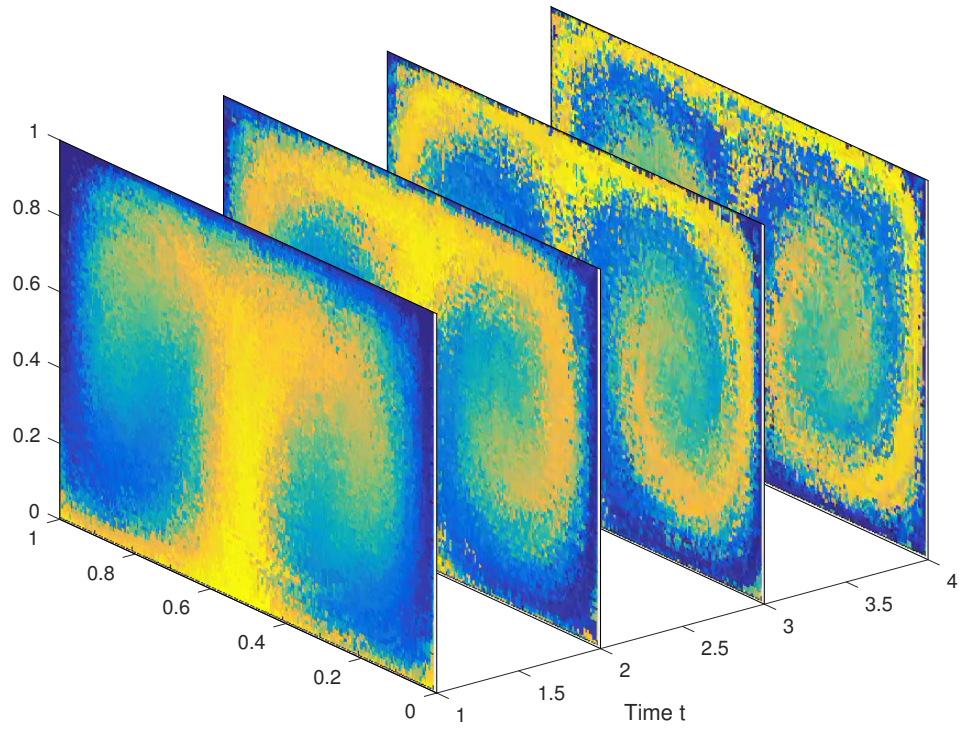

FIGURE 5.3: Double Gyre snapshots at different time steps.

can be robustly detected, such points are called interest points, and by matching two sets of interest points between two images, we can detect the translation and rotation that occur on the image.

Unfortunately, corner points and the general principle of features detection and tracking may completely fail when the object is not a rigid body. Moreover, observing fluid dynamics, we basically can not tell what is an object?!. Fig. (5.3) shows a set of movie frames for the double-gyre system, where in such dynamical system, there is no specific definition for an object, and there are no corner points that keep their shapes through time. Another example, if we look to the planet Jupiter as one object, then it is simple to detect and track it through images since it does not change its shape, and the overall view remains the same. But taking a closer view of the surface of Jupiter, which consisting of clouds of gases, we see a fluid dynamic consists of bands, tornadoes, and complex dynamics that continually changes shape on the "relativily" micro-scale.

On the other hand, it is important to distinguish between the concept of a feature that we

may notice in a single image and a feature that over time persists over several successive images. Persistence over time is more akin to coherency. So, we focus here on detecting such coherent structures which represent a persist feature in complex dynamic. This does not contradict with the rigid body view of an object since the rigid body moving according to a simple dynamic can be defined as a coherent structure.

5.2 Coherent Structures

In essentially all of the studies that have appeared in recent literature, no matter what the method, approach, or perspective, one starts with a dynamical system. From there follows the quantity to be analyzed. In other words, an underlying flow is assumed in the sense that generally a differential equation is required to proceed, whether explicitly or implicitly through observations of an experiment. For this, we will write,

$$\dot{\mathbf{x}} = \mathbf{F}(\mathbf{x}, t), \tag{5.1}$$

for a vector field, $\mathbf{F} : M \times \mathbb{R} \to M$, (typically $M \subset \mathbb{R}^2$ or perhaps $\mathbb{R}^3$), but this may be developed from a stream function from an underlying partial differential equation for example. In any case, then a flow mapping, $\mathbf{x}(t) = \varphi(\mathbf{x}_0, t_0, t)$ is inferred, even if this means numerical integration of the differential equation.

In recent work, aspects of advection and diffusion have been both involved in developing a better understanding of coherence [54, 72, 73], including for models of stochastic processes. We summarize that universally, previous work either begins with a model of the dynamical system, or at least attempted to empirically develop a model perhaps by optical flow [2, 14, 122], or similarly by other means [88], and recently by Koopman operator methods [1].

In contrast to all the mathematical formalism and machinery behind current studies of coherency, it can be said that people "recognize" coherent sets when they see them; consider that the Great Red Spot of Jupiter is clear to any and all that have seen it, as perhaps the most famous coherent set in the solar system. With this motivation, we will develop here an observer-based perspective of coherence.

FIGURE 5.4: Jupiter Portrait as viewed from the spaceship Cassini.

If we do not have a model, as the dynamical system is known only by remote sensing observations, then in practice the flow mapping, $\varphi(\mathbf{x}_0, t_0, t)$ is at best inferred, but generally not available, and often likewise nonlinear systems require numerical integration to infer the flow at sampled points. Here we will approach questions of coherence in the setting that we have only remote observations, but no model.

Developing a model of the flow either directly, $\varphi(\mathbf{x}_0, t_0, t)$, or as a model of the vector field (say by optic flow), may not always be practical or the best way to proceed.

Take as a case in point that the Great Red Spot (GRS) was observed and identified as persistent over many years without ever needing to develop a great deal of the formalism associated with our modern descriptions and algorithms of coherent sets. No transfer operators [4, 135], no Koopman operators [124, 200], and no vector field were required [122].

See Fig. 5.4, as seen in the year 2000 from the Cassini space probe, a joint NASA, European Space Agency (ESA), and Italian space agency Agenzia Spaziale Italiana (ASI) mission [138]. This true-color mosaic of Jupiter was constructed from images taken by the narrow-angle

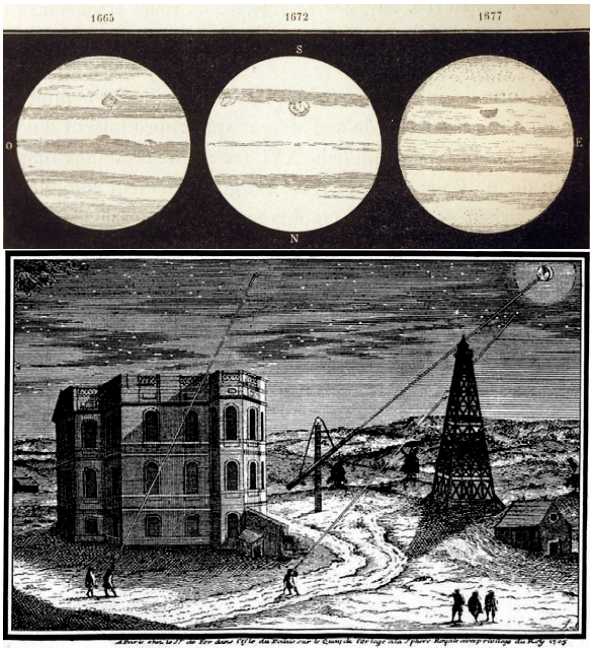

FIGURE 5.5: Jupiter as sketched by Giovanni Domenico Cassini (Top) in his own hand from 1665-1677, from the Memoires de l'Académie Royale des Sciences de Paris [66]. Note that North is drawn, and so labelled, on the bottom. We see that Cassini was seeing and sketching similar scenes over the several years, including apparently the large storm. (Bottom) A sketch of the observatory in Paris.

camera onboard NASA's Cassini spacecraft on December 29, 2000, during its closest approach to the giant planet at a distance of approximately 10 million kilometers (6.2 million miles) [139].

The solar system's largest and most persistent planetary hurricane storm, the vortex structure called the GRS is clearly visible in this image. There are also belts and zones as persistent latitudinal structures, as well as many other smaller storms, (but still massive by Earth standards). There are also other embedded objects, that are clearly present and notable by the naked eye, without ever needing a digital computational engine to identify.

It is as clear today to the casual observer of these modern images, as it was to the Renaissance era astronomer Giovanni Domenico Cassini himself that there are coherent structures on Jupiter [66]. See Fig. 5.5 where Cassini's sketches show some of the same structures as viewed across several years from 1665-1677, were clear enough that he was able to see them despite what were low-quality telescopes by any modern standard. Many of these structures persist today, hundreds of years later.

It is important to distinguish between the concept of a feature that we may notice in a single image as compared to a feature that persists over several successive images, over time. Persistence over time is more akin to what is meant by coherency; we will contrast image segmentation concepts versus motion segmentation concepts in the following section.

5.3 Directed Affinity Segmentation

We have developed the method of Directed Affinity Segmentation[2], which is introduced in our published work [5]. As shown in [5], our method demonstrates high performance in detecting coherent structures from movie data without the need for the intermediate stage of finding the vector field.

First, we review the static time problem of image segmentation, generally as clustering problems, and in the language of our image data from remote sensing. Consider clustering within a single scene, meaning a single frame of a movie. Suppose a grid of positions where color (or some other collection of pointwise measured quantities) is sampled, at each of $\{\mathbf{z}_i\}_{i=1}^{M}$ for M (usually uniformly spaced grid of) pixels over $\{\mathbf{z}_i\}_{i=1}^{M} \subset \mathcal{M} \subset \mathbb{R}^2$. So $\mathcal{M}$ is the framed image. At each of these, observe $\mathbf{c}(\mathbf{z}) : \mathcal{M} \to \mathbb{R}^d$, (generally say $d = 3$ colors at each position) to form an observation matrix,

$$X_{i,r} = \mathbf{c}(\mathbf{z}_i), \qquad 1 \leq r \leq d \text{ colors.} \tag{5.2}$$

Since $\mathbf{c}$ is a vector valued measured observation with d observations (colors), then X is $M \times d$. For many frames sampled across time, we will write, $X_{i,r}[k]$, and for each time t_k, $0 \leq k \leq N-1$.

Our goal is to partition the space based on a notion of coherency, across time and space. By spatial partition the space of sampled data, we mean, given data $\{\mathbf{z}_i\}_{i=1}^{M}$, there is an assignment into labels, $\mathcal{S} = \{S_l\}_{l=1}^{k}$ that serves as a function from the pixel positions to (colored) labels. How this assignment should be done appropriately is a matter we now discuss, and we describe how it should relate to how the measured $\mathbf{c}$ values vary across time.

[2]This work, including Jupiter example, was part of my Master degree in Applied Mathematics.

Perhaps the two most commonly useful image segmentation methods are called k-means [106], and spectral segmentation [144], respectively, which we reviewed in Appendix A, to keep our focus here in our directed partitioning approach.

Image segmentation may be formulated as a spectral graph partitioning problem [144], which we also review in Appendix A. However, these methods need a major adjustment when applied to image sequences (movies) for motion segmentation, despite that traditionally they have been applied to movies with some degree of success [170]. The key difference is what underlies a notion of coherent observations, remembering that the arrow of time has directionality.

We require *affinity matrices that are not symmetric,* and when considered as graphs, they are *directed graphs.* Therefore much of the theoretical underpinning of the standard spectral partitioning needs some adjustment since it relies on symmetric matrices and undirected graphs. We will need a graph Laplacian for weighted directed graphs.

Proceeding computationally, from spectral clustering with color alone, image segmentation may be formulated as a graph partitioning problem, and as such, doing so with color alone means formulating the data set; assign data set [144],

$$X = [X_{1,:}^T | X_{2,:}^T | ... | X_{M,:}^T], \tag{5.3}$$

So, for color alone, X is $d \times N$. Columns of X are the color channels at each pixel position $\mathbf{z}_i$, and we write $X_i = X_{i,:}^T$. If distance is based on color alone, we write a pairwise distance function:

$$D_{i,j} = \|X_i - X_j\| = \sqrt{\sum_{l=1}^{d} (X_{i,k} - X_{j,k})^2}, \tag{5.4}$$

which describe a matrix of distance function values across the sample of points, for distance function, $d(\mathbf{z}_i, \mathbf{z}_j)$, and $d : M \times M \to \mathbb{R}^+$. It is these specific choices, of data X, and distance, that will lead to a time based symmetric spectral method as discussed in Appendix A.

In order to develop a directed affinity matrix, we first replace X in Eq. 5.3 with,

$$X(t) = [X_{1,:}(t)^T | X_{2,:}(t)^T | ... | X_{M,:}(t)^T], \tag{5.5}$$

where $X_{i,:}(t)$ denotes the column vector of d colors at $\mathbf{z}_i$, pixel location i, at time t in the movie sequence. Generally the colors at pixel i will be changing over time. Then let,

$$\begin{aligned} D_1(i,j,a,\tau) &= \sum_{l=0}^{\tau-1} \|X_i(t+la) - X_j(t+(l+1)a)\| \\ &= \sum_{l=0}^{\tau-1} \sqrt{\sum_{l=1}^{d} (X_{i,k}(t+(l)a) - X_{j,k}(t+(l+1)a))^2}. \end{aligned} \tag{5.6}$$

This compares the scene at pixel position i, through τ-time instances starting from time t, $l = 0, a, 2a, ..., (\tau-1)a$, to the scene at pixel j through τ-time instances one step in the future, $l = 1, a, 2a, ..., \tau a$. Note that the norm, the inner sum, is the same as the color measuring norm in Eq. 5.4.

Now we measure the spatial distance between the pixels, as they appear naturally in the scenes represented by the images. Let,

$$D_2(i,j)^2 = \|\mathbf{z}_i - \mathbf{z}_j\|^2, \tag{5.7}$$

be the standard spatial Euclidean distance between pixel positions. Adding these two norms defines a spatial and time delayed color distance function,

$$D(i,j,a,\tau)^2 = D_1(i,j,a,\tau)^2 + \alpha D_2(i,j)^2. \tag{5.8}$$

Note that $D_1(i,j,0,0)$ is identical to the distance function in Eq. 5.4 used for image segmentation, but including $a > 0$, $\tau > 0$ allows us to consider motion segmentation and so coherency. Finally an affinity matrix follows,

$$\mathcal{W}_{i,j} = e^{-D(i,k,a,\tau)^2/2\sigma^2}. \tag{5.9}$$

Notice we have suppressed including all the parameters in writing $\mathcal{W}_{i,j}$, and that besides time parameters a and τ that serve as sampling and history parameters, together the parameters α and σ serve to balance spatial scale and resolution of color histories.

In Eq. 5.9, we see the *asymmetric* matrix reflecting the arrow of time. Such a difference from the standard symmetric affinity is fundamental and naturally must be included in any concept

of coherence. Clustering in this setting then reflects the concept of coherence, as a scene that retains its "appearance," but for now, we continue with the idea that maintaining appearance is a sensible idea of coherence.

