ABSTRACT

CONTINUUM MODELING OF ACTIVE NEMATICS VIA DATA-DRIVEN EQUATION DISCOVERY

by
Connor Robertson

Data-driven modeling seeks to extract a parsimonious model for a physical system directly from measurement data. One of the most interpretable of these methods is Sparse Identification of Nonlinear Dynamics (SINDy), which selects a relatively sparse linear combination of model terms from a large set of (possibly nonlinear) candidates via optimization. This technique has shown promise for synthetic data generated by numerical simulations but the application of the techniques to real data is less developed. This dissertation applies SINDy to video data from a bio-inspired system of mictrotubule-motor protein assemblies, an example of nonequilibrium dynamics that has posed a significant modelling challenge for more than a decade. In particular, we constrain SINDy to discover a partial differential equation (PDE) model that approximates the time evolution of microtubule orientation. The discovered model is relatively simple but reproduces many of the characteristics of the experimental data. The properties of the discovered PDE model are explored through stability analysis and numerical simulation; it is then compared to previously proposed models in the literature.

Chapter 1 provides an introduction and motivation for pursuing a data driven modeling approach for active nematic systems by introducing the Sparse Identification of Nonlinear Dynamics (SINDy) modeling procedure and active nematic systems. Chapter 2 lays the foundation for modeling of active nematics to better understand the model space that is searched. Chapter 3 gives some preliminary considerations for using the SINDy algorithm and proposes several approaches to mitigate common errors. Chapter 4 treats the example problem of rediscovering a governing partial

differential equation for active nematics from simulated data including some of the specific challenges that arise for discovery even in the absence of noise. Chapter 5 details the procedure for extracting data from experimental observations for use with the SINDy procedure and details tests to validate the accuracy of the extracted data. Chapter 6 presents the active nematic model extracted from experimental data via SINDy, compares its properties with previously proposed models, and provides numerical results of its simulation. Finally, Chapter 7 presents conclusions from the work and provides future directions for both active nematic systems and data-driven modeling in related systems.

CONTINUUM MODELING OF ACTIVE NEMATICS VIA DATA-DRIVEN EQUATION DISCOVERY

by
Connor Robertson

A Dissertation
Submitted to the Faculty of
New Jersey Institute of Technology and
Rutgers, The State University of New Jersey – Newark
in Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy in Mathematical Sciences

Department of Mathematical Sciences
Department of Mathematics and Computer Science, Rutgers-Newark

May 2023

APPROVAL PAGE

CONTINUUM MODELING OF ACTIVE NEMATICS VIA DATA-DRIVEN EQUATION DISCOVERY

Connor Robertson

Dr. Anand Oza, Dissertation Co-Advisor Date
Assistant Professor of Mathematical Sciences, NJIT

Dr. Travis Askham, Dissertation Co-Advisor Date
Assistant Professor of Mathematical Sciences, NJIT

Dr. David Shirokoff, Committee Member Date
Associate Professor of Mathematical Sciences, NJIT

Dr. Zuofeng Shang, Committee Member Date
Assistant Professor of Mathematical Sciences, NJIT

Dr. Jörn Dunkel, Committee Member Date
Professor of Mathematics, Massachussetts Institute of Technology, Boston, MA

BIOGRAPHICAL SKETCH

Author: Connor Robertson

Degree: Doctor of Philosophy

Date: May 2023

Undergraduate and Graduate Education:

- Doctor of Philosophy in Mathematical Sciences,
  New Jersey Institute of Technology, Newark, NJ, 2023
- Bachelor of Science in Applied and Computational Mathematics,
  Brigham Young University, Provo, UT, 2018

Major: Mathematical Sciences

Publications:

Connor Robertson, Jared Wilmoth, Scott Retterer, and Miguel Fuentes-Cabrera, "Video frame prediction of microbial growth with a recurrent neural network". *Front. Microbiol. Systems Microbiology* 05 January 2023

Jonathan Sakkos, Joe Weaver, Connor Robertson, Bowen Li, Denis Taniguchi, Ketan Maheshwari, Daniel Ducat, Paolo Zuliani, Andrew Stephen McGough, Tom Curtis, and Miguel Fuentes-Cabrera, "Investigating the growth of an engineered strain of Cyanobacteria with an Agent-Based Model and a Recurrent Neural Network". *bioRxiv* 12 October 2021

Tyler Jarvis, Jordan Clough, Jane Cox, Konnor Peterson, Mitchell Sailsbury, Connor Robertson, Tyler Moncur, Katie Palmer, and Darren Lund, "Using survey data and mathematical modeling to prioritize water interventions in developing countries". *Water Resources Management* 13 January 2021

Presentations:

Connor Robertson, Anand Oza, and Travis Askham, "Data-driven continuum modeling of active nematics via sparse identification of nonlinear dynamics" *Society for Industrial and Applied Mathematics Conference on Computational Science and Engineering*, Amsterdam, Netherlands, February 2023.

Connor Robertson, Anand Oza, and Travis Askham, “Data-driven continuum modeling of active nematics via sparse identification of nonlinear dynamics” *Annual Meeting of the American Physical Society Division of Fluid Dynamics*, Indianapolis, Indiana, November 2022.

Connor Robertson, Anand Oza, and Travis Askham, “Data-driven continuum modeling of active nematics via sparse identification of nonlinear dynamics” *Annual Meeting of the American Physical Society*, Chicago, Illinois, March 2022.

Connor Robertson, “Neural networks for function approximation and data-driven modeling” *Machine Learning and Optimization Seminar - Department of Mathematical Sciences NJIT*, Newark, New Jersey, October 2021.

Connor Robertson and Mitchell Sailsbury, “Facility location using Markov Chains” *College of Physical and Mathematical Sciences Student Research Conference - Brigham Young University*, Provo, Utah, March 2018.

Connor Robertson and Mitchel Sailsbury, “Efficiency of water distribution in water poor areas of the world” *Student Days - Annual Meeting of the Society for Applied and Computational Mathematics*, Pittsburgh, Pennsylvania, July 2017.

Posters:

Connor Robertson, “Continuum modeling of active nematics via data-driven equation discovery” *Dana Knox Student Research Showcase*, New Jersey Institute of Technology, April, 2023.

Connor Robertson, “Data-driven discovery of PDEs for active nematic systems” *National Academy of Inventors - NJIT Chapter Workshop*, New Jersey Institute of Technology, October, 2022.

Connor Robertson, “Discovering governing equations of an active nematic system using PDE-Find” *Gesellschaft fur Angewandte Mathematik und Mechanik Juniors’ Summer School*, Magdeburg, Germany (virtual), May 2020.

For Indigo and Robin, who are just beginning to discover.

ACKNOWLEDGMENT

I have been the recipient of an enormous amount of support and guidance through the process of researching for this dissertation.

I would like to first thank my extremely patient advisors Anand Oza and Travis Askham, who have provided their direction, perspectives, opportunities, and even tutoring to help me bring this project to fruition. They have spent countless hours working alongside me and providing direction when I was lost. Even when errors were discovered or results were in question, they stuck with me and were encouraging and supportive voices in the chaos. Thank you for your friendship and interest in my success.

My committee members Dave Shirokoff, Zuofeng Shang, and Jörn Dunkel have also provided invaluable insights in all of our interactions, and I'm grateful for the time and effort they have put into reviewing and evaluating my work. Dr. Shirokoff and Dr. Shang's numerical experience has been the inspiration for several new approaches I tried, and Dr. Dunkel has been an invaluable aid in navigating data-driven modeling for active nematics.

I would like to thank Matthew Golden, Prof. Roman Grigoriev, Chaitanya Joshi, Prof. Aparna Baskaran, and Prof. Michael Hagan for discussions that helped to shape the direction of this work. I'd also like to thank Rohit Supekar and Dan Messenger, who each took time to talk through challenges in data-driven modeling on experimental data with me and whose input was very helpful.

I am thankful to the Department of Mathematical Sciences and most especially the administrative staff for financial and personal support to accomplish my goals. They were always considerate of my requests and looked for ways to provide the growth and opportunity that I needed.

I'd like to thank all of my fellow graduate students for their friendship and conversations which made the work bearable. I'd especially like to thank Erli Wind-Andersen for his mentorship and friendship. He spent more time listening to my wild ideas and providing great feedback than I could ask of anybody. Axel Turnquist and Binan Gu provided additional mentorship and advice as I navigated the program, and I am grateful for them. I'd like to thank Kosuke Sugita, Diego Rios, Yasser Almoteri, and Subhrashish Chakraborty for being amazing classmates and friends. To put it bluntly, their support and friendship have kept me sane and smiling.

I owe a special debt of gratitude to Soheil Saghafi, who has been a stalwart companion and friend in the program. He has been helpful at every milestone, and our conversations inspired me to try for more than I had planned.

I'd like to acknowledge that it was my good friend Mitchell Sailsbery who convinced me that pursuing a PhD was the right choice for me. I put up a big fight, but he was persistent and convincing, and I deeply appreciate his effort. The path he helped me get on has led me here, and I know I wouldn't have had the same opportunities or growth without his influence.

So many thanks to my parents, Kent and Tania Robertson, who didn't blink an eye when I told them I would be forgoing comfortable employment to pursue research and additional schooling. They have been unconditionally encouraging and loving and have spared no effort in helping my family thrive while we have been here.

Finally, I'd like to thank my wife, Heidi Robertson, who has sacrificed more than anyone else (including myself) to make this happen. She has listened to my complaints, patiently counseled me when I was burdened with challenges, carried the weight of our family when I was absorbed in the work, and encouraged me at every single step of the process. Her contribution to this work is the most important and the most valuable to me.

TABLE OF CONTENTS

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

TABLE OF CONTENTS
(Continued)

Chapter **Page**

LIST OF TABLES

LIST OF TABLES (Continued)

Table **Page**

LIST OF FIGURES

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

LIST OF FIGURES (Continued)

Figure **Page**

LIST OF FIGURES (Continued)

Figure **Page**

LIST OF FIGURES
(Continued)

Figure **Page**

CHAPTER 1

INTRODUCTION

1.1 Background

Fitting a mathematical model to data is ubiquitous in the sciences and enables analytical machinery to be applied to the study of physical phenomena. Common terms for the practice include “statistical inference,” “system identification,” “inverse problems,” and “data-driven modeling.” For systems in which a continuum approximation for many interacting particles is appropriate, partial differential equations (PDEs) provide parsimonious models and powerful analytical tools. While fitting a PDE of known form to data is a classical problem, the discovery of a PDE of unknown form has only recently received significant attention. This is largely due to the fact that the process is sensitive to noise in the data and the space of possible models is large. Owing to recent computational and algorithmic advances and the availability of larger and higher quality data sets, sparse regression methods show significant promise for the discovery of PDEs of unknown form from data.

There has also been interest from the biology and physics communities in the field of “active matter,” which seeks to extend the framework of statistical mechanics to incorporate non-equilibrium phenomena. A celebrated example is the so-called *active nematic* system studied by Zvonimir Dogic and colleagues, which consists of rod-like filaments called microtubules coupled by motor proteins called kinesins [16, 46, 75]. The motor proteins consume energy and thus exert extensile forces on the microtubules, leading to self-organized collective motion. When the concentration of microtubules is relatively high, the microtubule-kinesin bundles form an active nematic phase characterized by dynamic creation and annihilation of topological defects. Many phenomenological models for this system have been proposed, some of which are inspired by the theory of liquid crystals [21,34,35,67,89,90] and others by kinetic theory from statistical mechanics [31,32,96]. However, it has proven difficult to determine which, if any, of these models provides the most accurate description for the available data.

This dissertation outlines the ideas behind current data-driven modeling strategies and the results of applying these strategies to the modeling of the active nematic system. The final outcome of this work is the construction of a relatively simple PDE model that reproduces key qualitative and quantitative features of the experimental data.

1.2 Sparse Identification of Nonlinear Dynamics

Scientific discovery often builds on the fundamental practice of comparing observations with proposed mathematical models. Recent improvements in data collection and computational power have enabled a more automated approach in which the form of the model is derived more directly from the data.

For the remainder of this dissertation, we will focus on the data-driven discovery method introduced for ordinary differential equations (ODEs) by Brunton et al. [9], named Sparse Identification of Nonlinear Dynamics (SINDy), and extended for PDEs by Rudy et al. [73], where it was called PDE-Find. This dissertation will use the term SINDy in reference to both the ODE and PDE formulations of the problem. This method is selected because of the interpretability of the results, which can be analyzed using PDE and simulation techniques. For an overview of other model discovery methods and the relationship of SINDy to these methods, see Appendix A.1.

We will consider the SINDy paradigm as it is applied to vector-valued data which depends on both space and time coordinates. Let our data be represented by $\boldsymbol{u}(\boldsymbol{x}, t) \in \mathbb{R}^n$, where $\boldsymbol{x} = (x_1, x_2, \ldots, x_d)$ are spatial coordinates and t is time. Suppose that we have N samples of the system $\boldsymbol{u}_\ell = \boldsymbol{u}(\boldsymbol{x}_\ell, t_\ell)$ for $\ell = 1, \ldots, N$. We then seek a model of the form:

$$\partial_t \boldsymbol{u}(\boldsymbol{x},t) = \boldsymbol{\Theta}(\boldsymbol{u},\nabla\boldsymbol{u},\ldots,\nabla^M\boldsymbol{u},\boldsymbol{x},t)\boldsymbol{\xi}\ , \tag{1.1}$$

where M is some maximum order for the spatial derivatives,

$$\begin{gathered}\boldsymbol{\Theta}(\boldsymbol{u},\nabla\boldsymbol{u},\ldots,\nabla^M\boldsymbol{u},\boldsymbol{x},t) = \\ \begin{bmatrix}\boldsymbol{f}_1(\boldsymbol{u},\nabla\boldsymbol{u},\ldots,\nabla^M\boldsymbol{u},\boldsymbol{x},t) & \boldsymbol{f}_2(\boldsymbol{u},\nabla\boldsymbol{u},\ldots,\nabla^M\boldsymbol{u},\boldsymbol{x},t) & \ldots & \boldsymbol{f}_m(\boldsymbol{u},\nabla\boldsymbol{u},\ldots,\nabla^M\boldsymbol{u},\boldsymbol{x},t)\end{bmatrix},\end{gathered} \tag{1.2}$$

and $\boldsymbol{\xi} \in \mathbb{R}^m$ are coefficients. We call $\boldsymbol{\Theta} \in \mathbb{R}^{(n\cdot N)\times m}$ the "library" of terms where each column is a (possibly nonlinear) function of the field $\boldsymbol{u}$ and its derivatives. For example, if $\boldsymbol{f}_j = \boldsymbol{u}\cdot\nabla\boldsymbol{u}$ is a nonlinear advection, as arises in the material derivative of $\boldsymbol{u}$, the ij^{th} entry of $\boldsymbol{\Theta}$ would be:

$$\boldsymbol{\Theta}_{ij}(\boldsymbol{u},\nabla\boldsymbol{u},\ldots,\nabla^M\boldsymbol{u},\boldsymbol{x},t) = [\boldsymbol{f}_j(\boldsymbol{u},\nabla\boldsymbol{u},\ldots,\nabla^M\boldsymbol{u},\boldsymbol{x},t)]_i = (\boldsymbol{u}\cdot\nabla\boldsymbol{u})_i\ .$$

Remark 1.2.1. *In a slight abuse notation, we will use $\boldsymbol{\Theta}$, without its dependence on the field and its derivatives, to refer to the $(n\cdot N)\times m$ matrix given by*

$$\boldsymbol{\Theta} = \begin{bmatrix}\boldsymbol{\Theta}(\boldsymbol{u}_1,\nabla\boldsymbol{u}_1,\ldots,\nabla^M\boldsymbol{u}_1,\boldsymbol{x}_1,t_1) \\ \vdots \\ \boldsymbol{\Theta}(\boldsymbol{u}_N,\nabla\boldsymbol{u}_N,\ldots,\nabla^M\boldsymbol{u}_N,\boldsymbol{x}_N,t_N)\end{bmatrix}.$$

Similarly, we will use $\partial_t\boldsymbol{u}$ to refer to the vector made up by stacking the values $\partial_t\boldsymbol{u}(\boldsymbol{x}_\ell,t_\ell)$ for $\ell = 1,\ldots,N$.

A visual summary of the method from [73] is shown in Figure 1.1. To summarize the process, we seek the optimal combination of nonlinear functions $\boldsymbol{f}_i$ that best correlates with the time derivative of our data, $\partial_t\boldsymbol{u}$. Ultimately, this yields a closed form PDE that models the data. This approach allows the modeler to incorporate terms inspired by known physics and directly observe their relevance. The PDE-Find methodology requires numerical derivatives of the data in time to form $\partial_t \boldsymbol{u}$, and in space to form $\boldsymbol{\Theta}$. Further discussion on numerical differentiation of data is included in Section 3.2.

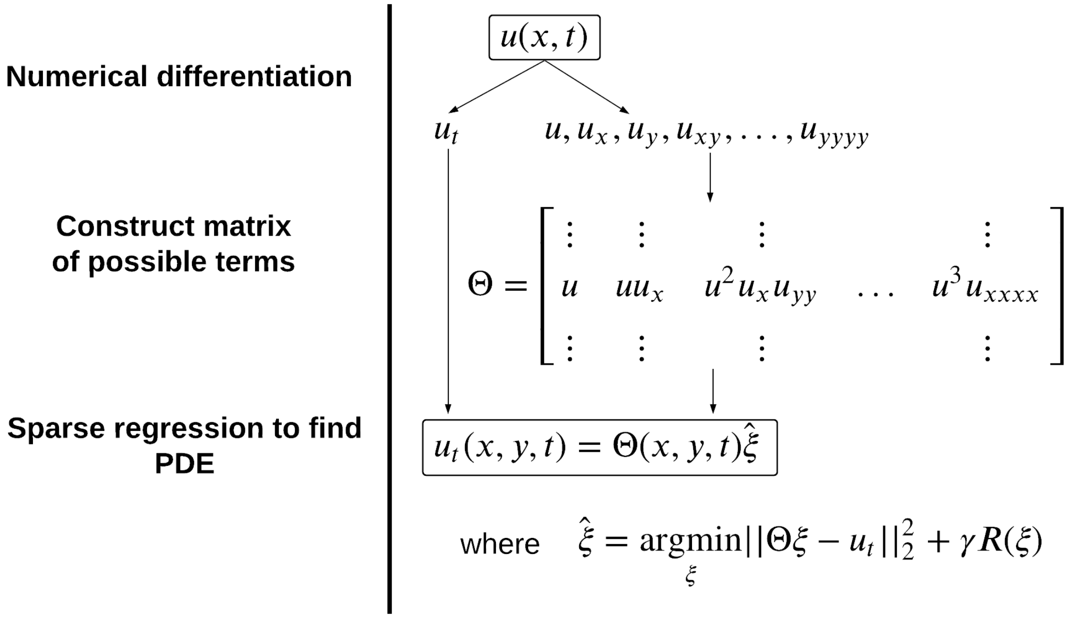

Numerical differentiation
$u(x,t)$
u_t
$u, u_x, u_y, u_{xy}, \ldots, u_{yyyy}$
Construct matrix of possible terms
$\Theta = \begin{bmatrix} \vdots & \vdots & \vdots & & \vdots \\ u & uu_x & u^2 u_x u_{yy} & \ldots & u^3 u_{xxxx} \\ \vdots & \vdots & \vdots & & \vdots \end{bmatrix}$
Sparse regression to find PDE
$u_t(x,y,t) = \Theta(x,y,t)\hat{\xi}$
where $\hat{\xi} = \underset{\xi}{\text{argmin}} ||\Theta\xi - u_t||_2^2 + \gamma \boldsymbol{R}(\xi)$

Figure 1.1 Flowchart of the PDE-Find algorithm [73].

Since we can have an arbitrary number of basis functions $\boldsymbol{f}_i$, we usually desire a sparse solution to $\boldsymbol{\xi}$, meaning that the majority of entries in $\boldsymbol{\xi}$ are zero. This can be accomplished by solving (or approximately solving) a regularized least squares problem of the form

$$\hat{\boldsymbol{\xi}} = \arg\min_{\boldsymbol{\xi}} \|\partial_t \boldsymbol{u} - \boldsymbol{\Theta}\boldsymbol{\xi}\|^2 + \gamma R(\boldsymbol{\xi}) \tag{1.3}$$

where R is a regularizing function (commonly a norm). Methods for solving this problem are discussed in Section 3.3.

The PDE-Find method has been shown to successfully recover the governing equation for simulated data with additive noise [25, 42, 73, 100] and for experimental data on colloidal microrollers [87]. However, its application to the active nematic system [36,44] brought to light several methodological hurdles that must be overcome in order to extract a successful model. Chapter 3 explores these challenges in more depth.

1.3 Microtubule-Kinesin Active Nematic System

Active nematics comprise an “active matter” system that is composed of “nematically ordered” constituents. In active matter, internal or external mechanisms contribute energy to a system, pushing it out of equilibrium [58,71]. For example, a flock of birds or a school of fish convert food into movement via chemical processes and demonstrate organized collective motion [11]. Nematic ordering is the state of having orientational order without positional order, wherein the constituents are aligned along an axis but are not arranged in a spatial pattern such as a grid. Systems with nematic order have been studied extensively in the context of liquid crystals [54], which are made up of elongated molecules suspended in a fluid and which themselves flow like a fluid. These molecules retain their orientational order either due to shape or external fields (e.g. magnetic fields). Figure 1.2 shows some observed liquid crystal phases.

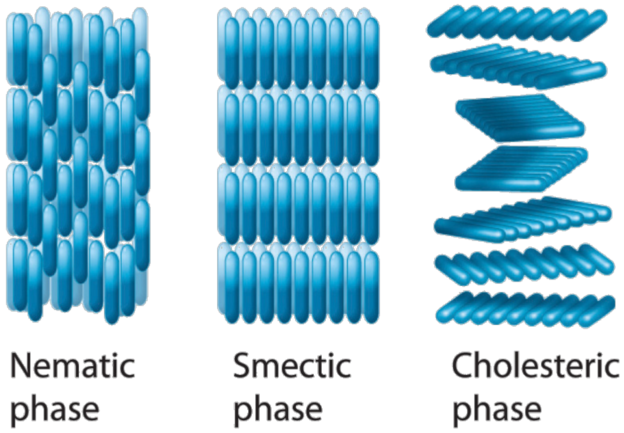


Figure 1.2 Schematic of observed liquid crystal phases. The nematic phase (leftmost) will be the focus of this dissertation.

While there have been studies of active nematic phases in elongated granular rods [1,63,81], this dissertation will focus on the biophysical system of microtubules (MTs) and motor proteins (kinesin) discovered by Zvonimir Dogic and colleagues [75]. In this system, pairs of MTs are connected via a kinesin molecule which uses adenosine triphosphate (ATP) to unbind and bind from the MTs, resulting in a motion that resembles walking. As it does, the MTs "slide" alongside one another in opposite directions. In isolation, this individual motion has little effect; however, in dense suspensions of MTs and motor proteins, bundles of MTs form and create long moving chains whose collective motion causes patterns to form at a scale much larger than the individual MTs [30]. MTs and motor proteins are both found in the cytoplasm of eukaryotic cells, and Dogic's experiments showed that these constituents can self-organize into an active liquid crystal capable of generating coherent fluid flows reminiscent of cytoplasmic streaming in cells [75].

Although active nematics share many properties with liquid crystals, the main difference is the consumption of energy by the constituent particles. This energy production causes global motion which can in turn spontaneously create or destroy discontinuities in the orientation of the molecules, breaking long range nematic ordering and causing chaotic behavior. These discontinuities are one of the defining features of active nematic systems and are known as "defects" (Figure 1.3). Defects are known to occur in liquid crystals and other materials [47,83]. Defects may be assigned a topological "charge," which denotes the angle through which the orientation field rotates along a small loop encircling the defect. In two dimensions $+1/2$ and $-1/2$ defects are observed, as shown in Figure 1.3a-c. Pairs of oppositely-charged defects may also annihilate when they meet (Figure 1.3d).

The pioneering experiments of [75] have been profitably extended by Dogic and collaborators to other configurations. Specifically, the MT-kinesin assembly was encapsulated inside a shape-changing lipid vesicle, resulting in exotic defect dynamics

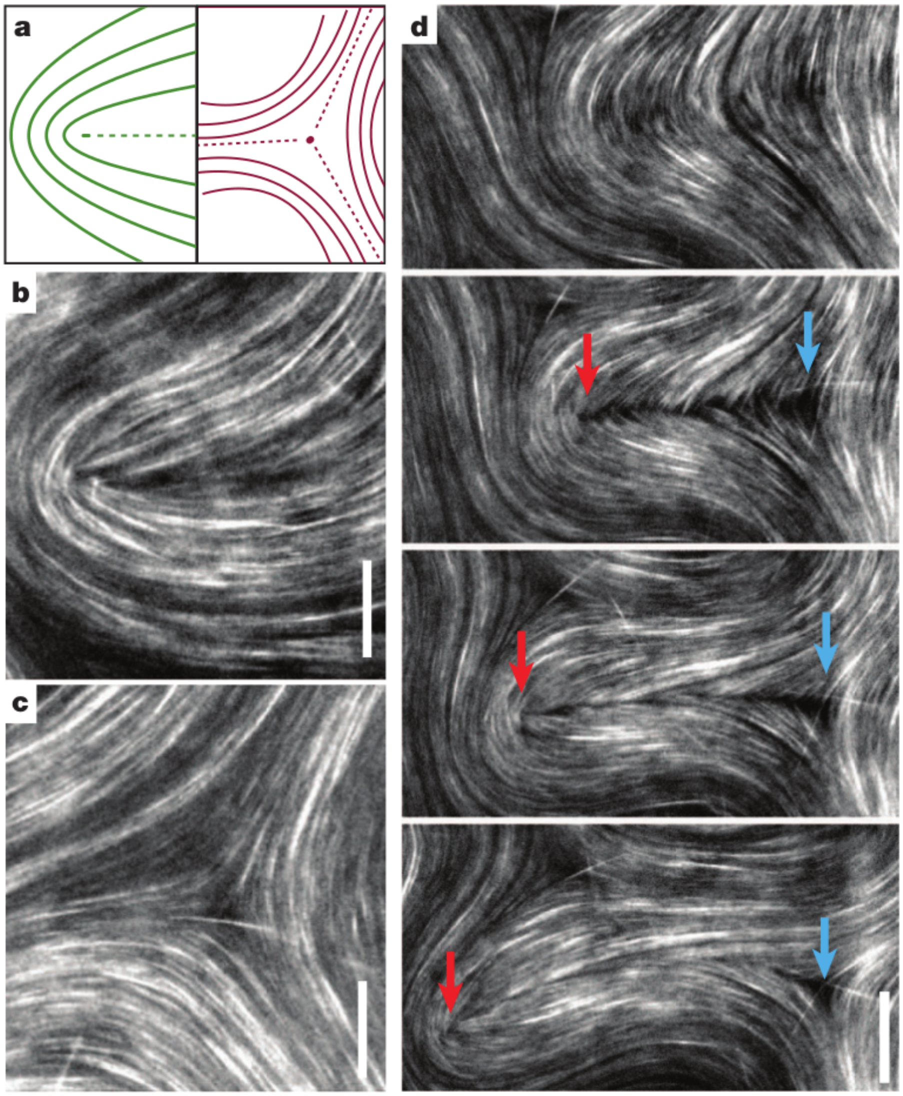


Figure 1.3 Example of the defects observed in the MT-kinesin active nematic system. (a) Diagrams of +1/2 (left) and −1/2 (right) defects. (b) +1/2 defect observed in experiment. (c) −1/2 defect observed in experiment. (d) Time sequence illustrating the creation of a positive-negative pair of defects, which is caused by the buckling of MT bundles.
Source: [75]

and the emergence of protrusions from the vescicle [46]. Defects were observed to exhibit long-range orientational order in 2D [16], and 2D confinement transformed the chaotic dynamics of MTs into regular patterns [65]. More recently, active nematics were studied in three dimensions, where it was observed that defect "loops" exhibit complex dynamics [22].

Several continuum models for this system have been proposed and are discussed in Chapter 2. The complexity of the multiscale and nonlinear interactions of the microtubules, proteins and fluid has made it difficult to arrive at a consensus for a model for this system. However, this complexity makes this system an excellent candidate for a data-driven modeling approach.

1.4 Notation

The mathematical notation used for the remainder of this dissertation will be as follows:

- Lowercase bold symbols represent vectors (order one tensors).

  Example: Vector $\boldsymbol{a}$

$$\boldsymbol{a} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \end{bmatrix}$$

- Uppercase bold symbols are tensors of order greater than one

  Example: Order 2 tensor $\boldsymbol{A}$

$$\boldsymbol{A} = \begin{bmatrix} A_{11} & A_{12} & \dots \\ A_{21} & \ddots & \\ \vdots & & \end{bmatrix}$$

- Contractions of tensors are written using Einstein summation notation

  Example: Contraction of tensor $\boldsymbol{A}$ with tensor $\boldsymbol{B}$ along dimension j

$$A_{ij}B_{jk} = \sum_j A_{ij} \cdot B_{jk}$$

- Indexing of data for tensors is written using subscripts of bold variables

  Example: Data points of tensor $\boldsymbol{A}$ at position i, j

$$\boldsymbol{A}_{ij}$$

- Gradients of tensors are derivatives along a new first dimension

  Example: Gradient of tensor $\boldsymbol{A}$

$$(\nabla \boldsymbol{A})_{kij} = \partial_k A_{ij}$$

- Divergences of tensors are derivatives along the first dimension

  Example: Divergence of tensor $\boldsymbol{A}$

$$(\nabla \cdot \boldsymbol{A})_j = \partial_i A_{ij}$$

- Tensor products represent concatenations

  Example: Concatenation of tensors $\boldsymbol{A}$ and $\boldsymbol{B}$

$$(\boldsymbol{A} \otimes \boldsymbol{B})_{ijkl} = A_{ij} B_{kl}$$

- Symmetrized terms are represented with a text superscript S

  Example: Symmetrized version of $\boldsymbol{A}$

$$\left(\boldsymbol{A}^{\mathrm{S}}\right)_{ij} = \frac{1}{2}\left(A_{ij} + A_{ji}\right)$$

- Terms that are made symmetric and traceless are represented with a text superscript ST

  Example: Symmetric and traceless version of $\boldsymbol{A}$

$$\boldsymbol{A}^{\mathrm{ST}} = \boldsymbol{A}^{\mathrm{S}} - \frac{1}{d}\mathrm{Tr}(\boldsymbol{A}^{\mathrm{S}})\boldsymbol{I}$$

  where d is the space dimension. Most of the results herein are presented for $d = 2$.

- Transpose of a second order tensor is written with a superscript $\intercal$

  Example: Transpose of tensor $\boldsymbol{A}$

  $$(\boldsymbol{A}^\intercal)_{ij} = A_{ji}$$

- Dot products of first and second order tensors are contractions over the first index

  Example: Dot product of vector $\boldsymbol{a}$ and tensor $\boldsymbol{A}$

  $$(\boldsymbol{a} \cdot \boldsymbol{A})_j = a_i A_{ij}$$

- Dot products of two second order tensors are contractions over the last and first indices respectively

  Example: Dot product of tensor $\boldsymbol{A}$ and tensor $\boldsymbol{B}$

  $$(\boldsymbol{A} \cdot \boldsymbol{B})_{ik} = A_{ij} B_{jk}$$

- Double dot products of two second order tensors are contractions over all indices

  Example: Double dot product of tensor $\boldsymbol{A}$ and tensor $\boldsymbol{B}$

  $$\boldsymbol{A} : \boldsymbol{B} = A_{ij} B_{ij}$$

- The tensor commutator is defined with square brackets

  Example: Commutator applied to $\boldsymbol{A}$ and tensor $\boldsymbol{B}$

  $$[\boldsymbol{A}, \boldsymbol{B}] = \boldsymbol{A} \cdot \boldsymbol{B} - \boldsymbol{B} \cdot \boldsymbol{A}$$

- The tensor anticommutator is defined with curly braces

  Example: Anticommutator applied to $\boldsymbol{A}$ and tensor $\boldsymbol{B}$

  $$\{\boldsymbol{A}, \boldsymbol{B}\} = \boldsymbol{A} \cdot \boldsymbol{B} + \boldsymbol{B} \cdot \boldsymbol{A}$$

- The support of a vector is the set of indices corresponding to non-zero entries, $\mathrm{supp}(\boldsymbol{a}) = \{i : a_i \neq 0\}$.

CHAPTER 2

MODELING PRELIMINARIES

As briefly outlined in Section 1.3, this work presents a data-driven modeling approach to active nematic systems using the SINDy framework. Though the method is a more automated approach than modeling from first principles, it requires identifying the most important state variables and physical constraints of the system in order to construct an over-complete library of nonlinear candidate terms. Section 2.1 introduces the Landau-de Gennes continuum theory for liquid crystals from which the microtubule-kinesin system can be approached. Section 2.2 presents an overview of previously proposed models of the system and discusses their similarities and differences. Section 2.3 utilizes the previous material in this chapter to outline a procedure for generating an overcomplete library of nonlinear tensor terms which could be included in an equation for the orientation or velocity evolution of the microtubules (MTs).

2.1 Microtubule Orientation

The celebrated Landau-de Gennes theory of liquid crystals [2,5,72,95] is a continuum field theory that describes the orientation of rod-like molecules in terms of the tensor order parameter $\boldsymbol{Q} = \boldsymbol{Q}(\boldsymbol{x}, t) \in \mathbb{R}^{d\times d}$, where $\boldsymbol{x} \in \mathbb{R}^d$. Specifically, let $\boldsymbol{p} \in \mathbb{R}^d$ be a unit vector describing the orientation of a "headless" molecule, and let $\mathbb{P}(\boldsymbol{p})$ be a probability measure on the unit sphere $\mathcal{S} = \{\boldsymbol{p} : |\boldsymbol{p}| = 1\}$. Since the molecules do not have head or tail, this probability measure must satisfy

$$\mathbb{P}(\boldsymbol{p}) = \mathbb{P}(-\boldsymbol{p}).$$

By this property, the first moment of the measure vanishes:

$$\int_{\mathcal{S}} \boldsymbol{p} \, \mathrm{d}\mathbb{P}(\boldsymbol{p}) = 0.$$

However, the second moment

$$\boldsymbol{D} = \int_{\mathcal{S}} \boldsymbol{p} \otimes \boldsymbol{p} \, \mathrm{d}\mathbb{P}(\boldsymbol{p}) \tag{2.1}$$

is generally nonzero, and is a symmetric positive semi-definite $d \times d$ matrix for which $\mathrm{Tr}(\boldsymbol{D}) = 1$. Note that, if the molecules are "isotropic" in orientation (equally oriented in all directions), then $\mathrm{d}\mathbb{P}_0(\boldsymbol{p}) = \frac{1}{|\mathcal{S}|} \mathrm{d}\boldsymbol{p}$ is the uniform measure, which yields

$$\boldsymbol{D}_0 = \frac{1}{|\mathcal{S}|} \int_{\mathcal{S}} \boldsymbol{p} \otimes \boldsymbol{p} \, \mathrm{d}\boldsymbol{p} = \frac{1}{d} \boldsymbol{I}.$$

The Landau-de Gennes tensor $\boldsymbol{Q}$ measures the deviation of $\boldsymbol{D}$ from its isotropic state:

$$\boldsymbol{Q} = \boldsymbol{D} - \boldsymbol{D}_0 = \int_{\mathcal{S}} \left(\boldsymbol{p} \otimes \boldsymbol{p} - \frac{1}{d} \boldsymbol{I} \right) \mathrm{d}\mathbb{P}(\boldsymbol{p}). \tag{2.2}$$

This "tensor order parameter" is a symmetric and traceless $d \times d$ matrix. In two dimensions ($d = 2$), $\boldsymbol{Q}$ can be written as [21]:

$$\boldsymbol{Q} = \begin{bmatrix} \lambda & \mu \\ \mu & -\lambda \end{bmatrix} = S \left(\boldsymbol{n} \otimes \boldsymbol{n} - \frac{\boldsymbol{I}}{2} \right), \tag{2.3}$$

where

$$S = 2\sqrt{\lambda^2 + \mu^2} \tag{2.4}$$

is the so-called "scalar order parameter" and

$$\boldsymbol{n} = \frac{1}{\sqrt{(\lambda + S/2)^2 + \mu^2}} \begin{bmatrix} \lambda + S/2 \\ \mu \end{bmatrix} \tag{2.5}$$

is the so-called "director," the average orientation of the rods [4]. The larger eigenvalue of $\boldsymbol{Q}$ is $S/2$, and the corresponding eigenvector is $\boldsymbol{n}$. Given the relation

$$\mathrm{Tr}(\boldsymbol{Q}^2) = S^2/2 \tag{2.6}$$

we use Equation (2.2) to deduce that

$$\begin{aligned}\frac{S^2}{2} = \operatorname{Tr}(\boldsymbol{Q}^2) &= \int_{\mathcal{S}}\int_{\mathcal{S}} \operatorname{Tr}\left[\left(\boldsymbol{p}\otimes\boldsymbol{p} - \frac{1}{2}\boldsymbol{I}\right)\left(\tilde{\boldsymbol{p}}\otimes\tilde{\boldsymbol{p}} - \frac{1}{2}\boldsymbol{I}\right)\right]\,\mathrm{d}\mathbb{P}(\boldsymbol{p})\mathrm{d}\mathbb{P}(\tilde{\boldsymbol{p}}) \\ &= \int_{\mathcal{S}}\int_{\mathcal{S}} (\boldsymbol{p}\cdot\tilde{\boldsymbol{p}})^2\,\mathrm{d}\mathbb{P}(\boldsymbol{p})\,\mathrm{d}\mathbb{P}(\tilde{\boldsymbol{p}}) - \frac{1}{2} \le \frac{1}{2}. \end{aligned} \tag{2.7}$$

That is, values of S near zero (unity) signify low (high) orientational order of the molecules. Note that $S = 0$ at a defect and $\boldsymbol{Q}$ is continuous across defects, while the director $\boldsymbol{n}$ is not.

