

Data-Driven Bayesian Uncertainty Quantification for Problems in Systems Biology

by

Karen R. Larson

B.Sc., Davidson College; Davidson, NC, 2012

M.Sc., Brown University; Providence, RI, 2016

A dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in The Division of Applied Mathematics at Brown University

PROVIDENCE, RHODE ISLAND

May 2020

© Copyright 2020 by Karen R. Larson

This dissertation by Karen R. Larson is accepted in its present form
by The Division of Applied Mathematics as satisfying the
dissertation requirement for the degree of Doctor of Philosophy.

Date_____

Anastasios Matzavinos, Ph.D., Advisor

Recommended to the Graduate Council

Date_____

Martin Maxey, Ph.D., Reader

Date_____

Sarah D. Olson, Ph.D., Reader

Approved by the Graduate Council

Date_____

Andrew G. Campbell, Dean of the Graduate School

Vita

EDUCATION

Brown University, Providence, Rhode Island, USA

M.S., Applied Mathematics, May 2016

Davidson College, Davidson, North Carolina, USA

B.S. Mathematics and Honors in Applied Mathematics, May 2012

PEER-REVIEWED JOURNAL ARTICLES

- Karen Larson, Clark Bowman, Costas Papadimitriou, Petros Koumoutsakos, and Anastasios Matzavinos. Detection of arterial wall abnormalities via Bayesian model selection. *Royal Society Open Science*, 2019.
- Karen Larson, Loukas Zagkos, Mark Mc Auley, Jason Roberts, Nikos I. Kaval-laris, and Anastasios Matzavinos. Data-driven selection and parameter estima-tion for DNA methylation and mathematical models. *Journal of Theoretical Biology*, 2019.

PEER-REVIEWED BOOK CHAPTERS

- Clark Bowman, Karen Larson, Alexander Roitershtein, Derek Stein, and Anas-tasios Matzavinos. Bayesian uncertainty quantification for particle-based sim-

ulation of lipid bilayer membranes. In *Cell Movement: Modeling and Applications*, eds. M. Stolarska and N. Tarfulea. Springer, 2018.

PAPERS UNDER REVIEW

- Karen Larson, Sarah D. Olson, and Anastasios Matzavinos, Bayesian uncertainty quantification for micro-swimmers with fully resolved hydrodynamics.
- Karen Larson, Clark Bowman, Zhizhong Chen, Panagiotis Hadjidoukas, Costas Papadimitriou, Petros Koumoutsakos, and Anastasios Matzavinos. Data-driven prediction and origin identification of epidemics in population networks.

TEACHING EXPERIENCE

- Co-instructor, APMA 1650: Statistical Inference I, Summer 2017
- Teaching Assistant, APMA 1650: Statistical Inference I, Spring 2016
- Teaching Assistant, APMA 1650: Statistical Inference I, Fall 2015

HONORS AND AWARDS

- 2020, SIAM Student Travel Award for 2020 SIAM Conference on Uncertainty Quantification
- 2018, SIAM Student Chapter Certificate of Recognition
- 2018, Landahl Travel Grant for 2018 SMB Annual Meeting

- 2018, SIAM Student Travel Award for 2018 SIAM Conference on Uncertainty Quantification

PRESENTATIONS

- Invited Talk: *Data-driven prediction and origin identification of epidemics in population networks*, Trinity College, Hartford, CT, January 2020
- Invited Talk: *Bayesian Uncertainty Quantification of Biological Networks* SIAM Conference on Dynamical Systems, Snowbird, UT, May 2017
- Contributed Talk: *Data-driven prediction and origin identification of epidemics in population networks*, SIAM Conference on Uncertainty Quantification, Garching, Germany, March 2020
- Contributed Talk: *Data-driven prediction and origin identification of epidemics in population networks*, Conference on Multiscale Modeling in Biology, Minneapolis, MN, May 2019
- Contributed Talk: *Uncertainty Quantification and Biological Models*, Applied Math Graduate Student Seminar, Brown University, Providence, RI, Oct 2018
- Contributed Talk: *Data-driven prediction and origin identification of epidemics in population networks*, 2018 Annual Meeting of the Society for Mathematical Biology, Sydney, Australia, July 2018
- Contributed Talk: *Bayesian Uncertainty Quantification for Epidemic Spread on Networks*, SIAM Conference on Uncertainty Quantification, Garden Grove, CA, April 2018
- Contributed Talk: *Origin identification and uncertainty quantification for epidemic spread on networks*, Applied Math Days, RPI, Troy, NY, April 2018

- Contributed Talk: *Bayesian Uncertainty Quantification of Dynamic Processes in Networks*, SIAM Annual Meeting, Pittsburgh, PA, July 2017
- Poster: *Bayesian Uncertainty Quantification of Biological Networks*, Statistical Perspectives of Uncertainty Quantification, Atlanta, GA, May 2017
- Contributed Talk: *Uncertainty Quantification in Blood Flow*, Applied Math Days, RPI, Troy, NY, April 2017
- Invited Talk: *DPD Simulation of a Membrane*, Math SLAM at Brown University, Providence, RI, Nov. 1, 2015

PROFESSIONAL ORGANIZATIONS

- Society for Mathematical Biology
- Society for Industrial and Applied Mathematics
- Association for Women in Mathematics

Acknowledgments

To my advisor Anastasios Matzavinos, I would like to thank you for your advice, guidance, and mentorship. It has been a pleasure discussing project ideas and new interesting applications. None of this work would have been completed without your guidance.

I would like to thank Professor Sarah Olson for taking the time to serve on my thesis committee and our collaboration on exploring the dynamics of micro-swimmers. I also wish to thank Professor Martin Maxey for his time and effort serving on my committee.

Thank you to Panagiotis Hadjidoukas and Costas Papadimitriou. Our meetings were invaluable for my research and I appreciate your patience and time when working on our projects. I also wish to thank Petros Koumoutsakos for his insights.

Thank you to my collaborators at the University of Chester: Loukas Zagkos, Mark Mc Auley, Jason Roberts and Nikos Kavallaris. Through our discussions, I learned a lot about DNA methylation. To Clark Bowman, thank you for your input and suggestions. I have enjoyed collaborating with you over the past few years.

I am fortunate to have made such great connections with so many people throughout my six years at Brown. Alexis, Anna, and Steph, thank you for your friendship and encouragement throughout these years. I honestly wouldn't have made it

through grad school without you all. Amanda, thank you for your mentorship; I have always valued your advice and time. Thank you to Ross and Rebecca for watching Star Trek with me when the code is taking too long to run. Anna G., Becky, Emily, Justin, Marshall, Pooja, and Rebecca thank you for time you have spent organizing events for the department. Thank you to Katie, Angela, Jack, Jeremy, Nate, Taco, Toby, Tony, Susan, and Rich for making graduate school a little bit more fun. I cannot thank Ansel enough for his support throughout this process.

Lastly, I want to thank my parents Jon and Susan and brother Kyle for their love, support, and tea supplies. Thank you for encouraging me to pursue mathematics and never trying to make me fit into any boxes.

This thesis is based upon work funded by the National Science Foundation under Grant Nos. DMS-1521266 and DMS-1552903. The simulations were conducted using computational resources and services at Brown University's Center for Computation and Visualization.

Contents

Vita	iv
Preface and Acknowledgments	viii
1 Introduction	1
2 Bayesian Uncertainty Quantification Methods	6
2.1 Bayesian Uncertainty Quantification	7
2.2 Markov Chain Monte Carlo (MCMC)	11
2.3 Transitional Markov Chain Monte Carlo (TMCMC)	15
3 Detection of arterial wall abnormalities via Bayesian model selection	21
3.1 Introduction	22
3.2 Nonlinear One-Dimensional Blood Flow Model	24
3.3 Results	27
3.3.1 Parameter Estimation	30
3.3.2 Locating Structural Flaws with Model Selection	32
3.3.3 Model Selection for Cross-Sectional Area	35
3.3.4 Locating Defects with Misspecified Models	37
3.4 Discussion	41
4 Data-driven prediction and origin identification of epidemics in population networks	45
4.1 Introduction	46
4.2 SIR model	47
4.2.1 Single population model	48
4.2.2 Epidemic model on graphs	49
4.3 Results	50
4.3.1 Network 1: the 20-barbell graph	51
4.3.2 Network 2: the three-group network	60
4.4 Discussion	69

5	Data-driven selection and parameter estimation for DNA methylation mathematical models	71
5.1	Introduction	72
5.2	Models and Methods	75
5.2.1	DNA methylation models	75
5.3	Results	80
5.3.1	Parameter Estimation	81
5.3.2	Model Selection	83
5.3.3	Two Standard Deviation Added-Noise and Nonlinear Model	86
5.3.4	All Parameters	91
5.4	Discussion	95
6	Bayesian uncertainty quantification for micro-swimmers with fully resolved hydrodynamics	99
6.1	Introduction	100
6.2	Micro-swimmer Model	103
6.2.1	Fluid Equations	103
6.3	Results	107
6.3.1	Single Micro-Swimmer	108
6.3.2	Single Swimmer and Stokes Flow	114
6.3.3	Two Micro-Swimmers	115
6.3.4	Two Micro-Swimmers, Two Standard Deviations	120
6.4	Discussion and Conclusions	122
7	Conclusion	130
7.1	Summary of Contributions	131
7.2	Future Directions	134
7.2.1	Classification of Patients	134
7.2.2	Graphical Model Estimation	136

List of Tables

3.1	Posterior means and uncertainties for parameter estimation on the 19-artery network for three cases O_i (noise level $\sigma = 0.01\eta$).	32
3.2	Numerical results for identification of a $\beta_6 = 0.5$ aneurysm using noisy data from the ends of arteries 1 and 7 (noise level $\sigma = 0.01\eta$).	34
3.3	Numerical results for identification of a $\beta_6 = 0.5$ aneurysm using data from the ends of arteries 1, 7, and 13 (noise level $\sigma = 0.01\eta$).	34
3.4	Numerical results for identification of a $\beta_6 = 0.5$ aneurysm using data from the ends of arteries 1, 7, 10, and 13 (noise level $\sigma = 0.01\eta$).	35
3.5	Numerical results for area-based identification of an $\alpha_6 = 1.5$ aneurysm using data from the ends of arteries 1, 7, 10, and 13 (noise level $\sigma = 0.01\eta$).	36
3.6	Numerical results for area-based identification of an $\alpha_6 = 0.5$ stenosis using data from the ends of arteries 1, 7, 10, and 13 (noise level $\sigma = 0.01\eta$).	37
3.7	Numerical results for area-based identification of an $\alpha_6 = 0.5$ stenosis using data from the ends of arteries 1, 7, 10, and 13 (noise level $\sigma = 0.05\eta$).	37
3.8	Numerical results for area-based identification of an $\alpha_6 = 0.8$ stenosis using data from the ends of arteries 1, 7, 10, and 13 (noise level $\sigma = 0.01\eta$).	38
3.9	Numerical results for area-based identification of an $\alpha_6 = 1.5$ aneurysm using data from the ends of arteries 1, 7, 10, and 13 (noise level $\sigma = 0.01\eta$) with misspecified cross-sectional areas (perturbed with noise level $\sigma_\alpha = 0.01$).	39
3.10	Numerical results for area-based identification of an $\alpha_6 = 0.5$ stenosis using data from the ends of arteries 1, 7, 10, and 13 (noise level $\sigma = 0.01\eta$) with misspecified cross-sectional areas (perturbed with noise level $\sigma_\alpha = 0.01$).	40
3.11	Numerical results for area-based identification of an $\alpha_6 = 0.5$ stenosis using data from the ends of arteries 1, 7, 10, and 13 (noise level $\sigma = 0.05\eta$) with misspecified model parameters (perturbed with noise level $\sigma_\alpha = 0.05$).	40
3.12	Numerical results for area-based identification of an $\alpha_6 = 0.8$ stenosis using data from the ends of arteries 1, 7, 10, and 13 (noise level $\sigma = 0.01\eta$) with misspecified model parameters (perturbed with noise level $\sigma_\alpha = 0.01$).	41
4.1	Numerical results for estimation of β , γ , T and σ for the 20-barbell graph	53