We proceed to cluster the system summarized by affinity $\mathcal{W}$ by interpreting the problem as random walks through the weighted *directed* graph, $G = (V, E)$ generated by $\mathcal{W}$ as a weighted adjaceny matrix. Stated equivalently, this is like a directed diffusion problem. So let,

$$\mathcal{P} = \mathcal{D}^{-1}\mathcal{W}, \tag{5.10}$$

where analogously to the symmetric case, $\mathcal{D}_{i,i} = \sum_j \mathcal{W}_{i,j}$, $\mathcal{D}_{i,j} = 0, i \neq j$. So $\mathcal{P}$ is a row stochastic matrix representing probabilities of a Markov chain through the directed graph G, where,

$$\mathcal{P}_{i,j} = p(j(t+a)|i(t)). \tag{5.11}$$

We may cluster the directed graph by concepts of spectral graph theory for directed graphs, following the weighted directed graph Laplacian described by Fan Chung [46], and a similar computation has been used for transfer operators in [72, 86] and as reviewed [24]. The Laplacian of the directed graph G is defined, [46],

$$\mathcal{L} = I - \frac{\Pi^{1/2}\mathcal{P}\Pi^{-1/2} + \Pi^{-1/2}\mathcal{P}^T\Pi^{1/2}}{2}. \tag{5.12}$$

See discussion of the symmetric spectral graph theory in Appendix A, and the ncut problem solution standard description by Courant-Fischer theory, and how that adapts to this weighted directed graph Laplacian case. Note that $\mathcal{P}$ is row stochastic implies that it row sums to one, or stated as the right eigenvector is the ones vector, $\mathcal{P}\mathbf{1} = \mathbf{1}$, but the left eigenvector corresponding to left eigenvalue 1 represents the steady state row vector of the long term distribution,

$$u = u\mathcal{P}, \tag{5.13}$$

which for example if $\mathcal{P}$ is irreducible, then $u = (u_1, u_2, ..., u_{pq})$ has all positive entries, $u_j > 0$ for all j, or say for simplicity $u > 0$. Let Π be the corresponding diagonal matrix,

$$\Pi = diag(u), \tag{5.14}$$

and likewise,

$$\Pi^{\pm 1/2} = diag(u^{\pm 1/2}) = diag((u_1^{\pm 1/2}, u_2^{\pm 1/2}, ..., u_{pq}^{\pm 1/2})), \tag{5.15}$$

which is well defined for either $\pm$ sign branch when $u > 0$. The the first smallest eigenvalue larger than zero, $\lambda_2 > 0$ such that,

$$\mathcal{L}v_2 = \lambda_2 v_2, \tag{5.16}$$

allows a bi-partition, by,

$$y = \Pi^{-1/2} v_2, \tag{5.17}$$

by sign structure. Analogously to the Ng-Jordan-Weiss symmetric spectral image partition method [144], the first k eigenvalues larger than zero, and their eigenvectors, can used to associate a multi-part partition, by assistance of k-means clustering these eigenvectors.

5.4 Numerical Results

5.4.1 Clouds of Jupiter, and the Great Red Spot

The results of partitioning using the directed affinity matrix $\mathcal{W}$ is shown in Fig. 5.6 from a scene of the GRS, and including the affinity matrix and a permutation that brings it to block structure as indicated by colors matching the colored scene. Fig. 5.7 again shows a scene of the GRS of Jupiter and its segmentation according to comparing the different methods of k-means to a single scene, a spectral method from a single scene, and finally our directed spectral method. We see that our method (d), the regions found by the directed method are most coherent in the sense of showing across time what is clearly visible in a movie, and perhaps difficult to fully appreciate in a static figure here.

Our raw images consist of 14 images taken by the narrow-angle camera onboard NASA's Cassini spacecraft. The images span 24 Jupiter rotations between October 31 and November 9, 2000. We forward the reader to NASA website, [140], to see how the scene change through the movie frames since it is hard to detect the dynamic through still images clearly.

In our result, we chose a primary number of clusters that maximize the mutual information (in information theory sense) between movie frames. Then, for each cluster, every connected object has been extracted as a separate cluster. We exclude three frames out of the 14 available

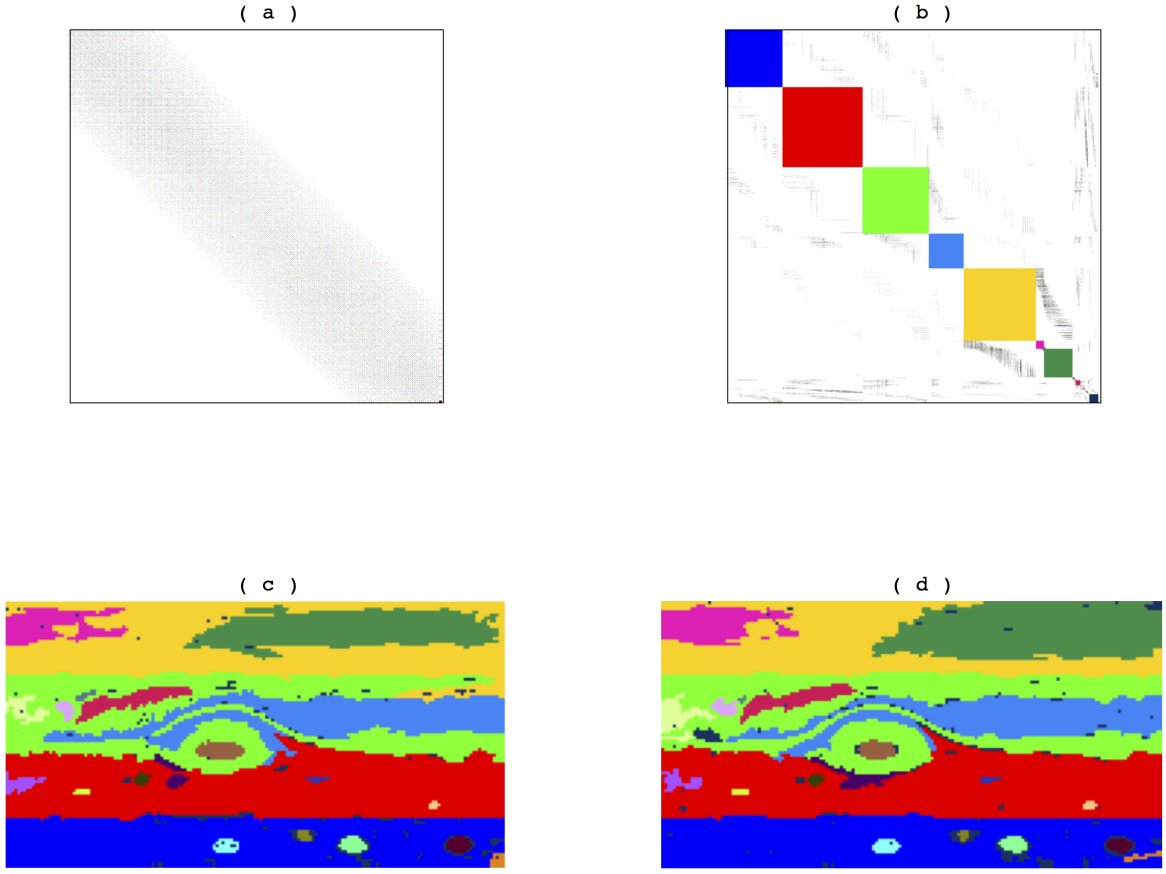

FIGURE 5.6: Given a small scene surrounding the Great Red Spot, and coarse-grained (for ease of computation and clarity of presentation in this figure), (a) The affinity matrix, $\mathcal{W}$, Eq. 5.9, (b) Affinity matrix sorted according to spectral partition by methods of Appendix A, Eq. 5.9-Eq. 5.17. (c) Coloring by each block of the sorted affinity matrix partitions the scene according to regions that are found in multiple frames. Eight Frames (1 to 8) used to create our directed affinity (d) The partitioned scene after $t = T = 3$ time. (Frames 4 to 11).

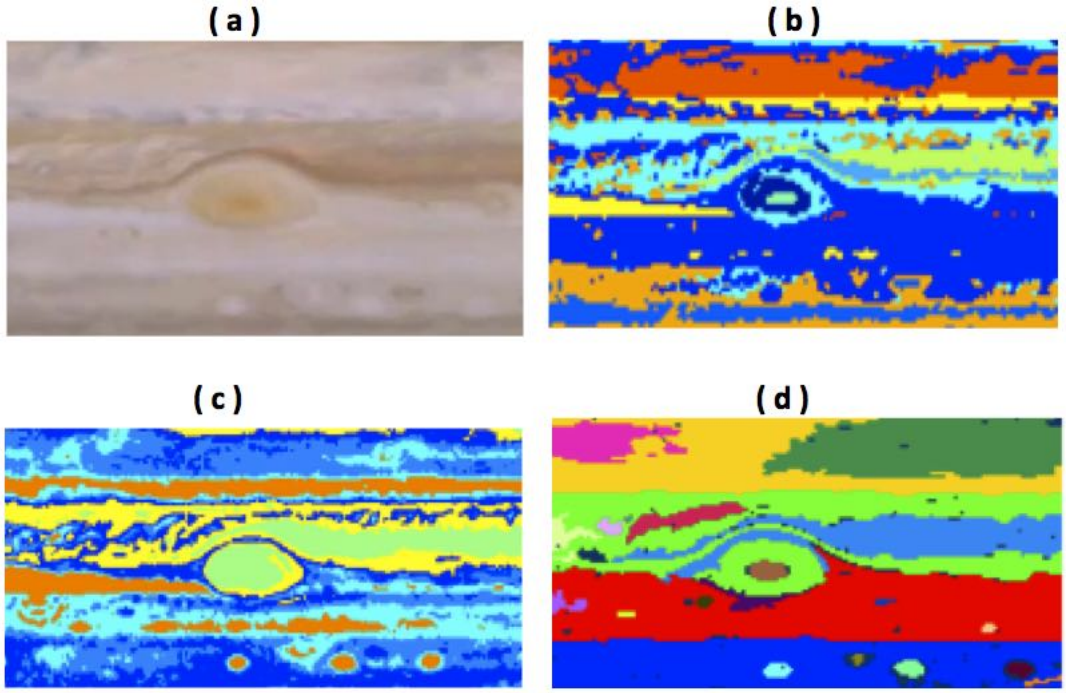

FIGURE 5.7: (a) A small scene surrounding the Great Red Spot, (b)A k-means clustering based on color only by affinity matrix. (c) Based on spectral partitioning with color alone affinity (d) Based on directed affinity matrix, as in Fig. 5.6.

frames because they have a sudden appearance for Jupiter moons. We choose a number of clusters that maximize the mutual information between frames.

The entire Cassini Jupiter data set is shown by directed spectral partition for coherence, as shown in Fig. 5.9. Most notable are the banded longitudinal structures, the many circular vortex storms, and the largest being the GRS.

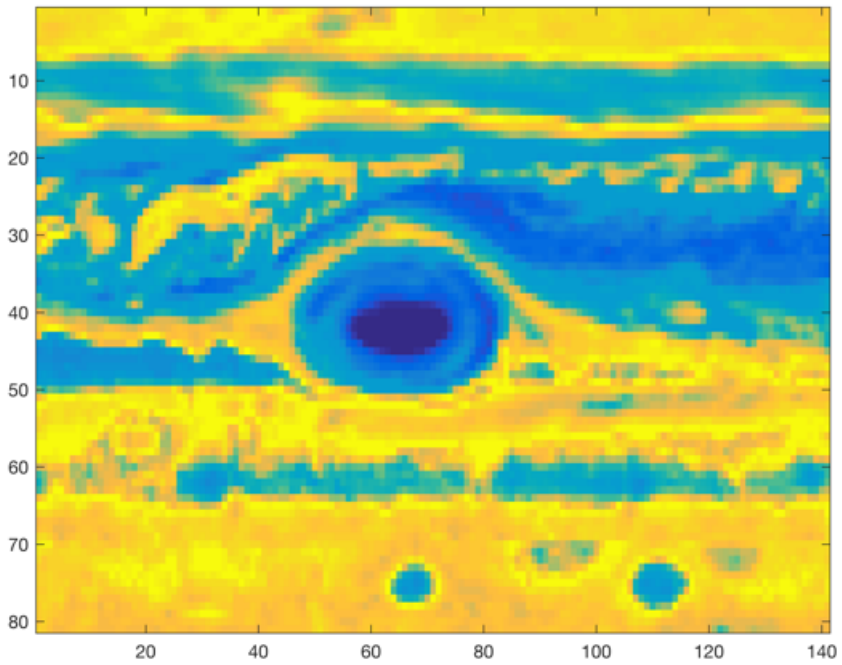

FIGURE 5.8: First Eigenvector of the directed affinity Matrix.

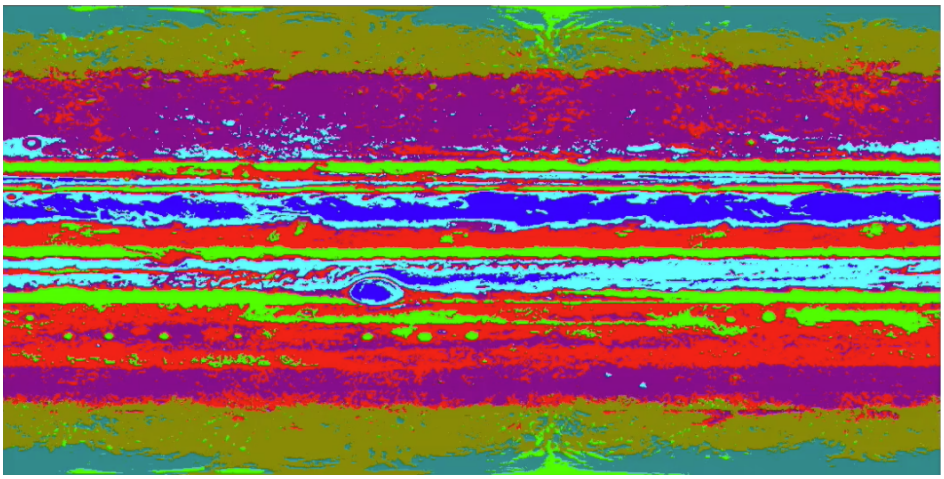

FIGURE 5.9: Directed spectral partition of Jupiter of the entire Cassini data set. Compare to Fig. 5.10.

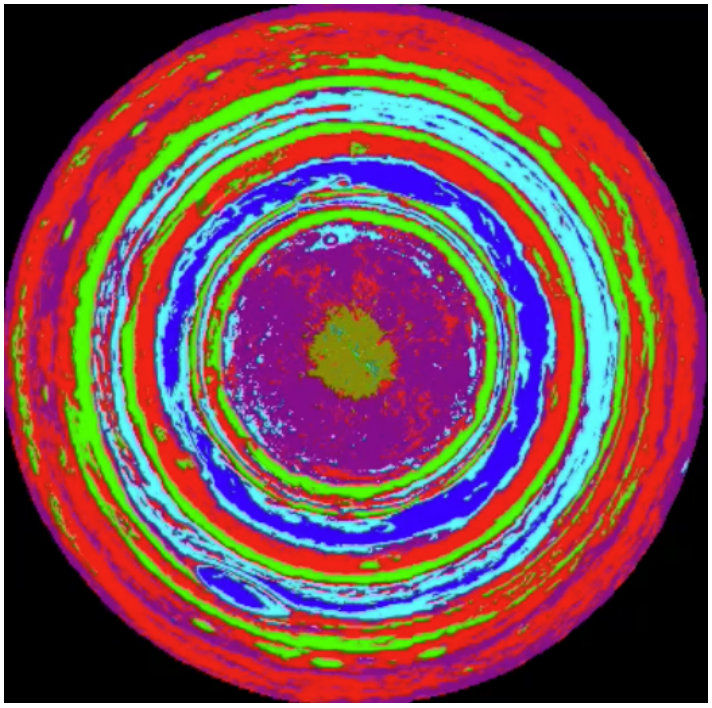

FIGURE 5.10: Directed spectral partition of Jupiter as shown on a projection as seen from the northern pole. Compare to Fig. 5.9.

5.4.2 Antarctic Ice Shelves

Global Warming causes global sea level to rise [133], the three main reasons for this are oceans expansion, ice sheets lose ice faster than it forms from snowfall, and glaciers at higher altitudes also melt. During the 20^{th} century, sea level rise has been dominated by the retreat of glaciers and expansion of the ocean, but this contribution starts to change in the 21^{st} century because of the ice shelves cracks. The ice sheets store most of the land ice (99.5%) [133], with a sea-level equivalent (SLE) of 7.4 m (24 ft) for Greenland and 58.3 m (191 ft) for Antarctica.

Ice sheets form in areas where snow that falls in winter does not melt entirely over the summer. Over thousands of years, the layers of snow pile up into thick masses of ice, growing thicker and denser as the weight of new snow and ice layers compresses the older layers.

Ice sheets are constantly in motion, slowly flowing downhill under their own weight. Near the coast, most of the ice moves through relatively fast-moving outlets called ice streams, glaciers, and ice shelves. As long as an ice sheet accumulates the same mass of snow as it loses to the sea, it remains stable.