2.2 Previously Proposed Models

Active nematic systems have drawn a range of modeling interest due to their nonequilibrium nature and their seemingly close relationship to passive liquid crystal systems. In fact, a large contingent of modeling approaches have looked to bridge the well understood dynamics of passive liquid crystals to these active systems. As a result, there exist strong similarities between previously proposed models, though it is challenging to reconcile the assumptions inherent in each model and their associated dimensionless parameters. However, all of the terms presented in these models can be included in the library used with the SINDy procedure.

2.2.1 The Beris-Edwards Model

A commonly used continuum model [14, 20, 21, 90, 102] for the MT-kinesin active nematic system is based on the Beris-Edwards equations for nematic liquid crystals [6]:

$$\frac{\partial}{\partial t}\boldsymbol{Q} + \boldsymbol{u}\cdot\nabla\boldsymbol{Q} - \boldsymbol{S} = \Gamma\boldsymbol{H}, \tag{2.8}$$

where $\boldsymbol{u}$ is the velocity of the fluid. In three dimensions ($d = 3$), the generalized co-rotation $\boldsymbol{S}$ has the form

$$\begin{aligned}\boldsymbol{S} &= (\varphi\boldsymbol{E} - \boldsymbol{\Omega})\cdot\left(\boldsymbol{Q} + \frac{\boldsymbol{I}}{3}\right) + \left(\boldsymbol{Q} + \frac{\boldsymbol{I}}{3}\right)\cdot(\varphi\boldsymbol{E} + \boldsymbol{\Omega}) - 2\varphi\left(\boldsymbol{Q} + \frac{\boldsymbol{I}}{3}\right)(\boldsymbol{Q} : \nabla\boldsymbol{u}) \\ &= [\boldsymbol{Q}, \boldsymbol{\Omega}] + 2\varphi\left(\frac{1}{3}\boldsymbol{E} - \boldsymbol{Q}(\boldsymbol{Q} : \nabla\boldsymbol{u}) + [\boldsymbol{E}\cdot\boldsymbol{Q}]^{\mathrm{ST}}\right), \end{aligned} \tag{2.9}$$

where

$$E_{ij} = \frac{1}{2}\left(\partial_i u_j + \partial_j u_i\right) \quad \text{and} \quad \Omega_{ij} = \frac{1}{2}\left(\partial_i u_j - \partial_j u_i\right) \tag{2.10}$$

are the rate of strain and vorticity tensors, respectively; and φ is the so-called *flow-alignment parameter*. We note that our definition of the vorticity in Equation (2.10) differs by a sign from that used by some authors (e.g. [21]). The molecular tensor $\boldsymbol{H}$ drives the relaxation of $\boldsymbol{Q}$:

$$\begin{aligned} \boldsymbol{H} &= -\frac{\delta \mathcal{F}}{\delta \boldsymbol{Q}} + \frac{\boldsymbol{I}}{3}\mathrm{Tr}\left(\frac{\delta \mathcal{F}}{\delta \boldsymbol{Q}}\right), \quad \text{where} \\ \mathcal{F} &= \int d\boldsymbol{x}\, \mathrm{Tr}\left(\frac{A}{2}\boldsymbol{Q}^2 + \frac{B}{3}\boldsymbol{Q}^3 + \frac{C}{4}\boldsymbol{Q}^4 + \frac{K}{2}(\nabla \boldsymbol{Q})^2\right) \end{aligned} \tag{2.11}$$

is the phenomenological Landau-de Gennes free energy. For $A < 0$ and $C > 0$, the free energy has a phenomenological double-well shape that captures the tendency of the system to depart from the isotropic state ($\boldsymbol{Q} = 0$) and converge to a nematically ordered state ($\boldsymbol{Q} \neq 0$). Distortions from the uniformly-aligned state are penalized for positive values of the elastic constant, $K > 0$.

The orientation evolution equation (2.8) is coupled with an incompressible Navier-Stokes equation:

$$\rho\left(\frac{\partial}{\partial t}\boldsymbol{u} + \boldsymbol{u}\cdot\nabla\boldsymbol{u}\right) = \nabla\cdot\boldsymbol{\sigma}, \quad \nabla\cdot\boldsymbol{u} = 0, \tag{2.12}$$

where ρ is the fluid density, and the stress tensor $\boldsymbol{\sigma} = \boldsymbol{\sigma}_{\text{viscous}} + \boldsymbol{\sigma}_{\text{elastic}} + \boldsymbol{\sigma}_{\text{active}}$ is made up of viscous, elastic and active contributions:

$$\begin{aligned} \boldsymbol{\sigma}_{\text{viscous}} &= 2\eta \boldsymbol{E}, \\ \boldsymbol{\sigma}_{\text{elastic}} &= -P\boldsymbol{I} + 2\varphi\left(\boldsymbol{Q} + \frac{\boldsymbol{I}}{3}\right)(\boldsymbol{Q}:\boldsymbol{H}) - \varphi\boldsymbol{H}\left(\boldsymbol{Q} + \frac{\boldsymbol{I}}{3}\right) - \varphi\left(\boldsymbol{Q} + \frac{\boldsymbol{I}}{3}\right)\boldsymbol{H} \\ &\qquad - \nabla\boldsymbol{Q}\frac{\delta\mathcal{F}}{\delta\nabla\boldsymbol{Q}} + \boldsymbol{Q}\boldsymbol{H} - \boldsymbol{H}\boldsymbol{Q}, \\ &= -P\boldsymbol{I} + [\boldsymbol{Q},\boldsymbol{H}] - 2\varphi\left(\frac{1}{3}\boldsymbol{H} - \boldsymbol{Q}(\boldsymbol{Q}:\boldsymbol{H}) + [\boldsymbol{Q}\cdot\boldsymbol{H}]^{\mathrm{ST}}\right) - \nabla\boldsymbol{Q}\frac{\delta\mathcal{F}}{\delta\nabla\boldsymbol{Q}}, \\ \boldsymbol{\sigma}_{\text{active}} &= -\zeta\boldsymbol{Q}. \end{aligned} \tag{2.13}$$

Here, η is the viscosity and ζ is the microtubule activity coefficient, with $\zeta > 0$ (< 0) signifying extensile (contractile) stresses [82]. It is important to note that for $\zeta = 0$, Equations (2.8)–(2.13) are the nematohydrodynamic equations of motion for passive nematic liquid crystals [21], as $\boldsymbol{\sigma}_{\text{elastic}}$ describes the stresses on the fluid due to the passive movement of the rods.

2.2.2 Incorporating Density Fluctuations

Equations (2.8)–(2.13) assume that the MT density is constant in space and time. However, variations in MT density can be substantial in experiments (seen for one experimental system in Figure 5.2). These pockets of low density also frequently correlate with defects and nearby "cracks" that are formed as they are created or annihilated. It is possible that the interplay of density, velocity, and orientation plays a significant role in the observed dynamics, as was conjectured by Giomi *et al.* [35]. Their model incorporates density fluctuations and has the form (for $d = 2$)

$$\begin{aligned}
\boldsymbol{Q}_t + \boldsymbol{u}\cdot\nabla\boldsymbol{Q} &= \varphi S\boldsymbol{E} + [\boldsymbol{Q},\boldsymbol{\Omega}] - \Gamma\boldsymbol{H},\\
\rho\boldsymbol{u}_t &= \nabla\cdot\boldsymbol{\sigma}, \quad \nabla\cdot\boldsymbol{u} = 0,\\
\rho_t + \boldsymbol{u}\cdot\nabla\rho &= \nabla\cdot[(D_0\boldsymbol{I} + D_1\boldsymbol{Q})\nabla\rho + \alpha_1\rho^2\nabla\cdot\boldsymbol{Q}],\\
\text{where}\quad \boldsymbol{\sigma} &= -P\boldsymbol{I} + 2\eta\boldsymbol{E} - \varphi S\boldsymbol{H} + [\boldsymbol{Q},\boldsymbol{H}] + \alpha_2\rho^2\boldsymbol{Q},\\
\boldsymbol{H} &= \left[\left(-A + \frac{1}{2}S^2C\right)\boldsymbol{Q} - K\Delta\boldsymbol{Q}\right].
\end{aligned} \tag{2.14}$$

In addition to incorporating density fluctuations and anisotropic diffusion (for $D_1 > 0$), Equation (2.14) differs from Equations (2.8)–(2.13) in a few ways. First, the flow-alignment parameter φ is multiplied by the scalar order parameter S wherever it appears. Second, in the velocity equation, the Reynolds number $\text{Re} \equiv \rho L|\boldsymbol{u}|/\eta$ is assumed to be small for MTs of length $L \sim 100\,\mu\text{m}$, so the convective term $\boldsymbol{u}\cdot\nabla\boldsymbol{u}$ is omitted. It should be noted that the molecular tensor $\boldsymbol{H}$ in Equation (2.14) is simply Equation (2.11) expressed in two dimensions ($d = 2$), since $\boldsymbol{Q}^2 = \frac{S^2}{4}\boldsymbol{I}$ in 2D.

Moreover, the terms $[\boldsymbol{E}\cdot\boldsymbol{Q}]^{\mathrm{ST}}$ and $[\boldsymbol{H}\cdot\boldsymbol{Q}]^{\mathrm{ST}}$ in Equations (2.9) and (2.13), respectively, vanish in 2D since $[\boldsymbol{A}\cdot\boldsymbol{B}]^{\mathrm{ST}} = 0$ for symmetric and traceless 2×2 matrices $\boldsymbol{A}$ and $\boldsymbol{B}$.

2.2.3 A Minimal Model in 2D

Oza & Dunkel [67] proposed a minimal model for active nematics that built off of the models described in Sections 2.2.1 and 2.2.2, but with a few notable changes. While the system is often modeled in two dimensions (2D) [20, 34, 35], the experiments are three-dimensional with a shallow depth. As such, upwellings or sinks in the MTs could invalidate the assumption of incompressibility ($\nabla \cdot \boldsymbol{u} = 0$). If this were the case, the incompressible Navier-Stokes equations would need to be replaced their compressible counterpart, and the orientation evolution equation (2.8) would have to be modified to read

$$\frac{\partial}{\partial t}\boldsymbol{Q} + \boldsymbol{u}\cdot\nabla\boldsymbol{Q} \to \frac{\partial}{\partial t}\boldsymbol{Q} + \nabla\cdot(\boldsymbol{u}\boldsymbol{Q}).$$

Furthermore, a common feature noted in experiments is the buckling of long strands of MT bundles. This particular feature is somewhat difficult to capture in continuum models. However, one possibility is to replace the gradient of the elastic energy term, $\Delta\boldsymbol{Q}$ from Equation (2.14), which penalizes inhomogeneities in the orientation field, with the higher-order term

$$-\gamma_2\Delta\boldsymbol{Q} + \gamma_4\Delta^2\boldsymbol{Q}.$$

The addition of such a term leads to patterns with a characteristic length scale $\sqrt{\gamma_4/\gamma_2}$. Similar terms have been used to describe buckling processes in elastic materials [84] and pattern formation in bacterial suspensions [23].

The minimal model [67] thus has the form

$$
\begin{aligned}
&\frac{\partial}{\partial t}\boldsymbol{Q} + \nabla\cdot(\boldsymbol{u}\boldsymbol{Q}) - \kappa[\boldsymbol{Q},\boldsymbol{\Omega}] = -\frac{\delta\mathcal{F}}{\delta\boldsymbol{Q}} = A\boldsymbol{Q} - C\boldsymbol{Q}^3 - \gamma_2\Delta\boldsymbol{Q} - \gamma_4\Delta^2\boldsymbol{Q},\\
&\boldsymbol{u} = -\frac{\zeta}{\nu}\nabla\cdot\boldsymbol{Q},\\
\text{where}\quad &\mathcal{F} = \int d\boldsymbol{x}\,\mathrm{Tr}\left\{-\frac{A}{2}\boldsymbol{Q}^2 + \frac{C}{4}\boldsymbol{Q}^4 - \frac{\gamma_2}{2}(\nabla\boldsymbol{Q})^2 + \frac{\gamma_4}{4}(\nabla\nabla\boldsymbol{Q})^2\right\}. \qquad (2.15)
\end{aligned}
$$

We note that Navier-Stokes equation for the velocity field $\boldsymbol{u}$ has been simplified to a relatively simple relation between $\boldsymbol{u}$ and $\boldsymbol{Q}$, which may be derived from the Hele-Shaw (thin-film) approximation applied to the Stokes equations (see Appendix A.5). Note that the velocity field is not divergence-free, as fluid exchange between the bulk and the quasi-2D active nematic layer is permitted (Figure 5.1b). Despite its relative simplicity, Equation (2.15) is able to capture defect creation and annihilation dynamics, and exhibits good agreement with experimental data on defect lifetimes and speed distributions [67].

2.2.4 A Kinetic Theory

Finally, a kinetic theory approach from the Smoluchowski equation for statistical mechanics have been proposed by Gao *et al.* [31, 32]. In its dimensionless form, the equations are

$$
\boldsymbol{D}^{\nabla} + 2\boldsymbol{E} : \boldsymbol{S}[\boldsymbol{D}] = 4\zeta\left(\boldsymbol{D}\cdot\boldsymbol{D} - \boldsymbol{S}[\boldsymbol{D}] : \boldsymbol{D}\right) + A\Delta\boldsymbol{D} - 2B\left(\boldsymbol{D} - \frac{\rho}{d}\boldsymbol{I}\right), \qquad (2.16)
$$

$$
\begin{aligned}
\nabla p - \Delta\boldsymbol{u} &= \nabla\cdot\sigma_B[\boldsymbol{D}],\\
\nabla\cdot\boldsymbol{u} &= 0,\\
\sigma_B[\boldsymbol{D}] &= \alpha\boldsymbol{D} + \beta\boldsymbol{S}[\boldsymbol{D}] : \boldsymbol{E} - 2\zeta\beta\left(\boldsymbol{D}\cdot\boldsymbol{D} - \boldsymbol{S}[\boldsymbol{D}] : \boldsymbol{D}\right),\\
\boldsymbol{S}[\boldsymbol{D}] &= \int_S \boldsymbol{p}\otimes\boldsymbol{p}\otimes\boldsymbol{p}\otimes\boldsymbol{p}\,\mathrm{d}\mathbb{P}(\boldsymbol{p}), \qquad (2.17)
\end{aligned}
$$

where $\boldsymbol{D}$ is the non-centered tensor order parameter defined in Equation (2.1), the upper-convective derivative is given as $\boldsymbol{D}^{\nabla} = \boldsymbol{D}_t + \boldsymbol{u}\cdot\nabla\boldsymbol{D} - (\nabla\boldsymbol{u}\cdot\boldsymbol{D} + \boldsymbol{D}\cdot\nabla\boldsymbol{u}^{\intercal})$ and $\boldsymbol{S}$ is the fourth moment of the probability density $\mathbb{P}$. The left hand side of the equation differs from the Beris Edwards approach in that it uses a transport mechanism based on microscopic modeling including the second term on the left hand side which arises from Jeffery's equation for ellipsoidal bodies in linear flow. The terms on the right hand side of the equation follow a similar structure as the Beris Edwards equations with differences in the diffusion coefficient for the $\Delta\boldsymbol{D}$ term and the nonlinearities governing the ordering ($\boldsymbol{D}$, $\boldsymbol{D}^2$, and $\boldsymbol{S}[\boldsymbol{D}]$ rather than $\boldsymbol{Q}$,$\boldsymbol{Q}^2$, and $\boldsymbol{Q}^3$). Though this formulation has a strong connection to the Beris-Edwards formulation, the model requires a closure assumption for evaluation of the fourth moment. As such, $\boldsymbol{S}$ can be approximated using the so-called Bingham closure [31, 32]. A fast algorithm for evaluating $\boldsymbol{S}$ under this closure assumption is described in [96].

2.3 Constructing a Library of Model Terms

The state variables present in the models of Section 2.2 are orientation $\boldsymbol{Q}, \boldsymbol{D}, \boldsymbol{S}$, velocity $\boldsymbol{u}$, and density ρ. Given the constant translation difference between tensor order $\boldsymbol{Q}$ and $\boldsymbol{D}$ in Equation (2.2), a complete library can be constructed without considering $\boldsymbol{D}$.

When constructing one dimensional (1D) systems for use with SINDy, it is common to consider polynomial interactions between state variables and their spatial derivatives [73]. However, generation of an over-complete library of polynomial terms complicates as the dimension of the system increases due to combinatorial complexity of tensor contractions. As such, a computational algebra system (CAS) is required to algorithmically generate a sufficiently large library of possibilities. Additionally, tensor equations can contain a variety of properties that need to be incorporated into the library structure (e.g. $\boldsymbol{Q}_t$ is symmetric and traceless).

With these constraints in mind, we propose a multistep process for generating a tensor library of order p:

1. Collect common state variables as a starting set of symbols for the library:
   E.g. $\boldsymbol{Q}, \boldsymbol{S}, \boldsymbol{u}$.

2. Expand the starting set of symbols by prepending derivative operators to the symbols, up to a given maximum derivative order:
   E.g. $\boldsymbol{Q}, \nabla\boldsymbol{Q}, \nabla\nabla\boldsymbol{Q}, \boldsymbol{u}, \nabla\boldsymbol{u}, \ldots$
   Note: In practice, each term has a "derivative order" d, e.g. $\boldsymbol{Q}$ has $d = 0$, while $\nabla\boldsymbol{Q}$ has $d = 1$. Given the commonly considered active stress relationship of $\boldsymbol{u} = -\nabla \cdot \boldsymbol{Q}$, it is assumed that $\boldsymbol{u}$ has $d = 1$.

3. Take all (unordered) outer-products of the expanded set of base terms with tensor order $2\ell + p$, $\ell \in \mathbb{N}_0$, up to a given total polynomial order in state variables:
   E.g. $\boldsymbol{Q}, \nabla\boldsymbol{u}, \boldsymbol{u} \otimes \nabla\boldsymbol{Q}, \boldsymbol{S} \otimes \nabla\boldsymbol{u}, \ldots$
   Note: In practice, we assign a "Q-order" q to each term. Specifically, $\boldsymbol{Q}^n$ has $q = n$, $\boldsymbol{u}$ has $q = 1$ and $\boldsymbol{S}$ has $q = 2$. The Q-order and derivative order of a composite term (e.g. $u \otimes \nabla\boldsymbol{Q}$) is the sum of the individual orders (e.g. $q = 2$ and $d = 2$). Terms with Q-order greater than a prescribed value $q_{\max} = 3$ or a derivative order greater than a prescribed value $d_{\max} = 2$ are eliminated from the library.

4. For any order $2\ell + p$ tensor produced in step 3, form all tensors of order p that can be obtained by successively contracting pairs of indices and permuting the remaining p:
   E.g. from $u_{i_1}\partial_{i_2}Q_{i_3 i_4}$ we obtain the second order tensors $\boldsymbol{u} \cdot \nabla\boldsymbol{Q}$, by contracting i_1, i_2, $(\partial_j Q_{\ell k})u_\ell$, by contracting i_1, i_3, $(\partial_j Q_{k\ell})u_\ell$, by contracting i_1, i_4, $\boldsymbol{u}\nabla\cdot\boldsymbol{Q}$, by contracting i_2, i_3, $\boldsymbol{u}\nabla \cdot (\boldsymbol{Q}^\intercal)$, by contracting i_2, i_4, and $\boldsymbol{u}\nabla\mathrm{Tr}(\boldsymbol{Q})$, by contracting i_3, i_4, as well the transposes of all of these by permuting the indices.

5. (Optional) Ensure that the terms satisfy physical constraints (symmetric, traceless, etc.). If the term does not immediately satisfy a condition, transform it to ensure it satisfies the necessary conditions:
   E.g. the term $\partial_j Q_{k\ell} u_\ell$ is not necessarily traceless or symmetric.

6. Compare terms to ensure symbolic uniqueness. For equivalent terms, keep the term with the shortest symbolic representation (keeping the first term generated in the case of a tie):
   E.g. the terms $\partial_j Q_{\ell k} u_\ell$ and $\partial_j Q_{k\ell} u_\ell$ are equivalent because $\boldsymbol{Q}$ is symmetric and the term $\boldsymbol{u}\nabla\mathrm{Tr}(\boldsymbol{Q})$ is zero because $\boldsymbol{Q}$ is traceless.

Examples of the terms yielded via this procedure for both orientation and velocity evolution equations are shown in Table 2.1. Following (2.15) we assume an overdamped Hele-Shaw limit (see Appendix A.5) in which the velocity $\boldsymbol{u}$ is determined directly by the divergence of the stress tensor.

Remark 2.3.1. *Note that due to the symmetries of the fourth moment tensor,* $\boldsymbol{S}$*, the contractions that occur within this term are removed from the library by hand.*

For example, $S_{ijkk} = D_{ij}$ is removed because the traceless version of this term is $\boldsymbol{Q}$. This redundancy would not be apparent in the symbolic library generating procedure outlined above but would be apparent in the actual library terms.

The procedure described above produces a large library of polynomial tensor interactions between state variables and their gradients. The library terms can be linearly combined to attain all common terms in the Beris-Edwards equations (2.8)–(2.13), a few of which are shown in Table 2.2. We note that some terms are not directly present in the library but can be constructed via linear combination of terms in the library. For example, $[\boldsymbol{Q}, \boldsymbol{\Omega}]$ can be obtained via the linear combination $[\boldsymbol{Q}, \boldsymbol{\Omega}] = 2[Q_{ki}\partial_k u_j]^{\mathrm{ST}} - 2[Q_{ki}\partial_j u_k]^{\mathrm{ST}}$. Since the $[\boldsymbol{Q}, \boldsymbol{\Omega}]$ term is present in many previously-proposed models (e.g. Equations (2.8), (2.14), (2.15)), the terms $2[Q_{ki}\partial_k u_j]^{\mathrm{ST}}$ and $2[Q_{ki}\partial_j u_k]^{\mathrm{ST}}$ are replaced in the library with their sum and difference:

$$2[Q_{ki}\partial_k u_j]^{\mathrm{ST}} - 2[Q_{ki}\partial_j u_k]^{\mathrm{ST}}$$

$$2[Q_{ki}\partial_k u_j]^{\mathrm{ST}} + 2[Q_{ki}\partial_j u_k]^{\mathrm{ST}}$$

Table 2.1 Terms generated via the procedure described in Section 2.3 for the $\boldsymbol{Q}$ evolution equation $\boldsymbol{Q}_t = \dots$ and velocity equation $\boldsymbol{u} = \nabla \cdot \boldsymbol{\sigma}$, where $\boldsymbol{\sigma}$ is the stress.

$\boldsymbol{Q}_t$ **Equation**	$\boldsymbol{u}$ **Equation (stress (σ) terms)**
State variables: $\boldsymbol{Q}, \boldsymbol{S}, \boldsymbol{u}$	State variables: $\boldsymbol{Q}$
Maximum derivative order: 2	Maximum derivative order: 2
Maximum order in $\boldsymbol{Q}$: 3	Maximum order in $\boldsymbol{Q}$: 3
Conditions: symmetric, traceless	Conditions: symmetric
Total number of terms: 46	Total number of terms: 52

Q_{ij}	Q_{ij}
$\partial_k \partial_k Q_{ij}$	$[\partial_k \partial_i Q_{kj}]^{\mathrm{S}}$
$[\partial_i u_j]^{\mathrm{ST}}$	$\partial_k \partial_k Q_{ij}$
$[Q_{ki} \partial_k \partial_l Q_{lj}]^{\mathrm{ST}}$	$Q_{ki} Q_{kj}$
$[Q_{kl} \partial_i \partial_j Q_{kl}]^{\mathrm{ST}}$	$[Q_{lj} \partial_k \partial_k Q_{il}]^{\mathrm{ST}}$
$Q_{kl} Q_{ijlk}$	$Q_{ji} \partial_k \partial_l Q_{kl}$
$[Q_{ki} \partial_j u_k]^{\mathrm{ST}}$	$Q_{lk} \partial_j \partial_i Q_{kl}$
$Q_{ij} \partial_k u_k$	$Q_{kj} Q_{lk} Q_{li}$
$[Q_{lk} S_{lijk}]^{\mathrm{ST}}$	$[Q_{mi} Q_{lk} \partial_j \partial_m Q_{kl}]^{\mathrm{S}}$
$[u_i u_j]^{\mathrm{ST}}$	$[Q_{lk} Q_{ml} \partial_m \partial_j Q_{ki}]^{\mathrm{S}}$
$\vdots$	$\vdots$

Common Term	Term in Library
$\boldsymbol{u} \cdot \nabla \boldsymbol{Q}$	$u_k \partial_k Q_{ij}$
$(\nabla \cdot \boldsymbol{u})\boldsymbol{Q}$	$\partial_k u_k Q_{ij}$
$\boldsymbol{E}^{\mathrm{ST}}$	$[\partial_i u_j]^{\mathrm{ST}}$
$(\boldsymbol{Q} : \nabla \boldsymbol{u})\boldsymbol{Q}$	$Q_{ij} Q_{kl} \partial_k u_l$
$\boldsymbol{Q}$	Q_{ij}
$\boldsymbol{Q}^3$	$Q_{ik} Q_{kl} Q_{lj}$
$\nabla^2 \boldsymbol{Q}$	$\partial_k \partial_k Q_{ij}$
$\vdots$	$\vdots$

Table 2.2 Table of commonly considered terms in the Beris-Edwards equation (2.8) for the orientational order parameter $\boldsymbol{Q}$. Note that $\boldsymbol{E}^{\mathrm{ST}} = \boldsymbol{E}$ for an incompressible velocity field, $\nabla \cdot \boldsymbol{u} = 0$.

CHAPTER 3

MODEL DISCOVERY PRELIMINARIES

As briefly outlined in Section 1.2, this dissertation applies the SINDy framework to discover a model from data. Given a library of nonlinear candidate terms, the original SINDy framework identifies the correct k term model as the solution of the k-sparse least squares problem:

$$\hat{\boldsymbol{\xi}} = \underset{\boldsymbol{\xi}, ||\boldsymbol{\xi}||_0 = k}{\operatorname{argmin}} ||\boldsymbol{\Theta}\boldsymbol{\xi} - \partial_t \boldsymbol{u}||_2^2 \ ,$$

where

$$||\boldsymbol{\xi}||_0 = |\{i : \xi_i \neq 0\}| = |\text{supp}(\boldsymbol{\xi})| \ .$$

This problem is equivalent to the regularized least squares problem:

$$\hat{\boldsymbol{\xi}} = \underset{\boldsymbol{\xi}}{\operatorname{argmin}} ||\boldsymbol{\Theta}\boldsymbol{\xi} - \partial_t \boldsymbol{u}||_2^2 + \gamma ||\boldsymbol{\xi}||_0 \ , \tag{3.1}$$

for some γ depending on k. The model is then given to be the linear combination of the corresponding (possibly nonlinear) library terms with non-zero coefficients in $\hat{\boldsymbol{\xi}}$.

There are several challenges in applying SINDy to real data that have been recognized in the literature [36, 44, 60, 74, 87]. The library must be constructed in a way that the target quantity, $\boldsymbol{u}_t$ above, is approximately in the range of the library matrix, $\boldsymbol{\Theta}$. Typically, the library is constructed using a standard basis, such as polynomials. An active nematic model is a tensor-valued PDE and thus requires a non-standard library construction. We outline a general approach to tensor-valued PDE library construction in Section 2.3 and show there that it contains the standard active nematic model terms. It is known that the quality of the data impacts SINDy's ability to recover a PDE model which captures the dynamics of the system [8]; we discuss the effect of data quality in more detail in Section 3.1. This is followed by an exploration of numerical differentiation of potentially noisy data in Section 3.2 which has been identified in the literature as a key challenge for SINDy [3,17,40,52,60,78]. Finally, the challenge of selecting a sparse subset from hundreds or possibly thousands of candidate terms is discussed in Section 3.3 even if the linear system does not satisfy the usual assumptions of linear regression (lack of multicollinearity, non-Gaussian residuals, etc. [79]).

3.1 Data Quality

It is known that data must be resolved to the scale of the underlying dynamics of the system in order to successfully recover the governing equation [8]. In fact, insufficiently dynamic data often yields a linear system which is not sparse, ultimately yielding a wide range of possible models that capture the dynamics [9]. This is demonstrated for simulated active nematic data in Section 4.2.3. However, once the resolution is sufficient to accurately compute derivatives of the state variables, the full dataset may be subsampled to reduce the computational burden. Several previous works applying SINDy to PDE discovery have used such sparsely sampled sets of the data library. This amounts to reducing the number of rows used in Equation (1.3). There have also been works demonstrating different sampling methods that can be applied to resolve multi-scale dynamics that may be present in the data [7, 12]. These techniques demonstrate accurate equation recovery but also demonstrate the fragility of SINDy to insufficient data and the danger of unknowingly overlooking small scale dynamics. These works also reference the need for “dynamic” data with sufficient variation in time to give stable numerical derivatives and a well-conditioned system [73,97].

In the context of experimental data, each of the above represents a need for accurate extraction of state variables from experimental observations. The procedure for this extraction will vary from application to application but will likely depend on the selection of several parameters in the extraction procedure. Working to ensure the extracted state variables satisfy physical constraints will help ensure accurate recovery of the governing equations. In this dissertation, we consider a physically motivated approach to selecting these parameters based on the conservation of mass and other physical constraints [87]. See Sections 5.3.1 and 5.3.2 for more details.

3.2 Fitting and Differentiating Noisy Data

There are a variety of well-established methods for numerical differentiation of data. However, guaranteeing the accuracy of derivatives without knowledge of the underlying generating function or the level of noise remains a challenge. Rudy *et al.* [73] identified that, in the absence of noise, classical numerical differentiation methods such as finite differences are adequate for the linear system in Equation (1.3). In the case of noisy data, these local methods are severely limited, and it is instead more practical to fit a differentiable basis of functions or first smooth the data. There are a few general principles from which most common methods for fitting noisy data emerge:

Differentiation via interpolation is not reasonable for noisy data due to the danger of over-fitting. As such, a new metric for "best fit" needs to be established. Usually, this amounts to considering the best L^2 fit of the data with least squares. Additionally, the least squares formulation can be augmented with regularization to reduce spurious higher order modes [48] or weighted to more closely fit certain segments of the data [13].

A natural (and popular) method is to locally fit a low order polynomial to data points using either least squares or weighted least squares in order to approximate the centermost point of the local grid. As the basis is known, the fitted polynomial can be differentiated prior to evaluation to approximate derivatives at that point. In econometrics and signal processing, this method is called LOESS (locally estimated scatterplot smoothing), LOWESS (locally weighted scatterplot smoothing), or the Savitsky-Golay filter [13, 76]. Due to the lower order of the derivatives and the lack of boundary conditions between neighboring polynomials, it is not guaranteed to have smooth derivatives [64]. Yet it has been previously shown to be effective for SINDy [73].

Alternatively, fitting noisy data with a global basis of functions can also be effective. Several popular and well-established methods for this exist including fitting orthogonal polynomials via least squares [33], smoothing splines [69], and truncated Fourier series [49]. Additionally, more modern methods such as fitting with neural networks made up of composed nonlinear functions have been shown to have promising robustness to noise [97]. This approach has also shown promise for performing SINDy on experimental data [87].

It was more recently proposed that differentiation of noisy data for SINDy could be partially avoided by considering instead a weak formulation of partial differential equation [60]. This approach has shown impressive noise robustness for simulated data and has also been applied for equation discovery in experimental settings [61, 62].

A final approach to consider is to first smooth data after which a simpler numerical differentiation method such as finite differences can be applied. Indeed, this method captures data variation and shape well and can be very computationally efficient. Furthermore, with an appropriately selected basis, it can be represented as a convolution of the original data, which strongly resembles the weak formulation of SINDy and thus presents great potential for noise robustness. However, it is hard to determine smoothing parameters which can guarantee accurate results. Given known physical conditions or characteristic features of the data, smoothing parameters can be chosen in order to recover a trustworthy and smooth data representation which can be accurately differentiated, an approach which is demonstrated in Section 5.3.

One global approach to smoothing the data is to apply a filter on the data after being transformed into Fourier space. This allows for the removal of highly oscillatory information from the data. Consider the spectral filter:

$$f(\boldsymbol{k}) = e^{-(\boldsymbol{k}/s)^2} = e^{-(k_1/s)^2} \otimes e^{-(k_2/s)^2} \otimes \ldots \otimes e^{-(k_n/s)^2} \tag{3.2}$$

where k is the wavenumber and s is a tunable smoothing parameter. Given data $\boldsymbol{u}(\boldsymbol{x}, t)$, the filtered data can then be written as

$$\tilde{\boldsymbol{u}}(\boldsymbol{x}, t) = \mathcal{F}^{-1}(f(\boldsymbol{k})\mathcal{F}(\boldsymbol{u}(\boldsymbol{k}, t)))$$

where $\mathcal{F}$ is the Fourier transform. For smaller values of s, this filter will increase the number of dampened high modes. This approach is similar to an integral approach to SINDy presented in [3, 77] which has a similar effect in smoothing the nonlinear library and improving coefficient estimation. Section 6.4 discusses the use of this filtering approach to assist in the use of non-periodic experimental data as initial conditions for forward simulation of a discovered model. As an additional benefit, this filtration opens up the possibility of using spectral differentiation for which the condition of a divergence free field can be imposed. This approach is described and used in Section 6.2.

3.3 Variable Selection and Sparse Regression

Equation (3.1) is a nonsmooth and nonconvex optimization problem which requires combinatorial effort to find a global minimizer; finding the true solution is only plausible if the library is relatively small. It is thus common to use a less computationally intensive method to find an approximate solution. We review some common methods below in Sections 3.3.1, 3.3.2, and 3.3.3.

It is also common to consider slight variations of the original SINDy framework. See 3.3.5 for a randomized variant and Section 3.3.4 for a variant based on total least squares.

3.3.1 Convex Relaxation and LASSO

It is common to consider a nonsmooth but convex approximation to Equation (3.1). Most common is to use the least absolute shrinkage and selection operator (LASSO) [91]:

$$\hat{\boldsymbol{\xi}} = \underset{\boldsymbol{\xi}}{\operatorname{argmin}} ||\boldsymbol{\Theta}\boldsymbol{\xi} - \partial_t \boldsymbol{u}||_2^2 + \gamma ||\boldsymbol{\xi}||_1. \tag{3.3}$$

This classical formulation has been shown to reduce some elements of $\hat{\boldsymbol{\xi}}$ to 0 due to the geometry of the L^1 penalty, ultimately reducing the size of the term library. The LASSO method is known to solve the original sparse regularized problem, Equation (3.1), under certain conditions [10]. However, these conditions are not met in most SINDy applications.

By varying the parameter γ, the LASSO method can return models which are less sparse (small γ) or more sparse (large γ). The range of models produced by varying γ are sometimes referred to as the LASSO path. See Section 6.1.1 where this approach was used on experimental data.

The LASSO regularization term, i.e. $\|\boldsymbol{\xi}\|_1$, biases all coefficients in $\hat{\boldsymbol{\xi}}$. Because the interest in using LASSO and other approximate methods for the solution of Equation (3.1) in SINDy is to discover the support of the optimal sparse library, the biased coefficients are usually discarded and new coefficients are computed via a standard least squares solve for the reduced library [15].

Suppose that $\hat{\boldsymbol{\xi}}$ is the vector of coefficients computed by the LASSO or another sparse solution procedure. Let $J = \operatorname{supp}(\hat{\xi})$ be the support of $\hat{\boldsymbol{\xi}}$. Unbiased versions of the non-zero coefficients can then be computed as an ordinary least squares solution of the system

$$\boldsymbol{\xi}^\star = \underset{\boldsymbol{\xi}}{\operatorname{argmin}} ||\boldsymbol{\Theta}(:, J)\boldsymbol{\xi} - \partial_t \boldsymbol{u}||_2^2 \, . \tag{3.4}$$

This two-stage procedure of first discovering the support of the reduced library and then computing the coefficients in the library can be applied to the other approximate methods below. It is also convenient in allowing for other post-processing methods; see Section 3.3.4.

3.3.2 Sequentially Thresholded Ridge Regression

Like the LASSO, the ridge regression is a convex relaxation of Equation (3.1) [101].

$$\hat{\boldsymbol{\xi}} = \underset{\boldsymbol{\xi}}{\text{argmin}} ||\boldsymbol{\Theta}\boldsymbol{\xi} - \partial_t \boldsymbol{u}||_2^2 + \gamma ||\boldsymbol{\xi}||_2 \ . \tag{3.5}$$

In contrast to the LASSO, this method rarely constrains elements of $\hat{\boldsymbol{\xi}}$ to 0 but instead reduces the size of coefficients for independent variables which don't contribute strongly to resolving $\partial_t \boldsymbol{u}$.

The original SINDy algorithm is known as Sequentially Thresholded Ridge Regression (STRidge). The idea of STRidge is to find an approximate solution to Equation (3.1) by iteratively solving the ridge regression regularized problem in Equation (3.5) and forcing any element of $\boldsymbol{\xi}$ below a threshold (τ) to 0 [9, 73]. See Algorithm 1 for the full procedure. STRidge is known to converge to a local solution of Equation (3.1); see [99].

Algorithm 1 `STRidge`$(\boldsymbol{\Theta}, \boldsymbol{U}_t, \gamma, \tau_{\text{init}}, M)$

```
$\tau \leftarrow \tau_{\text{init}}$
$n \leftarrow$ number of columns in $\boldsymbol{\Theta}$
$J \leftarrow \{1, \ldots, n\}$
for $i = 1, \ldots, M$ do
    $\hat{\boldsymbol{\xi}} \leftarrow \arg\min_{\boldsymbol{\xi}:\text{supp}(\boldsymbol{\xi})\subset J} \|\boldsymbol{\Theta}\boldsymbol{\xi} - \boldsymbol{U}_t\|_2^2 + \gamma\|\boldsymbol{\xi}\|_2^2$
    $J \leftarrow \{i : |\hat{\boldsymbol{\xi}}_i| \geq \tau\}$
    $\tau \leftarrow$ update as defined in [73]
end for
return $J$
```

STRidge requires the selection of the γ and τ parameters. A heuristic procedure to determine the optimal threshold parameter given a regularization parameter is presented in [73]. To do so, it separates a portion of the rows of the library $\mathbf{\Theta}$ as a "test set" and uses the remainder to find a sparse set of terms. Once the terms have been determined, it evaluates the linear model's efficiency to predict the values in the test set and updates the threshold parameter according to its performance when compared with the previous parameter.