4.2	Numerical results for estimation of β , γ , T and σ for the perturbed 20-barbell graph	59
4.3	Numerical results for estimation of β , γ , and T for the three group network with $\sigma = 0.01\alpha$	61
4.4	Numerical results for estimation of β , γ , S_0 , I_0 , and R_0 for the three group network with $\sigma = 0.01\alpha$	65
4.5	Subset of model selection results for the three-group network.	67
4.6	Subset of model selection results for the perturbed three-group network.	68
5.1	Subset of model selection results for DNA methylation models.	86
5.2	Model selection results for estimating four most sensitive transition rates for two model scenarios and two observed data sets for 5% added noise.	91
5.3	Model selection results for estimating all transition rates added noise for two model scenarios and two observed data sets for 1% added noise.	93
5.4	Model selection results for estimating all transition rates added noise for two model scenarios and two observed data sets for 5% added noise.	95
6.1	Posterior means and uncertainties for parameter estimation of resistance parameter α and noise level σ/β for the single swimmer using velocity data with varying noise levels	110
6.2	Posterior means and uncertainties for parameter estimation on resistance α and tail curvature stiffness K_C for the single swimmer using velocity data for varying noise levels.	113
6.3	Posterior means and uncertainties for parameter estimation of resistance α for two swimmers	117
6.4	Posterior means and uncertainties for parameter estimation on resistance α and curvature stiffness K_C for two swimmers	119
6.5	Posterior means and uncertainties for parameter estimation of resistance α and curvature stiffness K_C for two swimmers using two different noise levels	121
6.6	Parameters for swimmer model.	128
6.7	L_2 error between the observations and model outputs for a sample of α and K_C pairs.	129
7.1	Posterior means and uncertainties for parameter estimation on the three-node graph for three cases for the number of edges (noise level $\sigma = 0.01\alpha$).	137

List of Figures

2.1	Example of the Metropolis algorithm	13
2.2	Example of Simulated annealing	15
2.3	Implementation of TMCMC	20
3.1	Schematic of one-dimensional artery segment.	24
3.2	Schematic of Y-bifurcation in an arterial network.	26
3.3	Schematic of arterial network	27
3.4	Inflow boundary condition	28
3.5	Parameter estimation for blood viscosity μ , arterial stiffness β_2 , and noise level σ	31
3.6	Samples from eight generations of TMCMC	43
4.1	The 20-barbell graph	52
4.2	Parameter estimation results for the 20-barbell graph with infection rate $\beta = 0.02$, rate of removal $\gamma = 0.3$, noise level $\sigma = 0.01\alpha$, and time step $\Delta t = 0.005$	54
4.3	Time evolution of the susceptible, infective, removed, and total populations at nodes 3, 20, 21, and 27 of the 20-barbell graph	55
4.4	Parameter estimation results for the 20-barbell graph with infection rate $\beta = 0.02$, rate of removal $\gamma = 0.3$, time step $\Delta t = 0.005$, and noise level $\sigma = 0.01\alpha$	57
4.5	Parameter estimation results for the perturbed 20-barbell graph with infection rate $\beta = 0.02$, rate of removal $\gamma = 0.3$, time step $\Delta t = 0.005$, and noise level $\sigma = 0.01\alpha$ at time $T = 3$ using observations from nodes 3 and 24	59
4.6	The three-group network	60
4.7	Parameter estimation results for the three group network	62
4.8	Additional parameter estimation results for the three group network	64
4.9	Origin identification model evidence by node, three-group network	67
5.1	The three different states of a CpG dyad	76
5.2	Diagram for the methylation rates	76
5.3	Parameter estimation results using reference data from the model where $k_3 = k_{31} + k_{32}x_1^\gamma$ with a time history from $T = 0$ to $T = 10$	83
5.4	Parameter estimation results using reference data from model where $k_3 = k_{31} + k_{32}x_2^\gamma$ with a time history from $T = 0$ to $T = 10$	84
5.5	Parameter estimation results using the “incorrect” model on reference data, i.e. the model $k_3(x_1)$ on noisy data generated from $k_3(x_2)$ and similarly for model $k_3(x_2)$	85

5.6	Parameter estimation results for model $M_{1,3}$ using reference data $O_{1,3}$ with 5% added noise	88
5.7	Parameter estimation results for model M_2 using reference data O_2 with 5% added noise.	89
5.8	Parameter estimation results using a different model than the one that generated the reference data	90
5.9	Parameter estimation results for model $M_{1,3}$ using reference data $O_{1,3}$ with 1% added noise.	92
5.10	Parameter estimation results for model M_2 using reference data O_2 with 1% added noise.	93
5.11	Parameter estimation results $M_{1,3}$ using reference data $O_{1,3}$ with 5% added noise.	94
6.1	Representative computational model results for the micro-swimmer.	109
6.2	Parameter estimation results for a single swimmer with fluid resistance $\alpha = 1.0$ and tail curvature stiffness $K_C = 8.0$	111
6.3	Parameter estimation results for a single swimmer with fluid resistance $\alpha = 0$ (corresponding to Stokes flow), tail curvature stiffness $K_C = 8.0$, and tail spring stiffness $K_T = 10000$	115
6.4	Representative simulation results for two micro-swimmers	116
6.5	Parameter estimation results for two swimmers with fluid resistance $\alpha = 1.0$ and tail curvature stiffness $K_C = 8.0$	118
6.6	Parameter estimation results for two swimmers initially 0.5 apart with fluid resistance $\alpha = 1.0$ and curvature stiffness of tail $K_C = 8.0$ using two different noise levels for the tails and the heads	121
7.1	Parameter estimation results for the three stiffnesses β for the common femoral (CF), superficial femoral (SFA), and popliteal arteries and noise level σ	135
7.2	Three possible configurations for a three node graph.	137
7.3	The edge graph parameter estimation results for the three transition rates between the three nodes on the network $r_{1,2}$, $r_{1,3}$ $r_{2,3}$ and noise level σ	138
7.4	Two edge graph parameter estimation results for the three transition rates between the three nodes on the network $r_{1,2}$, $r_{1,3}$ $r_{2,3}$ and noise level σ	139
7.5	Single edge graph parameter estimation results for the three transition rates between the three nodes on the network $r_{1,2}$, $r_{1,3}$ $r_{2,3}$ and noise level σ	140

CHAPTER ONE

Introduction

Mathematical models are vital to the study of complex systems such as those in systems biology, including systems pathophysiology (1, 2, 3), epidemics (4, 5, 6), molecular biology (7, 8, 9), and cell mechanics (10, 11, 12). The ground truth is often unknown, and the sheer number of components may prove computational and analytic approaches to be infeasible, even when the underlying mechanics are well understood. In these scenarios, the comparative simplicity of models affords tractability at the cost of introducing uncertainty.

The most basic question that can be asked of a model is the forward problem: given a particular input, what output will be produced from the system? Supposing the model is deterministic, this question may seem straightforward – the model must simply be applied and the output observed. Yet there are many potential sources of uncertainty: the output may be subject to measurement error, or the model may depend on parameters which are themselves uncertain (see, e.g., (13, 14)), thereby propagating uncertainty forward. In the case where the model is itself stochastic (e.g., a branching process model of a disease), uncertainty may be a fundamental aspect of the system dynamics (15, 16). The broad field which describes the mathematical treatment of models in the presence of uncertainty is often referred to as *uncertainty quantification* (UQ) (17, 18, 19).

The mathematical treatment of uncertainty in this context may be trivial – for example, accounting for a measurement error in the output (e.g., a Gaussian noise term) may be as simple as executing the forward problem, then adding a Gaussian random variable. In this case, the forward problem is solved only once, so the resulting distribution is easily sampled from, even when the model itself has considerable complexity.

Conversely, many other problems become significantly more difficult when uncer-

tainty is introduced (16). The classical inverse problem is one such example: what input (or set of inputs) yields a particular output? Especially if the model is complex, the model itself is unlikely to be invertible; moreover, depending on the form of the uncertainty, the set of inputs which must be considered can be prohibitively large (20). To illustrate, consider that in the previous example of additive Gaussian noise, which is among the simplest models of measurement uncertainty, the entire input space could feasibly have produced any particular output, so long as the Gaussian noise happened to be exactly the difference between the true output and the target output. Even if the input space is finite, the number of times the model must be applied makes this “solution” impractical for all but the simplest and most efficient models.

The Bayesian approach to this parameter estimation problem, i.e., Bayesian uncertainty quantification, attempts to leverage tools from Bayesian statistics to make the problem more tractable (21, 14). There are a variety of approaches to sampling from high-dimensional and complex distributions in the Bayesian setting (22). As the goal of parameter estimation is the probability distribution corresponding to the relative likelihood of particular parameter values generating the observed output, this machinery is readily applied.

In this work, in Chapter 2 we introduce a particular recent approach to Bayesian uncertainty quantification which is practical even for comparatively expensive models, such as those necessary to study the dynamics of biological cells in atomistic detail. Rather than running the forward problem for a prohibitively large set of parameters, the algorithm uses a Markov chain Monte Carlo approach which evaluates the problem only at a discrete set of points. More importantly, the specific algorithm used, known as transitional Markov chain Monte Carlo (TMCMC), is massively parallelizable, and so the high cost of the model can be offset by the ability to execute

many copies of the model simultaneously (23, 24, 25, 17).

The main goal of this thesis is to explore various biological systems through Bayesian uncertainty quantification and numerical methods. In Chapters 3 - 6, we describe biological systems of interest and demonstrate the efficacy of using these Bayesian tools for estimating parameters and finding the best model to describe the observed data. Specifically, Chapter 3 studies an arterial blood flow network with aims to discover where abnormalities, such as aneurysms and stenoses, occur in the network (26). We use parameter estimation to identify the severity of these flaws and model selection to locate the abnormalities in the network.

Chapter 4 looks at epidemic spread on networks of populations (24). We investigate recovering parameters related to the infectiousness of the disease and consider identifying where the epidemic originated from. In addition, we consider cases where the population network is misspecified to discover whether the model parameters are still recoverable from the noisy observed data.

In Chapter 5, we apply the Bayesian methods to DNA methylation models in order to advance our understandings of the underlying dynamics (27). There is evidence that DNA methylation is related to healthspan and aging, and aberrant DNA methylation patterns are a hallmark of diseases such as cancer and Alzheimer's. We compare two prospective models for DNA methylation, which account for different dependencies on how the DNA transitioned to different states of methylation.

We perform parameter estimation on micro-swimmers in a viscous fluid in Chapter 6 (28). The models for micro-swimmers can be highly computationally complex, so parameter estimation has been prohibitively expensive. However, due to the parallel nature of our Bayesian framework, we can use it to better understand emer-

gent trajectories and swimming speeds of micro-swimmers. The Bayesian estimation method allows us to leverage our prior knowledge of micro-swimmers and the viscous fluid to inform the search of the parameter space and make inferences about how the material properties of the system are related. The result of this experiment, and those from Chapters 3 - 6 illustrate the feasibility of the framework for future applications to a variety of complex biological systems.

Chapter 7 summarizes the above results and introduces some future directions that we have considered for this work. For example, combining real medical data, the arterial blood flow model, and the framework to classify patients as either healthy or sick. We also describe a method for estimating the structure of a population network, i.e. how are the communities connected and what are the transition rates between the populations.

CHAPTER TWO

Bayesian Uncertainty Quantification Methods

2.1 Bayesian Uncertainty Quantification

To approach the parameter estimation and model selection problems using a Bayesian methodology such as transitional Markov chain Monte Carlo (TMCMC), we must first formally introduce the Bayesian formulation of uncertainty quantification (21). The hallmark of Bayesian statistics is the assumption of prior distributions for unknown parameters – in other words, probabilities are a representation of degrees of belief; any knowledge of the unknowns *a priori*, whether from previous experience or from physical limitations, informs belief in the parameter values even before a statistical model is applied (20).