Most of Antarctica has yet to see dramatic warming. However, the Antarctic Peninsula,

FIGURE 5.11: Map of Antarctica and surrounding islands. Source: wiki Files.

which juts out into warmer waters north of Antarctica, has warmed 2.5 degrees Celsius (4.5 degrees Fahrenheit) since 1950. A large area of the West Antarctic Ice Sheet is also losing mass, probably because of warmer water deep in the ocean near the Antarctic coast. In East Antarctica, no clear trend has emerged, although some stations appear to be cooling slightly. Overall, scientists believe that Antarctica is starting to lose ice, but so far the process has not become as quick or as widespread as in Greenland.

The icing of Antarctica began in the middle Eocene about 45.5 million years ago [70], and escalated during the Eocene–Oligocene extinction event about 34 million years ago. The Western Antarctic ice sheet declined somewhat during the warm early Pliocene epoch, approximately 5 to 3 million years ago; during this time the Ross Sea opened up [70, 204]. But there was no significant decline in the land-based Eastern Antarctic ice sheet.

The continent-wide average surface temperature trend of Antarctica is positive and significant at > 0.05 °C/decade since 1957 [74, 177], West Antarctica has warmed by more than 0.1 °C/decade in the last 50 years, and this warming is most active in winter and spring. Although this is partly offset by fall cooling in East Antarctica, this effect is restricted to the 1980s and 1990s [177].

Larsen Ice Shelf extends in a ribbon down the east coast of the Antarctic Peninsula from James Ross Island to the Ronne Ice Shelf, see Fig.5.11 and Fig.5.12. It consists of several distinct

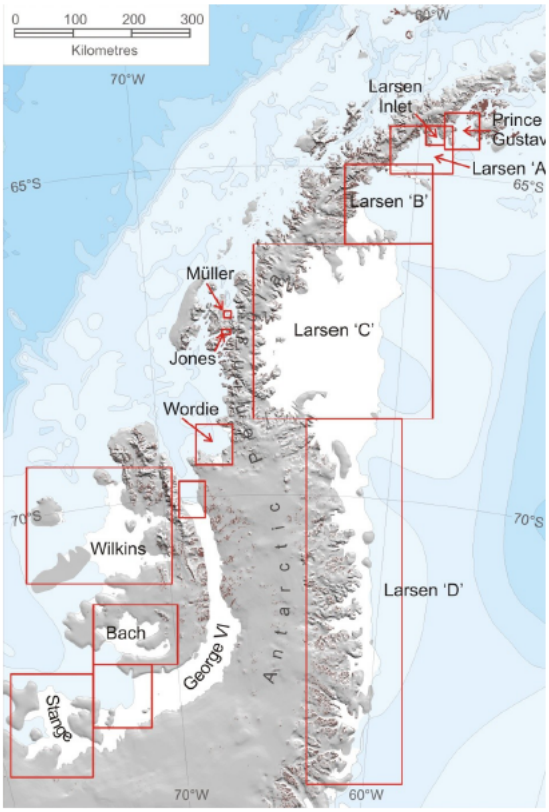

FIGURE 5.12: Location of ice shelves on the Antarctic Peninsula. Source [48].

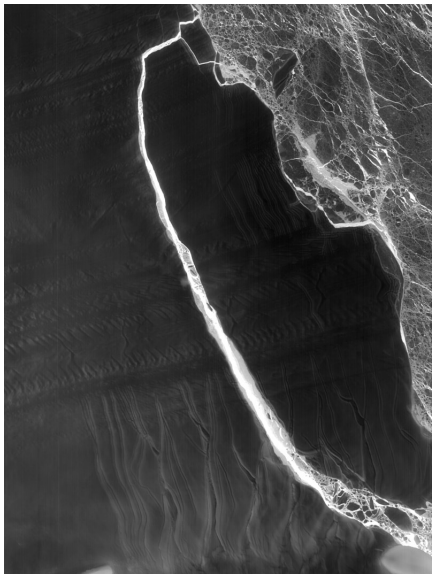

FIGURE 5.13: A-68 iceberg. The fractured berg and shelf are visible in these images, acquired on July 21, 2017, by the Thermal Infrared Sensor (TIRS) on the Landsat 8 satellite. Credit: NASA Earth Observatory images by Jesse Allen, using Landsat data from the U.S. Geological Survey.

ice shelves, separated by headlands. Larsen C ice crack already started in 2010, but it was very slow, and there were no signs of radical change according to Interferometry processing of the images. But, since October 2015, the ice crack of Larsen C has been growing fast until it finally collapses to result in A68, see Fig.5.13; the largest known iceberg ever known, more than 2,000 square miles in area or nearly the size of Delaware detached from one of the largest floating ice shelves in Antarctica and floated off in the Weddell Sea.

In [77], the authors derived structural glaciological description and analysis of surface morphological features of the Larsen C ice shelf from satellite images spanning the period 1963–2007. The results and conclusions of the research stated that: "*Surface velocity data integrated from the grounding line to the calving front along a central flow line of the ice shelf indicate that the residence time of ice (ignoring basal melt and surface accumulation) is 560 years. Based on the distribution of ice-shelf structures and their change over time, we infer that the ice shelf is likely to be a relatively stable feature and that it has existed in its present configuration for at least this length of time.*".

In [97], the authors modeled the flow of the Larsen C and northernmost Larsen D ice shelves using a model of continuum mechanics of ice flow, and the applied a fracture criterion to the simulated velocities to investigate the ice shelf's stability. The conclusion of the analysis shows that the Larsen C ice shelf is inferred to be stable in its current dynamic regime. This work published in 2010, and in the same year, Larsen C ice crack already existed, but for its slow-growing rate and according to analytic studies, there were no expectations for the fast-growing and collapse that happened for Larsen C.

One main technique to analyze and predict ice cracks is the interferometry. Interferometry [15, 116], is a family of techniques in which waves, usually electromagnetic waves, are superimposed, causing the phenomenon of interference, which is used to extract information. Interferometers are widely used in science and industry for the measurement of small displacements, refractive index changes, and surface irregularities.

Fig.5.14 shows the interferometry image as of April 20, 2017, and although it clearly shows the crack that already exists, it is absolutely can provide no information or indications about what can happen next. A couple of weeks after this image, Larsen C ice crack took different dynamic and divided into two branches, as shown in Fig.5.15.

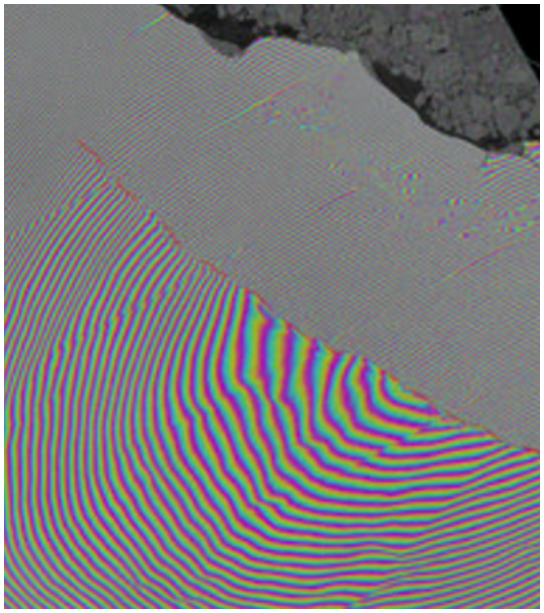

FIGURE 5.14: Interferometry (April 20, 2017). Two Sentinel-1 radar images from 7 and 14 April 2017 were combined to create this interferogram showing the growing crack in Antarctica's Larsen-C ice shelf. Polar scientist Anna Hogg said: "We can measure the iceberg crack propagation much more accurately when using the precise surface deformation information from an interferogram like this, rather than the amplitude (or black and white image) alone where the crack may not always be visible." Source [65].

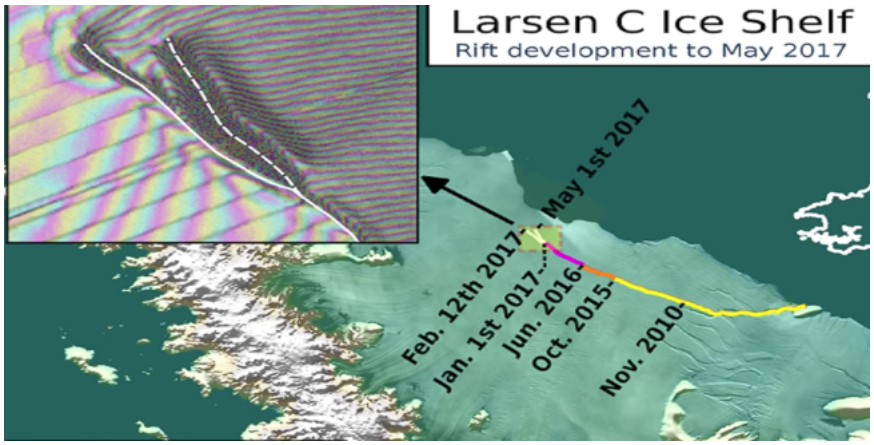

FIGURE 5.15: Lrasen C crack development (new branch) as of May 1, 2017. Labels highlight significant jumps. Tip positions are derived from Landsat (USGS) and Sentinel-1 InSAR (ESA) data. Background image blends BEDMAP2 Elevation (BAS) with MODIS MOA2009 Image mosaic (NSIDC). Other data from SCAR ADD and OSM. Credit: MIDAS project, A. Luckman, Swansea University.

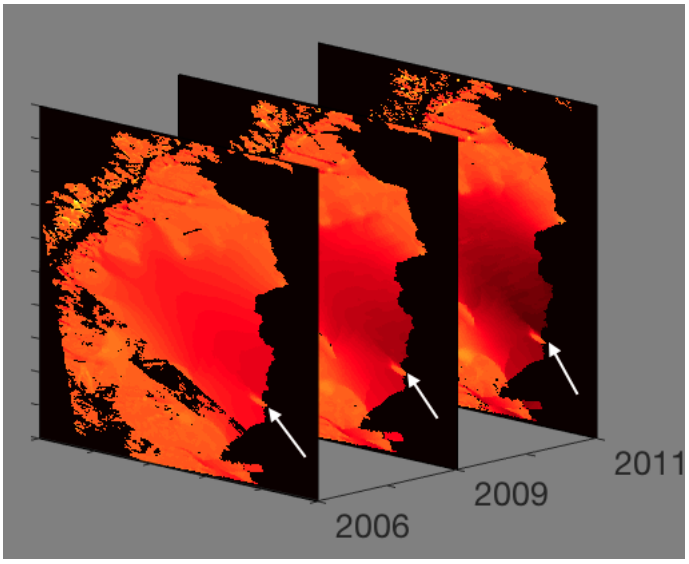

FIGURE 5.16: Ice surface velocity. The figure shows the data set for three different years around the very beginning of Larsen C ice crack in 2010. The data from the years 2007,2008, and 2010 have corrupted data on the region of interest, and then they are excluded. The color scale indicates the magnitude of the velocity from light red (low velocity) to dark red (high velocity), and the arrow points to the starting tip of the crack. Result of the directed partitioning is shown in Fig.5.17. Source of data: [158].

We will apply the directed affinity segmentation to both ice surface velocity field data, and satellite images only.

We will apply the directed affinity segmentation to satellite images of Larsen C ice shelf, and we will show that the directed affinity segmentation of spatiotemporal changes can work as early warning sign tool for critical transition in ice sheets. We will apply our "post-casting" experiments on images of Larsen C before splitting of the A68 iceberg and will compare our forecasting based on segmentation results to the actual event occurs.

In Fig.5.16 we see different snapshots of the ice surface velocity data set [136, 157, 158], which is part of the NASA Making Earth System Data Records for Use in Research Environments (MEaSUREs) Program, and it provides the first comprehensive [158], high-resolution, digital mosaics of ice motion in Antarctica assembled from multiple satellite interferometric synthetic-aperture radar systems. We apply the Directed affinity partitioning to the available data set, and the results are shown as a labeled image in Fig.5.17.

As shown on Fig.5.17 we note the following:

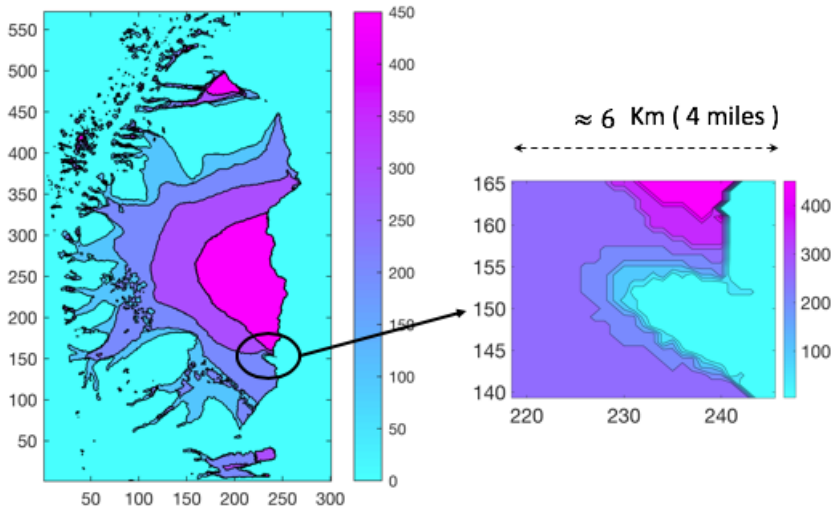

FIGURE 5.17: Directed Affinity result. (Left) The directed partitioning results for the ice surface velocity of the years 2006,2009,2011, and 2012. Noting that the ice shelf crack started in 2010. (Right) A narrow zoom to the region of interest that shows large varying in ice surface velocity within a small area. To give a clear view of the differences in speed, Fig.5.18 shows the surface plot for the same result.

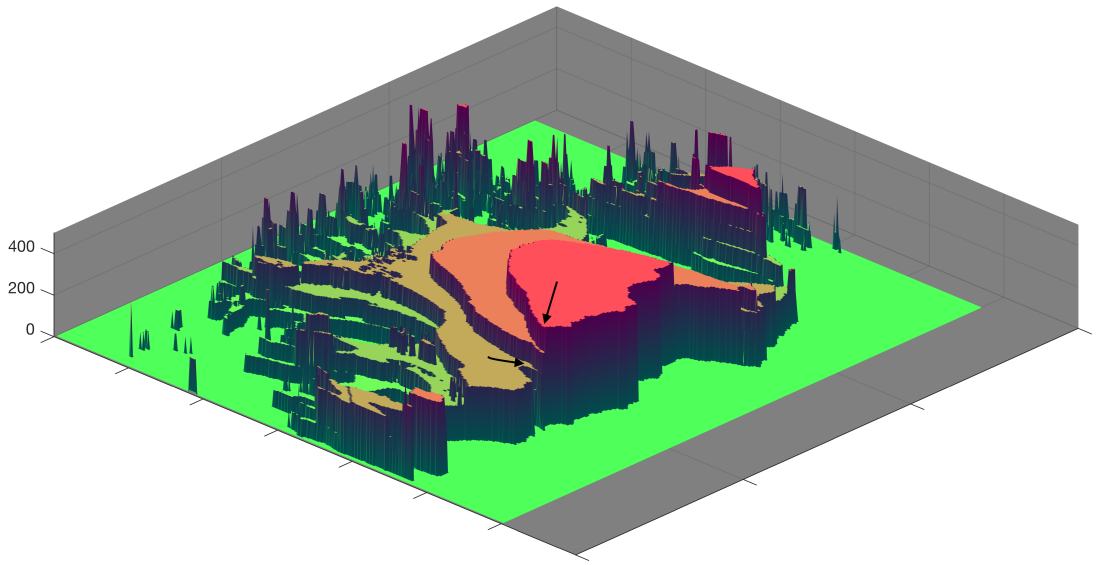

FIGURE 5.18: Directed affinity partitions with the mean velocity (speed) of the partition assigned for each label entries. The spatial distance between the arrows tips is less than two miles, while the difference in the speed is more than 200 m/year.

- The data collected from eight different sources [131, 158], with different coverage and different error range, and interpolating the data from different sources explains the smooth curves in segmentation around the region on interest.
- The directed partitioning shows the Larsen C ice shelf as coherent structures that contain each other.
- The zoom picture to the right shows the region where the Larsen C ice crack start, and we see that within narrow spatial distance (4 miles) there is a large change in velocity, more precisely, the outer boundaries of the different coherent sets become very close to contact (by considering the margin of error in the measurements [131], it has high probability that they contact).