As with the LASSO, new unbiased coefficients can be computed once the library has been selected. The regularization parameter, γ, can be adjusted in order to determine models of varying sizes, creating something similar to a LASSO path; this is done for real data in Section 6.1.1.

3.3.3 Forward Selection and Other Greedy Methods

Some other popular techniques to approximate the solution to Equation (3.1) are greedy methods. Examples include forward and backward variable selection and orthogonal basis pursuit, which were developed independently in the fields of statistics and signal processing [41, 68]. While these methods are known to recover exact solutions under some restrictive conditions on the library [19, 92], these conditions are not typically met in SINDy applications. Nonetheless, greedy methods have been shown to identify terms that correlate or capture the variance of the time evolution of the system and were effective in selecting the correct library for the synthetic data examples in Section 4.2.1 and showed good agreement with brute force solutions of Equation (3.1) for the real data examples in Section 6.1.

The majority of greedy methods iteratively add or remove terms from the library (or columns from the matrix of independent variables) by considering which term maximizes or minimizes some fitness or loss function. For forward selection, the most common success metric is to maximize the coefficient of determination R^2:

$$R^2 = 1 - \frac{\sum_i (u_i - \hat{u}_i)^2}{\sum_i (u_i - \overline{u})^2}, \tag{3.6}$$

where u_i is the data, $\hat{u}_i$ is the model's output, $\overline{u}$ is the sample mean of the data. For the R^2 metric, forward selection is mathematically equivalent to orthogonal matching pursuit. The forward selection process is detailed in Algorithm 2.

Algorithm 2 `ForwardSelection`($\boldsymbol{\Theta}, \boldsymbol{U}_t$,M)

```
$n \leftarrow$ number of columns in $\boldsymbol{\theta}$
$J \leftarrow \{\}$
for $i = 1, \ldots, \min(M, n)$ do
    $e_{\text{best}} \leftarrow \infty$
    $j_{\text{best}} \leftarrow \{\}$
    for $j \in \{1, \ldots, n\} \setminus J$ do
        $J_{\text{temp}} \leftarrow J \cup \{j\}$
        $\hat{\boldsymbol{\xi}} \leftarrow \arg\min_{\boldsymbol{\xi}} \|\boldsymbol{\Theta}(:, J_{\text{temp}})\boldsymbol{\xi} - \boldsymbol{U}_t\|_2^2$
        $e_{\text{temp}} \leftarrow \|\boldsymbol{\Theta}(:, J_{\text{temp}})\hat{\boldsymbol{\xi}} - \boldsymbol{U}_t\|_2^2$
        if $e_{\text{temp}} < e_{\text{best}}$ then
            $j_{\text{best}} \leftarrow j$
            $e_{\text{best}} \leftarrow e_{\text{temp}}$
        end if
    end for
    $J \leftarrow J \cup \{j_{\text{best}}\}$
end for
return $J$
```

Forward matching pursuit computes the solution on a reduced library as in Equation (3.4) at each step, so unbiased coefficients are computed as part of the process. However, new coefficients can still be computed for the sparse library found by forward selection using a different error metric, like total least squares. See Section 3.3.4.

Remark 3.3.1. *In some applications, both the support of the library and an optimal value for the level of sparsity (k) are determined using a single metric. One such metric is the Akaike Information Criterion (AIC):*

$$AIC = 2k - 2\ln(\hat{L}), \tag{3.7}$$

where k is the number of columns in the model, and $\hat{L}$ is the maximum likelihood estimator for the model. In the present work, we consider all models obtained as k ranges and use a more heuristic approach to determine k based on the improvement in the R^2 and the apparent uncertainty of the terms.

3.3.4 Errors-in-Variables Models

A variety of linear solution methods can be used to determine coefficients once the support of the reduced library has been determined using one of the methods above. One such approach which is appropriate for the application treated in this dissertation is an errors-in-variables approach [94]. In this formulation, error is assumed to be present for both the dependent and independent variables of the regression. Such methods are common in the field of system identification and inverse problems, and may be more appropriate for PDE discovery. Given that the linear system in SINDy is constructed using combinations of experimentally observed state variables and their derivatives, it is expected that there will be significant error in both sides of the linear relationship. The general form of an errors-in-variables model is

$$\boldsymbol{b} + \boldsymbol{\epsilon}_1 = (\boldsymbol{A} + \boldsymbol{\epsilon}_2)\boldsymbol{\xi}.$$

This can be compared to ordinary regression methods which assume a model form of $\boldsymbol{b} + \boldsymbol{\epsilon} = \boldsymbol{A}\boldsymbol{\xi}$.

The most common of these errors-in-variables models is called "total least squares", which can be formulated as the following optimization problem:

$$\begin{aligned} \min_{\hat{\boldsymbol{A}},\hat{\boldsymbol{b}},\boldsymbol{\xi}} \quad & \|(\boldsymbol{A},\boldsymbol{b}) - (\hat{\boldsymbol{A}},\hat{\boldsymbol{b}})\|_F \\ \text{subject to} \quad & \hat{\boldsymbol{A}}\boldsymbol{\xi} = \hat{\boldsymbol{b}}, \end{aligned}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. This problem amounts to finding the smallest L^2 shift of both the dependent and independent variables such that there is a set of coefficients $\boldsymbol{\xi}$ that exactly solves the shifted system.

In practice, total least squares has been shown to perform less well than ordinary least squares for prediction, but frequently superior for determining the correct coefficients of a linear model [59] due to the increased number of degrees of freedom of the method. However, this property also makes the method sensitive to the noise distribution of the terms in the term library. Appendix A.2.2 contains an example demonstrating that ill-posed systems with skewed error distributions cause an increase in the relative size of coefficients returned by total least squares, resulting in inflated coefficient magnitudes as compared to those returned by ordinary least squares. Although Appendix A.2 also contains a formulation for regularization of total least squares to mitigate these large coefficients, selecting a regularization parameter is not trivial in the context of model discovery and is thus not explored further in this dissertation.

3.3.5 Ensemble Methods

Recent work has demonstrated that considering ensembles of SINDy models can result in increased robustness to noise when selecting a sparse subset of terms [28]. This procedure uses the LASSO regression with a range of regularizing parameters γ on subsets of the data samples (rows $\boldsymbol{\Theta}$), sampled with replacement. The frequency of terms in the resulting collection of sparse models can then be used to determine the likelihood of that term's inclusion in the complete model. This procedure requires parameter selection for the regularization parameters of LASSO and the size and number of subsets as well as an averaging procedure to determine the term likelihoods.

Specifically, consider B random subsets of 10% of the rows of $\mathbf{\Theta}$ and a range $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_G)$ of regularization parameters in Equation (3.3) such that $||\boldsymbol{\xi}||_0 = 1$ for γ_1 and $\gamma_G = \gamma_1/10$. For random subset i and regularization parameter γ_j, compute the solution of the corresponding LASSO problem to obtain the vector $\boldsymbol{\xi}^{(ij)}$. After computing the $\boldsymbol{\xi}^{(ij)}$ vectors, the probability of inclusion for the k^{th} term P_k can be computed as an average

$$P_k = \frac{\left|\left\{(ij) : \xi_k^{(ij)} \neq 0\right\}\right|}{BG}.$$

The final subset of terms in the model is then selected using a predetermined threshold τ after which coefficients can be determined using an alternative method as was demonstrated for the LASSO in Equation (3.4).

CHAPTER 4

DISCOVERY ON SIMULATED DATA

This chapter describes the procedure for validating the use of the SINDy equation discovery algorithm for active nematic systems by first recovering the partial differential equation used to generate active nematic simulation data. Section 4.1 discusses the numerical methods used to simulate two models of active nematics. Section 4.2 outlines the results and insights gained by applying SINDy to that simulated data.

4.1 Simulating Active Nematic Systems

As a prototype for active nematic PDE models, and inspired by the discovery for experimental data discussed later in Chapter 6, consider the model:

$$\begin{aligned} \boldsymbol{Q}_t &= -\boldsymbol{u} \cdot \nabla \boldsymbol{Q} + [\boldsymbol{Q}, \boldsymbol{\Omega}] + c_3 \boldsymbol{E}^{\mathrm{ST}} - c_8 \Delta^2 \boldsymbol{Q}, \\ \boldsymbol{u} &= -D \nabla \cdot \boldsymbol{Q}. \end{aligned} \tag{4.1}$$

Note that the symmetric and traceless structure of $\boldsymbol{Q}$ as given in Equation (2.3) allows Equation (4.1) to be written in terms of two independent fields $\lambda = \lambda(\boldsymbol{x}, t)$ and $\mu = \mu(\boldsymbol{x}, t)$:

$$\begin{aligned} \lambda_t &= D(\lambda_x^2 - \lambda_y^2 + \lambda_x \mu_y + \lambda_y \mu_x) + D(\mu_{xx} - 2\lambda_{xy} - \mu_{yy})\mu - \frac{c_3 D}{2}(\lambda_{xx} + \lambda_{yy}) \\ &\quad - c_8(\lambda_{xxxx} + 2\lambda_{xxyy} + \lambda_{yyyy}), \\ \mu_t &= D(\lambda_x \mu_x + 2\mu_y \mu_x - \lambda_y \mu_y) + D(\mu_{xx} - 2\lambda_{xy} - \mu_{yy})\lambda - \frac{c_3 D}{2}(\mu_{xx} + \mu_{yy}) \\ &\quad - c_8(\mu_{xxxx} + 2\mu_{xxyy} + \mu_{yyyy}). \end{aligned} \tag{4.2}$$

The equations are solved on a 2D periodic domain. This system of equations is nonlinear and stiff due to the hyperdiffusion terms. A Fourier pseudospectral method is thus employed in space and a fourth order integrating factor Runge-Kutta (IFRK4) scheme in time to solve the initial value problem [29, 55]. Specifically, the integrating factor method solves the linear component of the evolution equation exactly and thus allows us to sidestep a severe (fourth-order) time-step restriction that would arise if using a Runge-Kutta scheme directly on Equation (4.2). To illustrate the method consider a generic PDE for the variable $u = u(\boldsymbol{x}, t)$ of the form

$$u_t = \mathcal{L}u + \mathcal{N}(u, t),$$

where $\mathcal{L}$ and $\mathcal{N}$ are linear and nonlinear operators, respectively. The spatially-discretized form of this equation is

$$u_t = \boldsymbol{L}u + \boldsymbol{N}(u, t). \tag{4.3}$$

We can then multiply Equation (4.3) by an integrating factor $e^{-\boldsymbol{L}t}$ to obtain

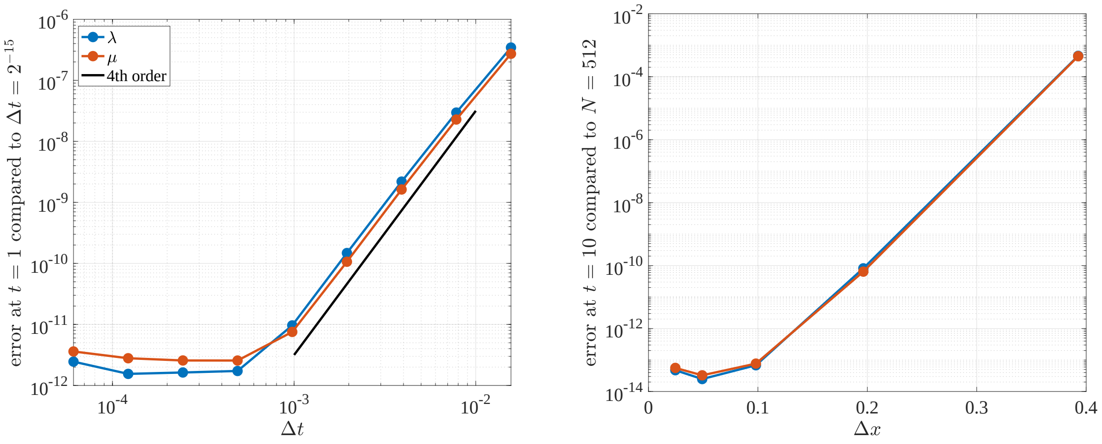
$$v_t = e^{-\boldsymbol{L}t}\boldsymbol{N}(e^{\boldsymbol{L}t}v, t), \tag{4.4}$$

where $v = e^{-\boldsymbol{L}t}u$ and $\boldsymbol{L}$ is the matrix form of the discretized linear operator $\mathcal{L}$. The integrating factor can be computed cheaply for a spectral discretization because it is diagonal in the Fourier basis. This reformed and discretized PDE can be evolved forward in time using a fourth order Runge-Kutta method:

$$\begin{aligned} a &= \boldsymbol{N}(v_n, t_n), \quad b = \boldsymbol{N}(v_n + a/2, t_n + \Delta t/2), \\ c &= \boldsymbol{N}(v_n + b/2, t_n + \Delta t/2), \quad d = \boldsymbol{N}(v_n + c, t_n + \Delta t), \\ v_{n+1} &= v_n + \frac{\Delta t}{6}(a + 2b + 2c - d). \end{aligned}$$

We implemented this method in MATLAB and verified that it exhibits fourth-order convergence in the time step Δt and spectral convergence in the number of grid points N (Figure 4.1).

We note that care must be taken to properly preserve the properties of odd and even derivatives in Fourier space [43]. Specifically, for a 1D spatial grid with



(a) Convergence in the time step Δt.

(b) Convergence in the grid size $\Delta x = 2\pi/N$.

Figure 4.1 Demonstration of the fourth order convergence in Δt and exponential convergence in the grid size Δx of the numerical method described in Section 4.1. The parameters are those given in Section 4.1.

N points, where N is even, the wavenumbers k assume integer values in the range $-N/2 + 1 \leq k \leq N/2$. The so-called "Nyquist mode" $k = N/2$ must be made zero for terms with an odd number of derivatives, while this mode is nonzero for terms with an even number of derivatives. The extension to 2D follows naturally.

It is well known that spectral schemes for nonlinear PDEs suffer from aliasing errors unless certain decay conditions are met for the Fourier coefficients [66]. For a quadratic nonlinearity, the Fourier coefficients should have decayed to zero (to some numerical tolerance) for modes with wave number $|k| > N/3$. Likewise, for cubic nonlinearities, the coefficients $|k| > N/4$ should be zero. This condition is sometimes imposed by filtering the coefficients as the simulation proceeds. Such filtering was not necessary for the simulations presented in this thesis; i.e. the decay conditions were met by the simulations without intervention. Filtering was applied in later sections to initial conditions obtained from the data but never as part of the simulation procedure; cf. Sections 6.4 and 6.5.

4.2 Rediscovering Active Nematic PDEs from Simulated Data

Although the SINDy method has been used on a variety of simulated data from canonical models, its applications on real data and more complex models are only beginning to appear [39, 42, 57]. A recent paper has successfully used the method to obtain a continuum PDE description of a driven colloidal suspension of Quincke rollers [87]. To get a sense of the ability and sensitivity of this method as applied to the active nematic system, the SINDy discovery was first tested on clean simulation data.

4.2.1 Accuracy of Term Recovery

An advantage of SINDy compared to other data-driven modeling approaches and parameter estimation methods is its ability to extract a closed form equation from inputted data. To verify that this would be effective for the active nematic system, governing PDEs equations were "rediscovered" from simulated data using multiple models: the model given in Equation (4.1) and the dimensionless form of the model given in Equation (2.15), the latter of which can be written as [67]

$$\begin{aligned} \boldsymbol{Q}_t + D\nabla\cdot(\boldsymbol{u}\boldsymbol{Q}) &= \boldsymbol{Q}\left(\frac{N_\gamma^2}{4} - N_\gamma^2\boldsymbol{Q}^2\right) - \gamma_2\Delta\boldsymbol{Q} - \frac{1}{N_\gamma^2}\Delta^2\boldsymbol{Q}, \\ \boldsymbol{u} &= -\nabla\cdot\boldsymbol{Q}. \end{aligned} \tag{4.5}$$

A random Fourier series for both λ and μ is used as a smooth initial condition for simulation. For each test, the procedure from Section 2.3 was used to construct a term library with state variables $\boldsymbol{Q}, \boldsymbol{u}$, a maximum derivative order of 2, and a maximum order of $\boldsymbol{Q}$ of 3. Additional checks for symbolic uniqueness were used after substituting $\boldsymbol{u}$ using the relationship $\boldsymbol{u} \propto -\nabla\cdot\boldsymbol{Q}$ which is present for both simulated models (4.1) and (4.5). As both models also include a particularly high order derivative in the form of the bilaplacian $\Delta^2\boldsymbol{Q}$, this term was manually added to the library of candidate terms. Ultimately, the library contained 24 terms which were numerically constructed using finite differences.

4.2.1.1 Model 1 We first considered the simplest model suggested by the data-driven discovery (Section 6.1), as given in Equation (4.1) and shown in component form in Equation (4.2). This model is simple in that it includes a relatively small number of terms: advection, vorticity and rate of strain. We note that $\boldsymbol{E}^{\mathrm{ST}} = -\frac{D}{2}\Delta\boldsymbol{Q}$ for this velocity equation, so the higher-order term proportional to c_2 is required to ensure linear well-posedness of the equation. This model is compared to previously proposed models in Chapter 6.

As we will see in Section 6.1, the discovery process on experimental data on a square domain of length $L_E = 312.4\,\mu\mathrm{m}$ yields coefficients $c_3 = 0.3$, $c_8 = 3.3 \times 10^4$ $\mu\mathrm{m}^4/\mathrm{s}$ and $D = 420$ $\mu\mathrm{m}^2/\mathrm{s}$. We non-dimensionalize Equation (4.1) according to the length scale $L = L_E/(2\pi)$ and time scale $T = L^2/D$, and thus simulate Equation (4.1) on a 2D domain of size $[0, 2\pi]^2$ with coefficients $c_3 = 0.3$, $c_8 \to c_8 L^4/T \approx 0.03$ and $D \to DT/L^2 = 1$. The simulations were conducted using the method detailed in Section 4.1, and the dynamics were sufficiently resolved with 256^2 points in space and $\Delta t = 2^{-10}$. Simulation data was collected at intervals of $\Delta t = 0.25$ until $t_{\max} = 100$. After constructing the full nonlinear library of candidate terms for the evolution of $\boldsymbol{Q}$ as described in Section 4.2.1, forward selection was used to determine potential candidate models as detailed in Algorithm 2. This approach was able to identify the correct model terms within the first ten selected, but was not able to cleanly identify the model as shown in Table 4.1. Specifically, while the advection ($u_k\partial_k Q_{ij}$) and vorticity ($[\boldsymbol{Q}, \boldsymbol{\Omega}]$) terms are identified correctly as the first two terms, the third and fourth terms are spurious. However, continuing the procedure allows us to recover the correct equation: the fifth ($\Delta^2\boldsymbol{Q}$) and sixth ($\Delta\boldsymbol{Q}$) terms are correct, the latter being proportional to $\boldsymbol{E}^{\mathrm{ST}}$ as noted above, and the coefficients of the spurious

R^2	$\boldsymbol{Q}_t =$	Coefficients					
0.19	$u_k \partial_k Q_{ij}$	-0.15	-0.58	-0.58	-0.56	-0.61	-0.97
0.62	$[\boldsymbol{Q}, \boldsymbol{\Omega}]$		0.53	0.53	0.52	0.56	0.97
0.68	$Q_{ml} \partial_k Q_{ij} \partial_m Q_{kl}$			-0.88	-1.26	-1.06	0.01
0.74	$Q_{kl} \partial_k Q_{ji} \partial_m Q_{ml}$				0.94	0.91	-0.00
0.76	$\Delta^2 Q$					-0.00	-0.03
0.99	$\partial_k \partial_k Q_{ij}$						-0.15

Table 4.1 Greedy forward selection results for data generated using simulations of Equation (4.1). Each row represents the next term added in order to maximize the R^2 of the equation.

Subset size	R^2	Terms
1	0.19	$u_k \partial_k Q_{ij}$
2	0.62	$[\boldsymbol{Q}, \boldsymbol{\Omega}]$, $u_k \partial_k Q_{ij}$
3	0.70	$[\boldsymbol{Q}, \boldsymbol{\Omega}]$, $u_k \partial_k Q_{ij}$, $Q_{ml} \partial_k Q_{ij} \partial_m Q_{kl}$
4	0.99	$[\boldsymbol{Q}, \boldsymbol{\Omega}]$, $\Delta^2 \boldsymbol{Q}$, $\partial_k \partial_k Q_{ij}$, $u_k \partial_k Q_{ij}$

Table 4.2 Best subset selection results for data generated using simulations of Equation (4.1). Row n represents the optimal collection of n terms in order to maximize the R^2 of the equation. The data was subsampled in time and space, so only 50% of the data was used, to make the problem computationally feasible.

terms are made small. Using the original sequential thresholding approach of SINDy corroborates these results.

Due to the limited size of the library (24 terms), the subset of terms that maximizes the R^2 of the recovered equation after determining the coefficients via ordinary least squares can be computed up to size four. The results of this best subset selection are shown in Table 4.2. As expected, the advection and vorticity terms are selected in the best subsets of size one and two, matching the results of the forward selection. While the best subset of size 3 contains a spurious cubic term, the best subset of size four indeed recovers the correct terms and their coefficients. Although this combinatorial procedure is feasible for libraries with a small number of terms, it is only reasonable after subsampling the term library in time and space. Our numerical experiments revealed that random subsampling has minimal effect on the model recovered, a result consistent with prior literature [73].

It is challenging to identify the exact cause of the emergence of spurious terms in the forward selection. However, this method relies on terms being easily distinguishable to avoid spurious correlations [10]. Unfortunately, the procedure in Section 2.3 generates a library which is complete but also full of correlated terms. Some of these correlations can be attributed to the particular matrix structure of $\boldsymbol{Q}$ and the relationship $\boldsymbol{u} = -D\nabla \cdot \boldsymbol{Q}$. This issue is discussed in more detail in Section 4.2.2.

The result of this correlation is that regression methods, including greedy selection methods, struggle to distinguish the contribution of individual terms and instead use incorrect or multiple correlated terms to best fit the time evolution. Though this challenge has not been fully discussed in the literature of SINDy or related methods, we note that it could be a generic property of a set of features generated from a common set of base data. This idea is more fully discussed in Section 4.2.2 for simulated data and for experimental results in Section 6.1.

4.2.1.2 Model 2 For another example, we apply the discovery procedure to the model proposed by Oza & Dunkel [67] (Equation (4.5)). To generate the data the parameters $\gamma_2 = 1.5, D = 1.5, N_\gamma = 3$ were used, and simulations were conducted on a domain $[0, 2\pi]^2$ with 256^2 points in space and time step $\Delta t = 2^{-10}$. Data was collected at intervals of $\Delta t = 0.125$ until a final time $t_{\max} = 50$.

The results of greedy forward selection and best subset selection for this model are shown in Tables 4.3 and 4.4, respectively. Specifically, Table 4.3 shows that forward selection misidentifies spurious terms. Though the terms $\partial_k \partial_k Q_{ij}$, $Q_{kl} Q_{lj} Q_{ik}$, $u_k \partial_k Q_{ij}$, $\Delta^2 \boldsymbol{Q}$, and Q_{ij} are identified, at first glance it appears that the advection term $Q_{ij} \partial_k u_k$ is missing. However, it should be noted that

$$[\boldsymbol{Q}, \boldsymbol{\Omega}] + 2[Q_{kl} \partial_l \partial_j Q_{ik}]^{\mathrm{ST}} - 2[Q_{lk} \partial_i \partial_j Q_{kl}]^{\mathrm{ST}} = (\nabla \cdot \boldsymbol{u})\boldsymbol{Q}.$$

Thus, the correct equation is recovered via this composition of alternative terms in the library. It should be noted that this composition is only possible given the relation $\boldsymbol{u} = -\nabla \cdot \boldsymbol{Q}$. Specifically, the level of linear dependence present in this library is increased by this equality and could be mitigated but not entirely resolved by generating terms as in Section 2.3 with a base variable of only $\boldsymbol{Q}$. As was the case for the previous example, the coefficients for the two truly spurious terms are zero when the set of terms includes the correct model.

In comparison, the best subset selection of size six was able to fully recover the model terms including the advection terms $u_k \partial_k Q_{ij}$ and $Q_{ij} \partial_k u_k$. It is evident that the best subsets of size $n = 1, 2$ do not contain terms that are present in the governing equation, an example of a spurious term being $[\boldsymbol{uu}]^{\mathrm{ST}}$. However, the best subset of size $n = 3$ does contain the correct terms, and a substantial increase in R^2 is seen when the best subset of size $n = 5$ is used which contains only terms present in the model.

There are several factors that contribute to this difficulty, including the increased number of terms and the increased correlation between correct terms. This additional correlation is demonstrated in Figure 4.3 and is further discussed in Section 4.2.2.

4.2.2 Correlation in Nonlinear Library

There are key challenges that arise when generating an overcomplete polynomial library of nonlinear terms which can hamper the success of the sparse regression or parameter estimation. Foremost among these is the potential for multicollinearity in

Table 4.3 Greedy forward selection results for data generated using simulations of Equation (4.5)). Each row represents the next term added in order to maximize the R^2 of the equation.

R^2	$\boldsymbol{Q}_t =$	Coefficients									
0.10	$2[u_i u_j]^{\mathrm{ST}}$	0.16	0.16	0.16	0.16	0.21	0.26	0.36	0.09	-0.00	0.00
0.16	$2[Q_{kl}\partial_l\partial_j Q_{ik}]^{\mathrm{ST}}$		-0.08	-0.08	-0.09	-0.08	-0.17	-0.23	-0.43	-1.28	-1.42
0.21	$[\boldsymbol{Q}, \boldsymbol{\Omega}]$			0.11	0.11	0.10	0.00	-0.07	-0.29	-1.28	-1.41
0.23	$Q_{ij}\partial_m Q_{lk}\partial_k Q_{lm}$				0.06	0.12	0.12	0.06	0.13	0.09	-0.00
0.26	$\Delta^2\boldsymbol{Q}$					-0.00	-0.01	-0.01	-0.03	-0.10	-0.11
0.29	$[Q_{lk}\partial_i\partial_j Q_{kl}]^{\mathrm{ST}}$						0.18	0.29	0.74	2.56	2.83
0.34	$\partial_k\partial_k Q_{ij}$							-0.13	-0.31	-1.48	-1.44
0.52	$u_k\partial_k Q_{ij}$								-0.52	-1.35	-1.41
0.96	$Q_{kl}Q_{lj}Q_{ik}$									-6.64	-8.47
0.99	Q_{ij}										2.03

Table 4.4 Best subset selection results using 50% of the data generated using simulations of Equation (2.15). Row n represents the optimal collection of n terms in order to maximize the R^2 of the equation. In order to reduce the computational burden, $Q_{ij}\partial_k u_k$ is manually included for the set of size 6.

Subset size	R^2	Terms
1	0.10	$[u_i u_j]^{\mathrm{ST}}$
2	0.16	$[Q_{kl}\partial_l\partial_j Q_{ik}]^{\mathrm{ST}}$, $[u_i u_j]^{\mathrm{ST}}$
3	0.21	$Q_{lk}\partial_l\partial_k Q_{ji}$, $Q_{ij}\partial_k u_k$, $[u_i u_j]^{\mathrm{ST}}$
4	0.41	$\Delta^2\boldsymbol{Q}$, $\partial_k\partial_k Q_{ij}$, $Q_{ij}\partial_k u_k$, $u_k\partial_k Q_{ij}$
5	0.94	$\Delta^2\boldsymbol{Q}$, $\partial_k\partial_k Q_{ij}$, $Q_{ij}\partial_k u_k$, $u_k\partial_k Q_{ij}$, $Q_{kl}Q_{lj}Q_{ik}$
6	0.99	$\Delta^2\boldsymbol{Q}$, Q_{ij} , $\partial_k\partial_k Q_{ij}$, $u_k\partial_k Q_{ij}$, $Q_{kl}Q_{lj}Q_{ik}$, $Q_{ij}\partial_k u_k$

the generated library. This is an issue for libraries constructed for scalar terms, but is potentially more salient in tensor libraries due to the range of possible contractions for each set of base tensor terms. Though care is taken to make sure that terms are not symbolically equivalent, numerical experimentation has demonstrated that there is a wide range of data for which the generated library will have numerically similar terms. The easiest approach to examine this similarity is to consider the Pearson correlation coefficient between two vectors x and y:

$$\psi_{x,y} = \frac{\mathrm{cov}(x,y)}{\sigma_x\sigma_y} = \frac{\mathbb{E}[xy] - \mathbb{E}[x]\mathbb{E}[y]}{\sqrt{\mathbb{E}[x^2] - (\mathbb{E}[x])^2}\sqrt{\mathbb{E}[y^2] - (\mathbb{E}[y])^2}}. \tag{4.6}$$

This standard correlation metric measures the correlation between two vectors relative to the size of their variance.

Figure 4.2 shows the Pearson correlation coefficient between all generated library terms, using values of $\boldsymbol{Q}$ obtained by simulating Equation (4.1). In the figure, only symbolically unique terms are kept after substituting $\boldsymbol{u} = -D\nabla\cdot\boldsymbol{Q}$. Figure 4.2 demonstrates the prevalence of correlated terms even for simulated data with no noise pollution. For example, the term $u_k\partial_k Q_{ij}$, which is in the equation and thus enclosed in a green box, is highly correlated with the term $\partial_l Q_{ji} \partial_k Q_{lk}$ which is not. Although this in and of itself does not invalidate the SINDy method, it is known that strong correlations in the set of independent variables of a linear regression cause issues in coefficient calculation and in turn with sparse selection [27]. Although there is no immediate fix for this issue in the context of generating or manipulating SINDy's library of terms, it can be acknowledged and accounted for when considering the discovery results.

Figure 4.3 demonstrates the correlations between the generated library terms using values of $\boldsymbol{Q}$ generated by simulating Equation (4.5). In agreement with the results of Figure 4.2, this equation also suffers from strong correlations between library terms. Note in particular that the majority of terms which are built using the base variables $\boldsymbol{Q} \otimes \nabla\nabla\boldsymbol{Q}$ are correlated strongly with each other. The same is true for those built from $\boldsymbol{u} \otimes \nabla\boldsymbol{Q}$ or $\boldsymbol{Q} \otimes \nabla\boldsymbol{Q} \otimes \nabla\boldsymbol{Q}$. These consistent patterns can be accounted for when examining the results of either the forward selection or best subset selection. A notable difference between the two sets of results is that the correct terms (boxed in green) in Figure 4.2 are correlated mainly with terms outside the correct set, whereas correct terms in Figure 4.3 are more strongly correlated with each other. As an example of the latter, term $\Delta\boldsymbol{Q}$ (which is in the governing equation) is strongly correlated with $\boldsymbol{Q}$, $\boldsymbol{Q}^3$ and $\Delta^2\boldsymbol{Q}$ which are also in the governing equation.

4.2.3 Accuracy of Parameter Estimation

If the correct terms in the governing PDE can be recovered, the second phase of the discovery would be to accurately determine the corresponding coefficients. Using data collected via simulation of Equation (4.5), the SINDy linear system was constructed using only the correct terms and derivatives computed with finite differences. The system was then solved using ordinary least squares in order to find the parameters of the model.

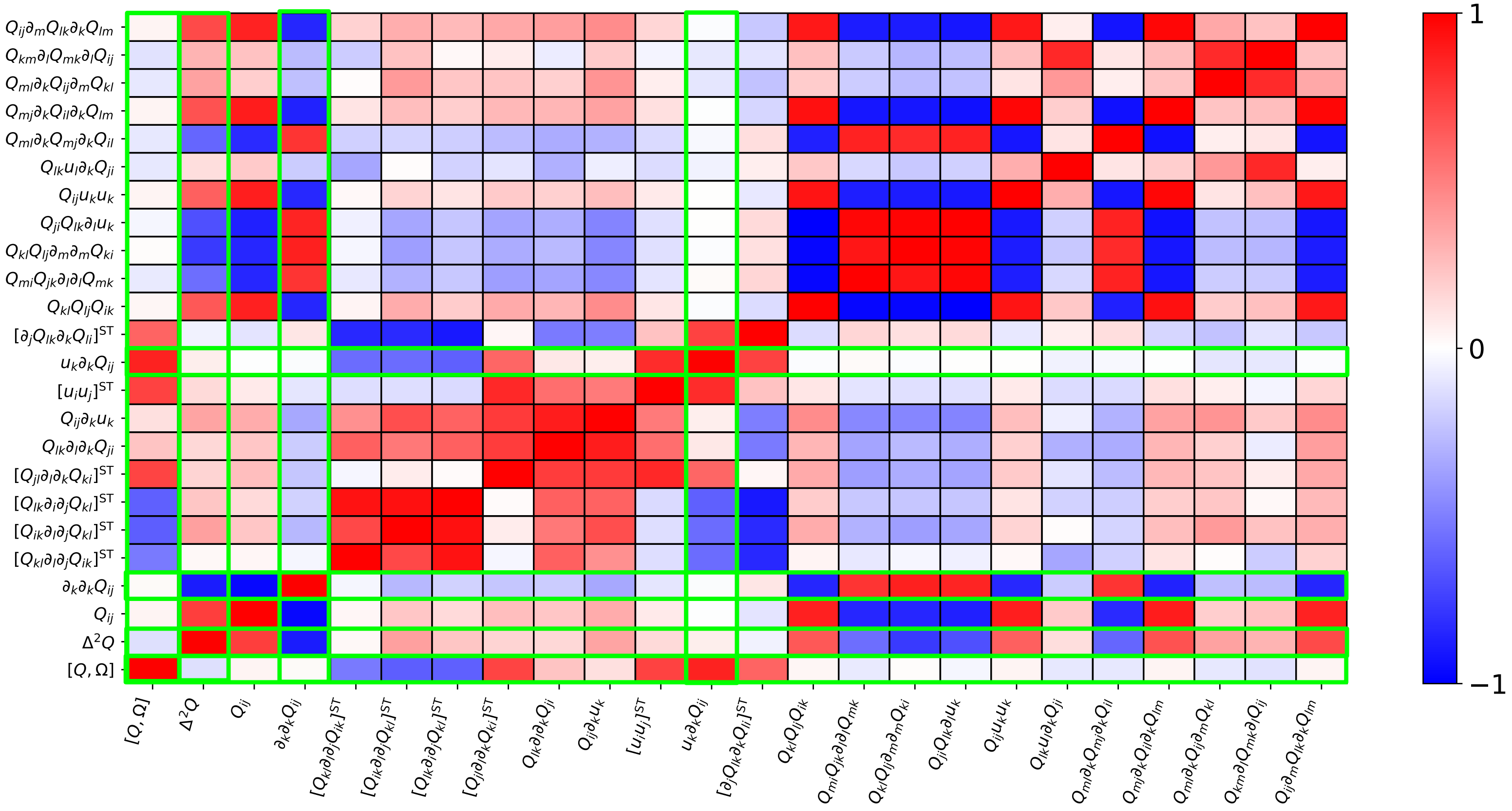


Figure 4.2 Pearson correlation coefficient (4.6) of all of the terms in the $\boldsymbol{Q}$-library, using data simulated from Equation (4.1). The correct terms are boxed in green.