In this work, we focus on two different, but related problems in uncertainty quantification. The first problem we study is parameter estimation, i.e. what set of parameters could generate observed data. An example of this problem would be trying to estimate the severity of an aneurysm in an arterial network. The second problem is model selection, i.e. what mathematical model best describes these observed data. An example of model selection is discovering which artery has an aneurysm (if any). These two problems are closely related to each other, as described in the problem setup below.

In the context of parameter estimation and model selection, the unknown parameters are the inputs to some mathematical model M (including not only model parameters, but also initial conditions, boundary conditions, etc.), which then produces a set of output quantities governed by the form of the model. A simple bacterial growth model, for example, might be parameterized by the probability p that a particular bacterium divides in a certain time period, but would also take as input the number X_0 of bacteria present at time 0. The Bayesian assumption is that the n

inputs, jointly parameterized by $\boldsymbol{\theta} \in \mathbb{R}^n$, are random variables with the multivariate, joint prior density $\pi(\boldsymbol{\theta}|M)$.

Given a particular set of inputs $\boldsymbol{\theta}$, the model M produces a set of m data, referred to jointly as $\mathbf{D} \in \mathbb{R}^m$, according to some likelihood $p(\mathbf{D}|\boldsymbol{\theta}, M)$ – this is the content of the model itself, i.e., the mapping from inputs to outputs, which is probabilistic owing to uncertainties and stochasticity in the model M . The parameter estimation problem instead targets the posterior $p(\boldsymbol{\theta}|\mathbf{D}, M)$, i.e., the mapping from the outputs back to the inputs, which is also informed by the prior beliefs $\pi(\boldsymbol{\theta}|M)$. In this setting, $\mathbf{D}$ is known and $\boldsymbol{\theta}$ is unknown, and so it should be stressed that $p(\mathbf{D}|\boldsymbol{\theta}, M)$ is purposefully referred to as a likelihood since it is *not* a distribution in $\boldsymbol{\theta}$.

Using Bayes' theorem, the posterior distribution can be computed as

$$p(\boldsymbol{\theta}|\mathbf{D}, M) = \frac{p(\mathbf{D}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)}{\rho(\mathbf{D}|M)}, \quad (2.1.1)$$

which additionally introduces the evidence $\rho(\mathbf{D}|M)$ of the model class, given by the multi-dimensional integral

$$\rho(\mathbf{D}|M) = \int_{\mathbb{R}^n} p(\mathbf{D}|\boldsymbol{\theta}, M)\pi(\boldsymbol{\theta}|M)d\boldsymbol{\theta}.$$

For the parameter estimation problem, it should be noted that ρ does not depend on $\boldsymbol{\theta}$, and so its role is simply to function as a normalization constant for the posterior. However, the evidence is essential for the model selection problem.

This scenario can be extended to the case case where the model M is one particular model in a parameterized class of models $\mathcal{M}$, the evidence $\rho(\mathbf{D}|M)$ functions as a measure of the degree to which the model matches the observed data, thereby

playing a critical role in model selection (29, 30). Let $Pr(M_i)$ be a prior distribution on models in the class $\mathcal{M}$. The posterior $Pr(M_i|\mathbf{D})$ can again be derived from Bayes' theorem:

$$Pr(M_i|\mathbf{D}) = \frac{\rho(\mathbf{D}|M_i)Pr(M_i)}{p(\mathbf{D}|\mathcal{M})}, \quad (2.1.2)$$

where $p(\mathbf{D}|\mathcal{M}) = \sum_i \rho(\mathbf{D}|M_i)$ is a normalization constant. Intuitively, $Pr(M_i|\mathbf{D})$ is a distribution which describes the probability of the data $\mathbf{D}$ having been generated from model M_i (as opposed to another model M_j) under the assumption that at least one model in $\mathcal{M}$ is the true model, i.e., was actually used to generate the data. If a uniform prior is assumed on models, this posterior is directly proportional to the evidence $\rho(\mathbf{D}|M_i)$, and so model selection is “free” when the evidence is already calculated for parameter estimation (29, 17, 31, 24).

To proceed further, assumptions must be made about the specific form of the likelihood $p(\mathbf{D}|\boldsymbol{\theta}, M)$. Here, we will assume that the measured data $\mathbf{D}$ are generated according to the following model prediction equation:

$$\mathbf{D} = \mathbf{g}(\boldsymbol{\theta}|M) + \mathbf{e}, \quad (2.1.3)$$

where $\mathbf{g}(\boldsymbol{\theta}|M)$ is the predicted output for a specific set of parameters $\boldsymbol{\theta}$ and $\mathbf{e}$ is a prediction error due to, e.g., measurement, computational, or modeling errors. Specifically, we will make the simplifying assumption that the error $\mathbf{e}$ is normally distributed with zero mean and covariance matrix $\boldsymbol{\Sigma}$, which may involve additional unknown parameters. Since $\mathbf{g}$ is assumed deterministic, it follows that $\mathbf{D}$ is also normally distributed, and so the likelihood $p(\mathbf{D}|\boldsymbol{\theta}, M)$ of the observed data can be

written as

$$p(\mathbf{D}|\boldsymbol{\theta}, M) = \frac{|\boldsymbol{\Sigma}(\boldsymbol{\theta})|^{-1/2}}{(2\pi)^{m/2}} \exp\left[-\frac{1}{2}J(\boldsymbol{\theta}, \mathbf{D}|M)\right], \quad (2.1.4)$$

where

$$J(\boldsymbol{\theta}, \mathbf{D}|M) = [\mathbf{D} - \mathbf{g}(\boldsymbol{\theta}|M)]^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}) [\mathbf{D} - \mathbf{g}(\boldsymbol{\theta}|M)] \quad (2.1.5)$$

is the weighted measure of fit between the model predictions and the measured data, $|\cdot|$ denotes determinant, and the parameter set $\boldsymbol{\theta}$ is augmented to include any unknown parameters in the covariance matrix $\boldsymbol{\Sigma}$ (29, 32). It follows that sets of parameters $\boldsymbol{\theta}$ that reduce $J(\boldsymbol{\theta}, \mathbf{D}|M)$, often called the “fitness” of the parameter set $\boldsymbol{\theta}$ to the data $\mathbf{D}$, generally increase the likelihood; thus, the most likely sets of parameters (before consideration of the prior $\pi(\boldsymbol{\theta}|M)$) can be intuitively recognized as those whose output most closely matches the observed data.

While this explicit form of the likelihood can be straightforwardly calculated for any chosen set of parameters $\boldsymbol{\theta}$, each evaluation must reference the model prediction function $\mathbf{g}$, which may be computationally intensive (e.g., a particle simulation). Since the parameters are assumed to be real-valued and the noise $\mathbf{e}$ spans the entire output space, computing the complete posterior $p(\boldsymbol{\theta}|\mathbf{D}, M)$ would then require an infinite number of evaluations of $\mathbf{g}$. By formulating the problem in a Bayesian context, however, we gain access to the Bayesian statistician’s toolkit, which contains a number of methods for approximately sampling such distributions: in particular, the class of Monte Carlo algorithms known as Markov chain Monte Carlo (MCMC).

2.2 Markov Chain Monte Carlo (MCMC)

As we have defined it, the major goal of parameter estimation and model selection is the recovery of the posterior distribution $p(\boldsymbol{\theta}|\mathbf{D}, M)$. Via (2.1.1), there are three distributions which must be known in order to achieve this goal. The first is the prior $\pi(\boldsymbol{\theta}|M)$ associated with the parameters $\boldsymbol{\theta}$ in a given model class – tautologically, this is “easy,” since it is (by definition) prior knowledge. The second component is the evidence $\rho(\mathbf{D}|M)$, but, as ρ does not depend on $\boldsymbol{\theta}$, it simply functions as a normalizing constant for parameter estimation. However, we need to compute it for model selection. The third component is more difficult: the likelihood $p(\mathbf{D}|\boldsymbol{\theta}, M)$ must also be known. As has been mentioned, the explicit form of the likelihood can only be poked at by observing specific realizations of the forward problem.

Given infinite time, it would be possible to recover the likelihood by evaluating the model prediction function $\mathbf{g}$ for every valid set of parameters $\boldsymbol{\theta}$. To approach the problem in a feasible finite time, we instead turn to a set of stochastic sampling methods which have become omnipresent as computers have become more efficient: Markov chain Monte Carlo, often referred to as MCMC (21, 20). Such methods rely on constructing Markov chains, i.e. a series of random samples where the probability of the next sample depends solely on the current sample, which have as their invariant distribution a specific distribution of interest (in the context of the parameter estimation problem, the posterior $p(\boldsymbol{\theta}|\mathbf{D}, M)$). There are a number of practical issues which surround the application of MCMC to parameter estimation: to understand the nature of these difficulties, and the ways in which they are addressed in the context of Bayesian UQ, we briefly review a subset of relevant MCMC algorithms.

The most fundamental such algorithm is the Metropolis algorithm (21). Let $p(\cdot)$

be a target distribution defined on a state space S , and suppose a function $f \propto p$ is known (this is not as absurd an assumption as it might sound at first glance – in the Bayesian framework, if p is a posterior distribution, then a likelihood function can serve the role of f). Begin with any initial guess $X_0 \in S$. The Metropolis algorithm functions by iteratively calculating a new value X_{k+1} given the current value X_k in such a way that, as $k \rightarrow \infty$, the distribution of X_k approaches p . In this way, simply calculating the values X_k up to some large k produces approximate samples from p .

The specific method for updating the guesses X_k is to make a proposal Y , chosen at random according to a symmetric proposal distribution $q(\cdot, X_k)$ (e.g., a multivariate Gaussian centered at X_k). The algorithm then chooses whether to accept the proposal (in which case $X_{k+1} = Y$) or reject it (in which case $X_{k+1} = X_k$). By choosing the acceptance probability to be the ratio $f(Y)/f(X_k)$, the invariant distribution of the chain becomes the target distribution p . Convergence to the stationary distribution can be proven easily if q is chosen carefully – in the case of a Gaussian, its infinite support makes possible the transition to any state x with $p(x) > 0$ (irreducibility), while the possibility of rejecting a proposal ensures aperiodicity (33).

Even this basic algorithm appears to have ready applications to the parameter estimation problem. Choosing the invariant distribution p to be the posterior $p(\boldsymbol{\theta}|\mathbf{D}, M)$, along with a standard proposal distribution q (a multivariate Gaussian will work for real-valued parameters), the only difficulty seems to be the calculation of some $f \propto p$ – this is the same problem as was encountered initially, where the likelihood function is not known, and it is not computationally feasible to estimate the likelihood over the entire parameter space. The key is that the Metropolis algorithm does not rely on the general form of f , but instead only relies on the calculation of f at various points chosen according to the proposal distribution q . Supposing the algorithm provides a guess Y for the set of parameters, the likelihood

$f(Y) = p(\mathbf{D}|Y, M)$ can be estimated as in (2.1.4), assuming output is generated according to the model prediction equation (2.1.3).

The main difficulty of this approach has been alluded to already – the convergence to the stationary distribution is asymptotic, and the time to convergence can vary greatly (depending on q , the dimensionality, and the form of the likelihood). Specifically, the algorithm may get “stuck” sampling from a region where the fit appears to be locally optimal for long times, as the probability of transitioning to other regions of high density may be extremely low (this is known as a probabilistic *bottleneck*). This problem can be seen in Figure 2.1, which uses the Metropolis algorithm to approach a simple toy problem in one dimension. The result is that approximate sampling from the posterior $p(\boldsymbol{\theta}|\mathbf{D}, M)$ may require running the chain for longer than is computationally feasible.