Directed partitioning gives us informative clustering, meaning that each cluster has homogeneous properties, such as the magnitude and the direction of the velocity. In general, for the coherent sets $A_1 \subset A_2 \subset ... \subset A_n$, physically, the set A_{i-1} keeps its coherence within A_i because of a set of properties (i.e., chemical or mechanical properties) that rules the interaction between them. However, the contact between the boundaries of the sets A_{i-1} and A_i, see Fig.5.19, means a direct interaction between A_{i-1} and A_{i+1}, which opens a "new possibilities" of a new kind of reactions different from we already have.

In the case of the ice velocity partitioning, we collect the observations and discuss them, however, since the sets boundaries are not completely contacted, and the direction of the velocity has no critical change, which we believe result from interpolating the data and smoothing the measurements, we state nothing more than such close interaction between coherent sets boundaries can be an early warning sign that should be considered and investigated by applying "what if" assumptions and analyzing the consequences from any change or any error in the measured data.

However, our directed partitioning method achieved better results using the satellite images [134]. To reduce the effect of noise (clouds and image variable intensity), we considered the average image over one month as a single snapshot for the directed affinity constructions.

Fig.5.20 the directed affinity partitioning for two time windows starting from December 2015. The directed partitioning began to detect the significant change in the Larsen C ice shelf on July 2016. In Fig.5.21 we see that by September 2016, we detect a structure very close in

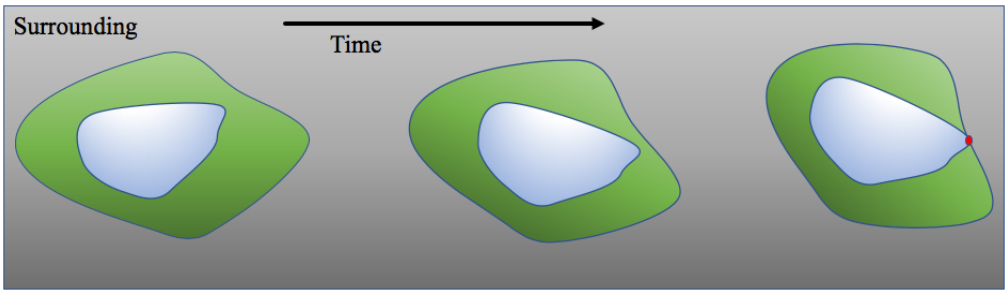

FIGURE 5.19: Two coherent sets dynamic. As the inner set contact the boundaries of the outer one, than give the chance for a new reactions that "may" cause critical transition.

shape to the iceberg A-68, which calved from Larsen C on July 2017. Moreover, by November 2016, see Fig.5.22, the boundaries of the detected partitions match the crack dividing into two branches that happened in later in May 2017 and shown in Fig.5.15.

We see that the directed affinity partitioning can be a useful early warning sign, that indicates the possibility critical spatiotemporal transitions, and it can help to bring the attention for specific regions to investigate different possible scenarios in the analytic study.

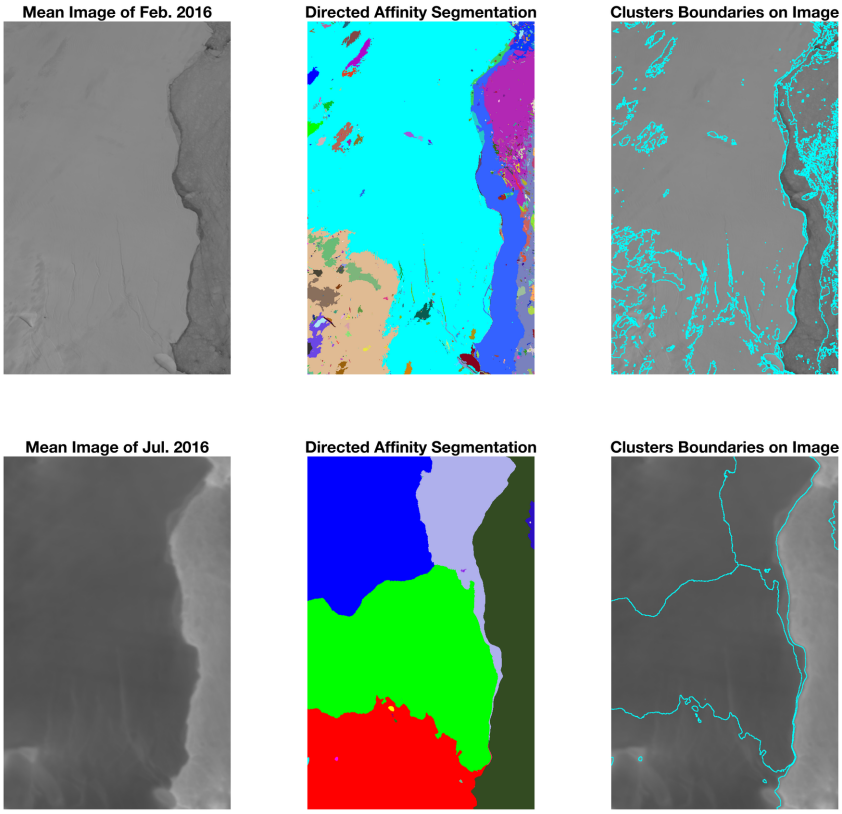

FIGURE 5.20: For two time windows (top and bottom), we see (Left) The mean image of the images included in the window. (Middle) The Directed Affinity Segmentation Labeled Clusters. (Right) Overlying of the directed affinity segmentation boundaries over the mean image of the window. We took these time windows of Feb. 2016 and July 2016 as a detailed example, and more time windows results are shown in Fig. 5.21. We see that during 2016, there was no significant change in Larsen C crack at the beginning of the year. In July 2016, the directed affinity segmentation propose a large change in the crack dynamics, and this change keeps going faster as Fig. 5.21 shows.

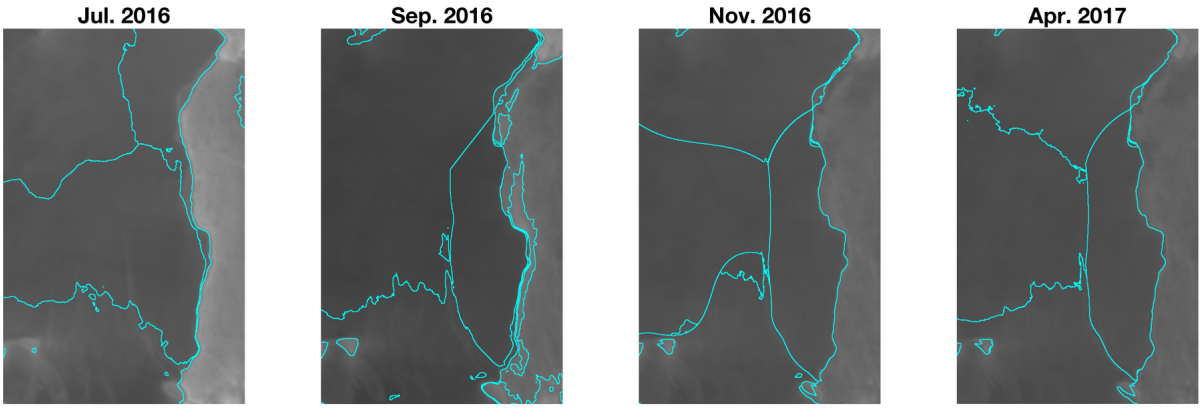

FIGURE 5.21: In analogy to Fig. 5.20-Right, this figure shows the Directed Affinity Segmentation boundaries for different time windows starting from July 2016 to April 2017.

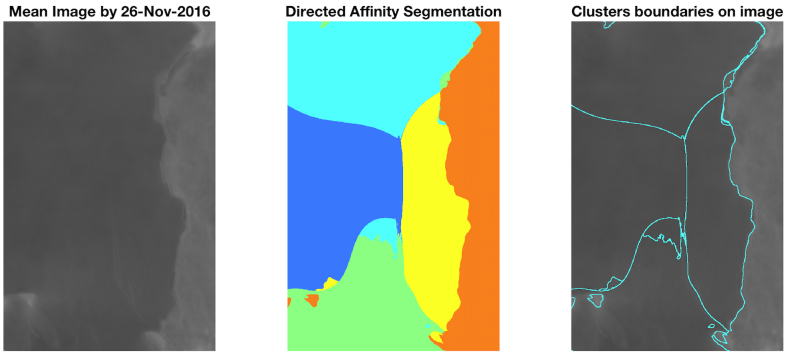

FIGURE 5.22: The mean image and the directed affinity partitioning as of November 2016. The results shows similar structure to the crack branching that occurred on May 2017 and shown in Fig.5.15, and similar structure the final iceberge that calved from Larsen C on July 2017.

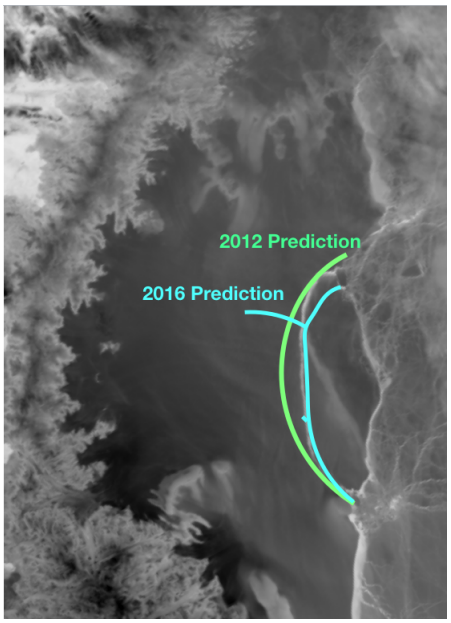

FIGURE 5.23: 2012 prediction based on ice surface velocity data, and 2016 prediction based only on satellite images. Compare to the actual crack (white curve between the two prediction curves) on July 2017, shown in Fig.5.13.

Chapter 6

Extensions and Future Directions

> The important thing is not to stop questioning. Curiosity has its own reason for existence. One cannot help but be in awe when he contemplates the mysteries of eternity, of life, of the marvelous structure of reality. It is enough if one tries merely to comprehend a little of this mystery each day.
>
> *Albert Einstein*
>
> *(1879 - 1955)*

In this chapter, we discuss some of extensions and future direction of our work, which is based on promising primary results achieved during our research.

6.1 Large Networks of Non-Identical Oscillators

We showed in our numerical results the robust result of ER in detecting the sparse structure of complex systems and coupled networks. Recovering the coupling structure (adjacency matrix) of coupled networks is one of the important problems in many fields.

In one of our ongoing research we are developing a general framework to reconstruct the coupling structure in large complex networks of nonidentical oscillators. Consider unified Lorenz system [161]:

$$\begin{cases} \dot{x} = (25\alpha + 10)(y - x) \\ \dot{y} = (25 - 35\alpha)x + (29\alpha - 1)y - xz \\ \dot{z} = xy - \frac{8+\alpha}{3}z \end{cases} \tag{6.1}$$

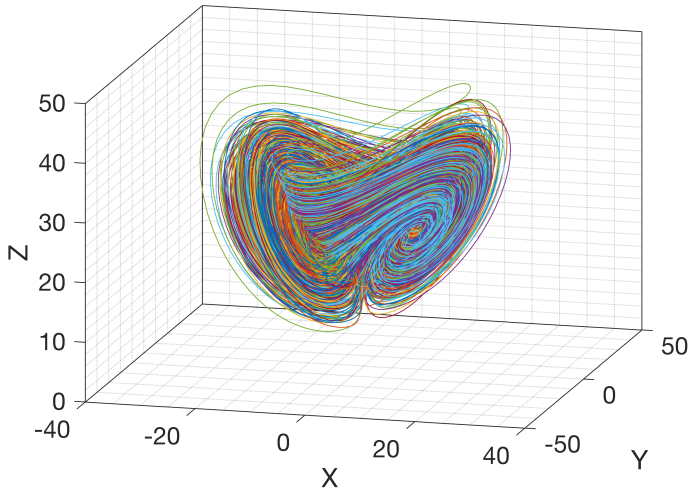

FIGURE 6.1: Coupled Oscillators of Unified Lorenz System. $n = 100$ nodes with unique value of α for each oscillator.

where $\alpha \in [0, 1]$. The system is chaotic for all $\alpha \in [0, 1]$, and when $\alpha \in [0, 0.8]$, the system reduces to the general Lorenz system, and when $\alpha = 0.8$, it becomes the general Lu system. When $\alpha \in (0.8, 1]$, the system is the general Chen system.

The main idea of the new approach, is to not give attention (temporary) to the individual parameters in network, instead, we write the oscillators in the form:

$$\Psi = [\psi_1, \psi_2, ..., \psi_n] \tag{6.2}$$

where $\psi_i \in \mathbb{R}^{N \times d}$ is the i^{th} d-dimensional oscillator, and N is number of observations, and $\Psi \in \mathbb{R}^{N \times nd}$ is the nodes library and n is number of nodes. Fig.6.1 shows a network of randomly coupled oscillators.

Similar to basic methodology discussed in Chapter 4, we compute the derivative $\dot{\psi}_i \equiv \boldsymbol{f}$. Now, instead of searching in the polynomial expansion for the function ϕ_j that maximize the mutual information with each individual dynamic of the node ψ_i, we search for node ψ_j which its polynomial expansion combined with ψ_i maximize the mutual information with the dynamic $\dot{\psi}_i$. Then, to find the nodes that influence the node i, the forward ER can be expressed as:

$$
\begin{aligned}
u_k &= arg \max_{j \in \mathcal{S}, j \notin s_{k-1}} I(\dot{\psi}_i; \mathcal{V}(\dot{\psi}_i, \Phi(\psi_j)) | \mathcal{V}(\dot{\psi}_i, \Phi(\Psi_{s_{k-1}}))), \\
s_k &= s_{k-1} + u_k
\end{aligned} \tag{6.3}
$$

where $\Phi(\Psi_{s_{k-1}})$ is the polynomial expansion for nodes indicated by indices in the set s_{k-1}. Different from ER in Chapter 4, we start here with the assumption $s_0 = \{i\}$, meaning that it is assumed that each node is surely influenced by itself.

The discussion above summarize the main idea, and we are following the algorithms engineering techniques to ensure the efficiency and scaleability of the algorithm, combined with other projects such as the dynamic dictionary learning and the efficient basis expansion.

Fig.6.2 shows a binary tree network created using identity coupling function (x coupled to x, y coupled to y, and z coupled to z on each coupled nodes), Fig.6.3 shows the adjacency matrix of coupled system, and the recovered adjacency matrix described above.

Note that for 100 nodes, even with the 2^{nd} order power polynomial expansion, we will have $\frac{302!}{300!2!} = 45,451$ candidate function, meaning that the measurement to basis ratio is about 4%.

6.2 Efficient Basis Construction

In Lorenz system example in Chapter 4, we assumed the 5^th expansion order since we are assuming the black box modeling, and to test the method under unstable conditions. However, the principle of causation entropy CSE [180], can efficiently detect the relevant features in the state variables $\boldsymbol{z}$ before applying the power polynomial expansion.