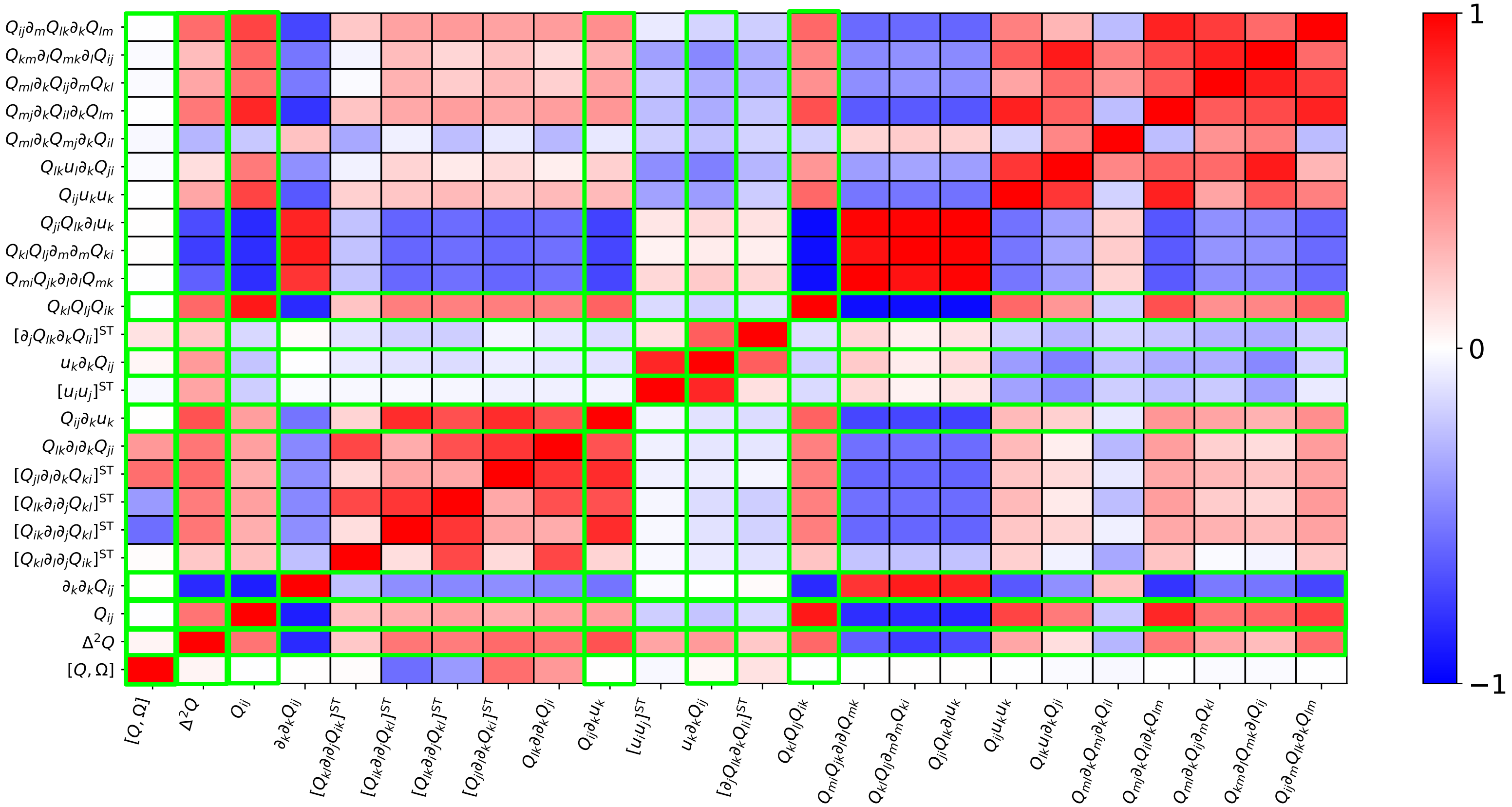
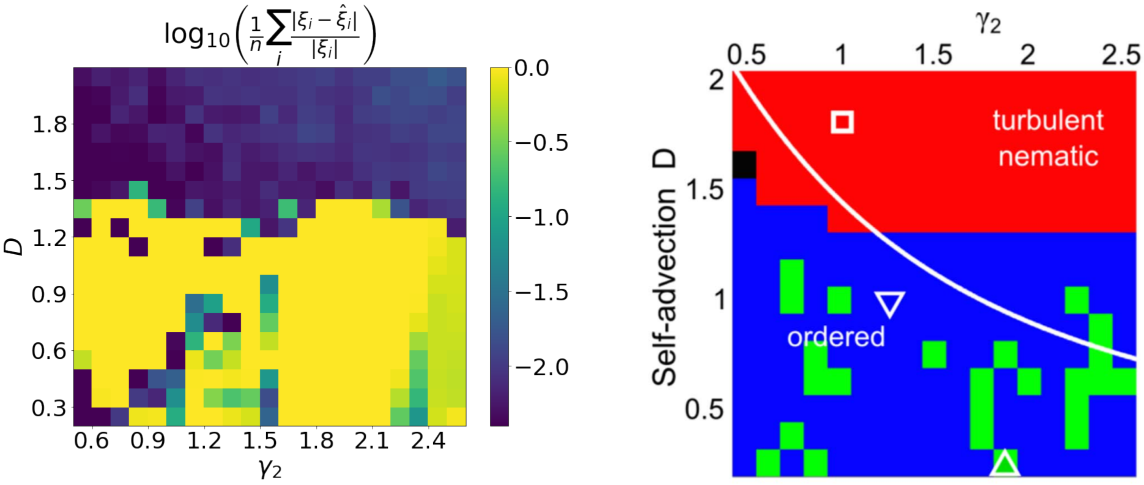


Figure 4.3 Pearson correlation coefficient (4.6) of all of the terms in the $\boldsymbol{Q}$-library, using data simulated from Equation (4.5). The correct terms are boxed in green.

The accuracy of the coefficients in the recovery as the model parameters D and γ_2 are varied is shown in Figure 4.4. From Figure 4.4(a), we observe that the coefficients are recovered more accurately for relatively large values of D. The phase diagram in Figure 4.4(b), which was obtained using numerical simulations in [67], shows that smaller (larger) values of D generally correspond to ordered (chaotic) states. Taken together, these panels show that the coefficients can be accurately recovered only if the data is sufficiently dynamic to present the key features of the system. In regimes in which the data converges to a steady or patterned state, the dynamics are not sufficiently apparent to achieve an accurate regression.



(a)

(b)

Figure 4.4 (a) Accuracy of coefficient recovery from simulated data of Equation (4.5) for different values of dimensionless model parameters D and γ_2. (b) Phase diagram from [67] that shows ordered states (blue and green) and chaotic states (red) for different choices of the dimensionless parameters D and γ_2. Comparing panels (a) and (b), we note that SINDy is less accurate when the system achieves a relatively static ordered state. The simulations are conducted for $N_\gamma = 3$ and 256^2 grid points in space.

CHAPTER 5

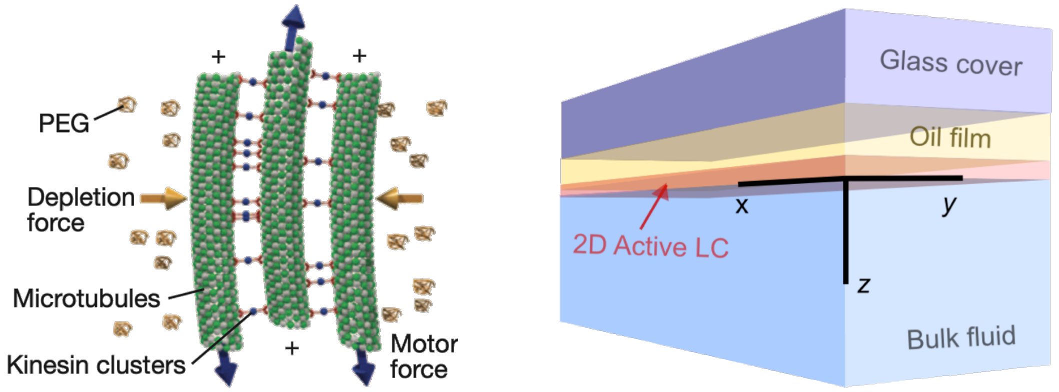
DATA EXTRACTION

This chapter treats the process of extracting the key state variables of the microtubule (MT)-kinesin system from experimental videos. Specifically, we extract the coarse-grained density, orientation and velocity of MTs. The approach is notable in that the videos are the only input and the extraction and subsequent smoothing is carefully tuned to satisfy known physical constraints. Section 5.1 describes the origin of the experimental videos used. Section 5.2 describes the procedure for approximating the state variables from pixel intensity of the video frames. In Section 5.3, the approximated state variables are smoothed and the resulting data is validated to ensure accuracy for the application of SINDy.

5.1 Experimental Data

The microtubule (MT)-kinesin active nematic system was created and studied by Zvonimir Dogic and colleagues in their pioneering experiments [16, 75]. In these experiments, extracted bovine brain cell MTs were bundled using a polymer (PEG) and spun onto an oil-water interface using a centrifuge. The polymer induces the so-called "depletion attraction" between MTs and thus binds them into bundles (Fig. 5.1a). ATP was then added to the mixture and the system was observed and photographed using fluorescence microscopy. A schematic of the experimental setup is shown in Figure 5.1b and some sample experimental images are shown in Figure 5.2. Analogous experimental platforms have also been developed by several other groups [24, 36, 88].

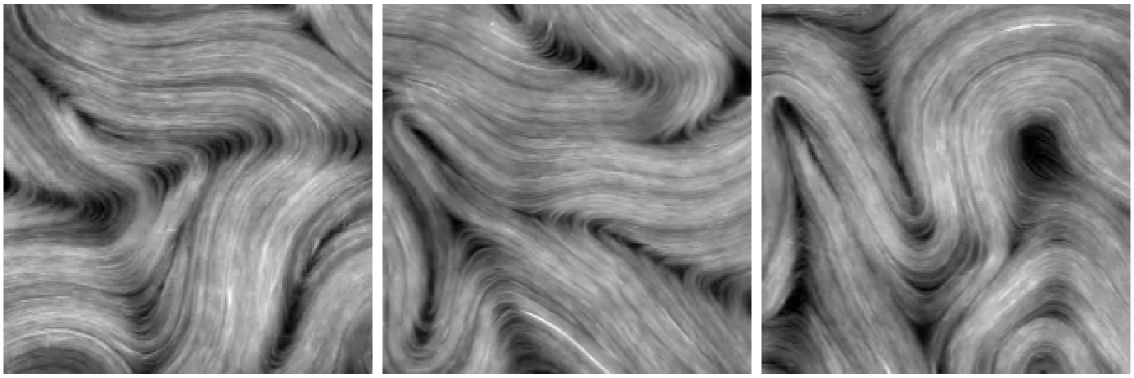
In this dissertation, the experimental video will be Supplemental Movie 1 in DeCamp *et al.* [16]. Key features of experimental observations for this system are:



(a)

(b)

Figure 5.1 (a) Diagram of MTs and kinesin clusters binding, reproduced from [75]. Kinesin clusters exert sliding forces between the MTs, and PEG polymers induce attractive "depletion" interactions between MTs. (b) Schematic of the experimental setup, reproduced from [67]. A thin oil film (thickness $\sim 3\ \mu$m) separates a 2D active nematic film (thickness 0.2–1.0 μm) at the oil-water interface from a solid glass cover.



(a) Time 10 seconds

(b) Time 100 seconds

(c) Time 190 seconds

Figure 5.2 Snapshots of Supplemental Video 1 in [16].

1. Topological defects (Figure 1.3) which are created and annihilated in pairs.
2. Chaotic flow patterns.
3. Strands in MT bundles are distinct enough to extract MT orientation using image processing techniques.
4. The intensity of the image roughly correlates with the density of MTs in that area, i.e. brighter areas generally have more MTs.
5. The total intensity is roughly constant in time, implying that MT mass is roughly conserved in this system.
6. The MT density is uniform in most areas away from defects.

5.2 Approximating State Variables from Image Intensity

The only direct observable from the experimental images is image intensity or brightness, which we denote I_{ijk} at frame i and pixel j, k. However, there are established approaches for extracting key state information such as orientation from these images. Below we describe our procedure for extracting density (Section 5.2.1), velocity (Section 5.2.2) and (Section 5.2.3).

5.2.1 Density

Given the dark background of the experimental setup and the reflective capacity of the microtubules, the image intensity can be viewed as a direct approximation to microtubule density. However, there are pitfalls to this approximation. For example, an experimental adjustment partway through collection in [16] caused an increase in brightness, which would translate to an illusion of increased density. This is illustrated in the left panel of Figure 5.3, which shows the spatially averaged image intensity over time for the experiment. We thus normalized the image intensity after the abrupt jump, resulting in the relatively constant average intensity depicted in the right panel of Figure 5.3. We thus conclude that the number of MTs remains roughly constant in the field of view for the duration of the experimental video.

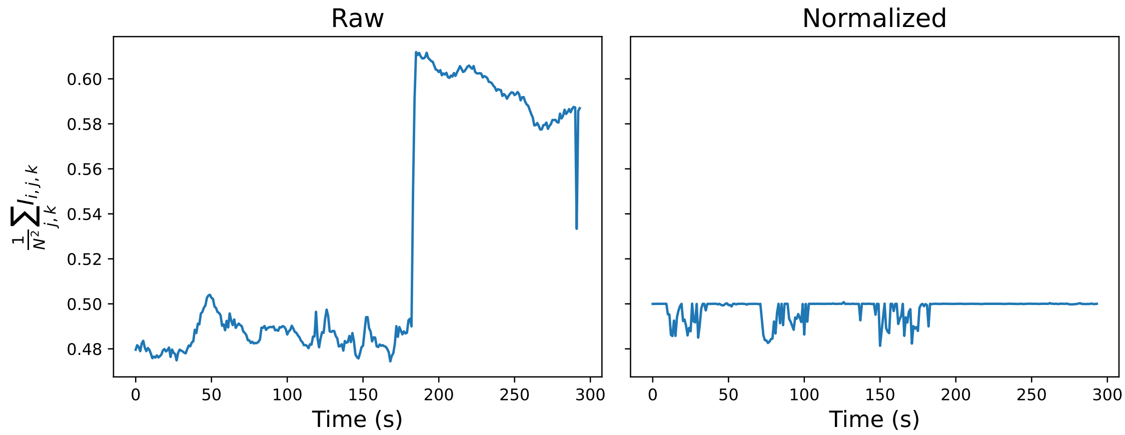


Figure 5.3 Left panel shows the image intensity, averaged over the domain, as a function of time. The right panel shows the same, but with the intensity normalized so as to remove the abrupt increase in brightness at time $t \approx 200$ s.

5.2.2 Velocity

The experimental video images contain a reasonable amount of variation in shadow and granularity, which provides an opportunity to approximate velocity using a patch-based particle image velocimetry (PIV) approach.

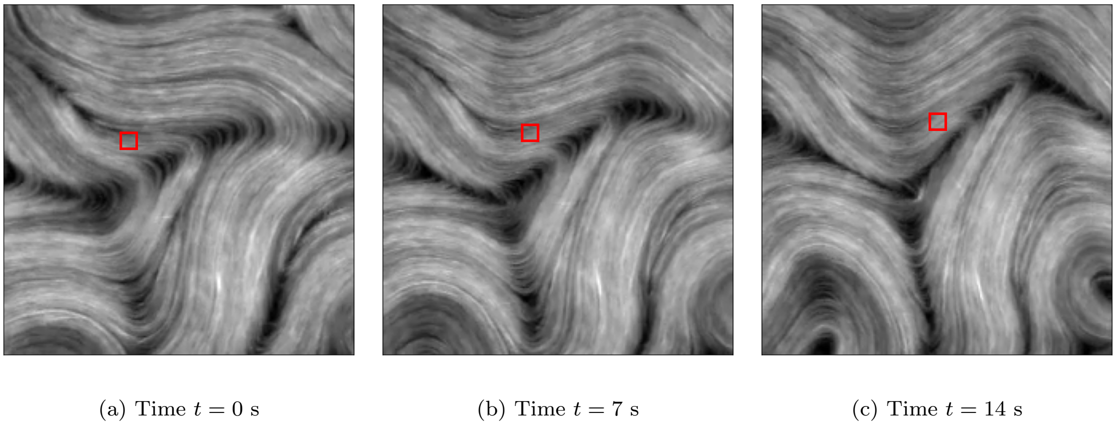
Consider M sequential 2D images (frames) equispaced in time by an interval Δt. The images have N pixels in both directions, with a spacing Δx between them. The velocity $\hat{\boldsymbol{v}}_{ijk}$ at a given pixel (j, k) at frame i can be approximated by comparing a "patch" of the image around that pixel with patches of the same size in the subsequent frame [80], as illustrated in Figure 5.4. Mathematically, the velocity approximation with a square patch of side length p can be written as:

$$\hat{\boldsymbol{v}}_{ijk} = \frac{1}{\Delta t}\left(\boldsymbol{x}_{(i+1)\hat{j}\hat{k}} - \boldsymbol{x}_{ijk}\right) = \frac{\Delta x}{\Delta t}\left(\hat{j} - j, \hat{k} - k\right), \tag{5.1}$$

$$\hat{j}, \hat{k} = \underset{1 \le m,n \le N}{\operatorname{argmin}} \sum_{r=-p/2}^{p/2} \sum_{s=-p/2}^{p/2} \left(I_{i(j+r)(k+s)} - I_{(i+1)(m+r)(n+s)}\right)^2 \tag{5.2}$$

for $1 \le i, j \le N$. The edges of the image are handled using reflections such that

$$I_{(-i)(-j)(-k)} = I_{ijk} \quad \text{and} \quad I_{(N+i)(N+j)(N+k)} = I_{(N-i)(N-j)(N-k)}. \tag{5.3}$$


(a) Time $t = 0$ s

(b) Time $t = 7$ s

(c) Time $t = 14$ s

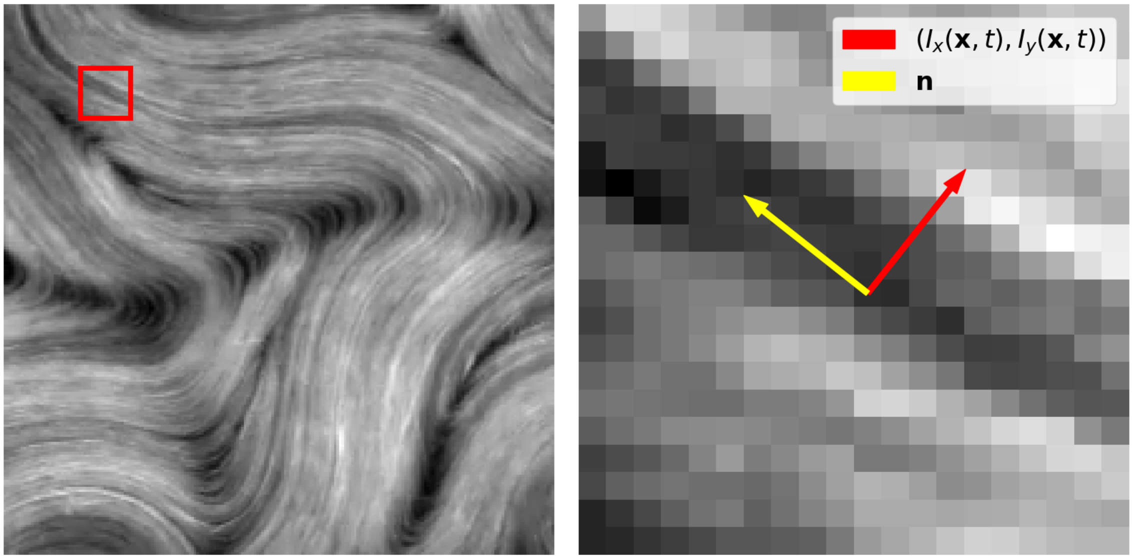
Figure 5.4 Particle image velocimetry via patch tracking over several time frames, as described in Section 5.2.2. The red box represents the image patch being tracked.

It is important to note that the velocity obtained through Equation (5.2) is based on the pixel locations; a discrete number of velocity vectors are thus permitted, making the velocity field nonsmooth. We remedy this issue by using Gaussian smoothing, as described in Section 5.3.2. The approximation in Equation (5.2) also assumes sufficiently fine experimental sampling; specifically, data must be sampled sufficiently frequently in time to minimize patch changes between frames while also sufficiently sampled in space to provide unique identifying detail. References to additional more rigorous methods from the study of "optical flow" are described in Appendix A.3.

5.2.3 Orientation

The experimental images considered herein include a reasonable amount of contrast between microtubule bundles due to shadowing and small non-uniformities in density. These contrasts allow for a local orientation to be computed using the intensity gradients of the image. This approach is standard and makes use of the image "structure tensor" [14, 50]

$$\boldsymbol{J}(\boldsymbol{x}, t) = \nabla I \nabla I^{\intercal} = \begin{pmatrix} I_x(\boldsymbol{x}, t)^2 & I_x(\boldsymbol{x}, t) I_y(\boldsymbol{x}, t) \\ I_x(\boldsymbol{x}, t) I_y(\boldsymbol{x}, t) & I_y(\boldsymbol{x}, t)^2 \end{pmatrix}, \tag{5.4}$$



(a)

(b)

Figure 5.5 The local MT orientation is extracted from experimental images by measuring the local gradient of the image intensity $I(\boldsymbol{x}, t)$, as described in Section 5.2.3. (a) The boxed segment demonstrates the zoomed area in (b). (b) The intensity gradient and its perpendicular component which is used to approximate the director $\hat{\boldsymbol{n}}$.

where subscripts denote partial derivatives of the image intensity $I(\boldsymbol{x}, t)$. Due to the symmetry of this matrix, the eigenvector $\boldsymbol{J}$ corresponding to the smaller eigenvalue represents the least intensity variation and hence an approximation to the director $\hat{\boldsymbol{n}}(\boldsymbol{x}, t)$. The eigenvectors of the structure tensor are shown for a zoom-in of a single experimental image in Figure 5.5. Figure 5.8(a) shows a plot of the resulting director $\hat{\boldsymbol{n}}(\boldsymbol{x}, t)$ for a single experimental image.

5.3 Smoothing the State Variables

As discussed in Section 3.2, SINDy suffers when the input data is not sufficiently smooth, as numerical derivatives of the data become overcome with noise. To circumvent this issue, smoothing is applied to the density, velocity $\hat{\boldsymbol{v}}$ and director $\hat{\boldsymbol{n}}$ obtained from experimental images as described in Section 5.2.1, Section 5.2.2 and Section 5.2.3, respectively. Inspired by kinetic theories of the form considered by Gao *et al.* [31,32], an appropriately smoothed density function $\Psi(\boldsymbol{x}, \boldsymbol{n}, t)$ that represents the density of MTs with a given director $\boldsymbol{n}$ at a given spacetime location $(\boldsymbol{x}, t)$ is constructed. The state variables of the system can then be recovered as the moments of the density function.

Specifically, let $I(\boldsymbol{x}, t)$, $\hat{\boldsymbol{n}}(\boldsymbol{x}, t) = (\cos\hat{\theta}, \sin\hat{\theta})$ and $\hat{\boldsymbol{v}}(\boldsymbol{x}, t)$ be the intensity, director and velocity field obtained from experimental data. We define Ψ as

$$\Psi(\boldsymbol{x}, \boldsymbol{n}(\theta), t) = \int_{-\infty}^{\infty} \mathrm{d}t' \int_{\mathbb{R}^2} \mathrm{d}\boldsymbol{x'}\, I(\boldsymbol{x}', t') g_{\boldsymbol{Q}}(\boldsymbol{x} - \boldsymbol{x}', t - t') \tau_{\sigma_n}(\theta - \hat{\theta}(\boldsymbol{x}', t')), \tag{5.5}$$

where $g_{\boldsymbol{Q}}$ is a spacetime smoothing function for the MT orientations to be specified in Section 5.3.1, and τ_{σ_n} is a modification of the so-called *wrapped Gaussian distribution* with standard deviation σ_n. As discussed in Appendix A.4, τ_{σ_n} is invariant under $\theta \to \theta + \pi$, the symmetry appropriate for nematics.

Remark 5.3.1. *While Eq. 5.5 involves an integral over all of space and time, the Gaussian density $g_{\boldsymbol{Q}}$ makes the integral effectively local. The result is that $\Psi(\boldsymbol{x}, \boldsymbol{n}(\theta), t)$ is an empirical distribution, based on the values of I and $\hat{\boldsymbol{n}}$ in a neighborhood of $(\boldsymbol{x}, t)$. Because $g_{\boldsymbol{Q}}$ is a Gaussian, Ψ is smooth in space-time.*

Using this formulation, the filament density is

$$\rho(\boldsymbol{x}, t) = \int_{S^1} \mathrm{d}\boldsymbol{n}\, \Psi(\boldsymbol{x}, \boldsymbol{n}, t) = \int_{-\infty}^{\infty} \mathrm{d}t' \int_{\mathbb{R}^2} \mathrm{d}\boldsymbol{x'}\, I(\boldsymbol{x}', t') g_{\boldsymbol{Q}}(\boldsymbol{x} - \boldsymbol{x}', t - t'). \tag{5.6}$$

The second moment tensor defined in Equation (2.1) is then

$$\begin{aligned} \boldsymbol{D}(\boldsymbol{x}, t) &= \int_{S^1} \mathrm{d}\boldsymbol{n} \Psi(\boldsymbol{x}, \boldsymbol{n}, t) \boldsymbol{n}\boldsymbol{n} \\ &= \int_{-\infty}^{\infty} \mathrm{d}t' \int_{\mathbb{R}^2} \mathrm{d}\boldsymbol{x'}\, I(\boldsymbol{x}', t') \hat{\boldsymbol{n}}\hat{\boldsymbol{n}}(\boldsymbol{x}', t') g_{\boldsymbol{Q}}(\boldsymbol{x} - \boldsymbol{x}', t - t'), \end{aligned} \tag{5.7}$$

where the second line follows from taking $\sigma_n = 0$. That is, the orientations are not directly smoothed; doing so would simply lead to a constant prefactor in Equation (5.7), as shown in Appendix A.4. The Q-tensor, or the centered second moment defined in Equation (2.2) is then

$$\boldsymbol{Q}(\boldsymbol{x},t) = \frac{\boldsymbol{D}(\boldsymbol{x},t)}{\rho(\boldsymbol{x},t)} - \frac{\boldsymbol{I}}{2}. \tag{5.8}$$

The velocity of the microtubules can be obtained from the filament flux

$$\boldsymbol{j}(\boldsymbol{x},t) = \int_{-\infty}^{\infty} \mathrm{d}t' \int_{\mathbb{R}^2} \mathrm{d}\boldsymbol{x}' I(\boldsymbol{x}',t')\hat{\boldsymbol{v}}(\boldsymbol{x}',t') g_v(\boldsymbol{x}-\boldsymbol{x}',t-t'), \tag{5.9}$$

where g_v is a spacetime smoothing function for the velocity field to be specified in Section 5.3.2. The smoothed filament velocity field can thus be expressed as

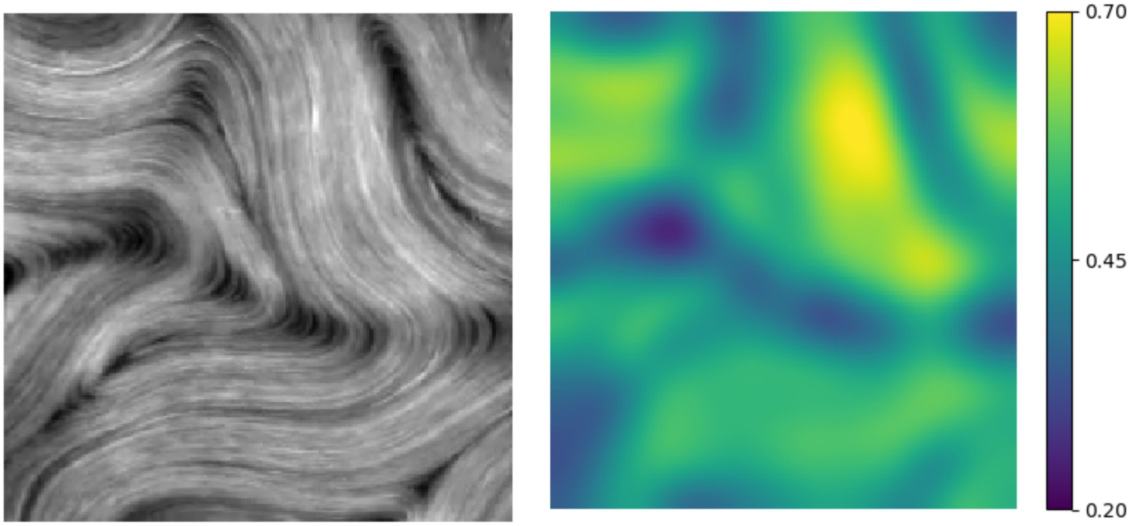
$$\boldsymbol{u}(\boldsymbol{x},t) = \frac{\boldsymbol{j}(\boldsymbol{x},t)}{\rho(\boldsymbol{x},t)} \tag{5.10}$$

These continuous representations provide a structure for approximating smoothed versions of the state variables $\rho(\boldsymbol{x},t)$, $\boldsymbol{Q}(\boldsymbol{x},t)$ and $\boldsymbol{u}(\boldsymbol{x},t)$ that will be ultimately used for the equation discovery in Chapter 6. Specifically, the image intensity I, microtubule orientation $\hat{\boldsymbol{n}}$, and velocity $\hat{\boldsymbol{v}}$ obtained from experimental images are made smooth through integration with the densities $g_{\boldsymbol{Q}}$ and g_v.

Remark 5.3.2. *The definition of the smoothed filament flux, Equation 5.9, uses a different weighting than the density and orientation use, g_v instead of $g_{\boldsymbol{Q}}$, to define the empirical distribution of the filament flux near $(\boldsymbol{x},t)$. It was found that these quantities needed different amounts of smoothing to effectively satisfy the physically motivated validation methods described below. While it is possible to achieve the same results using a consistent smoothing for the empirical distributions, this requires a pre-processing step in which the velocity data is first smoothed or sharpened. Equation 5.9 is preferred for the simplicity of the notation.*

5.3.1 Smoothing Density and Q-tensor

Given a frame i and a pixel location (j,k), we first wish to obtain a smooth density ρ_{ijk}. The discrete analogue of Equation (5.6) can be written for $1 \le i \le M$, $1 \le$



(a)

(b)

Figure 5.6 (a) The experimental image at time $t = 10$ seconds and (b) the corresponding smoothed density ρ.

$j, k \leq N$ as

$$\rho_{ijk} = \sum_{l=-w_{t_Q}}^{w_{t_Q}} \sum_{m=-w_{x_Q}}^{w_{x_Q}} \sum_{n=-w_{x_Q}}^{w_{x_Q}} g_{\boldsymbol{Q}}(l, m, n) I_{(i+l)(j+m)(k+n)}, \tag{5.11}$$

where

$$g_{\boldsymbol{Q}}(l, m, n) = G_{\sigma_{t_Q}}(l) G_{\sigma_{x_Q}}(m) G_{\sigma_{x_Q}}(n), \quad G_\sigma(i) = \frac{\exp\left(-\frac{i^2}{2\sigma^2}\right)}{\sum_{l=-w}^{w} \exp\left(-\frac{i^2}{2\sigma^2}\right)} \tag{5.12}$$

and $w = 4\sigma + 1/2$ is the window size. That is, the spacetime smoothing function g is assumed to be a product of 1D Gaussians G with possibly different standard deviations σ in space (σ_{x_Q}) and time (σ_{t_Q}). We note that reflections were used on the spatial and time boundaries, as specified by Equation (5.3). The extracted density at a single frame is compared with the experimental image in Figure 5.6.

We proceed by obtaining a smoothed Q-tensor $\boldsymbol{Q}_{ijk}$, reminding the reader that $\boldsymbol{Q}_{ijk}$ refers to the tensor $\boldsymbol{Q}$ evaluated at the ith time point and (j, k) pixel location as described in Section 1.4. Given the local director $\hat{\boldsymbol{n}}_{ijk}$, as obtained from experimental images using the procedure described in Section 5.2.3, Equation (5.7) can be discretized as

$$\boldsymbol{D}_{ijk} = \sum_{l=-w_{t_Q}}^{w_{t_Q}} \sum_{m=-w_{x_Q}}^{w_{x_Q}} \sum_{n=-w_{x_Q}}^{w_{x_Q}} g_{\boldsymbol{Q}}(l,m,n) I_{(i+l)(j+m)(k+n)} \hat{\boldsymbol{n}}_{(i+l)(j+m)(k+n)} \hat{\boldsymbol{n}}_{(i+l)(j+m)(k+n)}, \tag{5.13}$$

and Equation (5.8) implies that

$$\boldsymbol{Q}_{ijk} = \frac{\boldsymbol{D}_{ijk}}{\rho_{ijk}} - \frac{\boldsymbol{I}}{2}. \tag{5.14}$$

The accuracy of data-driven modeling is highly dependent on the accuracy of the underlying data. The most salient feature of the microtubule orientation field are the defects or discontinuities, at which $\boldsymbol{Q} = 0$. It has been proposed that the presence and behavior of these defects are the main driver behind the observed dynamics in the system [16]. We thus determined the smoothing parameters σ_{t_Q} and σ_{x_Q} by ensuring that the resulting smooth field $\boldsymbol{Q}$ faithfully captures the number of defects, and their creation and annihilation dynamics. Specifically, we explored a range of values for σ_{t_Q} and σ_{x_Q} between 0 and 20 (in units of pixels) and identified the defects by locating the intersections of the zero-contours of the components of $\boldsymbol{Q}$, as described in [67]. The considered experimental data includes between 2-8 defects at any given time. We found the optimal values to be $\sigma_{t_Q} = 1$ and $\sigma_{x_Q} = 10$, which yielded roughly 6 defects per frame. Moreover, the creation and annihilation dynamics were represented reasonably well. See Figure 5.7 for a comparison of several different smoothing values and the corresponding defects at a fixed time. Upon close inspection of this figure, it can be observed that in panel (a) there are too many defects identified and in panel (c) there is a missing defect in the top right, while the defects are identified correctly in panel (b). Plots of the resulting director $\boldsymbol{n}$ and order parameter S for a single experimental image are shown in Figure 5.8(b,c). As can be observed, the director field before smoothing displays wild and random fluctuations in many areas while the smoothed final result demonstrates good agreement with the underlying MTs. Additionally, the order parameter S accurately reflects the defect crack through the

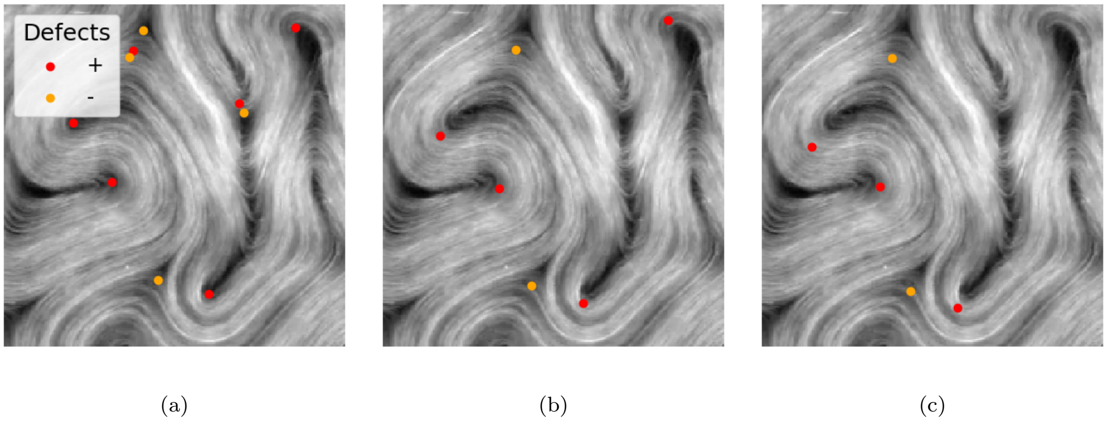


Figure 5.7 A comparison of the positive (red) and negative (orange) defects identified for several different levels of smoothing of the approximated orientation field at time $t = 290$ seconds. Defects for smoothing parameters (a) $\sigma_{t_Q} = 0, \sigma_{x_Q} = 4$, (b) $\sigma_{t_Q} = 1, \sigma_{x_Q} = 10$, and (c) $\sigma_{t_Q} = 2, \sigma_{x_Q} = 16$.

middle of the image; moreover, despite the low density at the top right of the image, there is no defect there, a feature correctly reflected in the plot of S.

5.3.2 Smoothing Velocity

We proceed by determining the smoothed velocity values $\boldsymbol{u}_{ijk}$ from the velocity data $\hat{\boldsymbol{v}}_{ijk}$ obtained using patch tracking, as described in Section 5.2.2. Equations (5.9) and (5.10) can be combined and written in the discrete form

$$\boldsymbol{u}_{ijk} = \frac{1}{\rho_{ijk}} \sum_{l=-w_{t_v}}^{w_{t_v}} \sum_{m=-w_{x_v}}^{w_{x_v}} \sum_{n=-w_{x_v}}^{w_{x_v}} g_v(l,m,n) I_{(i+l)(j+m)(k+n)} \hat{\boldsymbol{v}}_{(i+l)(j+m)(k+n)}, \tag{5.15}$$

To verify the effectiveness of the velocity extraction procedure, the parameters used to determine velocities $\boldsymbol{u}_{ijk}$, namely the standard deviations σ_{x_v} and σ_{t_v} implicit in Equation (5.15) and the patch window size p in Equation (5.2), were determined by their ability to the mass conservation law

$$\rho_t + \nabla \cdot (\rho \boldsymbol{u}) = 0. \tag{5.16}$$

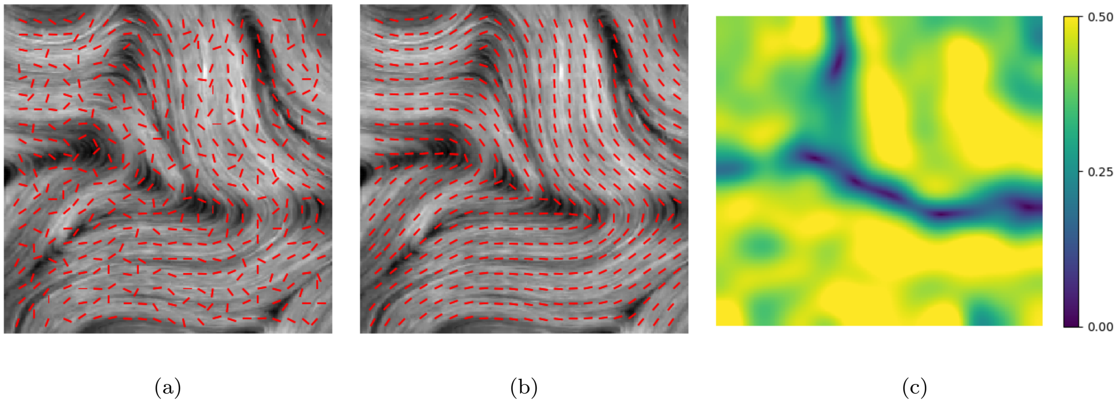


Figure 5.8 (a) Local director $\hat{\boldsymbol{n}}$ for a single experimental image, as obtained using the procedure described in Section 5.2.3. (b, c) The smoothed Q-tensor, obtained using the procedure described in Section 5.3.1, yields a smooth director field $\boldsymbol{n}$ (b) and order parameter S (c) through the formulas (2.5) and (2.4), respectively.

We considered a range of values of σ_{t_v} (0-5 pixels), σ_{x_v} (0-20 pixels), and p (1-15) and sought to minimize the quantity

$$\frac{\|\rho_t + \nabla \cdot (\rho \boldsymbol{u})\|}{\|\rho_t\|}, \tag{5.17}$$

thus obtaining the optimal values $\sigma_{t_v} = 2, \sigma_{x_v} = 12$, and $p = 3$. Centered second-order finite differences were used to evaluate both the time and space derivatives in Equation (5.17). After determining the parameters, we used ordinary least squares to find the optimal value of c in the equation

$$\rho_t = -c\nabla \cdot (\rho \boldsymbol{u}).$$

We obtained $c = 0.94$ which had an R^2 of 0.72. The obtained value of c is quite close to unity and the R^2 value is adequate, indicating that mass conservation is satisfied reasonably well by our extraction and smoothing procedure.