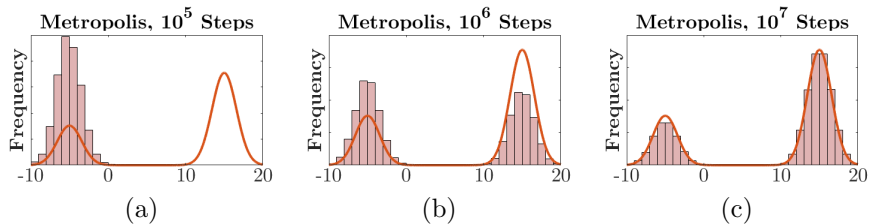


Figure 2.1: The Metropolis algorithm applied to the density $p(x) \propto 0.3e^{-0.2(x+5)^2} + 0.7e^{-0.2(x-15)^2}$ (solid curve). The proposal distribution used is a Gaussian with $\sigma = 4$. Shown are the most recent 10% of samples after the specific number of steps. (a) For short times, the chain gets “stuck” in the smaller peak, unable to cross the bottleneck to explore the full distribution. (b) For intermediate times, although the distribution is explored fully, the chain does not spend the correct proportion of time in each peak. (c) For large times, convergence to the invariant distribution is guaranteed via irreducibility and aperiodicity of the chain.

Naïvely, one way to address this problem would be to choose an extremely broad proposal distribution q , thereby allowing the chain to jump across bottlenecks in a single step. Unfortunately, this has the second effect of greatly increasing the number of proposals which are rejected. Instead, we borrow ideas from a second

MCMC method which is more robust to the problem of probabilistic bottlenecks. Simulated annealing is an approach to nonconvex optimization which introduces a *cooling schedule* to ensure that, in the short term, the chain can transition across regions where the likelihood may be extremely low (34, 35, 36, 33).

Simulated annealing bears a strong resemblance to the Metropolis algorithm; it functions by making iterative proposals with a carefully-chosen probability of acceptance. Unlike the Metropolis algorithm, the probability depends on the iteration k , such that the probability of transitioning to a region of low probability is much higher when k is small. The acceptance probability is often stated in terms of a temperature $T(k)$ which is monotonically decreasing, owing to its connections to the role of temperature in statistical mechanics (and begetting the term “cooling schedule”). The common explicit acceptance probability $\exp((\log f(Y) - \log f(X_k))/T(k))$ coincides with the Metropolis algorithm in the case that $T(k) = 1$, and so simulated annealing can be thought of as a generalization of Metropolis (34, 36, 20).

Since the transition probability depends on k , the chain $\{X_k\}$ no longer has an invariant distribution. Instead, each step of the chain has its own unique “invariant distribution” p_k . As k increases, $T(k)$ decreases, and if T falls to zero, the weak limit of the p_k should put its entire mass at the global maximum value of f (this is why simulated annealing is effective at optimization). Conversely, for small k , a large temperature greatly increases the probability of accepting a transition to a region where f is much smaller. If the cooling schedule is chosen to scale appropriately, it can thus be very effective at facilitating transitions across bottlenecks like that of Figure 2.1. A simple application of simulated annealing to the toy sampling problem can be seen in Figure 2.2.

The specific MCMC algorithm which will be applied to the Bayesian parameter

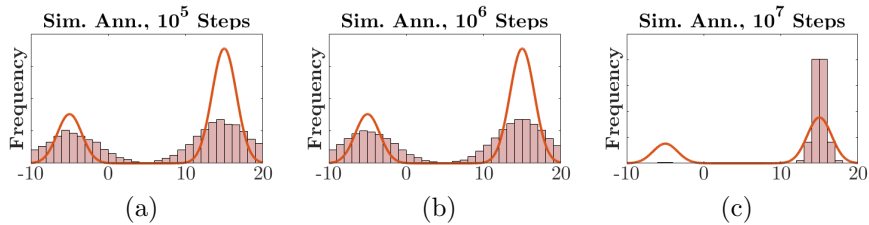


Figure 2.2: Simulated annealing with cooling schedule $T(k) = 5(1 - 10^{-7}k)$ for the toy problem of Figure 2.1. Again, the proposal distribution used is a zero-mean Gaussian with $\sigma = 4$. Shown are the most recent 10% of samples after the specific number of steps. (a, b) Even for short times, the high temperature smooths the target distribution, allowing sampling from both peaks. (c) As the temperature falls to zero, the distribution clusters around the global maximum.

estimation and model selection problems, known as transitional Markov chain Monte Carlo (TMCMC), makes use of the ideas central to both of these algorithms (17, 14). Each step of the chain uses the Metropolis algorithm to draw approximate samples from a target intermediate distribution, while the parent chain uses a cooling schedule to *transition* from the prior to the posterior, much as simulated annealing transitions from smoothed to peaked. The specifics of TMCMC, as well as its explicit implementation in the UQ framework, are detailed in Section 2.3.

2.3 Transitional Markov Chain Monte Carlo (TMCMC)

As mentioned in the previous section, MCMC algorithms are often used to generate approximate samples from high-dimensional and complex posterior distributions. While basic algorithms like Metropolis and simulated annealing will function in this regard, more efficient algorithms which improve the convergence rate are more useful in practice, since the convergence times of MCMC algorithms for some distributions (especially those with bottlenecks) can be quite long. Although there do exist more

efficient methods with proven convergence results (in particular, delayed rejection adaptive Metropolis (DRAM) (37) and differential evolution random sampling MC (DREAM) (38) are commonly used), we will use here the TMCMC algorithm, which has the computational advantage of being massively parallelizable due to its use of a large number of independent Markov chains (25, 17, 14).

Like simulated annealing, TMCMC attempts to avoid probabilistic bottlenecks by transitioning to the target distribution from a broad initial distribution (in this case, the prior $\pi(\boldsymbol{\theta}|M)$) that allows the chain to explore regions of low probability. To accomplish this, a series of intermediate probability densities are constructed iteratively:

$$f_j(\boldsymbol{\theta}) \sim [p(\mathbf{D}|\boldsymbol{\theta}, M)]^{q_j} \cdot \pi(\boldsymbol{\theta}|M), \quad j = 0, \dots, \lambda$$

$$0 = q_0 < q_1 < \dots < q_\lambda = 1.$$

Note that f_0 is simply the prior distribution, while f_λ is the posterior distribution (i.e., the product of the prior and likelihood, as in (2.1.1), up to the normalization constant $\rho(\mathbf{D}|M)$).

The basic idea of TMCMC is as follows: begin with a collection of samples from the prior distribution (which is easy, so long as the prior has a simple form such as a Gaussian or uniform distribution). Then, iteratively re-sample in such a way that the samples at each step j correspond to the intermediate density f_j . At the conclusion of the algorithm, the samples will be asymptotically distributed as f_λ (the posterior), with greater accuracy for small step sizes and large numbers of samples. During each iteration, the re-sampling procedure involves the use of a large number of independent Markov chains which are evolved using the Metropolis algorithm; since the chains are independent, the algorithm is easily parallelizable, and so the potentially expensive cost of evaluating the likelihood $p(\mathbf{D}|\boldsymbol{\theta}, M)$ is offset by the

ability to run multiple evaluations concurrently.

The explicit algorithm is summarized below as Algorithm 1:

Algorithm 1: TMCMC

```

1 Procedure TMCMC Ref. (17, 27)
2 BEGIN, SET  $j = 0, q_0 = 0$ 
3 Generate  $\{\boldsymbol{\theta}_{0,k}, k = 1, \dots, N_0\}$  from prior  $f_0(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|M)$  and compute
   likelihood  $p(\mathbf{D}|\boldsymbol{\theta}_{0,k}, M)$  for each sample.
4 loop:
5 while  $q_{j+1} \leq 1$  do
6   Analyze samples  $\{\theta_{j,k}, k = 1, \dots, N_j\}$  to determine  $q_{j+1}$ , weights
      $\bar{w}(\boldsymbol{\theta}_{j,k})$ , covariance  $\boldsymbol{\Sigma}_j$ , and estimator  $S_j$  of  $\mathbb{E}[w(\boldsymbol{\theta}_{j,k})]$ .
7   Resample based on samples available in stage  $j$  using the plausibility
     weights and the Metropolis algorithm in order to generate samples for
     stage  $j + 1$  and compute likelihood  $p(\mathbf{D}|\boldsymbol{\theta}_{j+1,k}, M)$  for each.
8   if  $q_{j+1} > 1$  then
9      $\perp$  BREAK,
10  else
11     $j = j + 1$ 
12    goto loop.
13  end
14 END

```

The algorithm begins by taking N_0 samples $\boldsymbol{\theta}_{0,k}$ (indexed first by generation, then by sample number) from the prior distribution $f_0(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|M)$. For each stage j of the algorithm, it uses the current set of samples to evaluate the plausibility weights $w(\boldsymbol{\theta}_{j,k})$ as

$$w(\boldsymbol{\theta}_{j,k}) = \frac{f_{j+1}(\boldsymbol{\theta}_{j,k})}{f_j(\boldsymbol{\theta}_{j,k})} = [p(\mathbf{D}|M, \boldsymbol{\theta}_{j,k})]^{q_{j+1}-q_j}.$$

Note that since the plausibility weights are calculated as the likelihood raised to a non-negative power, the observed data $\mathbf{D}$ is more likely to have come from sampled parameters $\boldsymbol{\theta}_{j,k}$ with higher plausibility weights (thus begetting the term). q_{j+1} has not yet been defined; the choice of q_{j+1} is what determines how smoothly the

prior distribution transitions to the posterior distribution (with a smoother transition coming at the cost of increased computational requirements). Recent literature suggests that q_{j+1} be chosen to make the covariance of the plausibility weights at stage j smaller than a tolerance covariance, recommending the specific value 1.0 (25, 17).

The algorithm then calculates the average S_j of the plausibility weights, the normalized plausibility weights $\bar{w}(\boldsymbol{\theta}_{j,k})$, and the scaled covariance $\boldsymbol{\Sigma}_j$ of the $\boldsymbol{\theta}_{j,k}$, which will be used to generate the next generation of samples $\boldsymbol{\theta}_{j+1,k}$:

$$S_j = \frac{1}{N_j} \sum_{k=1}^{N_j} w(\boldsymbol{\theta}_{j,k}),$$

$$\bar{w}(\boldsymbol{\theta}_{j,k}) = w(\boldsymbol{\theta}_{j,k}) / \sum_{k=1}^{N_j} w(\boldsymbol{\theta}_{j,k}) = w(\boldsymbol{\theta}_{j,k}) / (N_j S_j),$$

$$\boldsymbol{\Sigma}_j = \beta^2 \sum_{k=1}^{N_j} \bar{w}(\boldsymbol{\theta}_{j,k}) [\boldsymbol{\theta}_{j,k} - \boldsymbol{\mu}_j][\boldsymbol{\theta}_{j,k} - \boldsymbol{\mu}_j]^T.$$

Here, $\boldsymbol{\Sigma}_j$ is calculated using the sample mean $\boldsymbol{\mu}_j$ of the $\boldsymbol{\theta}_{j,k}$ and a scaling factor β .

As an aside, it is worth noting that the product of the S_j , i.e.,

$$S = \prod_{j=0}^{\lambda-1} S_j,$$

is an asymptotically unbiased estimator of the evidence $\rho(\mathbf{D}|M)$ (25), and so transitional Markov chain Monte Carlo is especially efficient for model selection (see Section 2.1).

The algorithm then independently generates N_{j+1} samples $\hat{\boldsymbol{\theta}}_{j+1,k}$ by choosing randomly among the current set of samples $\{\boldsymbol{\theta}_{j,k}\}$ to create leaders of Markov chains

$\hat{\boldsymbol{\theta}}_{j+1,\ell}$ such that $\hat{\boldsymbol{\theta}}_{j+1,\ell} = \boldsymbol{\theta}_{j,k}$ with probability $\overline{w}(\boldsymbol{\theta}_{j,k})$. Since the samples are chosen independently, there is the possibility of selecting the same sample from the current generation multiple times – call $n_{j+1,k}$ the number of times that $\boldsymbol{\theta}_{j,k}$ is chosen (which could range from 0 to N_{j+1}). Each unique sample is used as the starting point of an independent Markov chain whose behavior is governed by the Metropolis algorithm with target distribution f_j and a proposal distribution normally distributed with covariance $\boldsymbol{\Sigma}_j$ and centered at the current value. The scaling parameter β thus plays the role of controlling the rejection rate and jump size; recent literature recommends $\beta = 0.2$ (13, 25).