For example, for the same noise and outliers levels of the experiment shown in Fig.4.14, let $\boldsymbol{z}$ be the state variable with $N = 1000$ measurements. The pairwise CSE is given by

$$
C_{i,j} = I(\boldsymbol{z}_i^{t+1}; \boldsymbol{z}_j^t | \boldsymbol{z}_{s-j}^t) \tag{6.4}
$$

where $C_{i,j}$ is the causation entropy from the variable (column) $\boldsymbol{z}_j$ to $\boldsymbol{z}_i$, and $s = \{1, 2, 3\}$ is the set of all indices, and the superscript t indicate the measurement $1, 2, ..., N-1$, and $t+1$

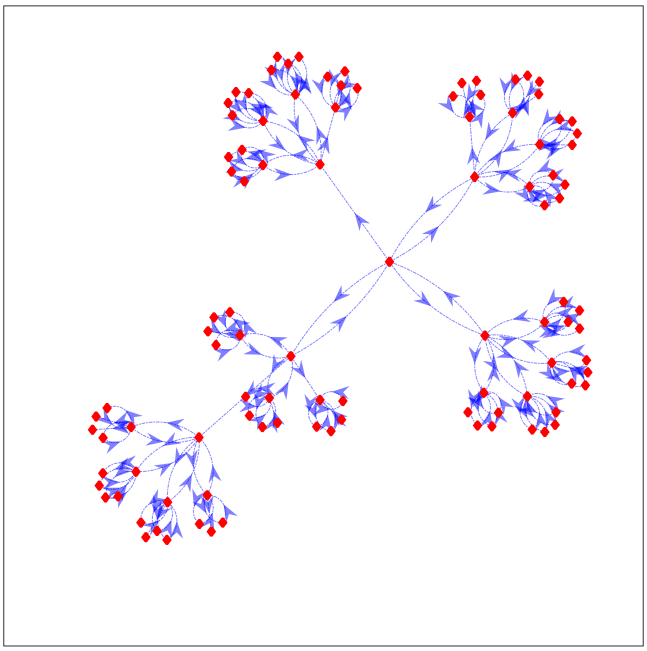

FIGURE 6.2: Network of unified Lorenz system with non-identical oscillators. (100 Nodes)

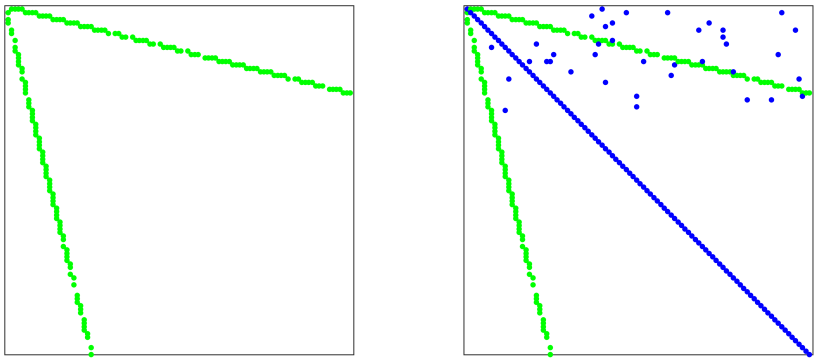

FIGURE 6.3: (left) True adjacency (right) Recovered adjacency by ER using 2000 measurements. Note that in generating the coupled network we neglect the self influence (the diagonal entries), however, it obvious to see the self influence in the recovered matrix since it is embedded influence in the ODE system itself. We see that our method achieved accurate recovery of the adjacency matrix.

0.0035	0.0031	2×10^{-14}
0.0053	0.0530	0.0056
0.0021	0.0030	0.3984

TABLE 6.1: CSE value for Lorenz system state variables

indicate the measurements $2, 3, ..., N$. The result of the matrix C averaged over 100 is shown in Table 6.1.

Our concern here is the low values of CSE, such as $C_{1,3} = 2 \times 10^{-14}$, which indicates the causation entropy from $\boldsymbol{z}_3$ to $\boldsymbol{z}_1$. Such value indicate that $\boldsymbol{z}_3$ has no influence to $\boldsymbol{z}_1$, and then, we can exclude it from the power expansion which highly reduce the computations complexity and the uncertainty of the regression. Moreover, similar approach also can be adopted to find the optimal expansion order.

Appendix A

On K-means and Spectral Clustering

A simple and common form of clustering that one might choose would be a k-means clustering of an image scene [106] based just on the pointwise measurements alone (say colors for example) as a solution to the partitioning problem, to find a partition $\mathcal{S}$ such that

$$\mathcal{S} = \underset{S}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{X_{j,:} \in S_i} \|X_{j,:} - \mu_i\|, \tag{A.1}$$

where $\|\cdot\|$ is the Euclidean norm of the color values, and μ_i are means in each color channel. We see the k-means method is a solution of a partitioning problem. An image such as that of the colors of Jupiter is shown in Fig. 5.7 (b and c) for an example of a static time segmentation of a Jupiter image with $d = 3$ colors $\mathbf{c}(\mathbf{z})$, measured pointwise where $\{\mathbf{z}_i\}_{i=1}^{N}$ are the pixel positions on the image. The k-means problem solution has a direct method of updating the cost function Eq. A.1 as membership of indexed values in each partition element is adjusted, thus shifting the group mean, until optimality is found sufficiently, as reviewed in many standard texts [9, 95, 106].

Beyond the k-means clustering concepts there are spectral clustering methods, and in fact even the k-means method has a spectral formulation,[95]. We will see that spectral methods seem to perform well from a clustering perspective alone.

Spectral descriptions of clustering will allow us to interpret our notions of motion segmentation more naturally in terms of coherence as analogies to spectral decompositions of transfer operators lead to dynamical systems concepts of coherence as already emphasized in the literature, [73, 122, 150]. So we proceed to recall the spectral concepts of clustering for image

segmentation.

There have been several complementary views of clustering by spectral methods, by graph cuts [105, 170], as random walkers, [125], and comparably as a diffusion process as described by diffusion map [162]. Many of these come back to some version of a max-flow min-cut algorithm [42], and in turn as related to the conductance also called Cheeger-constant as a measure of "bottleneckyness" of the underlying graph. In this section, we review the computations for the simpler case of weighted *undirected* graphs, appropriate for image segmentation.

Proceeding computationally, image segmentation may be formulated as a graph partitioning problem, and as such, doing so with color alone means formulating the data set; assign data set [144],

$$X = [X_{1,:}^T | X_{2,:}^T | ... | X_{M,:}^T], \tag{A.2}$$

So, for color alone, X is $d \times N$. Columns of X are the color channels at each pixel position $\mathbf{z}_i$, and we write $X_i = X_{i,:}^T$. If distance is based on color alone, and so as in Eq.. A.1, we write a pairwise distance function. Let

$$D_{i,j} = \|X_i - X_j\| = \sqrt{\sum_{l=1}^{d} (X_{i,k} - X_{j,k})^2}, \tag{A.3}$$

describe a matrix of distance function values across the sample of points, for distance function, $d(\mathbf{z}_i, \mathbf{z}_j)$, and $d : M \times M \to \mathbb{R}^+$. Next as done in many general spectral clustering methods, [144, 162] and as specialized to image segmentation [144, 170], a pairwise symmetric affinity matrix may be defined,

$$W_{i.j} = e^{-D_{i,j}^2 / 2\sigma^2}. \tag{A.4}$$

The value of $\sigma > 0$ may be chosen as a resolution parameter. It is convenient to emphasize the "practical" sparsity, by reassigning $W_{i,j} = 0$ if $W_{i,j} < \epsilon$ for a small threshold, $\epsilon > 0$. This can be interpreted as generating a weighted graph, $G = (V, E)$, where vertices $V = \{1, 2, ..., pq\}$ have edges between them whenever $W_{i,j} > 0$ and with weights accordingly.

A degree matrix, corresponding to the weighted symmetric directed graph is,

$$\mathcal{D}_{i,i} = \sum_j W_{i,j}, \ \mathcal{D}_{i,j} = 0, i \neq j, \tag{A.5}$$

Shi and Malik [170] noted that the max-cut is equivalent to,

$$\min_x ncut(x) = \min_y \frac{y^T(\mathcal{D} - W)y}{y^T\mathcal{D}y}, \tag{A.6}$$

where $ncut(x)$ is the normalized cut of the partition.

As can be proved through the Courant-Fischer theorem, [24], and [170] for the image segmentation setting. This then brings us to the generalized eigenvalue eigenvector problem,

$$(\mathcal{D} - W)y = \lambda\mathcal{D}y \tag{A.7}$$

where the second smallest eigenvalue and corresponding eigenvector solve the optimization problem. This could be written in terms of a symmetric normalized graph Laplacian, L, by noting that Eq. A.7 transforms into,

$$\mathcal{D}^{-1/2}(\mathcal{D} - W)\mathcal{D}^{-1/2}x = \lambda x, \tag{A.8}$$

or,

$$Lx = \lambda x, \tag{A.9}$$

if,

$$L = \mathcal{D}^{-1/2}(\mathcal{D} - W)\mathcal{D}^{-1/2}, \tag{A.10}$$

by substitution,

$$y = \mathcal{D}^{-1/2}x. \tag{A.11}$$

The affinity matrix eigenvalue problem has an interpretation as a stochastic matrix eigenvalue problem, by [125, 132],

$$P = \mathcal{D}^{-1}W. \tag{A.12}$$

Meila and Shi [125] noted that the affinity matrix W relates to random walks through a graph according to this stochastic matrix P, and this relates closely to a diffusive process underlying the diffusion map method, [137, 162].

Now the smallest eigenvalue of Eq. A.7 corresponds to the greedy partition (one element of the A-B partition is empty) so the second smallest eigenvalue corresponds to the Cheeger-balanced

partition, the best bi-partition. Then one could proceed by recursively bi-partitioning [126]. We follow the concept of [144] which is to choose the k smallest eigenvalues *after the zero eigenvalue* and corresponding eigenvectors and then to cluster these by use of k-means from there.

Appendix B

Misconceptions In Sparse System Identification

B.1 History of Carleman Linearization

There is a large confusion in literature about Carleman linearization, or also known as Carleman embedding. In this section we review the time line of Carleman linearization.

Kowalski and Steeb (1991), discussed and reviewed in their book introduction [113], the history and details on Carleman linearization in literature from the first publishing 1932 [40], to 1991, the year of their book publishing. We can summarize their findings as the following:

- From 1932 to 1970 (38 years), there was only one paper that mentioned the Carleman embedding technique in an exercise.
- From 1970 to 1981, there was 7 papers discussed Carleman embedding technique, where some of them did not reference Carleman, and one of them was re-discovery of Carleman Linearization.
- From 1982 - 1991, there was 9 papers discussed Carleman embedding technique, where some of them introduced extensions and reformulation of the technique.

We see that during around 60 years from Carleman embedding technique, only 17 papers was related to Carleman 1932, with many of them represent a reformulation and re-discovery of the technique. As consequence, Carleman linearization technique lost its identity, and start to appear in literature without being referenced as such. The reasons can be also the difficulty

of the original formulation, and the lake of discussion for the simple cases such as writing a quadratic function of two variables in the form:

$$f(x, y) = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 x^2 + \beta_4 xy + \beta_5 y^2. \tag{B.1}$$

In the following, we list few selected papers that are related to our work, where the power polynomial expansion adopted on them:

1. Epureanu and Dowell (1997) [63], introduced a technique to compute the linearized Poincaré map and the sensitivity vector required for the implementation of an Ott-Grebogi-Yorke (OGY) controller from experimental data.
2. Candes et al. (2005) [38], Introduced the compressive sensing using L_1 relaxation.
3. Yao and Bollt (2007) [205], developed a new method that adapts the least-squares best approximation by a Kronecker product representation and Carleman linearization to analyze underlying structure when it exists as a system of coupled oscillators.
4. Wang et al. [195], introduced a method for Predicting catastrophes in nonlinear dynamical systems by compressive sensing.
5. Brunton et al. (2015) [30], Combines the iterative hardthresholding with Carleman linearization for sparse regression SINDy.

B.2 Number of Measurements Vs. Time

in Table B.1 we list number of measurements, training data time span, levels of noise, and type of noise for recent work of SINDy and TW methods, and we can clearly see the following points:

1. In all approaches, number of measurements required to recover Lorenz system is mostly of order 4 and 5, with long time span for training data generation. The lowest number of measurements was $N = 2500$ in TW [186], and it was combined with assuming the second order power polynomial expansion (which gives 10 basis function), and assuming that the noise is only on a specific continuous blocks of the observations, and all other measurements are exact.

2. In the first example of SINDy, the noise is added only to the derivative $\boldsymbol{f}$, and the measurements $\boldsymbol{z}$, and then the basis functions matrix, are exact, and this was based on the assumption that both $\boldsymbol{f}$ and $\boldsymbol{z}$ are measurable, and the noise appear only on $\boldsymbol{f}$. Although we see that there is no reasonable justification of why we may measure noisy derivative while we have exact state variables, however, we see that number of observation used in the implementation was of order 5.

3. SINDy with control (SINDYc), is designed to work in low-data regime, and we see that it used number of measurements of order 4, and a trained the data over 20 time-unit length.

4. Compared to our results we discussed in Chapter 4, ER detect Lorenz system with 1500 measurements, which implies a training time ($[0, 0.75]$), with considering large noise ($\sigma = 0.2$) present on the state measurements $\boldsymbol{z}$.

In 2007, Yao and Bollt [205], proved the convergence of the least squares method with error on the parameters of the same order as the step size. Then, even the ordinary least squares will converge to accurate solution on the parameters sense, where the sparse parameters become easy to distinguished by dropping their values below the step size. The idea about this is how many measurements that will need?!, and using huge number of measurements for relatively small system sound like a misconception on the data-driven system identification.

The misconception is not just using huge number of measurements, but also not distinguishing between "measurements" and "information" in the sense of the oscillations and behavior of the dynamic. Fig.B.1 shows how 1000 measurements look like in Lorenz system with different step size.

It will be a subject of our future work to discuss this issue in more details, and to discuss the importance of attaching the step size to the number of measurements, since one of them is not clear information without the other.

N	training time	noise σ	noise type	reference
10^5	[0 100]	1	$\boldsymbol{f}$ only	SINDy [29, 30]
5×10^4	[0 50]	0.01	$\boldsymbol{z}$	SINDy [29, 30]
2×10^4	[0 20]	NA	NA	SINDYc [101, 102]
2500	[0 2.5]	0.0125	$\boldsymbol{z}$	TW [186]
4×10^4	[0 20]	0.0125	$\boldsymbol{z}$	TW [185, 186]

TABLE B.1: Number of measurements and noise levels used in recent literature for sparse system identification of the Lorenz system.

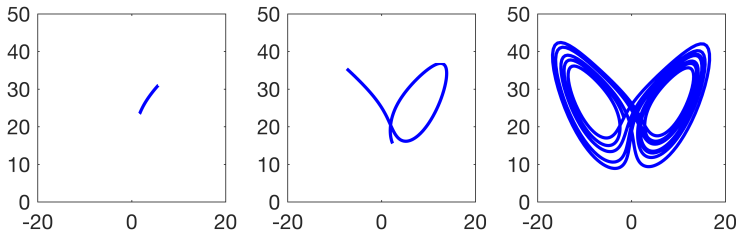

FIGURE B.1: 1000 sample points of Lorenz system. All the subplots shows 1000 measurements, however, from left to write we use $\tau = 10^{-4}, \tau = 10^{-3}$, and $\tau = 10^{-2}$ as a step size respectively. The oscillations and long dynamic behavior, especially in chaotic systems, represent unique feature which become easier to detect and estimated. However, in sparse system identification, we have embedded trade off (that seems to be absent in literature) between the accuracy of the sparse structure and the accuracy of the magnitude of the parameters. The accurate estimation of the sparse structure in the current state of the art methods requires more and more "dynamic behavior", which means "time" of sampling. In the other hand the accurate values of the parameters highly improved with smaller step size as proved by Yao and Bollt [205]. The trade off is then; increasing the step size to recover the sparse structure with lower measurements, comes with the cost of lower accuracy in parameters magnitude, and reducing the step size and to have higher accuracy in the parameters magnitude comes with the cost of large increase of number of measurements and computations complexity. And both cases have restrictions and bounds for the physical, experimental, and numerical feasibility of the step size from being too small or too large.

B.3 On Noise Addition

Traditionally, we used to write the linear system of equations, with the linear algebra in mind, as:

$$y = Ax + \eta. \tag{B.2}$$

where y is the dependent variable (or output), A is the independent variables (or input), x is the vector of parameters, and η is a vector of noise.

Naturally, after the modern matrix formulation of linearization of nonlinear dynamic appeared, it simply follow the linear algebra formulation, and written as:

$$\dot{z} = \Phi(z)\beta + \eta \tag{B.3}$$

where $\dot{z}$ is the vector field of the observed dynamic, z is the state variable, $\Phi(z)$ is the basis expansion of the state variable, β is the vector of parameters, and η is the noise.

However, the Eq.B.3 implies that the noise only appear in vector field $\dot{z}$ and the state variable is exact, and that is not true for most of system identification problems, and we discussed in Chapter 4 that the correct form is:

$$\dot{z} = \Phi(z + \eta_1)\beta + \eta_2 + \varepsilon \tag{B.4}$$

where $\Phi(z+\eta_1)$ is the basis expansion for the noisy state variable, η_2 is the noise in the vector field that can be seen as amplification function that depends on η_1, z, and the derivative estimation method, and ε is the computations error.