The final velocity extracted from the experimental data in [16] at several frames is shown in Figure 5.9. From this figure, we observe that the velocity extracted via particle image velocimetry (Figure 5.9a) contains some noise. The smoothed velocity in Figure 5.9b corrects for this randomness and demonstrates an accurate description of the movement of MTs over the three presented time steps.

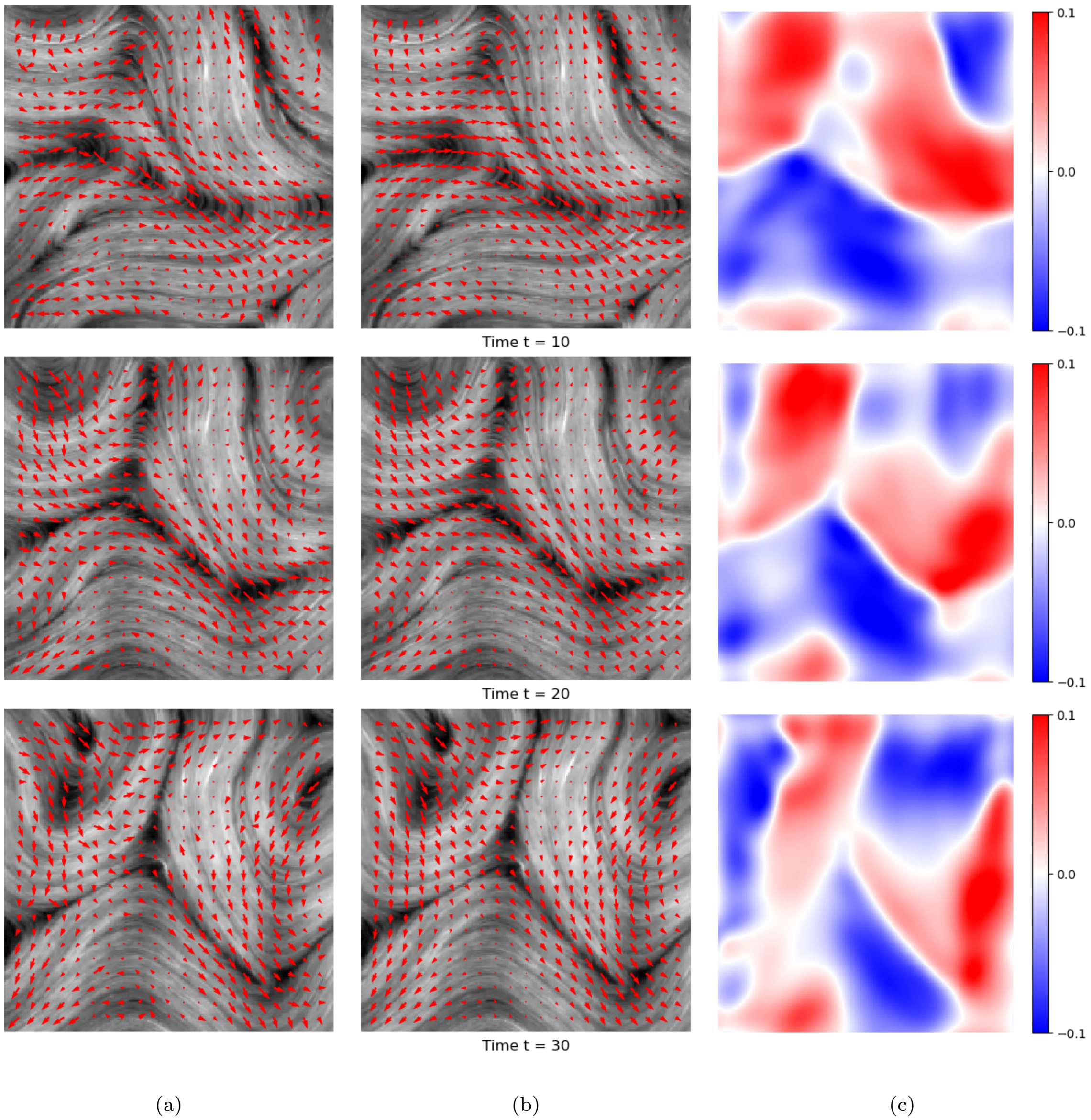


Figure 5.9 (a) Local velocity $\hat{\boldsymbol{v}}$ for a single experimental image, as obtained using the patch-tracking procedure described in Section 5.2.2. (b) The velocity field $\boldsymbol{u}$ obtained after smoothing, using the procedure described in Section 5.3.2. (c) The corresponding vorticity $\omega = \nabla \times \boldsymbol{u}$. The values of time are given in seconds.

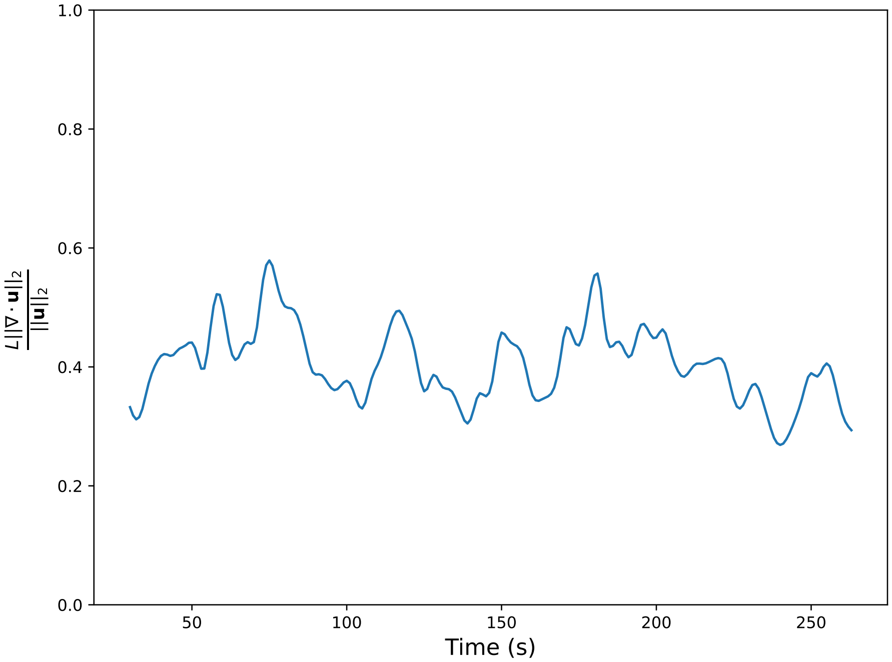


Figure 5.10 Plot of the relative size of the L^2-norms of $\nabla \cdot \boldsymbol{u}$ and $\boldsymbol{u}\sqrt{N/A}$, where A is the experimental area and $N = 8$ is the number of defects averaged over time.

It should be noted that the extracted velocity data can also be used to examine the approximate compressibility of the system. Specifically, $\nabla \cdot \boldsymbol{u}$ can be computed at each spacetime point using finite differences and compared against $\boldsymbol{u}\sqrt{N/A}$, where A is the experimental area and N the number of defects averaged over time. Figure 5.10 shows the ratio of the L^2 norms of these two quantities as a function of time. Note that it is fairly constant though not particularly small, implying that the velocity field is not divergence free, $\nabla \cdot \boldsymbol{u} \neq 0$ in general. As discussed in Section 2.2.3, while the system is compressible in 3D it may not be so in 2D due to upwelling or sinks caused by the exchange of fluid between the quasi-2D active nematic film and the bulk fluid underneath (Fig. 5.1). Practically, the fact that $\nabla \cdot \boldsymbol{u}$ is not small motivates the use of the traceless rate of strain $\boldsymbol{E}^{\mathrm{ST}}$ in the term library (Table 2.1) instead of the rate of strain $\boldsymbol{E}$, since $\mathrm{Tr}(\boldsymbol{E}) = \nabla \cdot \boldsymbol{u}$.

CHAPTER 6

DISCOVERY ON EXPERIMENTAL DATA

In this chapter we present the main result of this dissertation: a continuum PDE model for the active nematic system discovered from experimental video data. In Section 6.1 we describe a PDE discovered for the evolution of the Q-tensor. In Section 6.2 we describe a similarly discovered equation for the velocity field. In Section 6.3, we show that, taken together, the corresponding system of equations is linearly ill-posed and propose an augmentation to the system with a higher-order regularizing term. In Section 6.4 we present a way to recover the coefficients of some of the more uncertain terms in the proposed system by using a temporally nonlocal forecasting procedure which compares simulated results with the experimental data. In Section 6.5 we perform further numerical simulations of the discovered equation and qualitatively compare our results with the experiments.

Before proceeding we briefly comment on the units used in this section: for the sake of brevity dimensional quantities will typically not be written with their associated units. A parameter value c with dimensions of length to the power ℓ and time to the power τ shall be understood to mean $c \times 10^{\ell}$ microns$^{\ell}\cdot$ seconds$^{\tau}$. For example, a diffusivity $D = 4$ is understood to be $400\,\mu\text{m}^2/\text{s}$.

6.1 Orientation Evolution Equation for Microtubule-Kinesin System

Once validated state variable data has been extracted from the experimental video data, using the procedure detailed in Chapter 5, the library of terms is constructed using the procedure described in Section 2.3. The Gaussian smoothing used in the probability density formulation of the extraction (Section 5.3) yields smooth state variables, which allows for the use of centered finite differences instead of one of the more noise robust methods detailed in Section 3.2. However, data near boundaries was removed after differentiation to avoid inaccuracies incurred by the reflection of the Gaussian kernels in the data smoothing (see Equation (5.3)) or by the directional finite differences on the boundary points. Given the values of the smoothing parameters σ_{x_Q}, σ_{t_Q}, σ_{x_v} and σ_{t_v} determined in Sections 5.3.1 and 5.3.2, we found it sufficient to remove 30 edge pixels in both space and time. Finally, the constructed library was randomly sampled across space-time to improve computational efficiency.

Section 3.3 presented the idea of a two-stage regression process for approximating solutions of Equation (3.1); first, a sparse subset of the library terms is determined and second, the optimal (in some norm) coefficients are computed for that reduced library. Table 6.1 shows the results of using forward selection (see Algorithm 2) on the experimental data, in which terms were selected up to a library of size 10. After these k-sparse libraries were determined, for $1 \leq k \leq 10$, the coefficients for each sparse model were computed using ordinary least squares.

The results of forward selection with ordinary least squares (Table 6.1) exhibit several notable features. First, the coefficient of determination R^2 increases until reaching an inflection point at the fourth term, indicated by the dashed line, after which any additional terms add little to the reconstruction accuracy. This inflection point indicates that there is indeed a sparse subset of terms in the library that can account for the majority of the features of $\boldsymbol{Q}_t$. Second, the first two selected terms both have coefficients of magnitude roughly unity for models up to size ten. These terms, which correspond to advection and rotation, are included in almost all previously proposed models and are expected to have coefficients of -1 and $+1$, respectively. Finally, although most of the selected terms are present in the Beris-Edwards equation (2.8) for $\boldsymbol{Q}$, there is a notable absence of the elastic energy term $\Delta \boldsymbol{Q}$. Moreover, the alignment energy terms $\boldsymbol{Q}$ and $\boldsymbol{Q}^3$ are only selected after the inflection point, and the sign of the $\boldsymbol{Q}$ ($\boldsymbol{Q}^3$)-term is negative (positive), the opposite to what is expected from the phenomenological argument given below Equation (2.11).

Table 6.1 Forward selection results for the active nematic experimental video data [16].The experimental data is extracted using the procedure described in Chapter 5, and the library is generated using the procedure described in Section 2.3. Half of the data is sampled at random in spacetime for computational efficiency, and the coefficients are determined using ordinary least squares. The dashed line indicates the inflection point in R^2, after which adding more terms to the PDE causes the R^2 to increase only minimally.

R^2	$\boldsymbol{Q}_t =$	Coefficients									
0.32	$u_k \partial_k Q_{ij}$	-0.83	-0.85	-0.94	-1.05	-1.28	-1.13	-1.10	-1.15	-1.17	-1.17
0.57	$[\boldsymbol{Q}, \boldsymbol{\Omega}]$		0.70	0.84	1.19	1.18	1.19	1.12	1.12	1.13	1.15
0.64	$2[\partial_i u_j]^{\mathrm{ST}}$			0.08	0.13	0.14	0.14	0.14	0.18	0.18	0.18
0.77	$2[Q_{ik} Q_{lj} \partial_k u_l]^{\mathrm{ST}}$				-2.34	-2.30	-2.26	-2.15	-1.48	-1.34	-1.65
0.81	$2[u_k \partial_i Q_{kj}]^{\mathrm{ST}}$					0.26	0.16	0.14	0.16	0.18	0.17
0.82	$2[\partial_l Q_{ki} \partial_k Q_{lj}]^{\mathrm{ST}}$						0.70	0.76	0.68	0.73	0.66
0.83	$2[Q_{jl} \partial_l \partial_k Q_{ki}]^{\mathrm{ST}}$							-0.25	-0.26	-0.21	-0.23
0.83	Q_{ij}								-0.02	-0.03	-0.05
0.84	$Q_{ji} u_l \partial_k Q_{lk}$									-0.72	-0.87
0.84	$Q_{kl} Q_{lj} Q_{ik}$										0.47

As was noted in the numerical experiments to recover the governing PDE from simulation data (Chapter 4), the constructed library may be prone to correlations which can hinder the accuracy of the forward selection method [92, 93]. Figure 6.1 shows the Pearson correlation coefficient for the terms in the library, as computed with respect to the experimental data. There are strong correlations between several of the selected terms and other terms which are common to the Beris-Edwards equations. For example, the bulk alignment energy terms $\boldsymbol{Q}$ and $\boldsymbol{Q}^3$ are not included in the discovered equation, but are correlated with the discovered terms $\boldsymbol{E}^{\mathrm{ST}}$ and $[Q_{ik}Q_{lj}\nabla u_{kl}]^{\mathrm{ST}}$.

Given the size of the library for the orientation evolution (46 terms), a brute force approach to computing the true k-sparse solution of Equation (3.1) is feasible for $k \leq 6$ if a smaller portion of the data is considered. The results for a brute force computation of the best k-sparse library are presented in Table 6.2. The best subset of size 4 agrees with the forward selection results. The R^2 values for the subsets of size 5 and 6 are similar to that for size 4, indicating that little is gained in the R^2-metric from including more terms in the governing equation.

Comparing the forward selection (Table 6.1) and brute force (Table 6.2) results, we settled on an evolution equation with $k = 4$ terms:

$$\boldsymbol{Q}_t = -1.04\boldsymbol{u}\cdot\nabla\boldsymbol{Q} + 1.18[\boldsymbol{Q},\boldsymbol{\Omega}] - 0.26\boldsymbol{E}^{\mathrm{ST}} - 2.28\big(\boldsymbol{Q}(\boldsymbol{Q}:\nabla\boldsymbol{u}) - \boldsymbol{Q}^{\perp}(\boldsymbol{Q}^{\perp}:\nabla\boldsymbol{u})\big), \tag{6.1}$$

where we have expressed the cubic term in a somewhat compact form by introducing the notation

$$\boldsymbol{Q}^{\perp} = \boldsymbol{\epsilon}\boldsymbol{Q} = \begin{bmatrix} \mu & -\lambda \\ -\lambda & -\mu \end{bmatrix} \quad \text{for} \quad \boldsymbol{\epsilon} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}. \tag{6.2}$$

We note that the cubic flow-alignment term $\boldsymbol{Q}(\boldsymbol{Q} : \boldsymbol{\nabla}\boldsymbol{u})$ is also present in the Beris-Edwards equation (2.8); while this term is in our library, both the forward selection

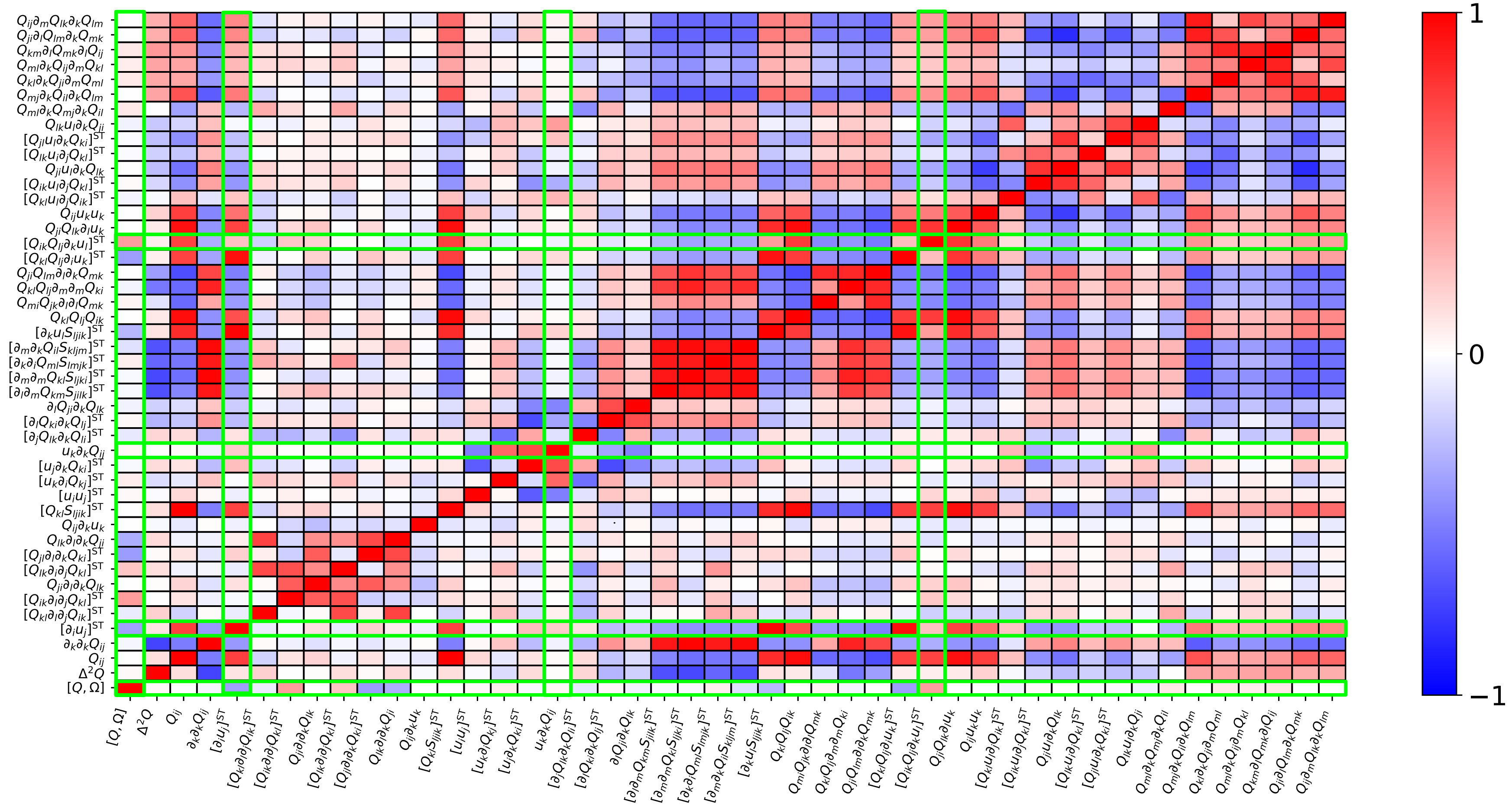


Figure 6.1 Pearson correlation coefficient of the library terms for the $\boldsymbol{Q}$-evolution equation (Table 2.1), evaluated on the experimental video data reported in [16]. The green boxes indicate terms that were selected using forward selection before the inflection point in R^2, as indicated by the dashed line in Table 6.1.

and best subset selection approaches select instead the composite term $\boldsymbol{Q}(\boldsymbol{Q} : \nabla \boldsymbol{u}) - \boldsymbol{Q}^{\perp}(\boldsymbol{Q}^{\perp} : \nabla \boldsymbol{u})$.

The coefficients in Equation (6.1) are determined using ordinary least squares. However, given the known error in both dependent and independent variables of the linear system $\boldsymbol{\Theta}$, an errors-in-variables approach (Section 3.3.4) may be more suited to determining the parameters of the model. The coefficients determined using total least squares are:

$$\boldsymbol{Q}_t = -1.22\boldsymbol{u} \cdot \nabla \boldsymbol{Q} + 1.49[\boldsymbol{Q}, \boldsymbol{\Omega}] - 0.37\boldsymbol{E}^{\mathrm{ST}} - 3.24\big(\boldsymbol{Q}(\boldsymbol{Q} : \nabla \boldsymbol{u}) - \boldsymbol{Q}^{\perp}(\boldsymbol{Q}^{\perp} : \nabla \boldsymbol{u})\big). \tag{6.3}$$

We recall that the coefficients of the advection and vorticity terms are expected to be -1 and $+1$, respectively, and we choose the remaining coefficients to be in between those given in Equations (6.1) and (6.3). We thus obtain the evolution equation for $\boldsymbol{Q}$ that we will use going forward:

$$\boldsymbol{Q}_t = -\boldsymbol{u} \cdot \nabla \boldsymbol{Q} + [\boldsymbol{Q}, \boldsymbol{\Omega}] - 0.3\boldsymbol{E}^{\mathrm{ST}} - 3\big(\boldsymbol{Q}(\boldsymbol{Q} : \nabla \boldsymbol{u}) - \boldsymbol{Q}^{\perp}(\boldsymbol{Q}^{\perp} : \nabla \boldsymbol{u})\big). \tag{6.4}$$

6.1.1 Alternative Sparse Regression Approaches

A variety of methods were presented in Chapter 3 for use in constructing a sparse solution to the linear system in Equation (6.9). Although all methods aim to approximate the same zero-norm regularized problem given in Equation (3.1), they are not guaranteed to yield equivalent sparse selections. However, in the case of the active nematic system considered herein, numerical experimentation showed that all of the common methods agreed on a few terms. Foremost among these common methods is brute force "best-subset" selection which truly finds the subset of a given size which maximizes the R^2 to give a "best-fit" model as seen in Table 6.2. As a comparison, Table 6.3 demonstrates sparse selection results using the more traditional LASSO method, see Equation (3.3), in which the regularization parameter γ is varied in order to find potential sparse models of different sizes. Table 6.4 demonstrates

Table 6.2 Best subset selection results on experimental data. A random subset of 5% of the data in spacetime is sampled for the sake of computational feasibility. Each row contains the model of that size which maximizes the R^2.

R^2	Terms
0.66	$[\partial_i u_j]^{\text{ST}},\ u_k \partial_k Q_{ij},\ [\boldsymbol{Q}, \boldsymbol{\Omega}]$
0.76	$[Q_{ik} Q_{lj} \partial_k u_l]^{\text{ST}},\ [\partial_i u_j]^{\text{ST}},\ u_k \partial_k Q_{ij},\ [\boldsymbol{Q}, \boldsymbol{\Omega}]$
0.76	$[\partial_i u_j]^{\text{ST}},\ [u_k \partial_i Q_{kj}]^{\text{ST}},\ [Q_{ik} Q_{lj} \partial_k u_l]^{\text{ST}},\ u_k \partial_k Q_{ij},\ [\boldsymbol{Q}, \boldsymbol{\Omega}]$
0.76	$[u_k \partial_i Q_{kj}]^{\text{ST}},\ [\partial_k u_l S_{ljik}]^{\text{ST}},\ Q_{kl} Q_{lj} \partial_m \partial_m Q_{ki},\ [Q_{ik} Q_{lj} \partial_k u_l]^{\text{ST}},$ $u_k \partial_k Q_{ij},\ [\boldsymbol{Q}, \boldsymbol{\Omega}]$

Table 6.3 LASSO path results using experimental data. A random subset of 50% of the data in spacetime is sampled. Each row represents the term added to the model as the LASSO regularization parameter γ is decreased.

γ	Terms
13.00	$u_k \partial_k Q_{ij}$
12.32	$u_k \partial_k Q_{ij},\ [u_j \partial_k Q_{ki}]^{\text{ST}}$
11.63	$u_k \partial_k Q_{ij},\ [u_j \partial_k Q_{ki}]^{\text{ST}},\ [\boldsymbol{Q}, \boldsymbol{\Omega}]$
10.95	$u_k \partial_k Q_{ij},\ [u_j \partial_k Q_{ki}]^{\text{ST}},\ [\boldsymbol{Q}, \boldsymbol{\Omega}],\ [Q_{jl} \partial_l \partial_k Q_{ki}]^{\text{ST}}$
10.26	$u_k \partial_k Q_{ij},\ [u_j \partial_k Q_{ki}]^{\text{ST}},\ [\boldsymbol{Q}, \boldsymbol{\Omega}],\ [Q_{jl} \partial_l \partial_k Q_{ki}]^{\text{ST}},\ [\partial_i u_j]^{\text{ST}}$
9.58	$u_k \partial_k Q_{ij},\ [u_j \partial_k Q_{ki}]^{\text{ST}},\ [\boldsymbol{Q}, \boldsymbol{\Omega}],\ [Q_{jl} \partial_l \partial_k Q_{ki}]^{\text{ST}},\ [\partial_i u_j]^{\text{ST}},$ $[\partial_l Q_{ki} \partial_k Q_{lj}]^{\text{ST}},\ [Q_{kl} Q_{lj} \partial_i u_k]^{\text{ST}},\ [Q_{ik} Q_{lj} \partial_k u_l]^{\text{ST}}$

Table 6.4 Sequentially thresholded ridge regression results using experimental data. The data is randomly sampled in spacetime (50%). Each row represents the terms selected by STRidge for the given regularization parameter γ.

γ	Terms
7.94	$[\boldsymbol{Q}, \boldsymbol{\Omega}]$
3.98	$[\boldsymbol{Q}, \boldsymbol{\Omega}]$, $u_k \partial_k Q_{ij}$
2.51	$[\boldsymbol{Q}, \boldsymbol{\Omega}]$, $[Q_{jl} \partial_l \partial_k Q_{ki}]^{\mathrm{ST}}$, $[u_j \partial_k Q_{ki}]^{\mathrm{ST}}$, $u_k \partial_k Q_{ij}$, $[\partial_l Q_{ki} \partial_k Q_{lj}]^{\mathrm{ST}}$
1.26	$[\boldsymbol{Q}, \boldsymbol{\Omega}]$, $[u_j \partial_k Q_{ki}]^{\mathrm{ST}}$, $u_k \partial_k Q_{ij}$

comparable selection results using the STRidge algorithm described in Section 3.3.2. Finally, Table 6.5 gives the results of the ensemble model approach described in Section 3.3.5.

Taking Tables 6.3-6.5 together, we observe that the advection $\boldsymbol{u} \cdot \nabla \boldsymbol{Q}$ and rotation $[\boldsymbol{Q}, \boldsymbol{\Omega}]$ terms are selected in each of the methods considered. Additionally, the flow alignment terms $\boldsymbol{E}^{\mathrm{ST}}$ and $\boldsymbol{Q}(\boldsymbol{Q} : \nabla \boldsymbol{u}) - \boldsymbol{Q}^{\perp}(\boldsymbol{Q}^{\perp} : \nabla \boldsymbol{u})$ are selected by LASSO (Table 6.3) and ensemble SINDy (Table 6.5), but not so by STRidge (Table 6.4). All three methods considered in this section discover the term $[u_i \partial_k Q_{kj}]^{\mathrm{ST}}$; this term is highly correlated with the advection $\boldsymbol{u} \cdot \nabla \boldsymbol{Q}$ (see Figure 6.1), as is the term $[u_k \partial_i Q_{kj}]^{\mathrm{ST}}$ that is part of the best subset of size 5 (Table 6.2). Again, although these methods do not provide a perfect consensus, they each lend support to the small set selected for inclusion in Equation (6.4).

It should also be noted that terms involving the second derivative of $\boldsymbol{Q}$ are rarely selected, or are selected at late stages of each of the sparse regression approaches considered. A similar phenomenon has been observed in prior data-driven discovery methods using experimental [87] and synthetic data [73]. It is yet unknown whether the relative absence of higher-order derivatives stems from inaccurate differentiation, sparse selection methods, or their lack of importance in the experimental system.

Table 6.5 Ensemble SINDy results using experimental data for discovery of terms in equation for $\boldsymbol{Q}$. Each row represents the probability of inclusion of that term using the procedure outlined in [28] and described in Section 3.3.5. We used $B = 250$ subsets and $G = 20$ regularization parameters on a dataset that is randomly sampled in spacetime (50%).

Probability of Inclusion	Term
0.971	$u_k \partial_k Q_{ij}$
0.9	$[\boldsymbol{Q}, \boldsymbol{\Omega}]$
0.9	$[u_j \partial_k Q_{ki}]^{\mathrm{ST}}$
0.6	$[Q_{jl} \partial_l \partial_k Q_{ki}]^{\mathrm{ST}}$
0.45	$[\partial_i u_j]^{\mathrm{ST}}$
0.399	$[\partial_l Q_{ki} \partial_k Q_{lj}]^{\mathrm{ST}}$
0.351	$[Q_{ik} Q_{lj} \partial_k u_l]^{\mathrm{ST}}$
0.176	$[Q_{ik} \partial_l \partial_j Q_{kl}]^{\mathrm{ST}}$
0.145	$[\partial_j Q_{lk} \partial_k Q_{li}]^{\mathrm{ST}}$
0.099	$[Q_{kl} Q_{lj} \partial_i u_k]^{\mathrm{ST}}$

6.2 Velocity Equation for Microtubule-Kinesin System

Although the evolution equation for $\boldsymbol{Q}$ has been the subject of the most active debate in the literature, there are some outstanding questions related to the stress terms present in the velocity evolution equation as noted in Chapter 2. The library of potential terms for the discovery of the velocity equation is outlined in Table 2.1, in which the procedure described in Section 2.3 is employed. The equation is postulated to have the overdamped Hele-Shaw flow form $\boldsymbol{u} = \nabla \cdot \boldsymbol{\sigma}$ described in Section A.5. The results of forward selection, see Algorithm 2, to maximize R^2 for this equation can be seen in Table 6.6 where coefficients are calculated using ordinary least squares.

An immediate observation is that the R^2 for the recovery of the velocity equation is lackluster as compared with that of the $\boldsymbol{Q}$-equation. This is likely due to poor extraction of the fluid velocity from the experimental images. However, the most common "active stress" term proportional to $\boldsymbol{Q}$, as originally derived by Simha & Ramaswamy [82], is identified as the most important and possibly the only term which correlates well with the velocity $\boldsymbol{u}$. Given the inflection in the R^2, it could be supposed that the second stress term $\nabla \cdot (\Delta \boldsymbol{Q})$ should also be included. However, the higher order of differentiation and low R^2 make the term particularly suspect. Validation using an ensemble method identifies a low probability for its inclusion as demonstrated in Table 6.7. Thus, it is concluded that the discovered velocity equation takes the form

$$\boldsymbol{u} = -D\nabla \cdot \boldsymbol{Q} \tag{6.5}$$

with $D > 0$ (Table 6.6), as would be expected for extensile nematics.

To further validate our result, we use spectral differentiation to compute the value of D in Equation (6.5), and its incompressible counterpart

$$\boldsymbol{u} = -D\nabla \cdot \boldsymbol{Q} - \nabla p, \quad \nabla \cdot \boldsymbol{u} = 0. \tag{6.6}$$

Table 6.6 Forward selection results for the velocity equation. The data is randomly sampled in spacetime (50%), and the coefficients are determined using ordinary least squares.

R^2	$\boldsymbol{u} =$	Coefficients									
0.27	$\partial_i(Q_{ij})$	-2.30	-3.20	-3.26	-3.37	-3.42	-3.43	-3.43	-2.78	-2.04	-2.02
0.36	$\partial_i(\partial_k\partial_k Q_{ij})$		-2.51	-2.70	-2.71	-2.30	-2.38	-2.40	-2.31	-2.46	-2.40
0.36	$2\partial_i(Q_{lk}\partial_j\partial_l Q_{ik}]^{\mathrm{S}})$			-1.72	-3.07	-3.15	-3.79	-2.10	-1.97	-1.81	-0.95
0.37	$\partial_i(Q_{ki}Q_{jk})$				-2.55	-2.56	-2.26	-2.28	-0.04	3.32	3.43
0.37	$\partial_i(Q_{mk}Q_{ki}\partial_l\partial_l Q_{jm})$					-15.73	-14.71	-14.71	-17.19	-13.44	-15.33
0.38	$2\partial_i(Q_{jl}\partial_k\partial_i Q_{lk}]^{\mathrm{S}})$						1.35	3.05	3.17	3.20	3.16
0.38	$2\partial_i(Q_{jl}\partial_k\partial_k Q_{il}]^{\mathrm{S}})$							-3.39	-3.52	-3.63	-4.05
0.38	$\partial_i(Q_{kl}S_{lkij})$								-2.23	-5.52	-5.60
0.38	$\partial_i(Q_{kl}Q_{lk}Q_{ij})$									2.32	2.42
0.38	$\partial_i(Q_{kl}\partial_l\partial_k Q_{ij})$										-1.51

Table 6.7 Ensemble SINDy results using experimental data for discovery of stress terms in the velocity equation. Each row represents the probability of inclusion of that term using the procedure outlined in [28] and described in Section 3.3.5 with $B = 250$ subsets and $G = 20$ regularization parameters on a dataset that is randomly sampled in spacetime (50%).

Probability of Inclusion	Term
0.972	$\partial_i(Q_{ij})$
0.273	$\partial_i(\partial_k\partial_k Q_{ij})$
0.185	$\partial_i(Q_{mk}Q_{ki}\partial_l\partial_l Q_{jm})$
0.079	$\partial_i([Q_{ml}Q_{mk}\partial_i\partial_l Q_{kj}]^{\mathrm{S}})$
0.062	$\partial_i(\partial_m\partial_m Q_{kl}S_{ijlk})$
0.0	$\partial_i([\partial_k\partial_i Q_{jk}]^{\mathrm{S}})$

Using a Gaussian filter in Fourier space as described in Equation (3.2), the data can be made pseudo-periodic and spectral differentiation can be used. Equations (6.5) and (6.6) can be solved straightforwardly in Fourier space, the latter as

$$\tilde{\boldsymbol{u}} = -\mathrm{i}D\left(\boldsymbol{I} - \hat{\boldsymbol{k}}\hat{\boldsymbol{k}}\right) \cdot \boldsymbol{k} \cdot \tilde{\boldsymbol{Q}}, \tag{6.7}$$

where tildes denote Fourier transformed quantities, $\boldsymbol{k}$ is the wavevector and $\hat{\boldsymbol{k}} = \boldsymbol{k}/|\boldsymbol{k}|$.

Figure 6.2 shows the best fit value of D and the corresponding R^2 using spectral differentiation. Ultimately, Figure 6.2 demonstrates a range of potential values for the parameter D and highlights the uncertainty in its value due to the poor reconstruction quality in the linear system. This uncertainty is explored further in Section 6.4. Additionally, the R^2 of the recovery and the consistency of the results supports the claim that incompressibility should not be enforced, as is also suggested from the experimental data (Figure 5.10).

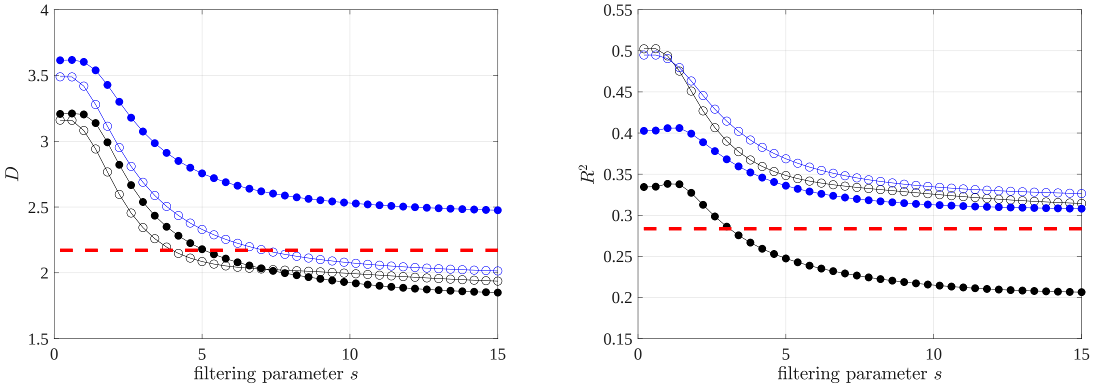


Figure 6.2 Dependence on the filtering parameter s (see Equation (3.2)) of the best fit values of the coefficient D in the velocity equations (6.5) (open circles) and (6.6) (filled circles). Black lines represent derivatives computed spectrally while blue lines represent the results using finite differences and subsequently filtering. The dashed red line marks the D and R^2 values obtained using finite differences without filtering (see the first row of Table 6.6).

6.3 Linear Stability of Discovered Equation

One of the main advantages of the SINDy methodology is that traditional analysis can be performed on the resulting equations. This facilitates analyzing the relationships between the physical quantities in the system and understanding the mechanisms which govern it. One such analysis is to consider the linear stability of the system. This is particularly important because the equations recovered via the SINDy procedure have no constraints to guarantee that the equation is linearly well-posed.