Finally, the samples $\boldsymbol{\theta}_{j+1,k}$ are drawn from the Markov chains, with each sample being the value obtained by stepping a chain forward once. $n_{j+1,k}$ samples are drawn from the chain which started at $\boldsymbol{\theta}_{j,k}$, yielding N_{j+1} samples total. At this point, the iteration j is complete, and the algorithm either moves forward to generation $j + 1$ or terminates if $q_{j+1} > 1$.

With the approach established, we apply the TMCMC approach to the Bayesian UQ setting in the context of several biological models on networks. In the following chapters, we describe five applications of using these methods to better understand the underlying mechanisms of biological networks.

Appendix: High-performance implementations

$\Pi4U$ (17) is a platform-agnostic task-based framework for uncertainty quantification that supports nested parallelism and automatic load balancing in large scale computing architectures. The software is open-source and includes HPC implementations

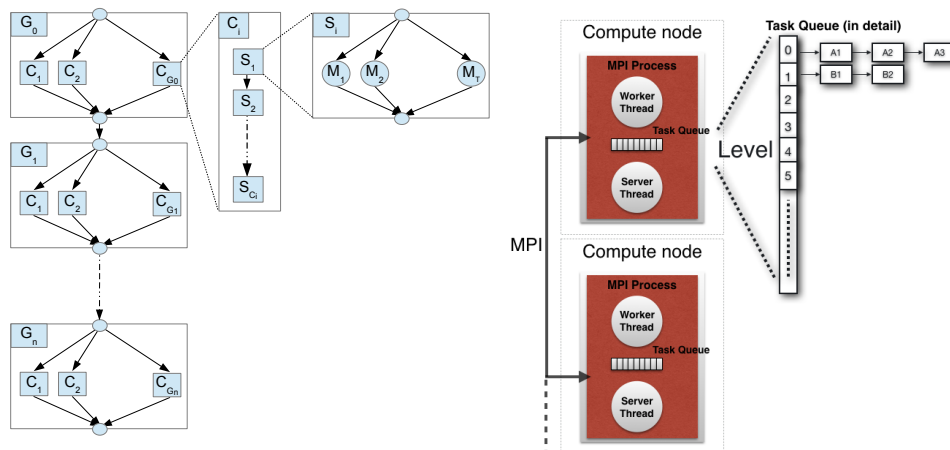


Figure 2.3: Task graph of the TMCMC algorithm (left) and parallel architecture of the TORC library (right).

for both multicore and GPU clusters of algorithms such as transitional Markov chain Monte Carlo (25, 13) and Approximate Bayesian Computational Subset-simulation (39). The irregular, dynamic and multi-level task-based parallelism of the algorithms (Fig 2.3, left) is expressed and fully exploited by means of the TORC runtime library (13). TORC is a software library for programming and running unaltered task-parallel programs on both shared and distributed memory platforms. TORC orchestrates the scheduling of function evaluations on the cluster nodes (Fig 2.3, right). The parallel framework includes multiple features, most prominently its inherent load balancing, fault-tolerance and high reusability. As a specific example, the TMCMC method within $\Pi 4U$ is able to achieve an overall parallel efficiency of more than 90% on 1024 compute nodes of Swiss supercomputer Piz Daint running hybrid MPI+GPU molecular simulation codes with highly variable time-to-solution between simulations with different interaction parameters.

CHAPTER THREE

Detection of arterial wall abnormalities via Bayesian model selection

3.1 Introduction

Mathematical models for hemodynamics trace back to the work of Euler, who described a one-dimensional treatment of blood flow through an arterial network with rigid tubes (40, 41); more sophisticated one-dimensional models are still used to study a variety of physio-pathological phenomena (42, 43, 44, 45, 46, 47, 48, 49). Computational advances have also allowed for the development of computationally intensive three-dimensional models (50, 51, 52, 53, 54, 55), which have been used to accurately simulate specific human arteries (e.g., the carotid arteries (56)) and model their material properties (e.g., of cerebral arterial walls (3)). There also exist multi-component models (57), which are amenable to applications such as modeling oxygen transport to solid tumors (1) and surgical tissue flaps (2, 58).

Despite the sophistication of these approaches, there remain a number of challenges in the creation of patient-specific models using individual medical data. In particular, computational expense usually limits arterial parameter estimation to the one-dimensional class of models (41, 51), which have nonetheless proven sufficiently robust to study fluid-structure interactions and viscoelasticity (48, 46) and create a patient-specific model for vascular bypass surgery (59). Several approaches exist for parameter estimation and uncertainty quantification for these models. Gradient descent has been used to estimate arterial compliance parameters (60), recovering single parameters assumed constant in space and time. Sensitivity analysis has also been used, successfully quantifying output sensitivity to various uncertainties in a stochastic flow network (61). More recently, computational methods based upon Bayesian optimization and multi-fidelity information fusion for model inversion have been explored (62).

The chief contribution of this work is to apply the Bayesian framework for uncertainty quantification from Chapter 2 to a bifurcating network of one-dimensional extensible arteries. The advantages of the approach are twofold. First, it utilizes transitional Markov chain Monte Carlo (TMCMC), a highly parallelizable algorithm for approximate sampling which allows practical uncertainty quantification even for large arterial networks (25, 63, 17); our high-performance implementation $\Pi 4U$ will be shown to simultaneously and efficiently estimate several unknown parameters in this setting. Second, the approach can practically be used for Bayesian model selection, allowing for evidence-based comparison between models with distinct physical assumptions. The approach thus represents a significant advance in fitting patient-specific hemodynamical flow models.

Specifically, we consider a branching network of 19 arteries in which a structural flaw (e.g., an aneurysm) has been introduced at an unknown location. Section 3.2 describes the one-dimensional blood flow model. In Section 3.3, we use the flawed model to simulate noisy observations of the flow velocity at fixed points in the network. We then use Bayesian model selection to probabilistically locate the defect within the network and accurately recover its structural properties, showing the approach to be effective even when parameters are corrupted with Gaussian noise. As the method readily admits clinical blood flow data, which have been shown to be measurable with non-invasive procedures (64, 65, 66, 67), it shows great potential in diagnosing patient-specific structural issues in the circulatory system.

3.2 Nonlinear One-Dimensional Blood Flow Model

We first introduce the one-dimensional blood flow model. While such models can be derived via a scaling of the Navier-Stokes equations for viscous flow (68), we use here the geometry- and conservation-motivated approach described by Sherwin et al. (41) and Peiró and Veneziani (51). In this approach, the viscous, incompressible flow is assumed to move only in the axial direction (i.e., along the one-dimensional artery), to exhibit axial symmetry, and to maintain constant internal pressure over orthogonal cross-sections. The artery is assumed to have low curvature and to be distensible in the radial direction. A schematic of the artery appears in Fig. 3.1.

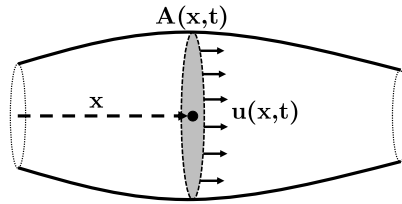


Figure 3.1: Schematic of one-dimensional artery segment.

The artery of constant length ℓ and position-dependent cross-sectional area $A(x, t)$ is filled with blood flowing at velocity $u(x, t)$ and with internal cross-sectional pressure $p(x, t)$, yielding the cross-sectional flux $Q(x, t) = A(x, t)u(x, t)$. Choosing u , A , and p as the independent variables, the partial differential equation governing the incompressible flow can be derived from conservation of mass and momentum:

$$\begin{cases} \frac{\partial A}{\partial t} + \frac{\partial(Au)}{\partial x} = 0, \\ \frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + \frac{1}{\rho} \frac{\partial p}{\partial x} + K_r \frac{u}{A} = 0, \end{cases} \quad (3.2.1)$$

where ρ is the flow density and K_r is a parameter representing viscous resistance per unit length, which is a function of the fluid viscosity μ (48, 47).

The system is closed using a constitutive law to relate pressure and area. Using the Laplace tube law and assuming that the arterial wall is purely elastic,

$$p = p_{ext} + \frac{\sqrt{\pi} E h}{(1 - \nu^2) \sqrt{A_0} \sqrt{A}} (\sqrt{A} - \sqrt{A_0}),$$

where p_{ext} is the external pressure, E is the Young's modulus of the wall, h is the wall thickness, A_0 is the relaxed cross-sectional area, and ν is the Poisson ratio, here taken to be $\frac{1}{2}$. For notational simplicity, we collect the coefficient into a single stiffness parameter B , yielding

$$p = p_{ext} + B(\sqrt{A} - \sqrt{A_0}).$$

Here, we make the assumption that $\frac{1}{\sqrt{A}}$ can be approximated by $\frac{1}{\sqrt{A_0}}$, so that B takes the form

$$B = \frac{\sqrt{\pi} E h}{(1 - \nu^2) A_0}.$$

Equation (3.2.1) can then be re-written in the form of a nonlinear hyperbolic conservation law (41):

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}}{\partial x} = \mathbf{S}(\mathbf{U}) \quad (3.2.2)$$

where

$$\mathbf{U} = \begin{bmatrix} A \\ u \end{bmatrix}, \quad \mathbf{F} = \begin{bmatrix} Au \\ \frac{u^2}{2} + \frac{p^e}{\rho} \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{p^v}{\rho} \end{bmatrix}, \quad \mathbf{S}(\mathbf{U}) = \begin{bmatrix} 0 \\ -K_r \frac{u}{A} \end{bmatrix},$$

in terms of the elastic component $p^e(x, t)$ and viscoelastic component $p^v(x, t)$ of the pressure.

The hyperbolic system is approximated numerically using a discontinuous Galerkin method. The one-dimensional domain $\Omega = (a, b)$ is discretized into N non-overlapping elements $\Omega_i = (x_i^L, x_i^R)$ such that $x_i^R = x_{i+1}^L$ and $\bigcup_{i=1}^N \bar{\Omega}_i = \bar{\Omega}$; discrete approximations to the corresponding weak formulation are found in terms of orthonormal Legendre polynomials of degree q (69, 70) (see, e.g., (69, 71) for the advantages of this approach). Inlet and outlet boundary elements use upwind flux, while a second-order Adams-Bashforth scheme (69) is used for time integration.

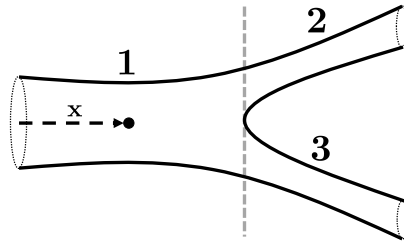


Figure 3.2: Schematic of Y-bifurcation in an arterial network.

To extend the model to a branching arterial network, multiple arteries are joined via coupled boundary conditions at bifurcations. An example of such a bifurcation appears in Figure 3.2. Boundary conditions are physically motivated; mass should be conserved through bifurcations, while momentum should be continuous at the boundary, i.e.,

$$\begin{aligned} A_1 u_1 &= A_2 u_2 + A_3 u_3, \\ p_1 + \frac{1}{2} \rho u_1^2 &= p_2 + \frac{1}{2} \rho u_2^2, \\ p_1 + \frac{1}{2} \rho u_1^2 &= p_3 + \frac{1}{2} \rho u_3^2, \end{aligned}$$

where p_i , A_i , and u_i correspond to the i th artery. Other branching configurations

appear in work by Sherwin (41).