B.4 Number of Basis Functions

The sentence: "Unfortunately, L_0 minimization is NP-Hard", is found in most sparse regression literature, including our work. However, we see sometimes a work that attempt to recover relatively low dimensional systems (such as Lorenz) by assuming low expansion order. For example, in [186], the authors assumed the second order power polynomial expansion for the

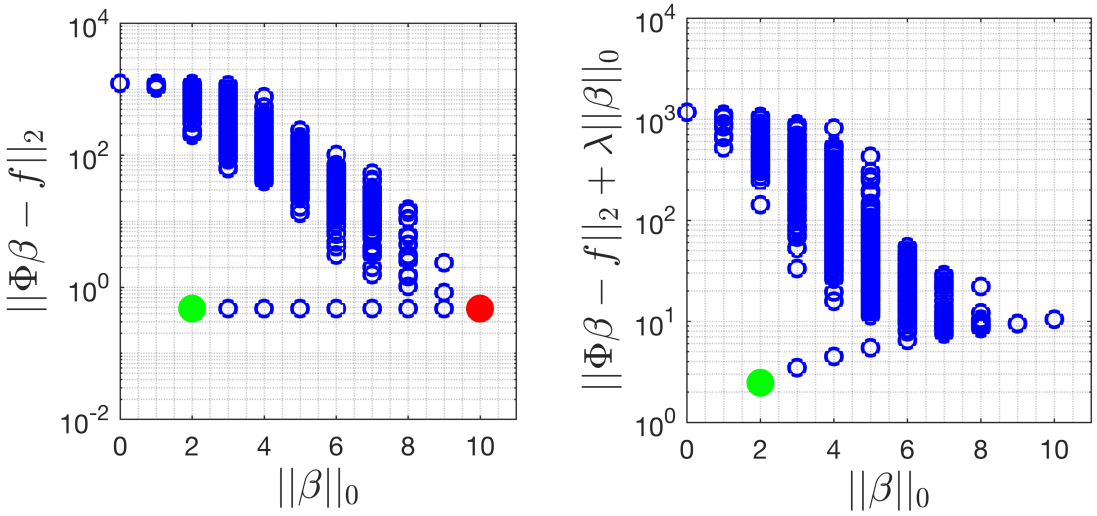

FIGURE B.2: Exhaustive search and the effect of penalty in L_0 minimization for Lorenz system, with s^{nd} order power polynomial expansion, 0.0005 step size, and 400 measurements. (Left) for 10 candidate functions Lorenz system $\boldsymbol{f} = \Phi\beta$, the blue markers shows $\|\boldsymbol{f} - \Phi\boldsymbol{\beta}\|_2$ for all possible combination of basis, the green marker shows the true solution, and the red marker shows the solution with minimum value. (Right) The value of objective function after adding a penalty term $\|\boldsymbol{f} - \Phi\boldsymbol{\beta}\|_2 + \lambda\|\boldsymbol{\beta}\|_0$. We assume $\lambda = 1$.

3-dimensional Lorenz system, which result with 10 candidate functions, and with step size 0.0005, it required 40,000 measurements to recover the sparse structure.

Fig.B.2 shows how all possible solutions of the system can easily reveal the true structure of the system with very low number of measurements. The point here is that assuming low number of candidate function contradict with the principle of the black box modeling, as discussed in Chapter 2, and the reasonable approach to assume higher order expansion.

In general, for robust and exact recovery of sparse parameters, exhaustive search can give robust solution with low measurements and reasonable time up to 10 dimensions, and it can keep the same robustness but with more expensive computations up to 13 or 14 dimension. Mixed integer programming may extended this robustness up to 30-40 dimensions. However, it is finally an NP-Hard problem.

Then, the robustness, efficiency, and applicability of any sparse regression problem, should be tested and evaluated with high dimensional systems.

Bibliography

[1] I Mezic A. Fabregat A.C. Poje. "Finite-time Partitions for Lagrangian Structure Identification in Gulf Stream Eddy Transport". In: *arxiv, 1606.07382* (2016).

[2] R Basnayake A. Luttman E.M. Bollt and S Kramer. "A Stream Function Approach to Optical Flow with Applications to Fluid Transport Dynamics". In: *Proceedings in Applied Mathematics and Mechanics* 11.1 (2012), pp. 855–856.

[3] Mehmet Eren Ahsen and Mathukumalli Vidyasagar. "Two new approaches to compressed sensing exhibiting both robust sparse recovery and the grouping effect". In: *2017 Indian Control Conference, ICC 2017 - Proceedings*. 2017, pp. 246–250. ISBN: 9781509017959. DOI: 10.1109/INDIANCC.2017.7846482.

[4] A.Lasota and J.A.Yorke. "Exact dynamical systems and the Frobenius-Perron operator". In: *Trans.Amer.Math.Soc.* 273 (1982), pp. 375–384.

[5] A.A.R. AlMomani and E. Bollt. "Go With the Flow, on Jupiter and Snow. Coherence from Model-Free Video Data Without Trajectories". In: *Journal of Nonlinear Science* (2018). ISSN: 14321467. DOI: 10.1007/s00332-018-9470-1.

[6] Abd AlRahman R. AlMomani, Jie Sun, and Erik Bollt. "How Entropic Regression Beats the Outliers Problem in Nonlinear System Identification". In: *Submitted.* (2019). URL: https://arxiv.org/abs/1905.08061.

[7] C. Anteneodo, A. M. Batista, and R. L. Viana. "Synchronization threshold in coupled logistic map lattices". In: *Physica D: Nonlinear Phenomena* (2006). ISSN: 01672789. DOI: 10.1016/j.physd.2006.10.001.

[8] Roberto Artuso et al. "Classical and Quantum Chaos". In: *Chaos* (2002). ISSN: 1063-651X.

[9] S. Arya and D. M. Mount. "Approximate Range Searching". In: *Computational Geometry: Theory and Applications* 17 (2000), pp. 135–163.

[10] Y Bai et al. *Computational methods for applied inverse problems*. Vol. 56. Walter de Gruyter, 2012.

[11] Afonso S Bandeira et al. "Certifying the restricted isometry property is hard". In: *IEEE Transactions on Information Theory* 59.6 (2013), pp. 3448–3450. ISSN: 00189448. DOI: 10.1109/TIT.2013.2248414.

[12] R G Baraniuk. "Compressive Sensing". In: *IEEE Signal Processing Magazine* 24.4 (2007), pp. 118–121. ISSN: 10535888. DOI: 10.1109/MSP.2007.4286571. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4286571.

[13] Adam B Barrett, Lionel Barnett, and Anil K Seth. "Multivariate Granger causality and generalized variance". In: *Physical Review E* 81.4 (2010), p. 041907.

[14] R. Basnayake and E.M. Bollt. "A Multi-Time Step Method to Compute Optical Flow with Scientific Priors for Observations of a Fluidic System". In: *BIRS Book Chapter, Springer Proceedings in Mathematics and Statistics* 70 (2014), pp. 59–79.

[15] Massimo Bassan. "Advanced interferometers and the search for gravitational waves". In: *Astrophysics and Space Science Library* 404 (2014), pp. 275–290.

[16] Arieh Ben-Naim. *A Farewell to Entropy: Statistical Thermodynamics Based on Information: S*. World Scientific, 2008.

[17] Jacques Bernoulli. "1713, Ars conjectandi, opus posthumum". In: *Accedit Tractatus de seriebus infinitis, et epistola gallice scripta de ludo pilae reticularis (Thurneysen Brothers, Basel, Switzerland)* (2005).

[18] SA Billings, MJ Korenberg, and S Chen. "Identification of non-linear output-affine systems using an orthogonal least-squares algorithm". In: *International Journal of Systems Science* 19.8 (1988), pp. 1559–1568.

[19] Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2008.

[20] Jeffrey D. Blanchard, Coralia Cartis, and Jared Tanner. "Compressed sensing: how sharp is the restricted isometry property". In: *CoRR* abs/1004.5.0602219 (2010), p. 21. ISSN: 0036-1445. DOI: 10.1137/090748160. URL: http://arxiv.org/abs/1004.5026.

[21] Thomas Blumensath. "Accelerated iterative hard thresholding". In: *Signal Processing* 92.3 (2012), pp. 752–756.

[22] Thomas Blumensath and Mike E Davies. "Iterative hard thresholding for compressed sensing". In: *Applied and computational harmonic analysis* 27.3 (2009), pp. 265–274.

[23] Thomas Blumensath and Mike E Davies. "Normalized iterative hard thresholding: Guaranteed stability and performance". In: *IEEE Journal of selected topics in signal processing* 4.2 (2010), pp. 298–309.

[24] E.M. Bollt and N. Santitissadeekorn. "Applied and Computational Measurable Dynamics". In: *Society for Industrial and Applied Mathematics* (2013).

[25] Erik M Bollt. "Model Selection, Confidence, and Scaling in Predicting Chaotic Time-Series." In: *International Journal of Bifurcation and Chaos (IJBC) in Applied Sciences and Engineering* 10.6 (2000), pp. 1407–1422.

[26] Dietrich Braess. "Nonlinear approximation theory". In: (1986).

[27] Joseph L. Breeden and Alfred Hbler. "Reconstructing equations of motion from experimental data with unobserved variables". In: *Physical Review A* 42.10 (1990), pp. 5817–5826. ISSN: 10502947. DOI: 10.1103/PhysRevA.42.5817.

[28] Reggie Brown, Nikolai F. Rulkov, and Eugene R. Tracy. "Modeling and synchronizing chaotic systems from time-series data". In: *Physical Review E* 49.5 (1994), pp. 3784–3800. ISSN: 1063651X. DOI: 10.1103/PhysRevE.49.3784.

[29] *Brunton*. https://faculty.washington.edu/kutz/page26/. SINDy: Matlab SINDy code base. Kutz Research Group Website, Open source code. Accessed: 2019-07-05.

[30] Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. "Discovering governing equations from data: Sparse identification of nonlinear dynamical systems". In: *arXiv* 1.609 (2015), pp. 1–26. ISSN: 0027-8424. DOI: 10.1073/pnas.1517384113. URL: http://arxiv.org/abs/1509.03580.

[31] Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. "Sparse identification of nonlinear dynamics with control (SINDYc)". In: *IFAC-PapersOnLine* 49.18 (2016), pp. 710–715.

[32] Stephen G Brush, Stephen G Brush, et al. *A history of modern planetary physics: nebulous Earth*. Vol. 1. Cambridge University Press, 1996.

[33] Robert P Bukata et al. *Optical properties and remote sensing of inland and coastal waters*. CRC press, 2018.

[34] Paul Butzer and François Jongmans. "PL Chebyshev (1821–1894): A guide to his life and work". In: *Journal of approximation theory* 96.1 (1999), pp. 111–138.

[35] E.J. Candes and M.B. Wakin. "An Introduction To Compressive Sampling". In: *IEEE Signal Processing Magazine* 25.2 (2008), pp. 21–30. ISSN: 1053-5888. DOI: 10.1109/MSP.2007.914731.

[36] Emmanuel J. Candès. "Compressive sampling". In: *Proceedings of the International Congress of Mathematicians* (2006), pp. 1433–1452. ISSN: 1095-9114. DOI: 10.4171/022-3/69.

[37] Emmanuel J Candès. "The restricted isometry property and its implications for compressed sensing". In: *Comptes Rendus Mathematique* 346.9-10 (2008), pp. 589–592. ISSN: 1631073X. DOI: 10.1016/j.crma.2008.03.014.

[38] Emmanuel J. Candès, Justin Romberg, and Terence Tao. "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information". In: *IEEE Transactions on Information Theory* 52.2 (2006), pp. 489–509. ISSN: 00189448. DOI: 10.1109/TIT.2005.862083.

[39] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. "Stable signal recovery from incomplete and inaccurate measurements". In: *Communications on Pure and Applied Mathematics* 59.8 (2006), pp. 1207–1223. ISSN: 00103640. DOI: 10.1002/cpa.20124. arXiv: 0503066v2 [math].

[40] Torsten Carleman. "Application de la théorie des équations intégrales linéaires aux systèmes d'équations différentielles non linéaires". In: *Acta Math.* 59 (1932), pp. 63–87. DOI: 10.1007/BF02546499. URL: http://dx.doi.org/10.1007/BF02546499.

[41] George Casella and Roger L Berger. *Statistical Inference*. 2002. ISBN: 0-534-24312-6. DOI: 10.1057/pt.2010.23.

[42] K Steiglitz C.H. Papadimitriou. "6.1 The Max-Flow, Min-Cut Theorem". In: *Combinatorial Optimization: Algorithms and Complexity. Dover.* (1998), pp. 120–128.

[43] S. Chen, S. a. Billings, and W. Luo. "Orthogonal least squares methods and their application to non-linear system identification". In: *International Journal of Control* 50.769892610 (1989), pp. 1873–1896. ISSN: 0020-7179. DOI: 10.1080/00207178908953472.

[44] Yilun Chen, Yuantao Gu, and Alfred O. Hero. "Sparse LMS for system identification". In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*. 2009. ISBN: 9781424423545. DOI: 10.1109/ICASSP.2009.4960286.

[45] F Christiansen, P Cvitanovi??, and V Putkaradze. "Spatiotemporal chaos in terms of unstable recurrent patterns". In: *Nonlinearity* (1997). ISSN: 09517715. DOI: 10.1088/0951-7715/10/1/004.

[46] F. Chung. "Laplacians and the Cheeger inequality for directed graphs". In: *Annals of Combinatorics* 9 (2005), pp. 1–19.

[47] *Climate change is supercharging a hot and dangerous summer*. https://www.washingtonpost.com/national/health-science/climate-change-is-supercharging-a-hot-and-dangerous-summer/2018/07/26/cf960ba8-905c-11e8-bcd5-9d911c784c38_story.html?utm_term=.b31c4f9a60a4. Accessed: 2019-06-27.

[48] Alison J Cook and David G Vaughan. "Overview of areal changes of the ice shelves on the Antarctic Peninsula over the past 50 years." In: *The cryosphere.* 4.1 (2010), pp. 77–98.

[49] Shane F Cotter, Kenneth Kreutz-Delgado, and Bhaskar D Rao. "Backward sequential elimination for sparse vector subset selection". In: *Signal Processing* 81.9 (2001), pp. 1849–1864.

[50] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. 2005. ISBN: 9780471241959. DOI: 10.1002/047174882X.

[51] Philip C Curtis et al. "n-parameter families and best approximation." In: *Pacific Journal of Mathematics* 9.4 (1959), pp. 1013–1027.

[52] Germund Dahlquist and A Bjork. "Equidistant interpolation and the Runge phenomenon". In: *Numerical Methods* (1974), pp. 101–103.

[53] PS De Laplace. "Mémoire sur la probabilité des causes par les événements". In: *Mém. de math. et phys. présentés à l'Acad. roy. des sci* 6 (1774), pp. 621–656.

[54] Andreas Denner, Oliver Junge, and Daniel Matthes. "Computing coherent sets using the Fokker-Planck equation". In: *arXiv preprint arXiv:1512.03761* (2015).

[55] Tom Dietterich. "Overfitting and undercomputing in machine learning". In: *ACM Computing Surveys* (1995). ISSN: 03600300. DOI: 10.1145/212094.212114.

[56] LEE DO Q. *Numerically efficient methods for solving least squares problems*. 2012.

[57] D L Donoho. "Compressed sensing". In: *IEEE Transactions on Information Theory* 52.4 (2006), pp. 1289–1306. ISSN: 00189448. DOI: Doi10.1109/Tit.2006.871582.

[58] D. L. Donoho and M. Elad. "Optimally sparse representation in general (nonorthogonal) dictionaries via 1 minimization". In: *Proceedings of the National Academy of Sciences* (2003). ISSN: 0027-8424. DOI: 10.1073/pnas.0437847100.

[59] David Leigh Donoho et al. *Sparse solution of underdetermined linear equations by stagewise orthogonal matching pursuit*. Department of Statistics, Stanford University, 2006.

[60] Edward Dougherty. *Mathematical morphology in image processing*. CRC press, 2018.

[61] Susanna S Epp. *Discrete mathematics with applications*. Cengage learning, 2010.