The discovered model, with a velocity determined by the active stress, is

$$\begin{aligned} \boldsymbol{Q}_t &= c_1 \boldsymbol{u} \cdot \nabla \boldsymbol{Q} + c_2 [\boldsymbol{Q}, \boldsymbol{\Omega}] + c_3 \boldsymbol{E}^{\mathrm{ST}} + c_6 \big(\boldsymbol{Q} (\boldsymbol{Q} : \nabla \boldsymbol{u}) - \boldsymbol{Q}^{\perp} (\boldsymbol{Q}^{\perp} : \nabla \boldsymbol{u}) \big) \ , \\ \boldsymbol{u} &= -D \nabla \cdot \boldsymbol{Q} \ , \end{aligned}$$

which can be expressed in terms of the matrix elements λ and μ as

$$\begin{aligned}\lambda_t &= -Dc_1(\lambda_x^2 + \lambda_x\mu_y - \lambda_y^2 + \lambda_y\mu_x) - Dc_2\left(-\mu_{xx} + \mu_{yy} + 2\lambda_{xy}\right)\mu \\ &\qquad - D\frac{c_3}{2}\left(\lambda_{xx} + \lambda_{yy}\right) - Dc_6\left((\lambda_{xx} + \lambda_{yy})(\lambda^2 - \mu^2) + 2\lambda\mu(\mu_{xx} + \mu_{yy})\right)\ , \\ \mu_t &= -Dc_1(\lambda_x\mu_x - \lambda_y\mu_y + 2\mu_x\mu_y) - Dc_2\left(\mu_{xx} - \mu_{yy} - 2\lambda_{xy}\right)\lambda \\ &\qquad - D\frac{c_3}{2}\left(\mu_{xx} + \mu_{yy}\right) - Dc_6\left((\mu_{xx} + \mu_{yy})(\mu^2 - \lambda^2) + 2\lambda\mu(\lambda_{xx} + \lambda_{yy}\right)\ . \qquad (6.8)\end{aligned}$$

Consider perturbations of $\boldsymbol{Q}$ around a uniformly aligned state of rods oriented with angle θ with constant order parameter $S_0 > 0$. That is, assume that the entries of $\boldsymbol{Q}$ are of the form

$$\lambda = \frac{S_0}{2}\cos(2\theta) + \epsilon\hat{\lambda}(t)\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}, \quad \mu = \frac{S_0}{2}\sin(2\theta) + \epsilon\hat{\mu}(t)\mathrm{e}^{\mathrm{i}\boldsymbol{k}\cdot\boldsymbol{x}}, \quad \boldsymbol{x} = x\cos(\phi) + y\sin(\phi)\ .$$

After substituting this form into Equation (6.8) and removing higher order terms in ϵ, we arrive at the system

$$\begin{bmatrix}\hat{\lambda}_t \\ \hat{\mu}_t\end{bmatrix} = \frac{Dk^2}{2}\begin{bmatrix}M_{11} & M_{12} \\ M_{21} & M_{22}\end{bmatrix}\begin{bmatrix}\hat{\lambda} \\ \hat{\mu}\end{bmatrix},$$

where $k = |\boldsymbol{k}|$ and

$$\begin{aligned}M_{11} &= S_0c_2\sin(2\phi)\sin(2\theta) + c_3 + \frac{S_0^2}{2}c_6\cos(4\theta), \\ M_{12} &= -S_0c_2\sin(2\theta)\cos(2\phi) + \frac{S_0^2}{2}c_6\sin(4\theta), \\ M_{21} &= -S_0c_2\sin(2\phi)\cos(2\theta) + \frac{S_0^2}{2}c_6\sin(4\theta), \\ M_{22} &= S_0c_2\cos(2\phi)\cos(2\theta) + c_3 - \frac{S_0^2}{2}c_6\cos(4\theta). \qquad (6.9)\end{aligned}$$

The eigenvalues of this system are given by

$$\begin{aligned}\gamma_1 &= \frac{D}{2}\left(c_3 + \frac{S_0^2}{2}c_6\right)k^2 \\ \text{and}\quad \gamma_2 &= \frac{D}{2}\left(c_3 + S_0c_2\cos(2(\theta - \phi)) - \frac{S_0^2}{2}c_6\right)k^2. \qquad (6.10)\end{aligned}$$

Note that for $c_2, D > 0$, as is the case in with our fitted values for these parameters (see Equation (6.4) and Table 6.6, respectively), γ_2 is largest for $\phi = \theta$, or when the perturbation is co-aligned with the rods. The maximum value of the eigenvalue is

$$\gamma_2^{\max} = \frac{D}{2}\left(c_3 + c_2 S_0 - \frac{S_0^2}{2}c_6\right)k^2. \tag{6.11}$$

For $c_6 < 0$, as in the discovered equation (6.4), $\gamma_2^{\max} > \gamma_1$ for all wavenumbers k. Since $\gamma_2^{\max} \to \infty$ as $k \to \infty$, we conclude that Equation (6.4) is linearly ill-posed.

As previously proposed in [67], the equation can be regularized by augmenting it with a higher-order derivative term $\Delta^2 \boldsymbol{Q}$, which will also prescribe a characteristic length scale. Doing so results in the equation

$$\begin{aligned}
&\boldsymbol{Q}_t = c_1 \boldsymbol{u}\cdot\nabla\boldsymbol{Q} + c_2[\boldsymbol{Q},\boldsymbol{\Omega}] + c_3\boldsymbol{E}^{\mathrm{ST}} + c_6\big(\boldsymbol{Q}(\boldsymbol{Q}:\nabla\boldsymbol{u}) - \boldsymbol{Q}^{\perp}(\boldsymbol{Q}^{\perp}:\nabla\boldsymbol{u})\big) - c_8\Delta^2\boldsymbol{Q},\\
&\boldsymbol{u} = -D\nabla\cdot\boldsymbol{Q}\,.
\end{aligned} \tag{6.12}$$

The eigenvalues (6.10) and (6.11) are thus modified to read

$$\begin{aligned}
\gamma_1 &= \frac{D}{2}\left(c_3 + \frac{S_0^2}{2}c_6\right)k^2 - c_8 k^4,\\
\gamma_2^{\max} &= \frac{D}{2}\left(c_3 + c_2 S_0 - \frac{S_0^2}{2}c_6\right)k^2 - c_8 k^4.
\end{aligned} \tag{6.13}$$

Figure 6.3 shows the dependence of γ_1 and $\gamma_2^{\max}$ on k, both with and without the higher-order regularization.

From this analysis, the discovered Equation (6.4) can be augmented as:

$$\begin{aligned}
\boldsymbol{Q}_t = {} & -1.04\boldsymbol{u}\cdot\nabla\boldsymbol{Q} + 1.18[\boldsymbol{Q},\boldsymbol{\Omega}] + 0.26\boldsymbol{E}^{\mathrm{ST}}\\
& - 2.28\big(\boldsymbol{Q}(\boldsymbol{Q}:\nabla\boldsymbol{u}) - \boldsymbol{Q}^{\perp}(\boldsymbol{Q}^{\perp}:\nabla\boldsymbol{u})\big) - 0.01\Delta^2\boldsymbol{Q}
\end{aligned} \tag{6.14}$$

where the the coefficients are again determined via ordinary least squares, or

$$\begin{aligned}
\boldsymbol{Q}_t = {} & -1.22\boldsymbol{u}\cdot\nabla\boldsymbol{Q} + 1.50[\boldsymbol{Q},\boldsymbol{\Omega}] + 0.37\boldsymbol{E}^{\mathrm{ST}}\\
& - 3.23\big(\boldsymbol{Q}(\boldsymbol{Q}:\nabla\boldsymbol{u}) - \boldsymbol{Q}^{\perp}(\boldsymbol{Q}^{\perp}:\nabla\boldsymbol{u})\big) - 0.02\Delta^2\boldsymbol{Q}
\end{aligned} \tag{6.15}$$

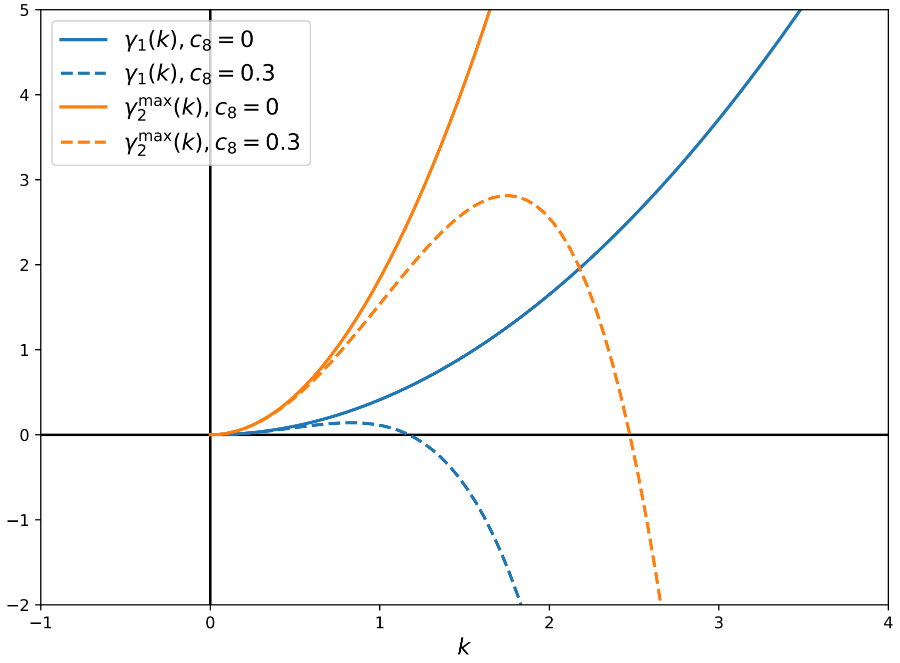


Figure 6.3 Eigenvalues $\gamma_1(k)$ and $\gamma_2^{\max}(k)$ of the linear stability problem (6.9) for the uniformly aligned state with constant order parameter $S_0 = 0.3$, both without (solid curves, Equation (6.11)) and with (dashed curves, Equation (6.13)) the higher-order regularization term $-c_8\Delta^2\boldsymbol{Q}$. The parameter values $c_2 = 1, c_3 = -0.3, c_6 = -3$ are taken from Equation (6.4), and $D = 3$ from Figure 6.2. Note that the eigenvalue γ_1 is subdominant to $\gamma_2^{\max}$.

where the coefficients are determined via total least squares. Some simple calculus can be used to determine the wavenumber of maximum growth $k_{\max}$, for which $\gamma_2^{\max}$ achieves its maximum value:

$$k_{\max} = \sqrt{\frac{D}{4c_8}\left(c_3 + c_2 S_0 - \frac{S_0^2}{2}c_6\right)}. \tag{6.16}$$

Using the previously considered parameters, $c_2 = 1, c_3 = -0.3, c_6 = -3, D = 3$ and including $c_8 = 0.02$ yields a dominant $k_{\max} = 2.25$. Given that the experimental domain is a square of side length $L_E = 31.24$, this would yield on average $\sim (k_{\max} L_E / 2\pi)^2 \approx 125$ defects in the domain at any given time. This is far beyond the number observed in experimental data and demonstrates that both regression approaches estimate an anomalously small coefficient for the bi-Laplacian term while the other coefficients remain virtually unchanged. The small coefficient can likely be explained by the fact that the bi-Laplacian must be approximated by taking several derivatives of the smoothed data and thus has high variance. An alternate method is thus necessary to determine an adequate value of c_8, which we peruse in Section 6.4.

6.4 Determining Parameters via Non-Local Penalties

SINDy has shown accuracy in recovering accurate coefficients for high order derivatives in the context of simulation data. However, several works have demonstrated that higher order derivative terms are often left out of discovered equations when working with highly noise polluted or experimental data [73,87]. As such, alternative methods should be used to estimate parameters in the discovered equation whose values are especially uncertain.

Specifically, in the augmented model (6.12), the coefficient of the bi-Laplacian c_8 in the orientation evolution equation and the activity coefficient D in the velocity equation are less certain than the other coefficients, which are in Equation (6.4). The analysis in Section 6.3 suggests that the value of c_8 selected by both ordinary and total least squares is not large enough to impose a reasonable characteristic length scale. Additionally, the low R^2 value in the discovery of the velocity equation and the sensitivity of D to data smoothing (see Figure 6.2) indicate a poor fit and less certainty. This poor fit brings into question the value of the coefficient D computed via regression techniques.

The SINDy methodology only seeks a PDE model that is accurate *locally in time*, i.e. it only imposes that the residual of the model is small at any instant, not that the model accurately simulates the data. This explains why the non-augmented model system shown in Equations (6.4) and (6.5) is selected by SINDy even though it is linearly ill-posed. To remedy this problem, we propose a more global approach to parameter estimation. Specifically, a temporally non-local penalty was devised that imposes that a simulation of Equation (6.12) with the parameters c_8 and D and an appropriate initial condition should approximately match the experimental data. Because the boundary conditions for the experiment are unknown, the PDE must be simulated with approximate boundary conditions. We elect to impose periodic boundary conditions in each space dimension, which allows us to simulate the equations using the simple and effective numerical apparatus described in Section 4.1. An initial condition must also be imposed, which we obtain from the experimental data.

In particular, the simulation domain is $[0, 2\pi]^2$ with periodic boundary conditions and the same resolution as the experimental images (220×220). As above, each frame of the video corresponds to one unit of time. The timestep for simulations was selected to be $\Delta t = 10^{-3}$. Filtered snapshots of the data were used as initial conditions: i.e. if $\lambda(\boldsymbol{x}_{ij}, t_0)$, $\mu(\boldsymbol{x}_{ij}, t_0)$ are the values at time t_0 obtained using the data extraction methods from Section 5.2.3, the initial conditions are obtained by filtering these values using the Gaussian filter in Equation (3.2) with parameter $s = 5$. This filtering helps ensure that the 1/3 aliasing law is enforced [66] and removes artifacts caused by the fact that the data is not periodic. The values of the simulation at simulation time

t_k starting with filtered data taken from the snapshot at experimental time t_0 are denoted by $\hat{\lambda}(\boldsymbol{x}_{ij}, t_k; t_0, c_8, D)$, $\hat{\mu}(\boldsymbol{x}_{ij}, t_k; t_0, c_8, D)$.

Because the periodic boundary condition is artificial, the simulation is compared only at the center-most 100 × 100 grid of the 220 × 220 experimental data. The set of indices ij corresponding to the points $\boldsymbol{x}_{ij}$ in this 100 × 100 grid is denoted by $\mathcal{I}$ below. The model error we propose is then based on the relative root mean squared error (RMSE) between the simulation and the snapshots of the data for these grid points:

$$\begin{aligned} E_{RMSE}(t_k; t_0, c_8, D) = \\ &\left(\sum_{ij\in\mathcal{I}} (\hat{\lambda}(\boldsymbol{x}_{ij}, t_k; t_0, c_8, D) - \lambda(\boldsymbol{x}_{ij}, t_0 + t_k))^2 \right. \\ &\left. + (\hat{\mu}(\boldsymbol{x}_{ij}, t_k; t_0, c_8, D) - \mu(\boldsymbol{x}_{ij}, t_0 + t_k))^2 \right)^{1/2} \Bigg/ \\ &\qquad \left(\sum_{ij\in\mathcal{I}} \lambda(\boldsymbol{x}_{ij}, t_0 + t_k)^2 + \mu(\boldsymbol{x}_{ij}, t_0 + t_k)^2\right)^{1/2} . \end{aligned} \tag{6.17}$$

Boundary information in the experiment propagates from the edges into $\mathcal{I}$ in fairly short time, so this error measure is only reasonable for moderate values of t_k. Specifically, a defect at the boundary can enter the comparison region within 12-16 seconds, so we only consider $t_k \leq 20$.

The parameter estimation problem for c_8 and D can then be phrased as:

$$\min_{c_8, D} E_{RMSE}(t_k; t_0, c_8, D) .$$

This problem is nonlinear and non-convex in c_8 and D. To obtain a rough estimate of the optimal c_8 and D, we compute the error for a 14 × 14 evenly spaced grid of c_8 and D values between 0.1 and 7.3. The resulting c_8 and D values are somewhat sensitive to t_0 and t_k as seen in Figure 6.4 but show a general trend toward increased hyperdiffusion for larger values of t_k. The large variation in optimal c_8 and D can also be attributed to the low variation in loss near the optimal points as is demonstrated in Figure 6.5. This figure shows the loss values for a range of parameter combinations of D and c_8 at three comparison times (t_k) averaged over the range of starting points (t_0) shown in Figure 6.4. It demonstrates that as time progresses, the loss landscape becomes flatter, thus making several parameters near-optimal.

Remark 6.4.1. *The non-local penalty above is related to some methods considered previously in the literature. A similar penalty was proposed in [74] for fitting neural network models to time series data. This penalty was also applied within the SINDy framework in [45] to identify sparse models and noise probability distributions for ordinary differential equation simulations with added noise. In [18], a penalty based on simulation is used, though the error is checked over relatively short horizons and reasonable boundary conditions are known. In [14], the horizon for comparison is explored via the Lyapunov time of the system.*

6.4.1 Comparison with Linear Stability

The results of the mean squared comparison of simulations using c_8 and D (Figure 6.4) can also be validated using the linear stability analysis presented in Section 6.3. Specifically, we may rearrange Equation (6.16) to obtain a relation between c_8 and the most unstable wavenumber $k_{\max}$:

$$c_8 = \frac{D}{4k_{\max}^2}\left(c_3 + S_0 c_2 - \frac{S_0^2}{2}c_6\right).$$

Substituting $k_{\max} = \frac{2\pi\sqrt{N_d}}{L_E}$ where L_E is the length of the experimental domain and N_d is the expected number of defects in the domain at any given point in time, we obtain a relation between c_8 and the expected number of defects:

$$c_8 = \frac{DL_E^2}{(4\pi)^2 N_d}\left(c_3 + S_0 c_2 - \frac{S_0^2}{2}c_6\right). \tag{6.18}$$

Figure 6.6 shows two lines in the (D, c_8)–plane as defined by Equation (6.18) for the values $N_d = 2$ and $N_d = 9$, where we use the fact that the experimental data has between two and nine defects at any given time. The optimal values of D and c_8

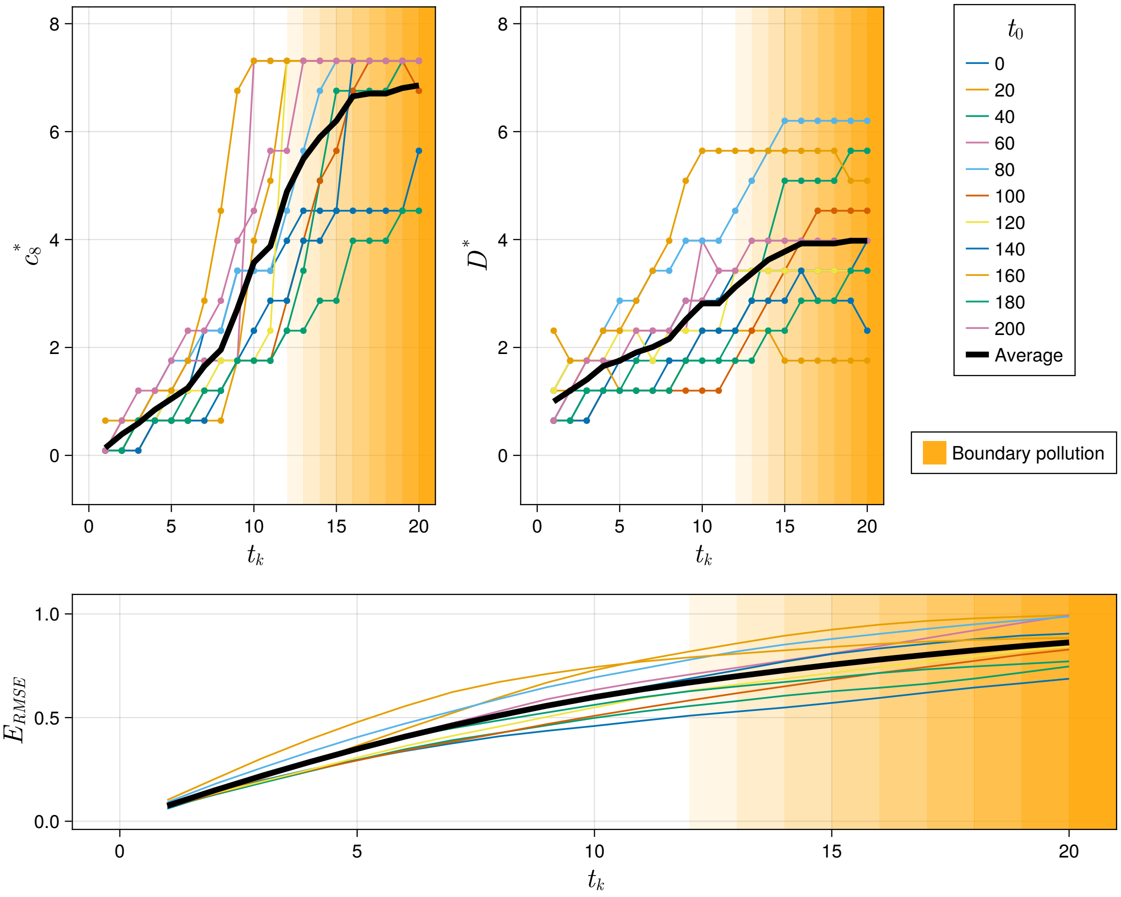


Figure 6.4 Optimal values of the bilaplacian coefficient c_8 and activity parameter D in the discovered equation (6.12). The remaining parameters in the $\boldsymbol{Q}$-equation are as in Equation (6.4). The optimal values are determined by the mean squared difference of the $\boldsymbol{Q}$-values between experiment and simulation, as written in Equation (6.17). The simulations were performed on the time interval $[t_0, t_0+t_k]$ across a range of values for both c_8 and D, using a range of spectrally filtered initial conditions obtained from the extracted experimental data at the times t_0 indicated in the legend (different colors). The black curves indicate the average of the colored curves, and the orange shaded region marks the time after which boundary information could pollute the comparison region.

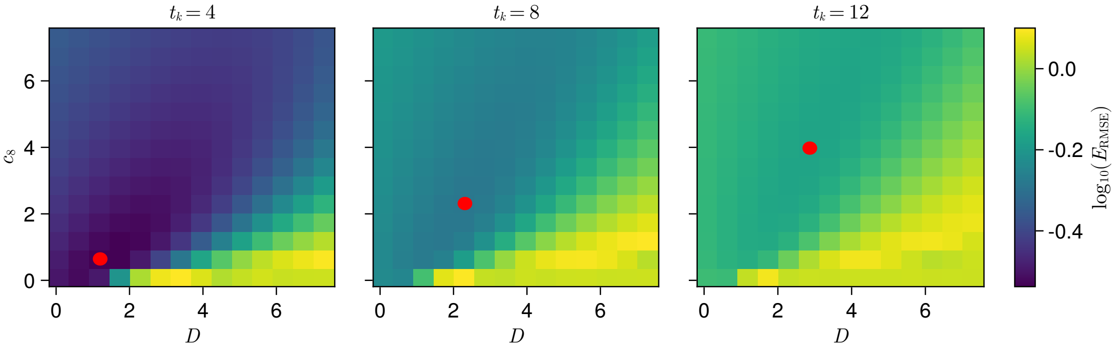


Figure 6.5 Root mean squared error difference between simulation and experiment for a range of parameter values (D, c_8) averaged over all starting times t_0 as shown in Figure 6.4. Red dots indicate the minimum loss for the parameter grid.

obtained from the black curve in Figure 6.4 are superimposed, and are found to be roughly consistent with the predictions of the linear stability analysis.

6.5 Numerical Experiments on the Discovered Equation

There are a range of metrics that can be used to demonstrate a model's ability to capture physical phenomenon as observed in active nematic experiments. These include the characteristic orientational order parameter S (Equation (2.4)) of the system and the expected number of defects. Though not comprehensive, these comparisons capture some of the key statistical features of the system.

The simulations were performed using Equation (6.12) using three pairs of values (D, c_8) which were chosen to be consistent with the optimal values presented in Figure 6.6 and the remaining coefficients from Equation (6.4). A filtered version of the first frame of experimental data using the filter described in Equation (3.2) was used as the initial condition. The simulations were run for $T = 300$, the duration of the experimental data, with a time step of $\Delta t = 10^{-3}$ using the scheme presented in Section 4.1 and used in Section 6.4.

Figure 6.7 shows the value of the spatially-averaged S over time, compared between experiment and simulations performed for three pairs of values (D, c_8). These

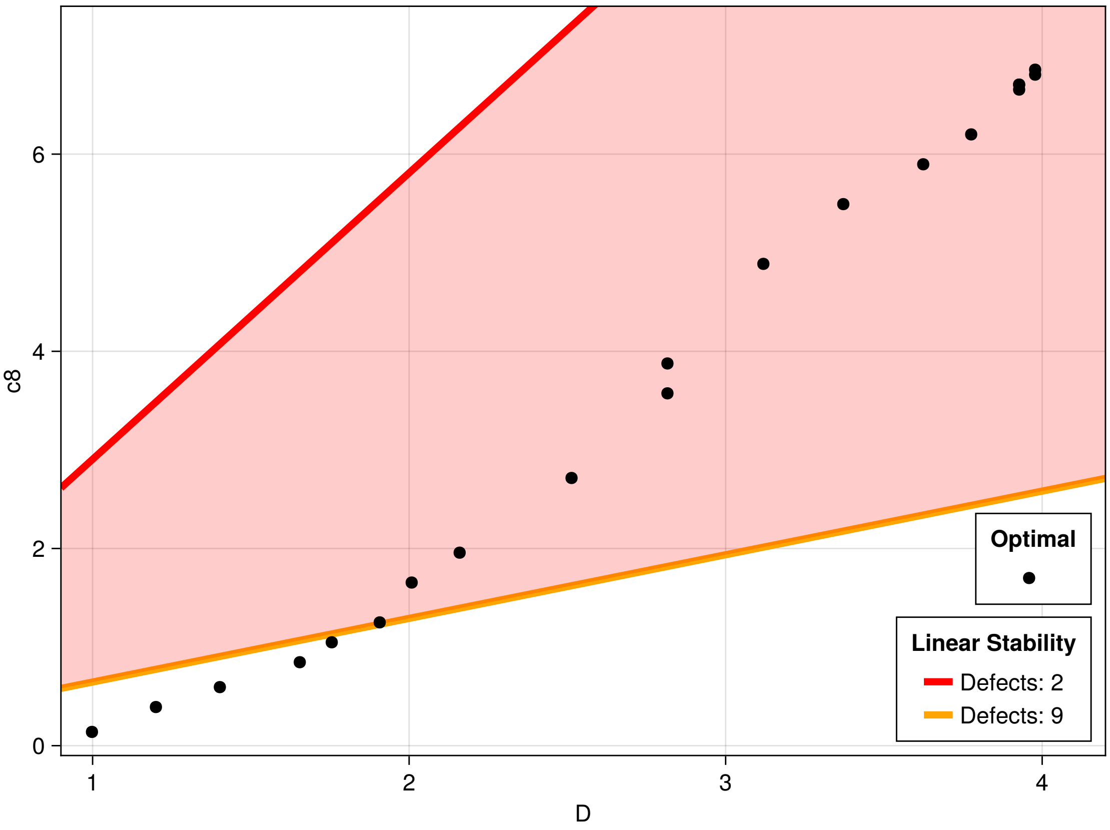


Figure 6.6 Validation of the inferred values of D and c_8 using linear stability analysis, as described in Section 6.4.1. The solid lines denote the prediction of the linear stability analysis (6.18) for $N_d = 2$ (red) and $N_d = 9$ (orange) defects. The black dots indicate the optimal values of c_8 and D, as obtained by averaging the mean squared difference between experiment and simulation (black curve in Figure 6.4).

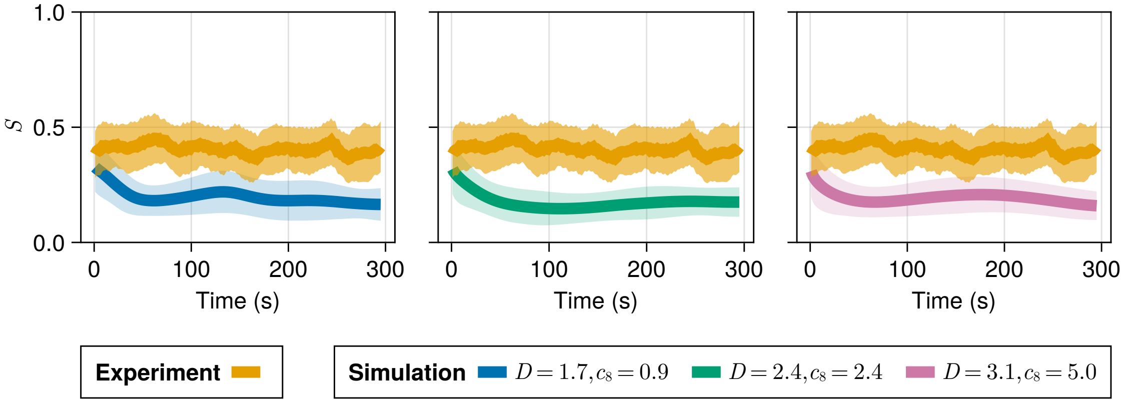


Figure 6.7 Comparison of the spatially averaged scalar order parameter S from simulations with several values of D and c_8 across the full experimental time. Shaded regions represent the standard deviation of S.

values are chosen to be consistent with the optimal values presented in Figure 6.6. It should be noted that variations in c_8 or D do not seem to affect the value of the scalar order parameter S strongly, and its spatially averaged value remains roughly steady throughout the simulation. Further insight is needed to understand the mechanism which prescribes this order as is discussed in Section 7.1.

To identify defects, we compute the intersections of the zero contours of λ and μ (the matrix elements of $\boldsymbol{Q}$), and determine its sign via contour integration [67]. Figure 6.8 demonstrates defects as they appear in both simulation and experimental data. Figure 6.9 shows the proportional of time for which a given number of defects is present, comparing the results from experiments and simulations with the three pairs of values (D, c_8) considered in Figure 6.7. We observe that, among the three pairs of values considered, the values $D = 2.4$, $c_8 = 2.4$ (green) capture the experimental data best. While the pair $D = 1.7$, $c_8 = 0.9$ (blue) better approximates the width of the distribution, the predicted average number of defects is about twice that observed in experiments. Similarly, the pair $D = 3.1$, $c_8 = 5.0$ (pink) predicts too few defects.

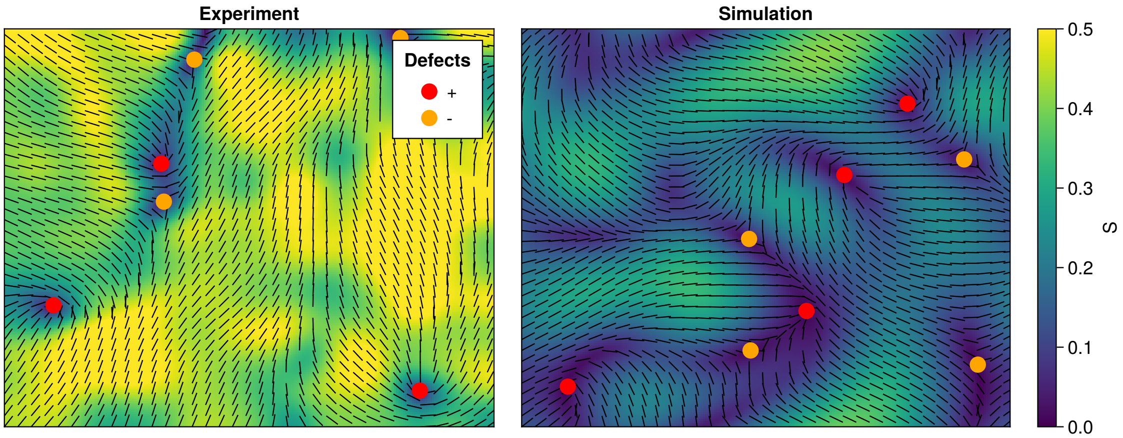


Figure 6.8 Defects in both experiment (left) and simulation (right) are identified by numerically computing the intersections of the zero-contours of the matrix elements of $\boldsymbol{Q}$, as detailed in [67].

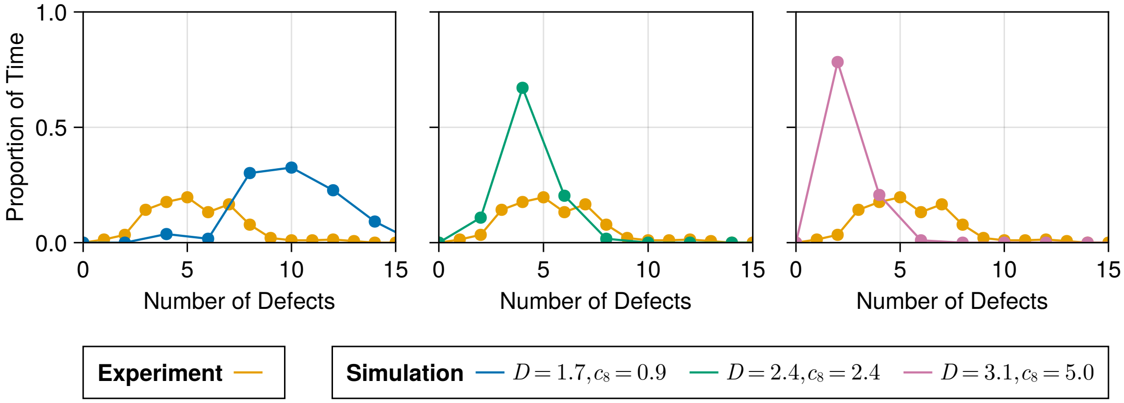


Figure 6.9 The proportion of time (vertical axis) for which a given number of defects is present (horizontal axis), in both experiments (orange) and simulations (blue, green, pink). The simulations are conducted for the three pairs of values (D, c_8) considered in Figure 6.7.

6.6 Discussion

We conclude by discussing the relationship between our augmented model (6.12), other models previously considered in the literature (Section 2.2) and those discovered recently by Joshi *et al.* [44] and Golden *et al.* [36] using data-driven equation discovery techniques. Beginning with the $\boldsymbol{Q}$-evolution equation, we note that the coefficients of the advection $\boldsymbol{u} \cdot \nabla \boldsymbol{Q}$ and vorticity $[\boldsymbol{Q}, \boldsymbol{\Omega}]$ terms are roughly -1 and $+1$, as posited in the Beris-Edwards equations (2.8). A similar result was obtained by both Joshi *et al.* [44] and Golden *et al.* [36]. While the velocity field $\boldsymbol{u}$ reconstructed from the experimental data is not divergence free, ($\nabla \cdot \boldsymbol{u} \neq 0$, Figure 5.10), our equation discovery process does not select the term $(\nabla \cdot \boldsymbol{u})\boldsymbol{Q}$ hypothesized by Oza & Dunkel [67]. Moreover, the terms $\boldsymbol{S} : \boldsymbol{Q}$ and $\boldsymbol{S} : \nabla \boldsymbol{u}$ that are present in previously proposed kinetic theories (Equation (2.17), [31,32,96]) are not discovered by forward selection (Table 6.1), although the latter is an element of the best subset of size six (Table 6.2).

As noted in Section 6.1, the bulk alignment energy terms $\boldsymbol{Q}$ and $\boldsymbol{Q}^3$ posited in the Beris-Edwards equations (2.8) are not included in the discovered equation, a result that is consistent with Joshi *et al.* [44] and Golden *et al.* [36]. Moreover, this result is consistent with the conjecture of Thampi *et al.* [89], who argued that the terms $\boldsymbol{Q}$ and $\boldsymbol{Q}^3$ could be discarded as the ordering of the system can arise naturally from its activity. We note that the bulk alignment energy terms are highly correlated with the flow alignment terms $\boldsymbol{E}^{\mathrm{ST}}$ and $\boldsymbol{Q}(\boldsymbol{Q} : \nabla \boldsymbol{u}) - \boldsymbol{Q}^{\perp}(\boldsymbol{Q}^{\perp} : \nabla \boldsymbol{u})$ (Figure 6.1), which are selected by both forward selection (Table 6.1) and best subset selection (Table 6.2). The cubic flow alignment term $\boldsymbol{Q}(\boldsymbol{Q} : \boldsymbol{\nabla u})$ is also present in the Beris-Edwards equation (2.8); while this term is in our library, both the forward selection and best subset selection approaches select instead the composite term $\boldsymbol{Q}(\boldsymbol{Q} : \nabla \boldsymbol{u}) - \boldsymbol{Q}^{\perp}(\boldsymbol{Q}^{\perp} : \nabla \boldsymbol{u})$. Interestingly, the term $\boldsymbol{Q}(\boldsymbol{Q} : \nabla \boldsymbol{u})$ is discovered by Joshi *et al.* [44], while cubic terms are absent from the equation discovered by Golden *et al.* [36]. We did simulate Equation (6.12) without the $\boldsymbol{Q}^{\perp}$ term and found that the system consistently converged to a defect-free striped state, suggesting that the $\boldsymbol{Q}^{\perp}$ term is necessary to drive the system away from this state and towards active turbulence.

While the flow alignment term $\boldsymbol{E}^{\mathrm{ST}}$ was neglected in the work of Oza & Dunkel (Equation (2.15), [67]), our results suggest that it plays a prominent role in the dynamics, a result consistent with those of Joshi *et al.* [44] and Golden *et al.* [36]. However, both Joshi *et al.* [44] and Golden *et al.* [36] obtain values $c_3 \approx 1$, which would be expected for slender (high aspect ratio) rods. We note that prior simulations of active nematics have used values of c_3 less than unity, e.g. $c_3 = 0.7$ in [20, 90]. The linear flow alignment term in our model can be expressed as $\tilde{c}_3 S_0 \boldsymbol{E}^{\mathrm{ST}}$, a form reminiscent of that proposed by Giomi *et al.* [34,35] (Equation (2.14)), where $\tilde{c}_3 = 0.74$ is closer to unity and $S_0 = 0.405$ is the orientational order parameter in experiment, averaged over time and space (Fig. 6.7). Moreover, since $\boldsymbol{E}^{\mathrm{ST}} = -\frac{D}{2}\Delta\boldsymbol{Q}$ provided $\boldsymbol{u} = -D\nabla\cdot\boldsymbol{Q}$, our discovered model (for $c_1 = -1$, $c_2 = 1$) is mathematically equivalent to the system

$$\begin{aligned} \boldsymbol{Q}_t &= -\boldsymbol{u}\cdot\nabla\boldsymbol{Q} + [\boldsymbol{Q},\boldsymbol{\Omega}] + \boldsymbol{E}^{\mathrm{ST}} + \frac{(1-c_3)D}{2}\Delta\boldsymbol{Q} \\ &\qquad + c_6\big(\boldsymbol{Q}(\boldsymbol{Q}:\nabla\boldsymbol{u}) - \boldsymbol{Q}^{\perp}(\boldsymbol{Q}^{\perp}:\nabla\boldsymbol{u})\big) - c_8\Delta^2\boldsymbol{Q}, \\ \boldsymbol{u} &= -D\nabla\cdot\boldsymbol{Q}\,. \end{aligned} \tag{6.19}$$

That is, our discovered equation (6.12) is equivalent to one in which the coefficient of $\boldsymbol{E}^{\mathrm{ST}}$ is unity, but augmented by a bulk elastic energy term proportional to $\Delta\boldsymbol{Q}$, as is typically assumed in Landau-de Gennes liquid crystal theory (see e.g. Equation (2.14)).