3.3 Results

We now apply the Bayesian framework of Section 2.3 to the blood flow model of Section 3.2. In particular, we study the example 19-artery network shown in Figure 3.3. The solution for our deterministic model, given by (3.2.2) and solved using a discontinuous Galerkin method with time step $\Delta t_1 = 0.00004$ seconds, plays the role of $\mathbf{g}$ in the model prediction equation (2.1.3). Measurements of the flow velocity are taken at N specified locations which vary by experiment and occur with a sampling period of $\Delta t_2 = 1600\Delta t_1 = 0.064$ seconds.

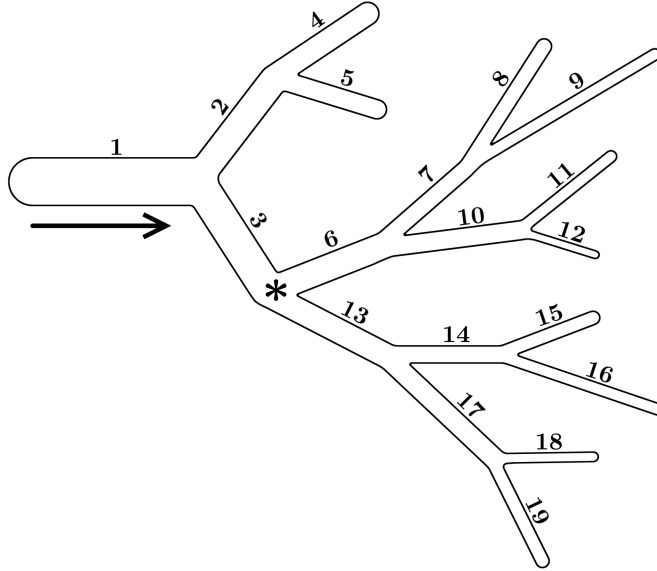


Figure 3.3: Schematic of arterial network (not to scale). 19 arteries have varied lengths (ranging from 0.026 to 0.17 m) and cross-sectional areas (ranging from 10^{-5} to 10^{-6} m^2). The star shows an example measurement location at a bifurcation.

Blood (viscosity $\mu^* = 0.0045$, with asterisks denoting reference values) in the network begins with zero velocity and is driven by a specified inflow velocity at the

beginning of the first artery: a sum of trigonometric polynomials, shown in Figure 3.4, which approximates the flow for three cardiac cycles (48). The length of three cycles (~ 3.3 seconds) allows for a total of 52 velocity data per measurement location using the sampling period Δt_2 , and so the output space of $\mathbf{g}$ has dimension $52N$. The outflow condition is a fully-absorbing boundary condition, described in more detail in (47).

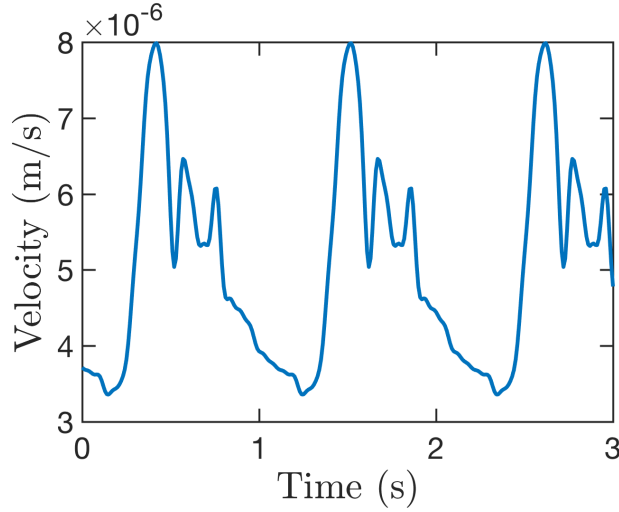


Figure 3.4: Inflow boundary condition for blood velocity (m/s) corresponding to three cardiac cycles.

We consider three free parameters: the blood viscosity μ , the arterial stiffness parameter B , and the relaxed cross-sectional area A_0 (the last two of which vary by artery). We note that when we estimate the parameters, we are only interested in the stiffness or cross-sectional area in one or two arteries. We fix the parameters for all arteries in the network other than the one or two that we study. A structural defect (e.g., an aneurysm or stenosis) can be modeled by varying the stiffness or relaxed area of a particular artery; to better emphasize the degree to which a flawed artery has been modified by its defect, results will use the scaled stiffness $\beta = B/B^*$ (with respect to the reference stiffness B^*) and scaled cross-sectional area $\alpha = A_0/A_0^*$ (with respect to the reference area A_0^*), and so arteries with no defect will have $\beta = \alpha = 1$.

We use our implementation of Bayesian uncertainty quantification to examine a number of questions in the context of this forward model, focusing in particular on the ability of uncertainty quantification to identify the location of structural flaws within the network using only noisy measurements of the flow velocity. To test the effectiveness of our implementation in these experiments thus requires noisy data $\mathbf{D}$ corresponding to a known truth; we use here synthetic data generated from the same model but with known, fixed parameters. It should be stressed that the approach is easily modified to admit real data and that there exist multiple practical methods for measuring blood flow velocities from *in vivo* arteries (64, 65, 66, 67). Section 3.3.4 will show that flaws can be located accurately even when the parameters used to generate the synthetic data are significantly perturbed from the parameters used to perform uncertainty quantification.

Explicitly, observed data $\mathbf{D}$ are generated as

$$D_k = v_k + \sigma \epsilon_k, \quad (3.3.1)$$

where D_k is the noisy observation at time t_k , v_k is the flow velocity at time t_k , ϵ_k is a zero-mean, unit-variance Gaussian random variable, and σ is the noise level. Here, we choose σ to be a fraction $\sigma = 0.01\eta$ (or sometimes 0.05η) of the standard deviation η of all velocity data v_k .

In the following results, we use our implementation of uncertainty quantification to generate 500 samples from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{D}, M)$ in a variety of scenarios. Posterior distributions are used for parameter estimation (Section 3.3.1) and to identify structural flaws via Bayesian model selection (Sections 3.3.2 and 3.3.4). Recovered posterior means, denoted with a hat (e.g., $\hat{\beta}$), are used as parameter estimates in our analysis.

3.3.1 Parameter Estimation

We first consider a basic case of parameter estimation to illustrate the feasibility of the approach. Specifically, we estimate the blood viscosity μ and the scaled stiffness β_2 of artery 2 (see Figure 3.3 for artery labels) assuming all other parameters are fixed to their reference values. The noisy data used, corrupted according to (3.3.1) with noise level $\sigma = 0.01\eta$, are sampled from a single location at the start of artery 6. We choose σ as an additional free parameter, requiring the approach to recover the noise level in addition to the target model parameters. A uniform distribution on $[0.5, 1.5] \times [0.5, 1.5] \times [0, 1]$ in the parameter space (μ, β_2, σ) is used as the parameter prior π .

To determine the effect of the choice of sampling location, we additionally consider separate cases using data obtained from the start of arteries 1 and 8; for notational clarity, we refer to O_i as the case of observing the upflow end of artery i .

The results for the case O_6 appear in Figure 3.5. μ and β_2 are positively correlated in the posterior, i.e., simultaneously raising or lowering both the blood viscosity and the stiffness of artery 2 yields qualitatively similar observed data. This could potentially indicate that in order to maintain a consistent rate of flow, a more viscous flow necessitates more rigid artery walls.

Numerical results for O_1 , O_6 , and O_8 are summarized in Table 3.1. The recovered posterior means of (μ, β_2, σ) were $(0.00410, 0.972, 0.00074)$, $(0.00449, 0.997, 0.00074)$, and $(0.00447, 0.994, 0.00077)$, respectively, closely matching the true values $\mu^* = 0.0045$ and $\beta_2^* = 1.0$ (σ^* differed by experiment due to differences in flow velocity by location: 0.00074, 0.00074, and 0.00077 for O_1 , O_6 , and O_8 , respectively). To quantify the degree of uncertainty in each parameter's posterior distribution, we

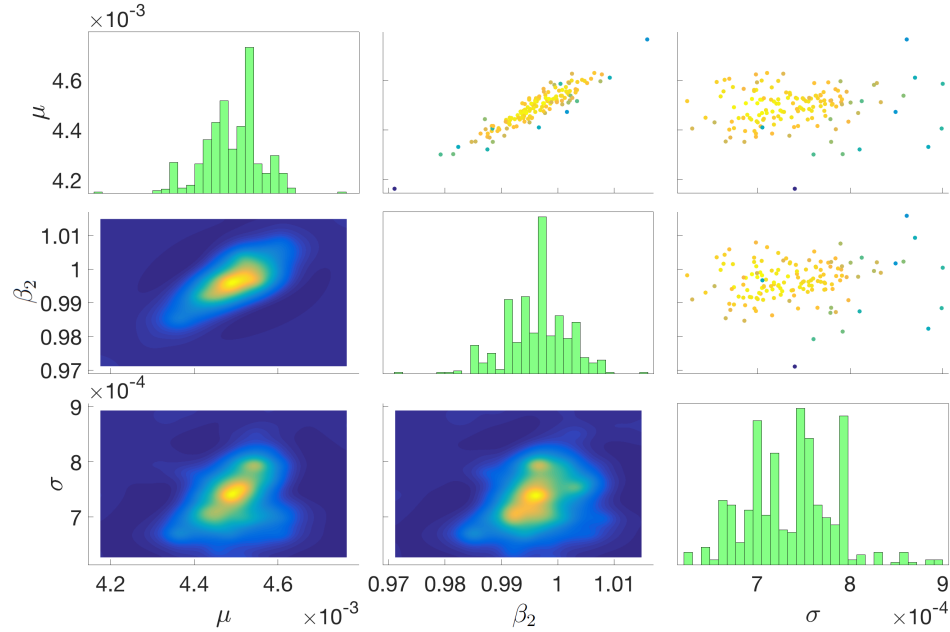


Figure 3.5: Parameter estimation results for blood viscosity μ , arterial stiffness β_2 for artery 2, and noise level σ using corrupted reference data from the beginning of artery 6 (O_6). Figures on the diagonal show histograms for each parameter. Subfigures below the diagonal show the marginal joint densities for each pair of parameters, while subfigures above the diagonal show the samples used in the final (convergent) stage of TMCMC. Colors correspond to likelihoods, with yellow likely and blue unlikely. For visual intuition on TMCMC convergence, see Appendix.

compute a coefficient of variation, defined as the ratio of the single-parameter posterior's standard deviation to its mean (denoting the results $u_{\hat{\mu}}$, $u_{\hat{\beta}_2}$, and $u_{\hat{\sigma}}$); here, O_6 and O_8 recover parameters with comparatively lower uncertainty than O_1 , whose measurements were largely dominated by the inflow boundary condition. Nonetheless, in all cases the reference values used to generate the synthetic data were within one standard deviation of the recovered posterior means.

Data Observations	$\hat{\mu}$	$u_{\hat{\mu}}$ (%)	$\hat{\beta}_2$	$u_{\hat{\beta}_2}$ (%)	$\hat{\sigma}$	$u_{\hat{\sigma}}$ (%)
O_1	0.004102	11.86	0.97175	3.48	0.00074218	5.87
O_6	0.004493	1.41	0.99663	0.51	0.00073577	6.30
O_8	0.004470	2.05	0.99367	0.70	0.00076524	7.65

Table 3.1: Posterior means and uncertainties for parameter estimation on the 19-artery network for three cases O_i (noise level $\sigma = 0.01\eta$).

3.3.2 Locating Structural Flaws with Model Selection

Given the practicality of parameter estimation and its intermediate estimation of the model evidence $\rho(\mathbf{D}|M)$, the Bayesian model selection framework described in Chapter 2.1 is a natural approach to locating structural flaws in the arterial network. Namely, define as M_i the model in which the scaled stiffness β of artery i has been perturbed from its reference value by an unknown amount, corresponding to, e.g., an aneurysm or stenosis. Parameter estimation as in Section 3.3.1 can be used to recover the perturbed stiffness which best matches the observed data, simultaneously yielding an estimate for the evidence $\rho(\mathbf{D}|M_i)$ of model M_i . Letting $\mathcal{M}$ be the collection of M_i for various arteries i in the network, the model selection distribution $Pr(M_i|\mathbf{D})$ is a probabilistic measure of the likelihood of the structural defect occurring in artery i (as opposed to a different artery j). The class $\mathcal{M}$ can easily be augmented with additional models; here, we also consider a model $M_{i,j}$ which freely varies the stiffness of two arteries i and j .