[62] Bogdan I Epureanu and Earl H Dowell. "System identification for the Ott-Grebogi-Yorke controller design". In: *Phys. Rev. E* 56.5 (Nov. 1997). URL: https://journals.aps.org/pre/pdf/10.1103/PhysRevE.56.5327.

[63] Bogdan I. Epureanu and Earl H. Dowell. "System identification for the Ott-Grebogi-Yorke controller design". In: *Phys. Rev. E* 56 (5 Nov. 1997), pp. 5327–5331. DOI: 10.1103/PhysRevE.56.5327. URL: https://link.aps.org/doi/10.1103/PhysRevE.56.5327.

[64] Paul Erdös and Eri Jabotinsky. "On analytic iteration". In: *Journal d'Analyse Mathematique* 8.1 (1960), pp. 361–376.

[65] *ESR: LARSEN C CRACK INTERFEROGRAM*. https://m.esa.int/spaceinimages/Images/2017/04/Larsen-C_crack_interferogram. Accessed: 2019-07-09.

[66] M. Falorni. "The discovery of the Great Red Spot of Jupiter". In: *Journal of the British Astronomical Association* 97 (4) (1987), pp. 215–219.

[67] J Doyne Farmer and John J Sidorowich. "Exploiting chaos to predict the future and reduce noise". In: *Evolution, Learning, and Cognition, World Scientific Press* (1988), p. 277.

[68] J. Doyne Farmer and John J. Sidorowich. "Predicting chaotic time series". In: *Physical Review Letters* 59.8 (1987), pp. 845–848. ISSN: 00319007. DOI: 10.1103/PhysRevLett.59.845.

[69] Willliam Feller. *An introduction to probability theory and its applications*. Vol. 2. John Wiley & Sons, 2008.

[70] Fabio Florindo and Pontus Lurcock. "Antarctic Climate History and Global Climate Changes". In: (2017).

[71] Simon Foucart. "Hard thresholding pursuit: an algorithm for compressive sensing". In: *SIAM Journal on Numerical Analysis* 49.6 (2011), pp. 2543–2563.

[72] Gary Froyland and Kathrin Padberg. "Almost-invariant sets and invariant manifolds - Connecting probabilistic and geometric descriptions of coherent structures in flows". In: *Physica D: Nonlinear Phenomena* 238.16 (2009), pp. 1507–1523. ISSN: 01672789. DOI: 10.1016/j.physd.2009.03.002.

[73] G. Froyland S. Lloyd and N Santitissadeekorn. "Coherent sets for nonautonomous dynamical systems". In: *Physica D* 239 (2010), pp. 1527–1541.

[74] ME Gagne, NP Gillett, and JC Fyfe. "Observed and simulated changes in Antarctic sea ice extent over the past 50 years". In: *Geophysical Research Letters* 42.1 (2015), pp. 90–95.

[75] Massimiliano Giona, Fabrizio Lentini, and Valerio Cimagalli. "Functional reconstruction and local prediction of chaotic time series". In: *Physical Review A* 44.6 (1991), pp. 3496–3502. ISSN: 10502947. DOI: 10.1103/PhysRevA.44.3496.

[76] ColetteM Girard. *Processing of remote sensing data*. Routledge, 2018.

[77] N. F. Glasser et al. "Surface structure and stability of the Larsen C ice shelf, Antarctic Peninsula". In: *Journal of Glaciology* (2009). ISSN: 00221430. DOI: 10.3189/002214309788816597.

[78] G H Golub and C F V Loan. *Matrix Computations*. 1996. ISBN: 0801854148. DOI: 10.1063/1.3060478.

[79] Gene Golub. "Numerical methods for solving linear least squares problems". In: *Numerische Mathematik* 7.3 (1965), pp. 206–216.

[80] Gene H Golub and Charles F Van Loan. "An analysis of the total least squares problem". In: *SIAM journal on numerical analysis* 17.6 (1980), pp. 883–893.

[81] Gene H Golub and Charles F Van Loan. *Matrix Computations (4th Ed.)* Baltimore, MD, USA: Johns Hopkins University Press, 2013. ISBN: 1-4214-0859-7.

[82] G Gouesbet. "Reconstruction of the vector fields of continuous dynamical systems from numerical scalar time series". In: *Phys. Rev. A* 43.10 (May 1991), pp. 5321–5331. DOI: 10.1103/PhysRevA.43.5321. URL: http://link.aps.org/doi/10.1103/PhysRevA.43.5321.

[83] C W J Granger. "Investigating Causal Relations by Econometric Models and Cross-spectral Methods". In: *Econometrica* 37.3 (1969), pp. 424–438. ISSN: 00129682. DOI:

10.2307/1912791. URL: http://links.jstor.org/sici?sici=0012-9682(196908)37:3%3C424:ICRBEM%3E2.0.CO;2-L%5Cnhttp://www.jstor.org/stable/1912791.

[84] Armin Grunwald. "Modes of orientation provided by futures studies: making sense of diversity and divergence". In: *European Journal of Futures Research* (2014). ISSN: 2195-4194. DOI: 10.1007/s40309-013-0030-5.

[85] Ryan N Gutenkunst et al. "Universally sloppy parameter sensitivities in systems biology models". In: *PLoS computational biology* 3.10 (2007), e189.

[86] Alireza Hadjighasem et al. "Spectral-clustering approach to Lagrangian vortex detection". In: *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* 93.6 (2016). ISSN: 15502376. DOI: 10.1103/PhysRevE.93.063107. arXiv: 1506.02258.

[87] Anders Hald. *A history of probability and statistics and their applications before 1750*. Vol. 501. John Wiley and Sons, 2003.

[88] George Haller. "Finding finite-time invariant manifolds in two-dimensional velocity fields." In: *Chaos* 10.1 (2000), pp. 99–108. ISSN: 1089-7682. DOI: 10.1063/1.166479. URL: http://www.ncbi.nlm.nih.gov/pubmed/12779366.

[89] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. "Statistical Learning with Sparsity: The Lasso and Generalizations". In: *Crc* (2015). ISSN: 0306-7734. DOI: 10.1201/b18401-1.

[90] R. H G Helleman and E. W. Montroll. "On a nonlinear perturbation theory without secular terms. I. Classical coupled anharmonic oscillators". In: *Physica* 74.1 (1974), pp. 22–74. ISSN: 00318914. DOI: 10.1016/0031-8914(74)90183-9.

[91] Edwin Hewitt and Robert E Hewitt. "The Gibbs-Wilbraham phenomenon: an episode in Fourier analysis". In: *Archive for history of Exact Sciences* 21.2 (1979), pp. 129–160.

[92] P C Hohenberg and Boris I Shraiman. "Chaotic behavior of an extended system". In: *Physica D: Nonlinear Phenomena* (1989). ISSN: 01672789. DOI: 10.1016/0167-2789(89)90121-8.

[93] Yi Hao Hsiao et al. "Real-world underwater fish recognition and identification, using sparse representation". In: *Ecological Informatics* (2014). ISSN: 15749541. DOI: 10.1016/j.ecoinf.2013.10.002.

[94] James M Hyman and Basil Nicolaenko. "The Kuramoto-Sivashinsky equation: A bridge between PDE'S and dynamical systems". In: *Physica D: Nonlinear Phenomena* (1986). ISSN: 01672789. DOI: 10.1016/0167-2789(86)90166-1.

[95] I.S. Dhillon Y. Guan and B Kulis. "Kernel k-means: spectral clustering and normalized cuts". In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.* (2004), pp. 551–556.

[96] Prateek Jain, Ambuj Tewari, and Purushottam Kar. "On Iterative Hard Thresholding Methods for High-dimensional M-estimation". In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1.* NIPS'14. Cambridge, MA, USA: MIT Press, 2014, pp. 685–693. URL: http://dl.acm.org/citation.cfm?id=2968826.2968903.

[97] D. Jansen et al. "Present stability of the Larsen C ice shelf, Antarctic Peninsula". In: *Journal of Glaciology* (2010). ISSN: 00221430. DOI: 10.3189/002214310793146223.

[98] John R Jensen and Kalmesh Lulla. "Introductory digital image processing: a remote sensing perspective". In: (1987).

[99] M S Jolly, I G Kevrekidis, and E S Titi. "Approximate inertial manifolds for the Kuramoto-Sivashinsky equation: Analysis and computations". In: *Physica D: Nonlinear Phenomena* (1990). ISSN: 01672789. DOI: 10.1016/0167-2789(90)90046-R.

[100] M S Jolly, R Rosa, and R Temam. *Accurate Computations on Inertial Manifolds.* 2001. DOI: 10.1137/S1064827599351738.

[101] *Kaiser.* https://github.com/eurika-kaiser/SINDY-MPC. Matlab: SINDy Model Predictive Control by Erika Kaiser. Accessed: 2019-07-05.

[102] Eurika Kaiser, J Nathan Kutz, and Steven L Brunton. "Sparse identification of nonlinear dynamics for model predictive control in the low-data limit". In: *Proceedings of the Royal Society A* 474.2219 (2018), p. 20180335.

[103] Nicholas Kalouptsidis et al. "Adaptive algorithms for sparse system identification". In: *Signal Processing* (2011). ISSN: 01651684. DOI: 10.1016/j.sigpro.2011.02.013.

[104] Kunihiko Kaneko. "Overview of coupled map lattices." In: *Chaos (Woodbury, N.Y.)* (1992). ISSN: 1089-7682. DOI: 10.1063/1.165869.

[105] Ravi Kannan, Santosh Vempala, and Adrian Vetta. "On Clusterings: Good, Bad and Spectral". In: *Journal of the ACM* 51.3 (2004), pp. 497–515. ISSN: 00045411. DOI: 10.1145/990308.990313. URL: http://dl.acm.org/citation.cfm?id=990313{\%}5Cnhttp://portal.acm.org/citation.cfm?doid=990308.990313.

[106] Tapas Kanungo et al. "An efficient k-means clustering algorithm: analysis and implementation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.7

(2002), pp. 881–892. ISSN: 0162-8828. DOI: 10.1109/TPAMI.2002.1017616. arXiv: arXiv:0711.0189v1.

[107] F Kaspar and H G Schuster. "Easily calculable measure for the complexity of spatiotemporal patterns". In: *Physical Review A* (1987). ISSN: 10502947. DOI: 10.1103/PhysRevA.36.842.

[108] M Korenberg et al. "Orthogonal parameter estimation algorithm for non-linear stochastic systems". In: *International Journal of Control* (1988). ISSN: 13665820. DOI: 10.1080/00207178808906169.

[109] Eric J. Kostelich and James A. Yorke. "Noise reduction in dynamical systems". In: *Physical Review A* 38.3 (1988), pp. 1649–1652. ISSN: 10502947. DOI: 10.1103/PhysRevA.38.1649.

[110] Krzysztof Kowalski and W-H Steeb. *Nonlinear dynamical systems and Carleman linearization*. World Scientific, 1991.

[111] LF Kozachenko and Nikolai N Leonenko. "Sample estimate of the entropy of a random vector". In: *Problemy Peredachi Informatsii* 23.2 (1987), pp. 9–16.

[112] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. "Estimating mutual information". In: *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* (2004). ISSN: 1063651X. DOI: 10.1103/PhysRevE.69.066138.

[113] Kowalski Krzysztof and Steeb Willi-hans. *Nonlinear Dynamical Systems and Carleman Linearization*. World Scientific, 1991.

[114] Y Kuramoto and T Tsuzuki. "Persistent Propagation of Concentration Waves in Dissipative Media Far from Thermal Equilibrium". In: *Progress of Theoretical Physics* (1976). ISSN: 0033-068X. DOI: 10.1143/PTP.55.356.

[115] Yoshiki Kuramoto. "Diffusion-Induced Chaos in Reaction Systems". In: *Progress of Theoretical Physics Supplement* (1978). ISSN: 0375-9687. DOI: 10.1143/PTPS.64.346.

[116] Claus Lämmerzahl, CW Francis Everitt, and Friedrich W Hehl. *Gyros, Clocks, Interferometers...: Testing Relativistic Gravity in Space*. Vol. 562. Springer Science & Business Media, 2001.

[117] Yueheng Lan and Predrag Cvitanović. "Unstable recurrent patterns in Kuramoto-Sivashinsky dynamics". In: *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics* (2008). ISSN: 15393755. DOI: 10.1103/PhysRevE.78.026208.

[118] Pierre L'ecuyer. "Pseudorandom Number Generators". In: *Encyclopedia of Quantitative Finance*. American Cancer Society, 2010. ISBN: 9780470061602. DOI: 10.1002/9780470061602.eqf13003. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9780470061602.eqf13003. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470061602.eqf13003.

[119] Pierre L'Ecuyer. "Uniform Random Number Generators". In: *International Encyclopedia of Statistical Science*. Ed. by Miodrag Lovric. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1625–1630. ISBN: 978-3-642-04898-2. DOI: 10.1007/978-3-642-04898-2_602. URL: https://doi.org/10.1007/978-3-642-04898-2_602.

[120] Lennart Ljung. "Perspectives on system identification". In: *Annual Reviews in Control*. 2010. ISBN: 9783902661005. DOI: 10.1016/j.arcontrol.2009.12.001.

[121] Edward N Lorenz. "Deterministic Nonperiodic Flow". In: *Journal of the Atmospheric Sciences* (1963). ISSN: 0022-4928. DOI: 10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2.

[122] Aaron Luttman et al. "A framework for estimating potential fluid flow from digital imagery." In: *Chaos (Woodbury, N.Y.)* 23.3 (2013), p. 033134. ISSN: 1089-7682. DOI: 10.1063/1.4821188. URL: http://www.ncbi.nlm.nih.gov/pubmed/24089970.

[123] AM Lyapunov. "Collected Works, vol. 3". In: *Akad. Nauk SSSR, Moscow* (1959).

[124] I Mezic M. Budisic R. M. Mohr. "Applied Koopmanism". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 22 (2012).

[125] J Shi M. Meila. "Learning segmentation by random walks". In: *Neural Information Processing Systems* 13 (2001).

[126] Tian Ma and Erik M. Bollt. "Relatively Coherent Sets as a Hierarchical Partition Method ". In: *International Journal of Bifurcations and Chaos* 23.7 (2013), p. 1330026.

[127] John C Mairhuber. "On Haar's theorem concerning Chebychev approximation problems having unique solutions". In: *Proceedings of the American Mathematical Society* 7.4 (1956), pp. 609–615.

[128] Niall M Mangan et al. "Model selection for dynamical systems via sparse regression and information criteria". In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 473.2204 (2017), p. 20170009.

[129] Andrey Andreyevich Markov. "Extension of the limit theorems of probability theory to a sum of variables connected in a chain". In: *Dynamic probabilistic systems* 1 (1971), pp. 552–577.

[130] C Masoller, Hugo L.D.de S Cavalcante, and J R.Rios Leite. "Delayed coupling of logistic maps". In: *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics* (2001). ISSN: 1063651X. DOI: 10.1103/PhysRevE.64.037202.

[131] *MEaSUREs InSAR-Based Antarctica Ice Velocity Map, Version 2*. https://nsidc.org/data/nsidc-0484/versions/2{#}data. Accessed: 2018-09-17.

[132] M. Meila et al. "A random walks view of spectral segmentation". In: *AI and Statistics (AISTATS)* 57 (2001), p. 5287. DOI: 10.1.1.33.1501. URL: http://scholar.google.com/scholar?hl=en{\%}7B{\&}{\%}7DbtnG=Search{\%}7B{\&}{\%}7Dq=intitle:A+Random+Walks+View+of+Spectral+Segmentation{\%}7B{\#}{\%}7D0{\%}5Cnhttp://scholar.google.com/scholar?hl=en{\&}btnG=Search{\&}q=intitle:A+Random+Walks+View+of+Spectral+Segmentation{\#}0.

[133] Matthias Mengel et al. "Future sea level rise constrained by observations and long-term commitment". In: *Proceedings of the National Academy of Sciences* (2016). ISSN: 0027-8424. DOI: 10.1073/pnas.1500515113. eprint: https://www.pnas.org/content/early/2016/02/17/1500515113.full.pdf. URL: https://www.pnas.org/content/early/2016/02/17/1500515113.

[134] *MODIS Antarctic Ice Shelf Image Archive*. http://nsidc.org/data/iceshelves_images/cgi-bin/modis_iceshelf_archive.pl. Accessed: 2018-09-17.