Our velocity equation $\boldsymbol{u} = -D\nabla\cdot\boldsymbol{Q}$ has a value $D > 0$, corresponding to extensile (as opposed to contractile) stresses [82]. Both Joshi *et al.* [44] and Golden *et al.* [36] also discovered velocity equations with extensile stresses. Joshi *et al.* [44] discovered an incompressible Stokes equation and regularized their equation with a Laplacian term $\Delta\boldsymbol{Q}$. A linear stability analysis of their equation, analogous to

Table 6.8 Forward selection results for the velocity equation, in which the library is augmented by the quadrupole term in Equation (6.20). The data is randomly sampled in spacetime (50%), and the coefficients are determined using ordinary least squares.

R^2	$\boldsymbol{u} =$	Coefficients		
0.27	$\partial_i(Q_{ij})$	-2.29	-3.18	-3.57
0.35	$\partial_i(\partial_k \partial_k Q_{ij})$		-2.50	-2.69
0.38	$Q_{ji}\partial_k Q_{ki}$			-4.08

that conducted in Section 6.3, yields a dominant eigenvalue of the form $\gamma(k) \sim a - bk^2$ for $a, b > 0$, implying that their equation does not predict the emergence of a characteristic length scale [31]. We plan to explore alternative velocity formulations in future work, as detailed in Section 7.1.2.

Moreover, Golden *et al.* [36] derive a constraint equation $\boldsymbol{E} : \boldsymbol{Q} =$ constant for the velocity and thus do not conduct simulations of their model. We note that this constraint is equivalent to saying that $\boldsymbol{E}$ and $\boldsymbol{Q}$ are highly correlated, which we also observe (Figure 6.1). We conclude by noting that recent work [56] has suggested that a "quadrupolar force" may be important in 2D active nematic systems. Sultan *et al.* [85] studied the influence of such a force by considering the velocity equation

$$\boldsymbol{u} = -D_1 \nabla \cdot \boldsymbol{Q} - D_2 \boldsymbol{Q} \cdot (\nabla \cdot \boldsymbol{Q}). \tag{6.20}$$

The D_2-term is not the divergence of a stress and thus is not part of our library. However, after manually adding it to the library we found that, while it was selected as the third-most important term, it increases the R^2-value by a small amount (Table 6.8).

CHAPTER 7

CONCLUSIONS

Though many models have been proposed for active nematic systems, capturing and understanding the full dynamics has remained a challenge. This dissertation has explored the use of the Sparse Identification of Nonlinear Dynamics (SINDy) modeling approach and presented a novel and concise model for the microtubule-kinesin active nematic system. This data-driven result has presented a new perspective with which to view previous models and lent a more objective opinion as to the form of a governing partial differential equation.

Chapters 1, 2, and 3 laid the foundations which motivate the need for a data-driven approach, the previous modeling work that inspires the approach, and the unique challenges that can arise when using the SINDy approach. Notably, special considerations are required when expanding the SINDy method to higher-dimensional systems and a new procedure is proposed for generating overcomplete nonlinear terms in higher dimensions using tensor concatenations followed by tensor contractions of state variables and their derivatives. Popular methods for sparse linear regression are reviewed, including a common two-stage approach to model discovery in which the selection of the sparse terms and the estimation of their parameters are separated into two individual calculations. This two stage approach allows sufficient freedom to combat issues caused by multicollinearity and noisy derivative values in the library of terms, as explored in the later chapters.

Chapter 4 presents the specific considerations that arise when performing SINDy on simulated active nematic systems. Most importantly, it is discovered that correlations in the generated nonlinear library are a fundamental feature of the model discovery process and must be accounted for in both the sparse selection and parameter estimation phases of modeling. Notwithstanding this challenge, it is demonstrated that the correct governing equations can be recovered for active nematics given sufficiently dynamic and representative data and with careful consideration of multicollinearity.

Chapters 5 and 6 present the main contribution of this work: appropriately extracting and validating state variable data and using this data to discover a PDE model for the microtuble-kinesin active nematic system. The linear stability of the model as returned by the original SINDy method is analyzed, suggesting the addition of a bi-Laplacian hyper-diffusion term. Parameters for this augmented model are then determined using a simulation-based metric for goodness of fit. The proposed model is compact and has parallels with previously proposed models. Numerical experiments confirm that the proposed model reproduces large-scale qualitative features of the experimental data, including the characteristic order parameter and number of topological defects.

This work functions as an additional demonstration that data-driven modeling can provide new insights and perspectives for complex systems which have posed significant challenges for previous modeling attempts [36, 44, 87].

7.1 Future Directions

The work presented in this dissertation raises a few questions about the nature of active nematic models and highlights new considerations for data-driven modeling with SINDy.

7.1.1 Characteristic Orientation for the Discovered Model

Foremost among these is the question of enforcing or encouraging a characteristic orientation of the system. The phenomenological bulk energy terms $\boldsymbol{Q}$ and $\boldsymbol{Q}^3$ which are present in the majority of proposed models, $\boldsymbol{Q}_t = a\boldsymbol{Q} - b\boldsymbol{Q}^3 + \dots$ for $a, b > 0$ (see

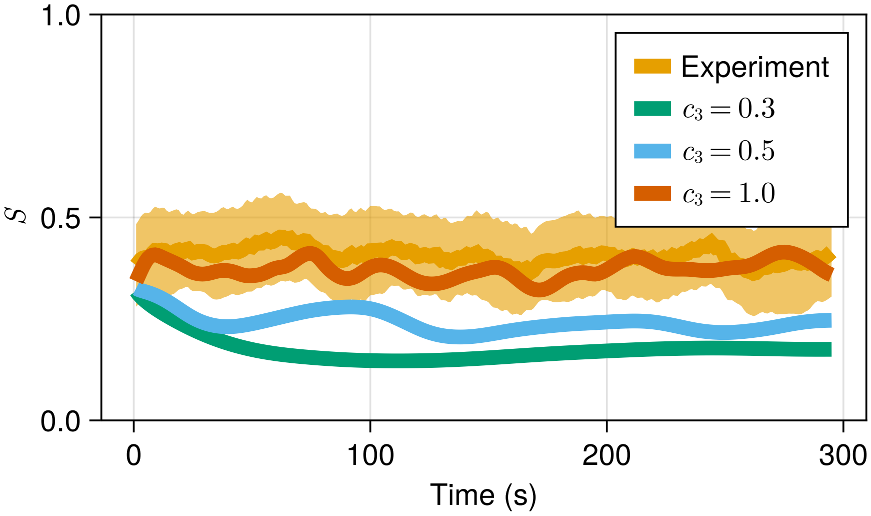


Figure 7.1 The spatially averaged value of the nematic order parameter S in experiment (orange) and simulations (green, blue and red) for three different values of the coefficient c_3 in Equation (6.12).

Equations (2.8), (2.14), and (2.15), for example) were introduced in order to drive the system away from the isotropic state with $S = 0$ to a nematically aligned state with $S = 2\sqrt{a/b}$. In contrast, the discovered equation presented in Equation (6.12) does not have an apparent mechanism for determining this characteristic order parameter. However, numerical experimentation has shown that the average value of S is controlled via the coefficient c_3 in Equation (6.12). Indeed, Figure 6.7 shows that the spatially averaged value of S is relatively consistent across the three pairs of coefficients (D, c_8). Figure 7.1 shows the spatially averaged value of S as the coefficient c_3 from Equation (6.12) is varied in simulations. Note that increasing the value of c_3 also increases the spatially averaged value of S, which is roughly constant through the duration of the simulation. Some analysis could be pursued to identify a potentially nonlinear mechanism in the model (6.12) which can explain this phenomenon.

7.1.2 Comparing Velocity Equation Forms

Though the SINDy method is meant to provide an automated tool for modeling, this dissertation has demonstrated that it requires some physical intuition in order to set up the problem appropriately. A finite library of nonlinear candidate terms must be determined *a priori*, including the selection of appropriate state variables (e.g. $\boldsymbol{Q}$, $\boldsymbol{u}$, and $\boldsymbol{S}$ in our setting) and the order of derivative operators and polynomial degree used to generate the library. More subtly, the left-hand-side that is selected for the regression problem has implications for the model physics. Though some adjustments to SINDy have been proposed to avoid this prescription [38], it is not clear whether this step of the process can be avoided. In this dissertation, a notable assumption was used in that the velocity equation was assumed to be an overdamped Hele-Shaw flow as described in Appendix A.5 and used in Section 6.2. However, other data-driven approaches have determined that the velocity form should instead be that of a Stokes flow [44], and it has also been proposed that the equation should impose a velocity form which incorporates interfacial friction [32].

While both compressible and incompressible overdamped Hele-Shaw flows are considered in Figure 6.2, in the future we plan to consider the incompressible Stokes flow and interfacial friction models. Specifically, the two models with extensile active stresses $\boldsymbol{\sigma} = -D\nabla \cdot \boldsymbol{Q}$ may be written in Fourier space as

$$\tilde{\boldsymbol{u}}_{\text{Stokes}} = -\mathrm{i}D_{\text{Stokes}}\frac{(\boldsymbol{I} - \hat{\boldsymbol{k}}\hat{\boldsymbol{k}})}{k^2} \cdot (\boldsymbol{k} \cdot \tilde{\boldsymbol{Q}}) \tag{7.1}$$

and

$$\tilde{\boldsymbol{u}}_{\text{int}} = -\mathrm{i}D_{\text{int}}\frac{(\boldsymbol{I} - \hat{\boldsymbol{k}}\hat{\boldsymbol{k}})}{2k} \cdot (\boldsymbol{k} \cdot \tilde{\boldsymbol{Q}}). \tag{7.2}$$

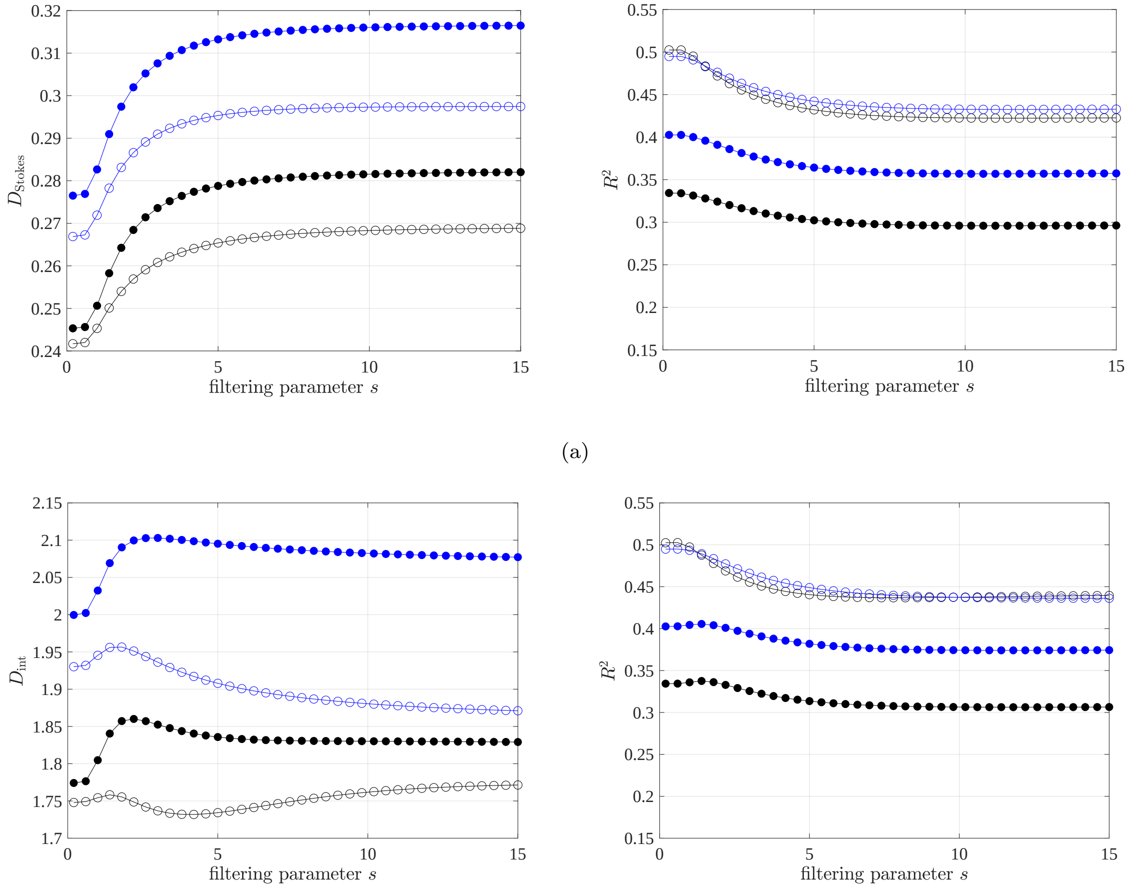
These forms, though minimal changes for the SINDy procedure, can have vastly different physical implications. As noted by Gao *et al.* [31], linear stability analysis of the models associated with both Equations (7.1) and (7.2) can be regularized using a Laplacian term $\Delta \boldsymbol{Q}$, as opposed to the higher-order $\Delta^2 \boldsymbol{Q}$ required in our work. However, (7.1) does not produce a characteristic length scale while (7.2) does.

We undertook a preliminary investigation of Equations (7.1) and (7.2) and their compressible counterparts, as obtained by removing the term $(\boldsymbol{I} - \hat{\boldsymbol{k}}\hat{\boldsymbol{k}})$. Specifically, we calculated the best fit values of the coefficients D_{Stokes} and D_{int}, respectively, and their associated R^2-values. Using filtered experimental values for λ and μ and spectral differentiation as was used for Figure 6.2, we obtain the results shown in Figure 7.2. These figures demonstrate a level of uncertainty in the form of the velocity equation as they all produce a reasonable fit (as determined by R^2). While the overdamped Hele-Shaw model is the conceptually simplest model, we conclude that the resolution of the data is insufficient to truly distinguish which model form should be imposed. Further efforts could be focused on collecting and examining data which could assist in distinguishing the results of these varied model forms.

7.1.3 Enforcing Well-posedness in Discovered Equations

This dissertation has demonstrated that though SINDy is capable of determining insightful nonlinear PDEs, it is not constrained to provide well-posed models. Unfortunately, this attribute of a discovered model cannot always be directly identified until the equation is subject to theoretical or numerical exploration. In this work, Section 6.3 described an analysis of the discovered PDE which revealed the instability and clearly provided a method to stabilize the equation. Section 6.4 then explored a temporally non-local method for estimating the parameters of the imposed stabilizing term and other uncertain terms.

It appears that SINDy is limited in its ability to enforce well-posedness of discovered models due to its temporal locality. More specifically, Equation (6.9) enforces that the data should approximately satisfy a PDE at any instant rather than approximating the solution of a PDE. To our knowledge, this shortcoming has



(a)

(b)

Figure 7.2 The resulting coefficients D and their corresponding R^2 using ordinary least squares with assumed velocity equation forms of (a) Stokes flow (7.1) and (b) an interfacial friction model (7.2). As in Figure 6.2, the dependence on the filtering parameter s is shown (see Equation (3.2)). Incompressibility enforced in the filled points and not in the open ones, the latter obtained by removing the $(\boldsymbol{I} - \hat{\boldsymbol{k}}\hat{\boldsymbol{k}})$–prefactor in Equations (7.1) and (7.2). Black lines represent derivatives computed spectrally while blue lines represent the results using finite differences and subsequently filtering.

.

not been fully addressed in applications of SINDy to PDE models. However the parameter estimation method in Section 6.4 suggests that incorporating simulation into the discovery procedure could result in more accurate parameters and more stable discovered models. Similar improvements were observed by other authors using similar penalties [45, 74].

As observed in [45], automatic differentiation (or a simpler procedure) allows for the simulation-based penalty to be incorporated directly into a sparse linear regression solver. However, there are a number of challenges in extending this idea. The first is that the simulation-based penalty simply increases the computational burden of the regression procedure, particularly in higher dimensions. Further, in many applications, appropriate boundary conditions are unknown, so there may not be a well-defined PDE boundary value problem to simulate. Similarly, it is likely that the simulation-based penalty will be evaluated for intermediate PDE models which are unstable or otherwise difficult to simulate. These challenges suggest that the design of effective and robust algorithms for simulation-based penalties are an interesting frontier in data-driven model discovery.

APPENDIX

A.1 Overview of Data-driven Discovery Methods

Recent advances in computational power and machine learning have brought a renewed interest in modeling physical phenomena directly from data. Classically, data is incorporated into models as a means to determine the correct model parameters for a given model form. In comparison, modern data-driven modeling and discovery methods aim to automatically determine both the model form and coefficients in order to match the data. Prediction has long been a goal for traditional scientific simulation and these methods present the potential for both data interpolation and extrapolation, especially in time. They also promise to facilitate new understanding of physical phenomena via novel model forms and new insights into previous models. Some common data-driven modeling methods are:

1. **Sparse Identification of Nonlinear Dynamics (SINDy)** [9] [73], Figure A.1

   curve fitting, linear regression

   This method is the method of choice for this dissertation and is outlined in Section 1.2. As a brief overview, data is transformed into various "terms" which are usually combinations of polynomials and numerical derivatives. A linear regression is then performed to quantify the contribution of these terms to the numerical time derivative of the data. Thus, this method aims to give a parsimonious closed form ordinary or partial differential equation governing the system.

2. **Weak SINDy** [60]

   weak curve fitting, linear regression

   This method follows the same general procedure as SINDy but instead of numerically differentiating the data it uses the concept of a weak PDE solution by multiplying the terms by test functions and then integrating by parts to transfer derivatives to the test functions. Thus, direct numerical derivatives are avoided in exchange for some obfuscation of the interpretation of each term in the system.

3. **Deep learning PDE (DL-PDE, DLG-PDE)** [97] [98], Figure A.2

   deep feed-forward neural networks, linear regression

   This method also follows the general outline of the SINDy method but substitutes the data fitting and numerical differentiation with a neural network. First, it aims

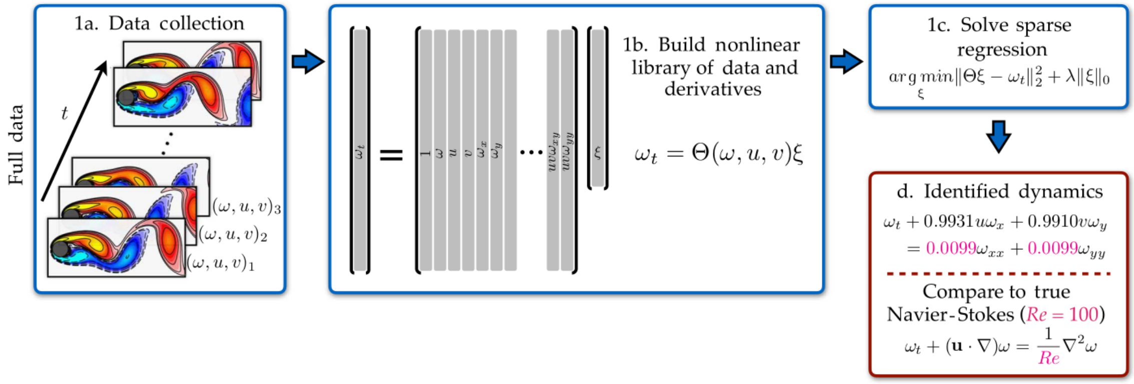

1a. Data collection
Full data
t
$(\omega, u, v)_3$
$(\omega, u, v)_2$
$(\omega, u, v)_1$
1b. Build nonlinear library of data and derivatives
$\omega_t = \Theta(\omega, u, v)\xi$
1c. Solve sparse regression
$\arg\min_{\xi} \|\Theta\xi - \omega_t\|_2^2 + \lambda\|\xi\|_0$
d. Identified dynamics
$\omega_t + 0.9931 u\omega_x + 0.9910 v\omega_y = 0.0099\omega_{xx} + 0.0099\omega_{yy}$
Compare to true Navier-Stokes ($Re = 100$)
$\omega_t + (\mathbf{u} \cdot \nabla)\omega = \frac{1}{Re}\nabla^2\omega$

Figure A.1 Outline of the SINDy/PDE-Find method.
Source: [73]

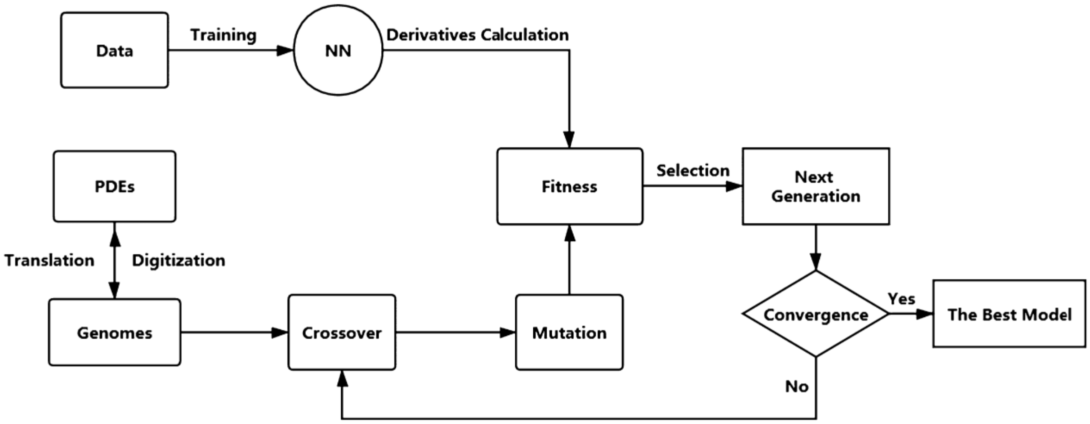


Figure A.2 Overview of DLGA-PDE Method.
Source: [98]

to fit a neural network to the raw data and uses this representation to construct the terms by relying on automatic differentiation of the network itself to get the numerical derivatives. It also can use genetic algorithms to determine the exact form for terms rather than explicitly constructing them. Overall, the aim is to provide a more flexible basis of possible term contributions for the regression procedure at the expense of more computation.

4. **PDE Net** [52], Figure A.3

   specialized convolutional neural networks

   Another neural network variation on the concepts introduced in SINDy, this method constructs a network which combines all aspects of the process: differentiating the data, creating various possible terms, and determining a sparse set of probable terms. It does so by using constrained but tunable convolutional filters which by connection to wavelets are guaranteed to approximate derivatives. After passing the data through these "derivative" filters, there are specially designed layers that can

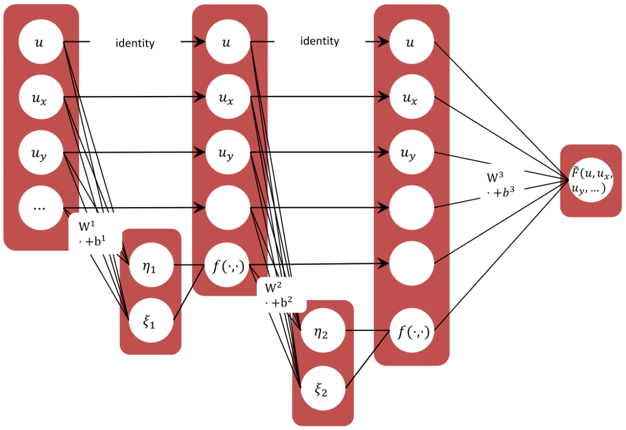


Figure A.3 Overview of symbolic regression network in PDE-Net.
Source: [52]

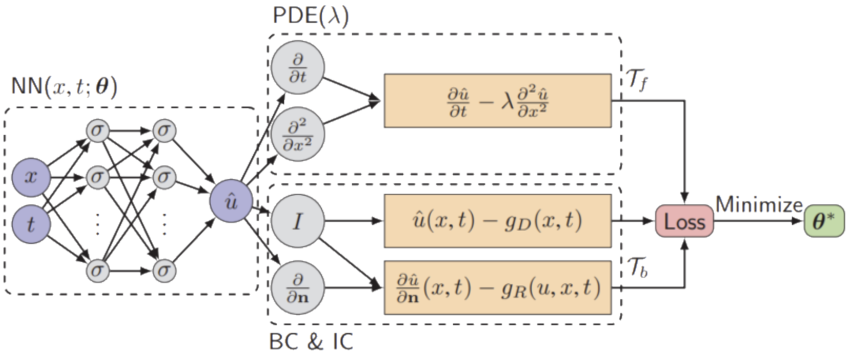

PDE(λ)
NN($x,t;\boldsymbol{\theta}$)
x
t
σ
$\hat{u}$
$\frac{\partial}{\partial t}$
$\frac{\partial^2}{\partial x^2}$
$\frac{\partial \hat{u}}{\partial t} - \lambda \frac{\partial^2 \hat{u}}{\partial x^2}$
$\mathcal{T}_f$
I
$\hat{u}(x,t) - g_D(x,t)$
$\frac{\partial}{\partial \mathbf{n}}$
$\frac{\partial \hat{u}}{\partial \mathbf{n}}(x,t) - g_R(u,x,t)$
$\mathcal{T}_b$
Loss
Minimize
$\boldsymbol{\theta}^*$
BC & IC

Figure A.4 Overview of PINNs.
Source: [53]

produce combinations of the terms and ultimately combine them. By these constraints and architecture, the network layers can be examined and interpreted to determine what constructed terms ultimately contribute to the final output.

5. **Physics informed neural networks (PINNs)** [70], Figure A.4

   deep feed-forward neural networks, specialized loss functions

   This method is less focused on discovering a new model structure and more on fitting the parameters of a given model. It aims to fit a neural network to given data representing the solution to some PDE and is trained by verifying that the neural network satisfies a given PDE form in its loss function. As such, it is fitting a highly nonlinear function to data but imbuing it with physical constraints.

A distinctive feature of SINDy is its level of interpretability. As an example, a neural network model fitted to data outputs a function made up of compositions of nonlinear functions and linear transformations which each have tuned parameters. The parameter space is usually so large as to make the network unwieldy to analysis and render the network a “black-box” transformation that gives almost no information beyond its output, which makes it difficult to interpret. On the other hand, SINDy results in a linear combination of known functions that can be analyzed to determine the contribution and sensitivity of each function with respect to the output, making the method very interpretable.

A.2 Total Least Squares

The TLS problem looks for the smallest L^2 shifts $\hat{A}$ and $\hat{b}$ that give an exact solution to the linear problem $(A+\hat{A})x = b+\hat{b}$, where $A \in \mathbb{R}^{N\times M}$, $x \in \mathbb{R}^M$ and $b \in \mathbb{R}^N$. The problem can then just be viewed as a projection of the original data onto the convex set $Ax = b$. This can be written in the form of an optimization problem as:

$$\begin{aligned} \underset{\hat{A},\hat{b},x}{\text{minimize}} \quad & \big\|[\hat{A},\hat{b}]\big\|_{\mathrm{F}} \\ \text{subject to} \quad & (A+\hat{A})x = b+\hat{b}. \end{aligned} \tag{A.1}$$

where $[\hat{A},\hat{b}]$ is the horizontal concatenation of the $\hat{A}$ matrix with the $\hat{b}$ vector and $\big\|\cdot\big\|_{\mathrm{F}}$ is the Frobenius norm:

$$\big\|A\big\|_{\mathrm{F}} = \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{M}|A_{ij}|^2} = \sqrt{\mathrm{Tr}(AA^\intercal)}.$$

Now this optimization problem has a closed form solution which can be concisely found by using the singular value decomposition of $[\hat{A},\hat{b}] = U\Sigma V^\intercal$: If we consider the SVD of the concatenated matrix $[A,b]$ (which has dimension $N\times(M+1)$):

$$[A,b] = U\Sigma V^\intercal,$$

we know that the closest projection onto an $N\times M$ matrix can be found by setting the smallest eigenvalue of Σ to 0. Thus,

$$[A+\hat{A}, b+\hat{b}] = U\hat{\Sigma}V^\intercal.$$

where $\Sigma = \hat{\Sigma}$ except that $\hat{\Sigma}_{M,M} = 0$. Then,

$$\begin{aligned} [\hat{A}, \hat{b}] &= [A, b] + [\hat{A}, \hat{b}] - [A, b] = [A + \hat{A}, b + \hat{b}] - [A, b] \\ &= U\hat{\Sigma}V^\intercal - U\hat{\Sigma}V^\intercal = -u_M \sigma_M v_M^\intercal. \end{aligned}$$

where u_M is the M^{th} column of U and σ_M is the corresponding singular value. Now, we can see that

$$\begin{aligned} &[A, b] = U\Sigma V^\intercal \quad \implies \quad [A, b]V = U\Sigma \quad \implies \quad [A, b]v_M = u_M \sigma_M \\ \text{and} \quad &[\hat{A}, \hat{b}] = -u_M \sigma_M v_M^\intercal = -[A, b]v_M v_M^\intercal. \end{aligned}$$

This gives:

$$\begin{aligned} [A + \hat{A}, b + \hat{b}] &= [A, b] - [A, b]v_M v_M^\intercal \\ \implies [A + \hat{A}, b + \hat{b}]v_M &= [A, b]v_M - [A, b]v_M = \mathbf{0} \\ \implies (A + \hat{A})v_{1:M-1,M} &= -v_{M,M}(b + \hat{b}), \end{aligned}$$

where $v_{1:M-1,M}$ is the M^{th} column of V with rows up to the $(M-1)^{\text{th}}$ row. This gives us a solution of

$$x_{\text{TLS}} = \frac{1}{v_{M,M}} v_{1:M-1,M},$$

where $v_{1:M-1,M}$ is the M^{th} column of V with rows up to the $(M-1)^{\text{th}}$ row.

A.2.1 Regularized Total Least Squares

Regularization is added when a problem is somewhat ill-posed and enforced regularity can stabilize the solution. Total least squares can suffer from this problem because, when compared with the OLS solution, the TLS solution has inherently more variance, which can be expressed as

$$\mathbb{E}[x_{\text{TLS}} - \mathbb{E}[x_{\text{TLS}}]]$$

where $\mathbb{E}$ is the expected value. This translates to a tendency of the solution to have unreasonably large numbers in certain ill-posed scenarios.

A modification to the standard TLS problem in Equation A.1 can be derived by adding a constraint on the size of the coefficients $\|Lx\|_2 < \delta$ (where L is a weighting matrix) which is usually called "Tikhonov regularization." Our problem then becomes:

$$\begin{aligned} \underset{\hat{A},\hat{b},x}{\text{minimize}} \quad & \big\|[\hat{A},\hat{b}]\big\|_{\mathrm{F}} \\ \text{subject to} \quad & (A+\hat{A})x = b+\hat{b}, \\ & \|Lx\|_2 \leq \delta. \end{aligned} \tag{A.2}$$

This problem can help to control the size of the elements in the solution but also introduces a significant computational challenge as inequality constraints are much more difficult to work with than equality constraints.

However, this problem can also be expressed as the solution to a nonlinear equation by using a few substitutions which was given as a Theorem in [37]. To show this, we first absorb the equality constraint into the objective function as:

$$\begin{aligned} & \big\|[\hat{A},\hat{b}]\big\|_{\mathrm{F}} \text{ and } (A+\hat{A})x = b+\hat{b} \\ \implies & \big\|[\hat{A},\hat{b}]\big\|_{\mathrm{F}} = \big\|[A+\hat{A}, b+\hat{b}] - [A,b]\big\|_{\mathrm{F}} = \big\|[A+\hat{A},(A+\hat{A})x] - [A,b]\big\|_{\mathrm{F}} \\ & \qquad = \big\|[\hat{A},(A+\hat{A})x - b]\big\|_{\mathrm{F}}. \end{aligned}$$

We now square our objective function and write the KKT conditions. First, the gradient with respect to $\hat{A}$ is:

$$\begin{aligned} 0 &= \nabla_{\hat{A}}\Big(\big\|[\hat{A},(A+\hat{A})x-b]\big\|_{\mathrm{F}}^2 + \lambda(\|Lx\|_2 - \delta)\Big) \\ &= \nabla_{\hat{A}}\Big(\big[\sum_{i=1}^{N}\sum_{j=1}^{M}\hat{A}_{ij}^2 + (A_{ij}x_j + \hat{A}_{ij}x_j - b_i)^2\big] + \lambda(\|Lx\|_2 - \delta)\Big) \\ &= \Big[\sum_{i=1}^{N}\sum_{j=1}^{M} 2\hat{A}_{ij} + 2(A_{ij}x_j + \hat{A}_{ij}x_j - b_i)x_j\Big] \end{aligned}$$

$$\implies (b - Ax)\|x\|_2^2 = (1+\|x\|_2^2)\hat{A}x.$$

The gradient with respect to x is:

$$
\begin{aligned}
0 &= \nabla_x \Big(\big[\sum_{i=1}^{N} \sum_{j=1}^{M} \hat{A}_{ij}^2 + (A_{ij}x_j + \hat{A}_{ij}x_j - b_i)^2 \big] + \lambda((\sum_{i=1}^{N} \sum_{j=1}^{M} (L_{ij}x_j)^2) - \delta) \Big) \\
&= \Big[\sum_{i=1}^{N} \sum_{j=1}^{M} 2(A_{ij}x_j + \hat{A}_{ij}x_j - b_i)(A_{ij} + \hat{A}_{ij}) + 2\lambda L_{ij}x_j L_{ij} \Big] \\
\implies & A^\intercal b = A^\intercal Ax + \hat{A}^\intercal Ax + \hat{A}^\intercal(Ax - b) + A^\intercal \hat{A}x + \hat{A}^\intercal \hat{A}x + \lambda L^\intercal Lx.
\end{aligned}
$$

Now, combining these:

$$
\begin{aligned}
A^\intercal b &= A^\intercal Ax + \hat{A}^\intercal Ax + \hat{A}^\intercal(Ax - b) + \hat{A}^\intercal \hat{A}x + A^\intercal \hat{A}x + \lambda L^\intercal Lx \\
&= A^\intercal Ax + \hat{A}^\intercal Ax + \hat{A}^\intercal \left(Ax - b + \|x\|_2^2 \frac{b - Ax}{1 + \|x\|_2^2} \right) + A^\intercal \hat{A}x + \lambda L^\intercal Lx \\
&= A^\intercal \left(Ax + \frac{(b - Ax)\|x\|_2^2}{1 + \|x\|_2^2} \right) + \hat{A}^\intercal Ax + \hat{A}^\intercal \left(\frac{Ax - b}{1 + \|x\|_2^2} \right) + \lambda L^\intercal Lx \\
\frac{(1 - \|x\|_2^2)A^\intercal b}{1 + \|x\|_2^2} &= \frac{A^\intercal Ax}{1 + \|x\|_2^2} + \hat{A}^\intercal Ax + \hat{A}^\intercal \left(\frac{Ax - b}{1 + \|x\|_2^2} \right) + \lambda L^\intercal Lx \\
A^\intercal b &= A^\intercal Ax + \hat{A}^\intercal Ax(1 + \|x\|_2^2) + \hat{A}^\intercal Ax - \hat{A}^\intercal b + A^\intercal b\|x\|_2^2 + \lambda(1 + \|x\|_2^2)L^\intercal Lx \\
& \boxed{A^\intercal b = A^\intercal Ax - \frac{\|b - Ax\|_2^2}{1 + \|x\|_2^2} Ix + \lambda(1 + \|x\|_2^2)L^\intercal Lx}.
\end{aligned}
$$

This is a surprising result since the Tikhonov regularized OLS solution can be written as:

$$A^\intercal b = A^\intercal Ax + \lambda L^\intercal Lx.$$

Thus, in principle, we see that they are not that far apart.

A.2.2 Computational Examples

For a simple computational example, we consider the system (Figure A.5):

$$y = c_1 x^5 + c_2 x^4, \;\; x \in (-1.1, 0.5), \tag{A.3}$$

which gives $A = \begin{pmatrix} x^5 & x^4 \end{pmatrix}$. Setting up this problem, we can see that TLS far outperforms OLS when noise is added to the matrix A (Figure A.6), which

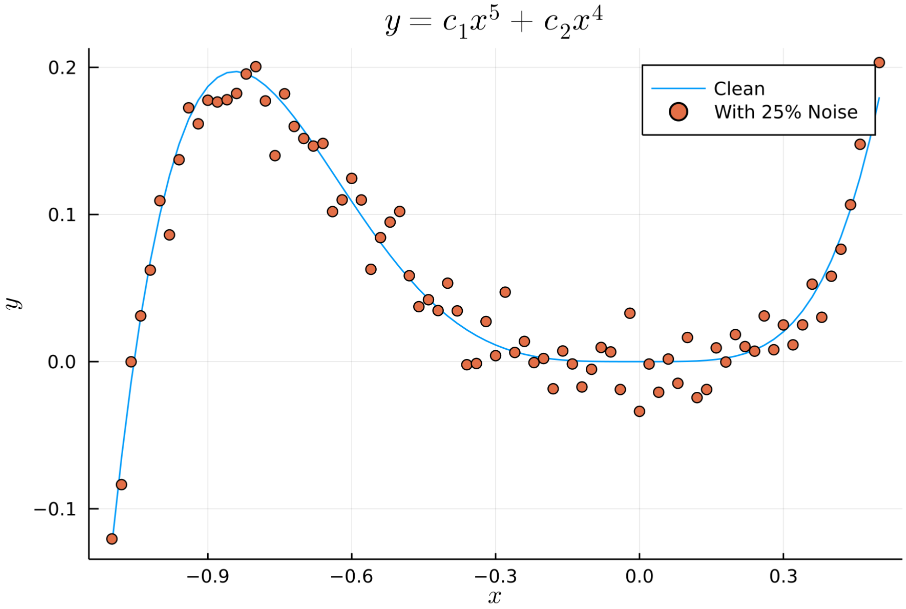


Figure A.5 Example function for feature identification with total least squares as described in Equation A.3 where $c_1 = 1.85$, $c_2 = 1.95$.