We first consider generating data $\mathbf{D}$ from a reference model using $\beta_6 = 0.5$ and $\beta_i = 1, i \neq 6$, i.e., a model in which the scaled stiffness of artery 6 has been halved from its reference value. We consider three cases for data collection: a two sensor

configuration using data from the end of arteries 1 and 7, a three sensor configuration using data from the end of arteries 1, 7, and 13, and a four sensor configuration using velocity data from the end of arteries 1, 7, 10, and 13. In each case, no sampling locations are adjacent to the damaged artery. The noise level is again chosen as $\sigma = 0.01\eta$, i.e., 1% Gaussian noise added to the velocity data, and we employ the same uniform prior on $[0, 3] \times [0, 1]$ for the parameters (β, σ) .

Table 3.2 presents numerical results for six potential flaw-location models $M_3, M_6, M_7, M_{11}, M_{13}$, and $M_{6:7}$ when taking flow measurements from the ends of arteries 1 and 7. Models M_6 and $M_{6:7}$, both of which include the correct defect location in artery 6, are assigned the largest probabilities under the model selection posterior ($Pr(M_j|\mathbf{D}) = 0.99949, 0.00042$, respectively); recovered parameter estimates $\hat{\beta}_6 = 0.499$ (M_6) and $\hat{\beta}_6 = 0.501$ ($M_{6:7}$) for the wall stiffness of the damaged artery were accurate to within one standard deviation. Though $M_{6:7}$ assumes a second defect in artery 7, the posterior mean estimated the stiffness to be similar to the reference value ($\hat{\beta}_7 = 1.036$). M_3, M_7, M_{11} , and M_{13} are not able to accurately match the observed data, and so require a significantly higher noise level σ to explain differences between the evaluated and observed velocities. For this reason, these models are assigned negligible mass by $Pr(M_i|\mathbf{D})$.

Note that the model $M_{6:7}$ contains M_6 in the sense that it can predict any combination of parameter values which M_6 can predict. In this light, the relatively higher evidence for M_6 over the broader error model $M_{6:7}$ is in keeping with theoretical results available for Bayesian model class selection wherein over-parameterized model classes are penalized due to Occam's factor (29).

Table 3.3 illustrates the corresponding results for a three-sensor configuration using blood flow velocity data from the ends of arteries 1, 7, and 13. The additional

Model	$\hat{\beta}$	$u_{\hat{\beta}}$ (%)	$\hat{\sigma}$	$u_{\hat{\sigma}}$ (%)	Log Evid.	$Pr(M_j \mathbf{D})$
M_3	0.741	2.82	0.000786	6.31	583.2	0.00009
M_6	0.499	6.61	0.000729	3.73	592.5	0.99949
M_7	1.055	3.14	0.000923	4.27	567.7	~ 0
M_{11}	1.022	0.76	0.000892	5.96	566.3	~ 0
M_{13}	0.587	5.81	0.000867	5.04	578.3	~ 0
$M_{6:7}$	[0.501, 1.036]	[3.75, 2.15]	0.000719	5.16	584.7	0.00042

Table 3.2: Numerical results for identification of a $\beta_6 = 0.5$ aneurysm using noisy data from the ends of arteries 1 and 7 (noise level $\sigma = 0.01\eta$).

data collected from artery 13 significantly reduce the (already small) probabilities assigned to models other than M_6 and $M_{6:7}$. Interestingly, leveraging information from the end of artery 13, which is in a parallel tree (rather than directly upstream or downstream) from the damaged artery, has the effect of shifting mass from M_6 to $M_{6:7}$ in the model posterior, finding $Pr(M_6|\mathbf{D}) = 0.708$ and $Pr(M_{6:7}|\mathbf{D}) = 0.292$. Nonetheless, both M_6 and $M_{6:7}$ estimate the damaged stiffness β_6 accurately (0.501 and 0.505, respectively), with $M_{6:7}$ again finding the stiffness β_7 of the undamaged artery 7 to be largely unchanged (1.031).

Model	$\hat{\beta}$	$u_{\hat{\beta}}$ (%)	$\hat{\sigma}$	$u_{\hat{\sigma}}$ (%)	Log Evid.	$Pr(M_j \mathbf{D})$
M_3	0.722	2.29	0.000811	4.81	875.6	0.00001
M_6	0.501	4.31	0.000770	5.23	886.3	0.70756
M_7	1.056	2.19	0.000956	4.64	853.1	~ 0
M_{11}	1.005	0.60	0.000949	5.02	848.8	~ 0
M_{13}	0.901	1.91	0.000887	5.06	858.3	~ 0
$M_{6:7}$	[0.505, 1.031]	[3.45, 2.33]	0.000770	3.58	885.5	0.29242

Table 3.3: Numerical results for identification of a $\beta_6 = 0.5$ aneurysm using data from the ends of arteries 1, 7, and 13 (noise level $\sigma = 0.01\eta$).

Finally, Table 3.4 shows numerical results when velocity data are sampled at four monitoring locations: at the ends of arteries 1, 7, 10, and 13. The additional data from the end of artery 10 drive the model probabilities assigned to M_3, M_7, M_{11} , and M_{13} down further ($< 10^{-8}$), rendering them orders of magnitude smaller than the probabilities assigned to M_6 and $M_{6:7}$ (0.9996 and 0.0004, respectively). The

estimated scaled stiffness remains accurate to within one standard deviation, with M_6 and $M_{6:7}$ finding $\hat{\beta}_6 = 0.520$ and 0.521 , respectively, and $M_{6:7}$ again estimates the stiffness of artery 7 to be only slightly perturbed ($\hat{\beta}_7 = 1.027$).

Model	$\hat{\beta}$	$u_{\hat{\beta}}$ (%)	$\hat{\sigma}$	$u_{\hat{\sigma}}$ (%)	Log Evid.	$Pr(M_j \mathbf{D})$
M_3	0.760	2.60	0.000817	3.93	1169.9	~ 0
M_6	0.520	2.57	0.000762	3.48	1187.3	0.9996
M_7	1.046	2.72	0.000926	3.86	1144.5	~ 0
M_{11}	0.996	0.19	0.000932	3.58	1134.1	~ 0
M_{13}	0.890	1.83	0.000902	3.47	1148.8	~ 0
$M_{6:7}$	[0.521, 1.027]	[3.98, 2.49]	0.000771	3.63	1179.5	0.0004

Table 3.4: Numerical results for identification of a $\beta_6 = 0.5$ aneurysm using data from the ends of arteries 1, 7, 10, and 13 (noise level $\sigma = 0.01\eta$).

Taken together, these configurations support two conclusions about Bayesian model selection for flaw identification: first, that increasing the number of locations at which data are sampled reduces the probabilities assigned to incorrect models, and second, that model selection can accurately determine the defect location and magnitude for a variety of sensor configurations, including configurations which do not sample from at or near the defect location.

3.3.3 Model Selection for Cross-Sectional Area

As previously suggested, aneurysms and stenoses can also be modeled by adjusting the initial cross-sectional area of an artery rather than its stiffness. Ideally, the Bayesian framework for model selection should provide similar results when the stiffnesses β are fixed and models M_i instead allow the scaled cross-sectional area α of the defective artery to be perturbed. In what follows, we examine similar scenarios to the above in the case where, rather than reducing its wall stiffness, the relaxed cross-sectional area of artery 6 is altered. A uniform prior on $[0, 3] \times [0, 1]$ is used for

the parameters (α, σ) .

We first consider the case $\alpha_6 = 1.5$, i.e., an aneurysm in which the defective artery (again, artery 6 in the reference model) has become enlarged by 50%. Noisy flow velocity data are collected from the ends of arteries 1, 7, 10, and 13, as in the final case of the previous section; results appear in Table 3.5. M_6 and $M_{6:7}$ are again the most likely models ($Pr(M_j|\mathbf{D}) = 0.998$ and 0.002 , respectively), suggesting that the previous results do not rely on the specific choice of the parameter β . Other models were assigned negligible probabilities. Similarly to the results for reduced stiffness, both M_6 and $M_{6:7}$ accurately recover the defect magnitude ($\hat{\alpha}_6 = 1.508, 1.507$, respectively), and $M_{6:7}$ finds artery 7 to be unchanged ($\hat{\alpha}_7 = 1.000$).

Model	$\hat{\alpha}$	$u_{\hat{\alpha}}$ (%)	$\hat{\sigma}$	$u_{\hat{\sigma}}$ (%)	Log Evid.	$Pr(M_j \mathbf{D})$
M_3	1.126	26.29	0.00554	3.40	781.0	~ 0
M_6	1.508	0.45	0.00078	4.33	1183.6	0.998
M_7	0.976	0.17	0.00415	4.15	827.9	~ 0
M_{11}	1.048	0.48	0.00471	3.41	810.0	~ 0
M_{13}	1.014	0.17	0.00518	3.18	793.0	~ 0
$M_{6:7}$	[1.507, 1.000]	[0.52, 0.031]	0.00077	3.24	1177.6	0.002

Table 3.5: Numerical results for area-based identification of an $\alpha_6 = 1.5$ aneurysm using data from the ends of arteries 1, 7, 10, and 13 (noise level $\sigma = 0.01\eta$).

We then consider the same scenario for a reduction $\alpha_6 = 0.5$ in the cross-sectional area of artery 6, i.e., a stenosis in which the defective artery has narrowed by 50%. Results are summarized in Table 3.6. M_6 and $M_{6:7}$ recover the reduced area accurately ($\hat{\alpha}_6 = 0.500, 0.501$, respectively) and are assigned the highest model evidence ($Pr(M_j|\mathbf{D}) \approx 1.00$ and $\sim 10^{-4}$, respectively).

Table 3.7 shows results for the same magnitude stenosis ($\alpha_6 = 0.5$) with increased observational noise level $\sigma = 0.05\eta$. The log evidence of models M_6 and $M_{6:7}$ is sharply reduced compared to Table 3.6, though M_6 and $M_{6:7}$ remain the most probable models under the model selection posterior, with $Pr(M_j|\mathbf{D}) = 0.983$ and

Model	$\hat{\alpha}$	$u_{\hat{\alpha}}$ (%)	$\hat{\sigma}$	$u_{\hat{\sigma}}$ (%)	Log Evid.	$Pr(M_j \mathbf{D})$
M_3	0.072	8.00	0.023	4.04	478.9	~ 0
M_6	0.500	0.05	0.00076	3.16	1178.0	1.00
M_7	1.127	0.78	0.0195	4.83	513.3	~ 0
M_{11}	0.790	2.37	0.0215	3.72	493.8	~ 0
M_{13}	0.941	0.73	0.0238	3.33	475.2	~ 0
$M_{6:7}$	[0.501, 1.000]	[0.08, 0.03]	0.00076	3.30	1164.6	~ 0

Table 3.6: Numerical results for area-based identification of an $\alpha_6 = 0.5$ stenosis using data from the ends of arteries 1, 7, 10, and 13 (noise level $\sigma = 0.01\eta$).

0.017, respectively. Both models additionally recover the reduced area accurately ($\hat{\alpha}_6 = 0.502$ and 0.503 , respectively) despite the increased noise.