[135] M Mori. "On the convergence of the spectrum of Perron-Frobenius operators". In: *Tokyo J. Math.* 17.1-19 (1994).

[136] Jeremie Mouginot, Bernd Scheuchl, and Eric Rignot. "Mapping of Ice Motion in Antarctica Using Synthetic-Aperture Radar Data". In: *Remote Sensing* 4.9 (Sept. 2012), 2753–2767. ISSN: 2072-4292. DOI: 10.3390/rs4092753. URL: http://dx.doi.org/10.3390/rs4092753.

[137] Boaz Nadler et al. "Diffusion Maps, Spectral Clustering and Eigenfunctions of Fokker-Planck operators". In: *Applied and Computational Harmonic Analysis* 21 (2005), pp. 5–30. ISSN: 10635203. DOI: 10.1016/j.acha.2006.04.006. URL: http://arxiv.org/abs/math/0506090.

[138] NASA. "http://saturn.jpl.nasa.gov/mission/quickfacts/". In: (). URL: http://saturn.jpl.nasa.gov/mission/quickfacts/.

[139] NASA. "http://www.nasa.gov/mission_pages/cassini/multimedia/pia04866.html". In: (). URL: http://www.nasa.gov/mission_pages/cassini/multimedia/pia04866.html.

[140] NASA. "JPL, NASA : https://photojournal.jpl.nasa.gov/catalog/PIA02863". In: (). URL: https://photojournal.jpl.nasa.gov/catalog/PIA02863.

[141] Richard E Neapolitan. *Probabilistic reasoning in expert systems: theory and algorithms*. CreateSpace Independent Publishing Platform, 2012.

[142] Deanna Needell and Roman Vershynin. "Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit". In: *Foundations of Computational Mathematics* 9.3 (2009), pp. 317–334. ISSN: 16153375. DOI: 10.1007/s10208-008-9031-3.

[143] John von Neumann. "Various techniques used in connection with random digits". In: *John von Neumann, Collected Works* 5 (1963), pp. 768–770.

[144] Andrew Y Ng, Michael I Jordan, and Yair Weiss. "On spectral clustering: Analysis and an algorithm". In: *Advances in neural information processing systems* 2 (2002), pp. 849–856. ISSN: 1049-5258. DOI: 10.1.1.19.8100.

[145] Henrik Aalborg Nielsen and Henrik Madsen. "Modelling the heat consumption in district heating systems using a grey-box approach". In: *Energy and Buildings* 38.1 (2006), pp. 63–71.

[146] Henrik Aalborg Nielsen, Henrik Madsen, et al. *Predicting the heat consumption in district heating systems using meteorological forecasts*. Citeseer, 2000.

[147] Shipra Ojha and Sachin Sakhare. "Image processing techniques for object tracking in video surveillance-A survey". In: *2015 International Conference on Pervasive Computing (ICPC)*. IEEE. 2015, pp. 1–6.

[148] *Online Open source*. https://en.wikipedia.org/wiki/Chebyshev_polynomials. Wikipedia: Chebyshev polynomials. Accessed: 2019-06-27.

[149] *Online open source*. https://en.wikipedia.org/wiki/Legendre_polynomials. Wikipedia: Legendre polynomials. Accessed: 2019-06-27.

[150] K. Onu, F. Huhn, and G. Haller. "LCS Tool: A computational platform for Lagrangian coherent structures". In: *Journal of Computational Science* 7 (2015), pp. 26–36. ISSN: 18777503. DOI: 10.1016/j.jocs.2014.12.002. arXiv: arXiv:1406.3527v1.

[151] Yagyensh Chandra Pati, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad. "Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition". In: *Proceedings of 27th Asilomar conference on signals, systems and computers*. IEEE. 1993, pp. 40–44.

[152] Dilcia Pérez and Yamilet Quintana. "A survey on the Weierstrass approximation theorem". In: *Divulgaciones Matemáticas* 16.1 (2008), pp. 231–247. URL: http://emis.ams.org/journals/DM/v16-1/art14.pdf.

[153] Luc Pronzalo and Eric Walter. *Identification of parametric models from experimental data*. 1997. ISBN: 3540761195.

[154] Markus Quade et al. "Sparse identification of nonlinear dynamics for rapid model recovery". In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 28.6 (2018), p. 063116.

[155] S Ramdani et al. "Slow manifolds of some chaotic systems with applications to laser systems". In: *International Journal of Bifurcation and Chaos in Applied Sciences and Engineering* (2000). ISSN: 02181274.

[156] Jeffrey Lee Rasmussen. "Evaluating outlier identification tests: Mahalanobis D squared and Comrey Dk". In: *Multivariate Behavioral Research* 23.2 (1988), pp. 189–202.

[157] E. Rignot, J. Mouginot, and B. Scheuchl. "Ice Flow of the Antarctic Ice Sheet". In: *Science* 333.6048 (2011), pp. 1427–1430. ISSN: 0036-8075. DOI: 10.1126/science.1208336. eprint: https://science.sciencemag.org/content/333/6048/1427.full.pdf. URL: https://science.sciencemag.org/content/333/6048/1427.

[158] *Rignot, E., J. Mouginot, and B. Scheuchl. 2017. MEaSUREs InSAR-Based Antarctica Ice Velocity Map, Version 2. [subset:2006-2011]. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center.* https://nsidc.org/data/nsidc-0484/versions/2. Accessed: 2018-09-17.

[159] Jorma Rissanen. *Stochastic Complexity and Modeling*. 1986. DOI: 10.1214/aos/1176350051. URL: http://www.jstor.org/stable/3035559%5Cnhttp://www.jstor.org/stable/pdfplus/10.2307/3035559.pdf?acceptTC=true.

[160] J C Robinson. *Infinite-Dimensional Dynamical Systems: An Introduction to Dissipative Parabolic PDEs and the Theory of Global Attractors*. Cambridge Texts in Applied Mathematics. Cambridge University Press, 2001. ISBN: 9780521635646. URL: https://books.google.com/books?id=zviPUpYUCWEC.

[161] Guo Rong-Wei and U. E Vincent. "Control of a Unified Chaotic System via Single Variable Feedback". In: *Chinese Physics Letters* 26.9 (Sept. 2009), p. 090506. DOI: 10.1088/0256-307x/26/9/090506. URL: https://doi.org/10.1088%2F0256-307x%2F26%2F9%2F090506.

[162] S Lafon R.R. Coifman. "Diffusion maps". In: *Appl. Comput. Harmon. Anal.* 21 (2006), pp. 5–30.

[163] Carl Runge. "Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten". In: *Zeitschrift für Mathematik und Physik* 46.224-243 (1901), p. 20.

[164] Barry Saltzman. "Finite Amplitude Free Convection as an Initial Value Problem—I". In: *Journal of the Atmospheric Sciences* (1962). ISSN: 0022-4928. DOI: 10.1175/1520-0469(1962)019<0329:FAFCAA>2.0.CO;2.

[165] Hiroki Sayama. *Introduction to the Modeling and Analysis of Complex Systems*. SUNY Binghamton. SUNY Open Textbooks, 2015. ISBN: 9781942341062. URL: http://textbooks.opensuny.org/introduction-to-the-modeling-and-analysis-of-complex-systems/.

[166] T Schreiber. "Measuring information transfer". In: *Physical review letters* 85.2 (2000), pp. 461–4. ISSN: 1079-7114. DOI: 10.1103/PhysRevLett.85.461. URL: http://www.ncbi.nlm.nih.gov/pubmed/10991308.

[167] Eugene Seneta. "The central limit problem and linear least squares in pre-revolutionary Russia: The background". In: *Mathematical scientist* 9 (1984), pp. 37–77.

[168] Claude E Shannon. "A mathematical theory of communication". In: *The Bell System Technical Journal* 27.July 1928 (1948), pp. 379–423. ISSN: 07246811. DOI: 10.1145/584091.584093. URL: http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf.

[169] Oscar Sheynin. *History of Statistics*. NG Verlag Berlin, 2012. ISBN: 978-3-942944-20-5.

[170] Jianbo Shi and Jitendra Malik. "Normalized cuts and image segmentation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.8 (2000), pp. 888–905. ISSN: 01628828. DOI: 10.1109/34.868688. arXiv: 0703101v1 [cs].

[171] Frank Y Shih. *Image processing and mathematical morphology: fundamentals and applications*. CRC press, 2017.

[172] Dragos Simandan. "Wisdom and foresight in Chinese thought: Sensing the immediate future". In: *Journal of Futures Studies* (2018). ISSN: 10276084. DOI: 10.6531/JFS.2018.22(3).00A35.

[173] G I Sivashinsky. "Nonlinear analysis of hydrodynamic instability in laminar flames-I. Derivation of basic equations". In: *Acta Astronautica* (1977). ISSN: 00945765. DOI: 10.1016/0094-5765(77)90096-0.

[174] Vladimir Ivanovich Smirnov. "Biography of AM Lyapunov". In: *International journal of control* 55.3 (1992), pp. 775–784.

[175] Suvrit Sra, Sebastian Nowozin, and Stephen J Wright. *Optimization for machine learning*. Mit Press, 2012.

[176] W-H Steeb and F Wilhelm. "Non-linear autonomous systems of differential equations and Carleman linearization procedure". In: *Journal of Mathematical Analysis and Applications* 77.2 (1980), pp. 601–611.

[177] Eric J Steig et al. "Warming of the Antarctic ice-sheet surface since the 1957 International Geophysical Year". In: *Nature* 457.7228 (2009), p. 459.

[178] Stephen M Stigler. "Gauss and the Invention of Least Squares". In: *Ann. Statist.* 9.3 (1981), pp. 465–474. DOI: 10.1214/aos/1176345451. URL: http://dx.doi.org/10.1214/aos/1176345451.

[179] Marshall H Stone. "The generalized Weierstrass approximation theorem". In: *Mathematics Magazine* 21.5 (1948), pp. 237–254.

[180] J Sun, D Taylor, and E Bollt. "Causal Network Inference by Optimal Causation Entropy". In: *SIAM Journal on Applied Dynamical Systems* 14.1 (2015), pp. 73–106. DOI: 10.1137/140956166. URL: https://doi.org/10.1137/140956166.

[181] Jie Sun, Carlo Cafaro, and Erik M Bollt. "Identifying the coupling structure in complex systems through the optimal causation entropy principle". In: *Entropy* (2014). ISSN: 10994300. DOI: 10.3390/e16063416.

[182] Robert Tibshirani et al. "Sparsity and smoothness via the fused lasso". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.1 (2005), pp. 91–108.

[183] Andrei Nikolaevich Tikhonov et al. *Numerical methods for the solution of ill-posed problems*. Vol. 328. Springer Science & Business Media, 2013.

[184] Andreas M Tillmann and Marc E Pfetsch. "The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing". In: *IEEE Transactions on Information Theory* 60.2 (2014), pp. 1248–1259. ISSN: 00189448. DOI: 10.1109/TIT.2013.2290112.

[185] *Tran.* https://github.com/GiangTTran/ExactRecoveryChaoticSystems. Matlab code: Exact Recovery of Chaotic Systems, by Giang Tran. Accessed: 2019-07-05.

[186] Giang Tran and Rachel Ward. "Exact Recovery of Chaotic Systems from Highly Corrupted Data". In: *Multiscale Modeling & Simulation* 15 (2017), pp. 1108–1129.

[187] S Vaegler et al. "SU-D-116-01: A Novel Reconstuction Framework of Prior Image Constrained Compressed Sensing (PICCS) Enabling the Use of Prior Images with Major Deviations". In: *Medical Physics*. 2013. ISBN: 0094-2405. DOI: 10.1118/1.4814054.

[188] Sven Vaegler et al. "Incorporation of local dependent reliability information into the Prior Image Constrained Compressed Sensing (PICCS) reconstruction algorithm". In: *Zeitschrift fur Medizinische Physik* (2015). ISSN: 18764436. DOI: 10.1016/j.zemedi.2015.09.002.

[189] Palghat P. Vaidyanathan and Piya Pal. "System identification with sparse coprime sensing". In: *IEEE Signal Processing Letters* (2010). ISSN: 10709908. DOI: 10.1109/LSP.2010.2060331.

[190] Jaap Van Brakel. "Some remarks on the prehistory of the concept of statistical probability". In: *Archive for history of exact sciences* 16.2 (1976), pp. 119–136.

[191] Martin Vejmelka and Milan Paluš. "Inferring the directionality of coupling with conditional mutual information". In: *Physical Review E* 77.2 (2008), p. 026214.

[192] Helen M Walker. "Degrees of Freedom". In: *Journal of Educational Psychology* (1940). ISSN: 0022-0663. DOI: 10.1037/h0054588.

[193] Liang Wang and Reza Langari. "Building Sugeno-type models using fuzzy discretization and orthogonal parameter estimation techniques". In: *IEEE Transactions on Fuzzy Systems* 3.4 (1995), pp. 454–458.

[194] Wen-Xu Wang, Ying-Cheng Lai, and Celso Grebogi. "Data based identification and prediction of nonlinear and complex dynamical systems". In: *Physics Reports* 644 (2016), pp. 1–76. ISSN: 0370-1573. DOI: http://dx.doi.org/10.1016/j.physrep.

2016.06.004. URL: http://www.sciencedirect.com/science/article/pii/S037015731630134X.

[195] Wen Xu Wang et al. "Predicting catastrophes in nonlinear dynamical systems by compressive sensing". In: *Physical Review Letters* 106.15 (2011). ISSN: 00319007. DOI: 10.1103/PhysRevLett.106.154101.

[196] Joshua J Waterfall et al. "Sloppy-model universality class and the Vandermonde matrix". In: *Physical review letters* 97.15 (2006), p. 150601.

[197] Karl Weierstrass. "Uber continuirliche functionen eines reellen arguments, die fur keinen worth des letzteren einen bestimmten differentailqutienten besitzen, Akademievortrag". In: *Math. Werke* (1872), pp. 71–74.

[198] Karl Weierstrass. "Uber die analytische Darstellbarkeit sogenannter willkurlicher Functionen einer reellen Veranderlichen". In: *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin* 2 (1885), pp. 633–639. URL: https://www.math.auckland.ac.nz/hat/fpapers/wei4.pdf.

[199] Henry Wilbraham. "On a certain periodic function". In: *The Cambridge and Dublin Mathematical Journal* 3 (1848), pp. 198–201.

[200] Matthew O. Williams, Ioannis G. Kevrekidis, and Clarence W. Rowley. "A Data-Driven Approximation of the Koopman Operator: Extending Dynamic Mode Decomposition". In: *Journal of Nonlinear Science* 25.6 (2015), pp. 1307–1346. ISSN: 14321467. DOI: 10.1007/s00332-015-9258-5. arXiv: 1312.3019.

[201] Edward O Wilson. *Consilience: The unity of knowledge*. Vol. 31. Vintage, 1999.

[202] Edwin Bidwell Wilson. "First and second laws of error". In: *Journal of the American Statistical Association* 18.143 (1923), pp. 841–851.

[203] JWT Wimpenny. "The validity of models". In: *Advances in dental research* 11.1 (1997), pp. 150–159.

[204] Ji-Woong Yang et al. "Surface temperature in twentieth century at the Styx Glacier, northern Victoria Land, Antarctica, from borehole thermometry". In: *Geophysical Research Letters* 45.18 (2018), pp. 9834–9842.

[205] Chen Yao and Erik M. Bollt. "Modeling and nonlinear parameter estimation with Kronecker product representation for coupled oscillators and spatiotemporal systems". In: *Physica D: Nonlinear Phenomena* 227.1 (2007), pp. 78–99. ISSN: 01672789. DOI: 10.1016/j.physd.2006.12.006.

[206] Smilka Zdravkovska and Peter L Duren. *Golden years of Moscow mathematics*. American Mathematical Soc., 2007.

[207] Long Zhang and Kang Li. "Forward and backward least angle regression for nonlinear system identification". In: *Automatica* 53 (2015), pp. 94–102.

[208] Long Zhang et al. "Two-stage orthogonal least squares methods for neural network construction". In: *IEEE transactions on neural networks and learning systems* 26.8 (2014), pp. 1608–1621.

[209] Tong Zhang. "Adaptive forward-backward greedy algorithm for learning sparse representations". In: *IEEE transactions on information theory* 57.7 (2011), pp. 4689–4708.