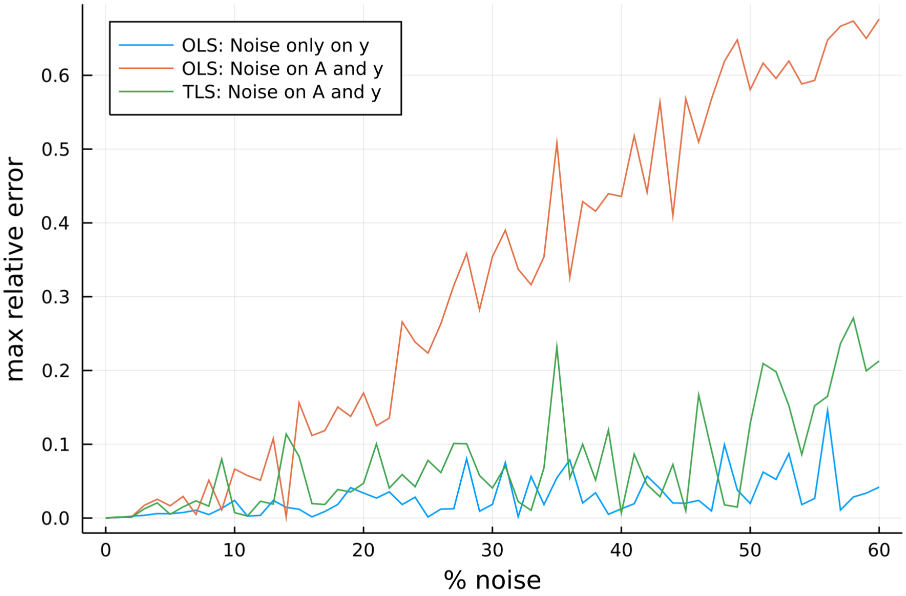


Figure A.6 Demonstration of the improved accuracy of total least squares for cases when noise is present in the independent variables.

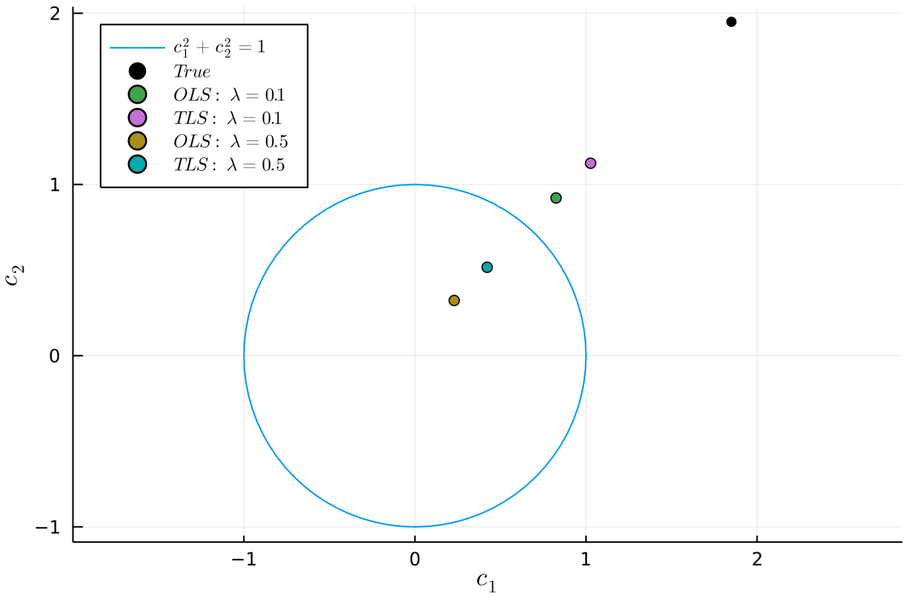


Figure A.7 Comparison of regularized versions of ordinary and total least squares to identify the coefficients described in Figure A.5.

is the reason that TLS was originally formulated. However, using the Newton's implementation from the previous section, we can see that the regularized TLS solution behaves very similarly to the regularized OLS solution as we constrain the size of the coefficients (Figure A.7).

Now, originally, we considered this problem because we noticed that in ill-posed scenarios when the correct terms (usually high order) were either partially represented or not present in A, the OLS coefficients tended toward zero and the TLS coefficients tended toward infinity. In order to recreate this setting, we consider an incorrect matrix $A = \begin{pmatrix} x^7 & x^6 \end{pmatrix}$. Although these terms are incorrect, they represent (roughly) the same dynamics (Figure A.8).

As expected, using these incorrect terms in our regression biases the OLS coefficients toward zero while biasing the TLS coefficients toward infinity (Figure A.9). Since the OLS coefficients are get smaller with more noise, we cannot use regularization to bring them closer to the true solution. However, we can use our regularized TLS method to bring the total least squares coefficients toward the correct

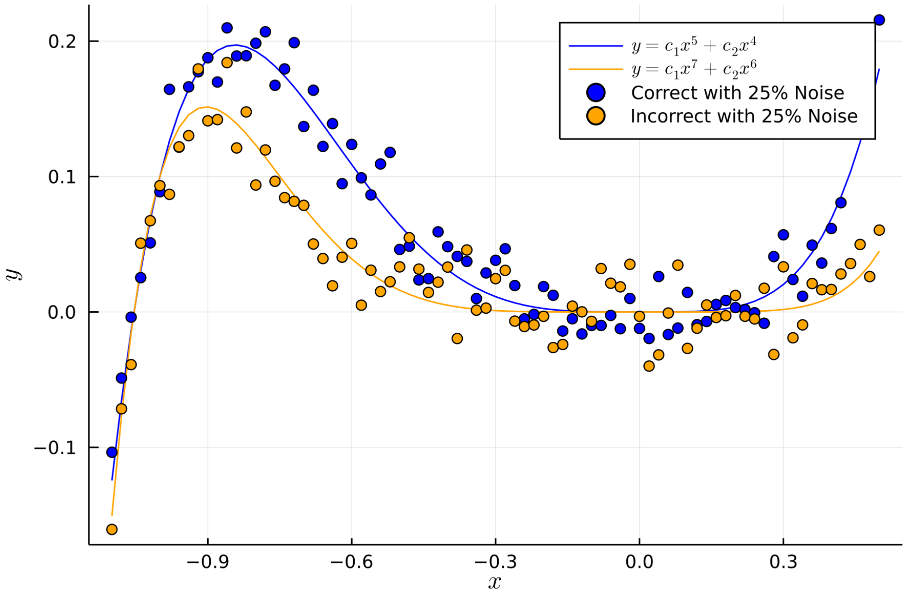


Figure A.8 Comparison of the example function given in Equation A.3 with an equation which is qualitatively similar but incorrect.

solution (Figure A.10). Thus, we see that total least squares can be used as a potential solution for ill-posed system identification or parameter recovery problems.

A.3 Optical Flow

Optical flow is tracking the flow of a feature in an image through time. Generally, you can consider pixel intensity as the most specific feature we want to watch flow through time: $I(x, y, t)$. If it is unchanging, we can have:

$$\begin{aligned} I(x,y,t) &= I(x+\Delta x, y+\Delta y, t+\Delta t) \\ &= I(x,y,t) + \frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t \\ \implies 0 &= \frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t \\ \frac{\partial I}{\partial t} &= -\nabla I \cdot V. \end{aligned}$$

This general form is underdetermined as we need to solve for two components of velocity $V = \begin{pmatrix} u \\ v \end{pmatrix}$ with only information from one value, pixel intensity: $I(x, y, t)$.

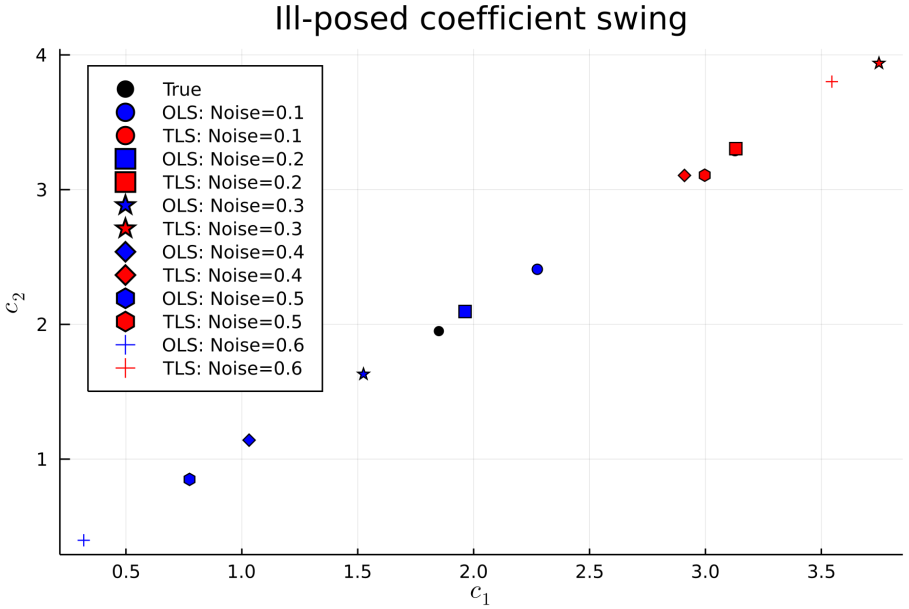


Figure A.9 The variance of calculated coefficients for the ill-posed least squares problem in which the correct term is not present.

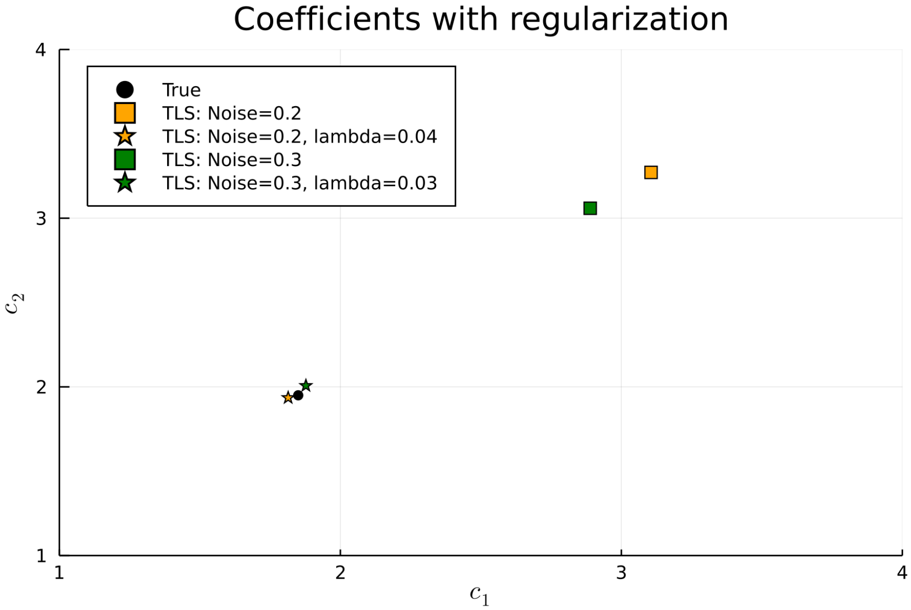


Figure A.10 Improvements to coefficient accuracy using regularization on ill-posed problem.

As a result, we need to make an assumption to close the system. Some common ideas are:

- Using least squares to solve the underdetermined problem above
- Horn Schunck method: Assume smoothness in the flow and minimize variation
- Account for lack of temporal smoothness [26]
- Scalar invariant feature transform which is the most common feature detection method [51]
- Tracking patches between frames [80]

A variety of methods are illustrated in the review in [86].

A.4 Circular Gaussian

Given that the domain of orientation $\boldsymbol{p}$ is over the half-circle in 2D, a standard Gaussian is replaced with a two-peaked variation of the wrapped Gaussian in terms of angle θ which relates to orientation $\boldsymbol{p}$ as $\boldsymbol{p} = (\cos\theta, \sin\theta)$:

$$\tau_{\sigma_p}(\theta) = \frac{1}{2\sigma_p\sqrt{2\pi}} \sum_{k=-\infty}^{\infty} \exp\left[\frac{-(\theta + \pi k)^2}{2\sigma_p^2}\right].$$

This distribution has peaks at $\theta = 0, \pi, -\pi$ and standard deviation σ_p. Now, since the distribution is a multiplication of Gaussians $\boldsymbol{\Psi}(\boldsymbol{x}, \boldsymbol{p}) = \Psi(\boldsymbol{x})\tau(\theta)$, we have:

$$\begin{aligned}
\boldsymbol{Q}(\boldsymbol{x}) &= \Psi(\boldsymbol{x}) \int \tau_{\sigma_p}(\theta - \hat{\theta}) \left(\begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta \end{bmatrix} - \frac{\boldsymbol{I}}{2} \right) d\theta \\
&= \frac{1}{2}\Psi(\boldsymbol{x}) \int \tau_{\sigma_p}(\theta - \hat{\theta}) \left(\begin{bmatrix} \cos 2\theta & \sin 2\theta \\ \sin 2\theta & -\cos 2\theta \end{bmatrix} \right) d\theta \\
&= \frac{1}{4}\Psi(\boldsymbol{x}) \int \tau_{\sigma_p}(\theta - \hat{\theta}) \left(\begin{bmatrix} e^{i2\theta} + e^{-i2\theta} & \frac{1}{i}(e^{i2\theta} - e^{-i2\theta}) \\ \frac{1}{i}(e^{i2\theta} - e^{-i2\theta}) & -(e^{i2\theta} + e^{-i2\theta}) \end{bmatrix} \right) d\theta \\
&= \frac{1}{2}\Psi(\boldsymbol{x}) e^{-2\sigma_p^2} \left(\begin{bmatrix} \cos 2\hat{\theta} & \sin 2\hat{\theta} \\ \sin 2\hat{\theta} & -\cos 2\hat{\theta} \end{bmatrix} \right) \\
&= \Psi(\boldsymbol{x}) e^{-2\sigma_p^2} \left(\hat{\boldsymbol{p}}\hat{\boldsymbol{p}} - \frac{\boldsymbol{I}}{2} \right)
\end{aligned}$$

where $\hat{\theta} = \theta(\boldsymbol{x})$ which is the observed orientation at position $\boldsymbol{x}$ and thus $\hat{\boldsymbol{p}} = (\cos\hat{\theta}, \sin\hat{\theta})$.

A.5 Thin Film Approximation

If the flow velocity in a thin plate is written as $(\boldsymbol{u}, w)$ where $\boldsymbol{u} \in \mathbb{R}^2$, $\boldsymbol{x} \in [0, L]^2$, $z \in [0, H]$, $H \ll L$, the Stokes equations can be written as:

$$\begin{aligned}
0 &= \boldsymbol{\nabla} \cdot \boldsymbol{\sigma} + \eta \left(\Delta \boldsymbol{u} + \boldsymbol{u}_{zz} \right), \\
0 &= \eta \left(\Delta w + w_{zz} \right), \\
0 &= \boldsymbol{\nabla} \cdot \boldsymbol{u} + w_z, \\
\boldsymbol{u}, w &= 0 \quad \text{on} \quad z = 0, H.
\end{aligned} \tag{A.4}$$

Here, the gradient $\boldsymbol{\nabla} = (\partial_x, \partial_y)$ and $\Delta = \partial_{xx} + \partial_{yy}$ are defined to be in the plane with the assumption that there is no stress acting in the z–direction.

Averaging in the z-direction by integrating the equations in the z-direction from $z = 0$ and $z = H$ and dividing by H, the equations are written:

$$\begin{aligned}
0 &= \boldsymbol{\nabla} \cdot \hat{\boldsymbol{\sigma}} + \eta \left(\Delta \hat{\boldsymbol{u}} + \frac{\boldsymbol{u}_z(\boldsymbol{x}, H) - \boldsymbol{u}_z(\boldsymbol{x}, 0)}{H} \right) \\
0 &= \eta \left(\Delta \hat{w} + \frac{w_z(\boldsymbol{x}, H) - w_z(\boldsymbol{x}, 0)}{H} \right) \\
0 &= \boldsymbol{\nabla} \cdot \hat{\boldsymbol{u}}, \qquad \text{(A.5)}
\end{aligned}$$

where hats denote quantities that are averaged in the z-direction. These equations can be solved using the ansatz:

$$\boldsymbol{u}(\boldsymbol{x}, z) = 6 \hat{\boldsymbol{u}}(\boldsymbol{x}) \frac{z}{H} \left(1 - \frac{z}{H} \right), \quad \boldsymbol{w} = 0, \qquad \text{(A.6)}$$

which is valid for $H \ll L$. Velocity in the plane, $\hat{\boldsymbol{u}}$, can then be written:

$$\begin{aligned}
0 &= \boldsymbol{\nabla} \cdot \hat{\boldsymbol{\sigma}} + \eta \left(\Delta \hat{\boldsymbol{u}} - \frac{12}{H^2} \hat{\boldsymbol{u}} \right) \\
0 &= \boldsymbol{\nabla} \cdot \hat{\boldsymbol{u}}. \qquad \text{(A.7)}
\end{aligned}$$

Thus, the depth-averaged equation for the horizontal velocity is the Stokes equation with a friction term proportional to $\hat{\boldsymbol{u}}$. In the limit $L \gg H$, L being the characteristic lengthscale of velocity fluctuations, Eq. (A.7) reduces to an equation in which $\boldsymbol{u}$ is directly determined by the stress $\hat{\boldsymbol{\sigma}}$, as proposed in [67].

REFERENCES

[1] Comment on “Long-Lived Giant Number Fluctuations in a Swarming Granular Nematic”. *Science*, 320(5876):612, 2008.

[2] David W Allender and Michael A Lee. Landau theory of biaxial nematic liquid crystals. *Molecular Crystals and Liquid Crystals*, 110(1-4):331–339, 1984.

[3] E Paulo Alves and Frederico Fiuza. Data-driven discovery of reduced plasma physics models from fully kinetic simulations. *Physical Review Research*, 4(3):033192, 2022.

[4] John Ball. The Q-tensor theory of liquid crystals. *Lecture slides, Benin Summer School*, 28, 2010.

[5] John M Ball and Apala Majumdar. Nematic liquid crystals: from Maier-Saupe to a continuum theory. *Molecular crystals and liquid crystals*, 525(1):1–11, 2010.

[6] Antony N Beris and Brian J Edwards. *Thermodynamics of flowing systems: with internal microstructure*. Number 36. Oxford University Press on Demand, 1994.

[7] Steven L Brunton and J Nathan Kutz. Methods for data-driven multiscale model discovery for materials. *Journal of Physics: Materials*, 2(4):044002, 2019.

[8] Steven L Brunton and J Nathan Kutz. *Data-driven science and engineering: Machine learning, dynamical systems, and control*. Cambridge University Press, 2022.

[9] Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016.

[10] Emmanuel J Candes. The restricted isometry property and its implications for compressed sensing. *Comptes rendus mathematique*, 346(9-10):589–592, 2008.

[11] Andrea Cavagna and Irene Giardina. Bird flocks as condensed matter. *Annu. Rev. Condens. Matter Phys.*, 5(1):183–207, 2014.

[12] Kathleen P Champion, Steven L Brunton, and J Nathan Kutz. Discovery of nonlinear multiscale systems: Sampling strategies and embeddings. *SIAM Journal on Applied Dynamical Systems*, 18(1):312–333, 2019.

[13] William S Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836, 1979.

[14] Jonathan Colen, Ming Han, Rui Zhang, Steven A Redford, Linnea M Lemma, Link Morgan, Paul V Ruijgrok, Raymond Adkins, Zev Bryant, Zvonimir Dogic, et al. Machine learning active-nematic hydrodynamics. *Proceedings of the National Academy of Sciences*, 118(10):e2016708118, 2021.

[15] Brian M de Silva, Kathleen Champion, Markus Quade, Jean-Christophe Loiseau, J Nathan Kutz, and Steven L Brunton. PySINDy: a python package for the sparse identification of nonlinear dynamics from data. *arXiv preprint arXiv:2004.08424*, 2020.

[16] Stephen J DeCamp, Gabriel S Redner, Aparna Baskaran, Michael F Hagan, and Zvonimir Dogic. Orientational order of motile defects in active nematics. *Nature materials*, 14(11):1110–1115, 2015.

[17] Charles B Delahunt and J Nathan Kutz. A toolkit for data-driven discovery of governing equations in high-noise regimes. *IEEE Access*, 10:31210–31234, 2022.

[18] Ricardo A Delgadillo, Jingwei Hu, and Haizhao Yang. Multiscale and nonlocal learning for PDEs using densely connected RNNs. *arXiv preprint arXiv:2109.01790*, 2021.

[19] David L Donoho. Compressed sensing. *IEEE Transactions on information theory*, 52(4):1289–1306, 2006.

[20] A. Doostmohammadi, M. F. Adamer, S. P. Thampi, and J. M. Yeomans. Stabilization of active matter by flow-vortex lattices and defect ordering. *Nat. Commun.*, 7:10557, 2016.

[21] Amin Doostmohammadi, Jordi Ignés-Mullol, Julia M Yeomans, and Francesc Sagués. Active nematics. *Nature communications*, 9(1):3246, 2018.

[22] Guillaume Duclos, Raymond Adkins, Debarghya Banerjee, Matthew S. E. Peterson, Minu Varghese, Itamar Kolvin, Arvind Baskaran, Robert A. Pelcovits, Thomas R. Powers, Aparna Baskaran, Federico Toschi, Michael F. Hagan, Sebastian J. Streichan, Vincenzo Vitelli, Daniel A. Beller, and Zvonimir Dogic. Topological structure and dynamics of three-dimensional active nematics. *Science*, 367(6482):1120–1124, 2020.

[23] Jörn Dunkel, Sebastian Heidenreich, Markus Bär, and Raymond E Goldstein. Minimal continuum theories of structure formation in dense active fluids. *New Journal of Physics*, 15(4):045016, 2013.

[24] Perry W. Ellis, Daniel J. G. Pearce, Ya-Wen Chang, Guillermo Goldsztein, Luca Giomi, and Alberto Fernandez-Nieves. Curvature-induced defect unbinding and dynamics in active nematic toroids. *Nat. Phys.*, 14:85–90, 2018.

[25] Andrei V Ermolaev, Anastasiia Sheveleva, Goëry Genty, Christophe Finot, and John M Dudley. Data-driven model discovery of ideal four-wave mixing in nonlinear fibre optics. *Scientific Reports*, 12(1):12711, 2022.

[26] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image Analysis: 13th Scandinavian Conference, SCIA 2003 Halmstad, Sweden, June 29–July 2, 2003 Proceedings 13*, pages 363–370. Springer, 2003.

[27] Donald E Farrar and Robert R Glauber. Multicollinearity in regression analysis: the problem revisited. *The Review of Economic and Statistics*, pages 92–107, 1967.

[28] Urban Fasel, J Nathan Kutz, Bingni W Brunton, and Steven L Brunton. Ensemble-SINDy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control. *Proceedings of the Royal Society A*, 478(2260):20210904, 2022.

[29] Bengt Fornberg and Tobin A Driscoll. A fast spectral algorithm for nonlinear wave equations with linear dispersion. *Journal of Computational Physics*, 155(2):456–467, 1999.

[30] Peter J Foster. *Active Dynamics of Microtubule and Motor Protein Networks.* PhD thesis, Harvard University, Boston, MA, 2017.

[31] Tong Gao, Meredith D Betterton, An-Sheng Jhang, and Michael J Shelley. Analytical structure, dynamics, and coarse graining of a kinetic model of an active fluid. *Physical Review Fluids*, 2(9):093302, 2017.

[32] Tong Gao, Robert Blackwell, Matthew A Glaser, Meredith D Betterton, and Michael J Shelley. Multiscale polar theory of microtubule and motor-protein assemblies. *Physical review letters*, 114(4):048101, 2015.

[33] Walter Gautschi. Orthogonal polynomials, quadrature, and approximation: computational methods and software (in matlab). *Orthogonal Polynomials and Special Functions: Computation and Applications*, pages 1–77, 2006.

[34] Luca Giomi, Mark J. Bowick, Xu Ma, and M. Cristina Marchetti. Defect annihilation and proliferation in active nematics. 110(22):228101.

[35] Luca Giomi, L Mahadevan, Bulbul Chakraborty, and Michael Hagan. Banding, excitability and chaos in active nematic suspensions. *Nonlinearity*, 25(8):2245, 2012.

[36] M. Golden, R. Grigoriev, J. Nambisan, and A. Fernandez-Nieves. Physically-informed data-driven modeling of active nematics. *arXiv*, 2202.12853, 2022.

[37] Gene H Golub, Per Christian Hansen, and Dianne P O'Leary. Tikhonov regularization and total least squares. *SIAM journal on matrix analysis and applications*, 21(1):185–194, 1999.

[38] Daniel R Gurevich, Patrick AK Reinbold, and Roman O Grigoriev. Learning fluid physics from highly turbulent data using sparse physics-informed discovery of empirical relations (SPIDER). *arXiv preprint arXiv:2105.00048*, 2021.

[39] Benjamín Herrmann, Philipp Oswald, Richard Semaan, and Steven L Brunton. Modeling synchronization in forced turbulent oscillator flows. *Communications Physics*, 3(1):195, 2020.

[40] Seth M Hirsh, David A Barajas-Solano, and J Nathan Kutz. Sparsifying priors for bayesian uncertainty quantification in model discovery. *Royal Society Open Science*, 9(2):211823, 2022.

[41] Ronald R Hocking. A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics*, pages 1–49, 1976.

[42] Jonathan Horrocks and Chris T Bauch. Algorithmic discovery of dynamic models from infectious disease data. *Scientific reports*, 10(1):7061, 2020.

[43] Steven G Johnson. Notes on FFT-based differentiation. *MIT Applied Mathematics, Boston, MA*, 2011.

[44] Chaitanya Joshi, Sattvic Ray, Linnea M. Lemma, Minu Varghese, Graham Sharp, Zvonimir Dogic, Aparna Baskaran, and Michael F. Hagan. Data-driven discovery of active nematic hydrodynamics. *Phys. Rev. Lett.*, 129:258001, Dec 2022.

[45] Kadierdan Kaheman, Steven L Brunton, and J Nathan Kutz. Automatic differentiation to simultaneously identify nonlinear dynamics and extract noise probability distributions from data. *Machine Learning: Science and Technology*, 3(1):015031, 2022.

[46] F. C. Keber, E. Loiseau, T. Sanchez, S. J. DeCamp, L. Giomi, M. J. Bowick, M. C. Marchetti, Z. Dogic, and A. R. Bausch. Topology and dynamics of active nematic vesicles. *Science*, 345(6201):1135–1139, 2014.

[47] Maurice Kléman. Defects in liquid crystals. *Reports on Progress in Physics*, 52(5):555, 1989.

[48] Ian Knowles and Robert J Renka. Methods for numerical differentiation of noisy data. *Electron. J. Differ. Equ*, 21:235–246, 2014.

[49] Cornelius Lanczos. Evaluation of noisy data. *Journal of the Society for Industrial and Applied Mathematics, Series B: Numerical Analysis*, 1(1):76–85, 1964.

[50] He Li, Xia-qing Shi, Mingji Huang, Xiao Chen, Minfeng Xiao, Chenli Liu, Hugues Chaté, and HP Zhang. Data-driven quantitative modeling of bacterial active nematics. *Proceedings of the National Academy of Sciences*, 116(3):777–785, 2019.

[51] Tony Lindeberg. Scale-space theory: A basic tool for analyzing structures at different scales. *Journal of applied statistics*, 21(1-2):225–270, 1994.

[52] Zichao Long, Yiping Lu, and Bin Dong. PDE-net 2.0: Learning PDEs from data with a numeric-symbolic hybrid deep network. *Journal of Computational Physics*, 399:108925, 2019.

[53] Lu Lu, Xuhui Meng, Zhiping Mao, and George Em Karniadakis. Deepxde: A deep learning library for solving differential equations. *SIAM review*, 63(1):208–228, 2021.

[54] Geoffrey R Luckhurst and Timothy J Sluckin. *Biaxial nematic liquid crystals: theory, simulation and experiment*. John Wiley and Sons, 2015.

[55] Yvon Maday, Anthony T Patera, and Einar M Rønquist. An operator-integration-factor splitting method for time-dependent problems: application to incompressible fluid flow. *Journal of Scientific Computing*, 5:263–292, 1990.

[56] Ananyo Maitra, Pragya Srivastava, M. Cristina Marchetti, Juho S. Lintuvuori, Sriram Ramaswamy, and Martin Lenz. A nonequilibrium force can stabilize 2D active nematics. *Proceedings of the National Academy of Sciences*, 115(27):6934–6939, 2018.

[57] Angelika Manhart, Stefanie Windner, Mary Baylies, and Alex Mogilner. Mechanical positioning of multiple nuclei in muscle cells. *PLoS computational biology*, 14(6):e1006208, 2018.

[58] M. C. Marchetti, J. F. Joanny, S. Ramaswamy, T. B. Liverpool, J. Prost, Madan Rao, and R. Aditi Simha. Hydrodynamics of soft active matter. *Rev. Mod. Phys.*, 85:1143, 2013.

[59] Ivan Markovsky and Sabine Van Huffel. Overview of total least-squares methods. *Signal processing*, 87(10):2283–2302, 2007.

[60] Daniel A Messenger and David M Bortz. Weak SINDy for partial differential equations. *Journal of Computational Physics*, 443:110525, 2021.

[61] Daniel A Messenger and David M Bortz. Learning mean-field equations from particle data using WSINDy. *Physica D: Nonlinear Phenomena*, 439:133406, 2022.

[62] Daniel A Messenger, Graycen E Wheeler, Xuedong Liu, and David M Bortz. Learning anisotropic interaction rules from individual trajectories in a heterogeneous cellular population. *Journal of the Royal Society Interface*, 19(195):20220412, 2022.

[63] Vijay Narayan, Sriram Ramaswamy, and Narayanan Menon. Long-lived giant number fluctuations in a swarming granular nematic. *Science*, 317(5834):105–108, 2007.

[64] Dimitar Ninevski and Paul O'Leary. Detection of derivative discontinuities in observational data. In *Advances in Intelligent Data Analysis XVIII: 18th International Symposium on Intelligent Data Analysis, IDA 2020, Konstanz, Germany, April 27–29, 2020, Proceedings 18*, pages 366–378. Springer, 2020.

[65] Achini Opathalage, Michael M. Norton, Michael P. N. Juniper, Blake Langeslay, S. Ali Aghvami, Seth Fraden, and Zvonimir Dogic. Self-organized dynamics and the transition to turbulence of confined active nematics. 116(11):4788–4797.

[66] Steven A Orszag. On the elimination of aliasing in finite-difference schemes by filtering high-wavenumber components. *Journal of Atmospheric Sciences*, 28(6):1074–1074, 1971.

[67] Anand U Oza and Jörn Dunkel. Antipolar ordering of topological defects in active liquid crystals. *New Journal of Physics*, 18(9):093006, 2016.

[68] 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*, pages 40–44. IEEE, 1993.

[69] David L Ragozin. Error bounds for derivative estimates based on spline smoothing of exact or noisy data. *Journal of approximation theory*, 37(4):335–355, 1983.

[70] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

[71] S. Ramaswamy. The mechanics and statistics of active matter. *Annu. Rev. Cond. Mat. Phys.*, 1:323–345, 2010.

[72] Miha Ravnik and Slobodan Žumer. Landau–de Gennes modelling of nematic liquid crystal colloids. *Liquid Crystals*, 36(10-11):1201–1214, 2009.

[73] Samuel H Rudy, Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Data-driven discovery of partial differential equations. *Science advances*, 3(4):e1602614, 2017.

[74] Samuel H Rudy, J Nathan Kutz, and Steven L Brunton. Deep learning of dynamics and signal-noise decomposition with time-stepping constraints. *Journal of Computational Physics*, 396:483–506, 2019.

[75] Tim Sanchez, Daniel TN Chen, Stephen J DeCamp, Michael Heymann, and Zvonimir Dogic. Spontaneous motion in hierarchically assembled active matter. *Nature*, 491(7424):431–434, 2012.

[76] Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639, 1964.

[77] Hayden Schaeffer. Learning partial differential equations via data discovery and sparse optimization. 473(2197):20160446. Publisher: Royal Society.

[78] Hayden Schaeffer and Scott G McCalla. Sparse model selection via integral terms. *Physical Review E*, 96(2):023302, 2017.

[79] George AF Seber and Alan J Lee. *Linear regression analysis*, volume 330. John Wiley and Sons, 2003.

[80] Gregory Shakhnarovich. *Learning task-specific similarity.* PhD thesis, Massachusetts Institute of Technology, 2005.

[81] X.-Q. Shi and Y.-Q. Ma. Topological structure dynamics revealing collective evolution in active nematics. *Nat. Commun.*, 4:3013, 2013.

[82] R Aditi Simha and Sriram Ramaswamy. Hydrodynamic fluctuations and instabilities in ordered suspensions of self-propelled particles. *Physical review letters*, 89(5):058101, 2002.

[83] Jürgen Spitaler and Stefan K Estreicher. Perspectives on the theory of defects. *Frontiers in Materials*, 5:70, 2018.

[84] Norbert Stoop, Romain Lagrange, Denis Terwagne, Pedro M Reis, and Jörn Dunkel. Curvature-induced symmetry breaking determines elastic surface patterns. *Nature materials*, 14(3):337–342, 2015.

[85] Salik A. Sultan, Mehrana R. Nejad, and Amin Doostmohammadi. Quadrupolar active stress induces exotic patterns of defect motion in compressible active nematics. *Soft Matter*, 18:4118–4126, 2022.

[86] Deqing Sun, Stefan Roth, and Michael J Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *International Journal of Computer Vision*, 106:115–137, 2014.

[87] Rohit Supekar, Boya Song, Alasdair Hastewell, Gary PT Choi, Alexander Mietke, and Jörn Dunkel. Learning hydrodynamic equations for active matter from particle simulations and experiments. *Proceedings of the National Academy of Sciences*, 120(7):e2206994120, 2023.

[88] Amanda J Tan, Eric Roberts, Spencer A Smith, Ulyses Alvarado Olvera, Jorge Arteaga, Sam Fortini, Kevin A Mitchell, and Linda S Hirst. Topological chaos in active nematics. *Nature Physics*, 15(10):1033–1039, 2019.

[89] Sumesh P Thampi, Amin Doostmohammadi, Ramin Golestanian, and Julia M Yeomans. Intrinsic free energy in active nematics. *Europhysics Letters*, 112(2):28004, 2015.

[90] Sumesh P Thampi, Ramin Golestanian, and Julia M Yeomans. Velocity correlations in an active nematic. *Physical review letters*, 111(11):118101, 2013.

[91] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[92] Joel A Tropp. Greed is good: Algorithmic results for sparse approximation. *IEEE Transactions on Information theory*, 50(10):2231–2242, 2004.

[93] Joel A Tropp and Stephen J Wright. Computational methods for sparse solution of linear inverse problems. *Proceedings of the IEEE*, 98(6):948–958, 2010.

[94] Henning U Voss, Jens Timmer, and Jürgen Kurths. Nonlinear dynamical system identification from uncertain and indirect measurements. *International Journal of Bifurcation and Chaos*, 14(06):1905–1933, 2004.

[95] Wei Wang, Pingwen Zhang, and Zhifei Zhang. Rigorous derivation from Landau–de Gennes theory to Ericksen–Leslie theory. *SIAM Journal on Mathematical Analysis*, 47(1):127–158, 2015.

[96] Scott Weady, Michael J. Shelley, and David B. Stein. A fast Chebyshev method for the Bingham closure with application to active nematic suspensions. *Journal of Computational Physics*, 457:110937, 2022.

[97] Hao Xu, Haibin Chang, and Dongxiao Zhang. DL-PDE: Deep-learning based data-driven discovery of partial differential equations from discrete and noisy data. *arXiv preprint arXiv:1908.04463*, 2019.

[98] Hao Xu, Haibin Chang, and Dongxiao Zhang. DLGA-PDE: Discovery of PDEs with incomplete candidate library via combination of deep learning and genetic algorithm. *Journal of Computational Physics*, 418:109584, 2020.

[99] Linan Zhang and Hayden Schaeffer. On the convergence of the SINDy algorithm. *Multiscale Modeling and Simulation*, 17(3):948–972, 2019.

[100] Yanxia Zhang, Jinqiao Duan, Yanfei Jin, and Yang Li. Discovering governing equation from data for multi-stable energy harvester under white noise. *Nonlinear Dynamics*, 106:2829–2840, 2021.

[101] Peng Zheng, Travis Askham, Steven L. Brunton, J. Nathan Kutz, and Aleksandr Y. Aravkin. A unified framework for sparse relaxed regularized regression: SR3. 7:1404–1423. Conference Name: IEEE Access.

[102] Zhengyang Zhou, Chaitanya Joshi, Ruoshi Liu, Michael M Norton, Linnea Lemma, Zvonimir Dogic, Michael F Hagan, Seth Fraden, and Pengyu Hong. Machine learning forecasting of active nematics. *Soft matter*, 17(3):738–747, 2021.