Model	$\hat{\alpha}$	$u_{\hat{\alpha}}$ (%)	$\hat{\sigma}$	$u_{\hat{\sigma}}$ (%)	Log Evid.	$Pr(M_j \mathbf{D})$
M_3	0.070	4.98	0.0234	2.56	481.1	~ 0
M_6	0.502	0.37	0.00376	4.10	849.6	0.983
M_7	1.118	0.90	0.0199	2.78	512.5	~ 0
M_{11}	0.800	2.09	0.0221	4.33	492.5	~ 0
M_{13}	0.940	0.88	0.0234	3.97	474.8	~ 0
$M_{6:7}$	[0.503, 1.000]	[0.40, 0.23]	0.00381	4.69	845.6	0.017

Table 3.7: Numerical results for area-based identification of an $\alpha_6 = 0.5$ stenosis using data from the ends of arteries 1, 7, 10, and 13 (noise level $\sigma = 0.05\eta$).

Finally, Table 3.8 considers the case of a smaller-magnitude stenosis ($\alpha_6 = 0.8$). Results were similar to those of Table 3.7, with accurate recovery of location ($Pr(M_j|\mathbf{D}) = 0.99999$ and 0.00001 for M_6 , $M_{6:7}$, respectively) and magnitude ($\hat{\alpha}_6 = 0.801$ and 0.802). As in the previous area-modification scenarios, $M_{6:7}$ found artery 7 to be unaffected ($\hat{\alpha}_7 = 1.000$), thereby coinciding with the single defect model M_6 .

3.3.4 Locating Defects with Misspecified Models

Results have so far assumed the model selection framework is provided the reference values for all model parameters, i.e., the non-defective stiffness and area of

Model	$\hat{\alpha}$	$u_{\hat{\alpha}}$ (%)	$\hat{\sigma}$	$u_{\hat{\sigma}}$ (%)	Log Evid.	$Pr(M_j \mathbf{D})$
M_3	0.175	4.13	0.00493	3.71	809.0	~ 0
M_6	0.801	0.18	0.00075	4.68	1185.1	0.99999
M_7	1.024	0.15	0.00404	3.78	838.0	~ 0
M_{11}	0.955	0.45	0.00447	3.92	820.5	~ 0
M_{13}	0.987	0.17	0.00488	3.89	802.7	~ 0
$M_{6:7}$	[0.802, 1.000]	[0.16, 0.033]	0.00077	3.11	1173.7	0.00001

Table 3.8: Numerical results for area-based identification of an $\alpha_6 = 0.8$ stenosis using data from the ends of arteries 1, 7, 10, and 13 (noise level $\sigma = 0.01\eta$).

each artery are known. In a scenario using real-world data, these “known” values must themselves be estimated from noisy measurements. A final but crucial test of the robustness of the framework is thus to perform experiments in which the reference parameters used by the method are incorrect, and so no combination of free parameters is capable of reproducing the observed data.

We now revisit the cases of Section 3.3.3, beginning with the case of an $\alpha_6 = 1.5$ aneurysm in artery 6. In addition to corrupting observed flow velocities with additive Gaussian noise, we now additionally corrupt the parameters themselves: the initial cross-sectional area α_k for each artery k is noised as

$$\alpha_k = \alpha_k^* (1 + \sigma_\alpha \epsilon_k), \quad (3.3.2)$$

where α_k^* is the reference value, ϵ_k is again a standard normal random variable, and σ_α is the parameter noise level. The structural parameters used to generate the synthetic data (α_k from Eq. 3.3.2) thus differ from the fixed values used in the defect models M_i (α_k^*).

As before, Bayesian model selection is performed assuming the prediction equation (2.1.3), which is now misspecified (it assumes correctness of the reference parameters α_i^*). As a result, the $\hat{\sigma}$ estimated by posterior samples must now capture

the effects of both the true observational noise level σ and the parameter noise level σ_α .

Table 3.9 shows numerical results for Bayesian model selection in this setting. Despite the misspecification, M_6 and $M_{6:7}$ again dominate the model posterior, with $Pr(M_6|\mathbf{D}) = 0.907$ and $Pr(M_{6:7}|\mathbf{D}) = 0.093$, respectively. Both overestimate the defect magnitude ($\hat{\alpha}_6 = 2.047, 2.123$, respectively), though $M_{6:7}$ again estimates artery 7 to be unaffected ($\hat{\alpha}_7 = 1.002$). We note that some error in $\hat{\alpha}_6$ is expected, as it attempts to fit observations from the noised-parameter model and thus varies significantly depending on the particular values of α_k from Eq. (3.3.2). Despite this effect, identification of the location appears robust to perturbation of model parameters, with all other models assigned negligible probability ($Pr(M_i|\mathbf{D}) \sim 0$).

Model	$\hat{\alpha}$	$u_{\hat{\alpha}}$ (%)	$\hat{\sigma}$	$u_{\hat{\sigma}}$ (%)	Log Evid.	$Pr(M_j \mathbf{D})$
M_3	1.312	29.2	0.00761	3.39	711.6	~ 0
M_6	2.047	0.89	0.00217	3.98	972.9	0.907
M_7	0.970	0.23	0.00628	3.48	748.7	~ 0
M_{11}	1.075	0.56	0.00645	3.76	746.9	~ 0
M_{13}	1.016	0.29	0.00725	3.53	716.8	~ 0
$M_{6:7}$	[2.123, 1.002]	[3.27, 0.11]	0.00216	4.32	970.6	0.093

Table 3.9: Numerical results for area-based identification of an $\alpha_6 = 1.5$ aneurysm using data from the ends of arteries 1, 7, 10, and 13 (noise level $\sigma = 0.01\eta$) with misspecified cross-sectional areas (perturbed with noise level $\sigma_\alpha = 0.01$).

Turning to the second case ($\alpha_6 = 0.5$), model selection again successfully locates the defect despite the misspecification (Table 3.10), with M_6 assigned nearly all mass by the model selection posterior. In this case, parameter estimation recovers the defect magnitude accurately ($\hat{\alpha}_6 = 0.515$). In keeping with previous results, defect model $M_{6:7}$ finds a similar reduction in cross-sectional area for the damaged artery ($\hat{\alpha}_6 = 0.517$) and little change in the defect-free artery ($\hat{\alpha}_7 = 1.003$).

The third case repeated the $\alpha_6 = 0.5$ experiment with increased observational

Model	$\hat{\alpha}$	$u_{\hat{\alpha}}$ (%)	$\hat{\sigma}$	$u_{\hat{\sigma}}$ (%)	Log Evid.	$Pr(M_j \mathbf{D})$
M_3	0.816	7.29	0.0220	4.36	485.7	~ 0
M_6	0.515	0.17	0.0021	3.22	973.6	0.999
M_7	1.118	0.66	0.0181	3.44	529.0	~ 0
M_{11}	0.808	1.72	0.0207	3.41	505.7	~ 0
M_{13}	0.943	0.69	0.0221	3.14	490.5	~ 0
$M_{6:7}$	[0.517, 1.003]	[0.23, 0.001]	0.0021	3.87	966.5	0.001

Table 3.10: Numerical results for area-based identification of an $\alpha_6 = 0.5$ stenosis using data from the ends of arteries 1, 7, 10, and 13 (noise level $\sigma = 0.01\eta$) with misspecified cross-sectional areas (perturbed with noise level $\sigma_\alpha = 0.01$).

noise $\sigma = 0.05\eta$; results for the same case with parameter noise (now also increased to $\sigma_\alpha = 0.05$) are shown in Table 3.11. $M_{6:7}$ is significantly more likely than in previous cases ($Pr(M_{6:7}|\mathbf{D}) = 0.783$), though M_6 is still assigned all remaining posterior mass ($Pr(M_6|\mathbf{D}) = 0.217$). The recovered uncertainties $u_{\hat{\alpha}}$ are significantly higher than in Table 3.10 due to the higher level of noise, with both M_6 and $M_{6:7}$ underestimating the magnitude of the damage ($\hat{\alpha}_6 = 0.597, 0.609$, respectively).

Model	$\hat{\alpha}$	$u_{\hat{\alpha}}$ (%)	$\hat{\sigma}$	$u_{\hat{\sigma}}$ (%)	Log Evid.	$Pr(M_j \mathbf{D})$
M_3	1.668	45.9	0.0195	4.35	518.5	~ 0
M_6	0.597	0.80	0.0111	3.31	636.6	0.217
M_7	1.085	0.62	0.00076	4.82	558.5	~ 0
M_{11}	0.894	1.52	0.0185	3.82	527.1	~ 0
M_{13}	0.947	0.51	0.0175	3.62	534.9	~ 0
$M_{6:7}$	[0.609, 1.015]	[1.22, 0.44]	0.0108	2.22	637.9	0.783

Table 3.11: Numerical results for area-based identification of an $\alpha_6 = 0.5$ stenosis using data from the ends of arteries 1, 7, 10, and 13 (noise level $\sigma = 0.05\eta$) with misspecified model parameters (perturbed with noise level $\sigma_\alpha = 0.05$).

Finally, Table 3.12 shows results for the fourth case ($\alpha_6 = 0.8$) in the presence of $\sigma_\alpha = 0.01$ parameter noise. While model selection again recovers the correct defect location ($Pr(M_j|\mathbf{D}) = 0.969, 0.031$ for $M_6, M_{6:7}$, respectively), the smaller-magnitude stenosis proves more challenging for parameter estimation, with $\hat{\alpha}_6 = 0.865$ and 0.869 , respectively, notably underestimating the magnitude of the damage.

Model	$\hat{\alpha}$	$u_{\hat{\alpha}}$ (%)	$\hat{\sigma}$	$u_{\hat{\sigma}}$ (%)	Log Evid.	$Pr(M_j \mathbf{D})$
M_3	1.767	22.2	0.00388	3.57	852.4	~ 0
M_6	0.865	0.35	0.00204	3.36	974.9	0.969
M_7	1.017	0.13	0.00320	3.87	889.5	~ 0
M_{11}	0.979	0.24	0.00369	3.61	860.3	~ 0
M_{13}	0.988	0.12	0.00345	3.49	869.5	~ 0
$M_{6:7}$	[0.869, 1.003]	[0.50, 0.076]	0.00209	3.22	971.5	0.031

Table 3.12: Numerical results for area-based identification of an $\alpha_6 = 0.8$ stenosis using data from the ends of arteries 1, 7, 10, and 13 (noise level $\sigma = 0.01\eta$) with misspecified model parameters (perturbed with noise level $\sigma_\alpha = 0.01$).

3.4 Discussion

Taken together, the results describe a robust approach for uncertainty quantification in the context of arterial networks. The model selection posterior universally assigned the highest probabilities (by several orders of magnitude) only to those models which included the true defect location, even in cases where simulated data were sparse, noisy, and poorly located. The Bayesian uncertainty quantification framework thus appears a powerful tool for comparing and fitting models.

Though all results were generated using simulated noisy data, they simultaneously suggest that the approach would prove useful for real-world inference. The experiments outlined in Section 3.3.2 show the method to successfully recover parameter values (often within one standard deviation) and identify the defect location in a range of sampling cases which varied sensor numbers and locations, and so the approach is not reliant on a particular set of observed data which may not be realistically attainable. Results were also consistent when using alternative magnitudes and parametrizations of arterial defects (the scaled cross-sectional area α and boundary stiffness β) and using models which considered different numbers of defects (in particular, the two-defect model $M_{6:7}$ which consistently found the “defective” artery 7 to be largely unaltered). We note that the approach readily facilitates the incorpo-

ration of real data, which can be used in place of simulated data without otherwise altering the method. A natural extension of this work is therefore direct application to medical datasets.

It is worth emphasizing the role of model selection in our approach. The general Bayesian inverse problem, which simultaneously considers structural properties of all arterial segments, cannot be feasibly solved due to its dimensionality. In Section 3.3.2, we instead propose a set of simple models which fix most model parameters to “known” values, then select among these models to locate flaws; while this approach has a massive computational advantage, it also makes the strong assumption that the properties of healthy artery segments are known exactly. We investigate robustness of the approach to this assumption in Section 3.3.4, which found inference to remain effective even when “known” model parameters were assumed incorrectly.

While model selection has here served primarily as a computational tool, it is more commonly used to select between distinct physical models of a system given noisy data. Robustness to this form of model misspecification, i.e., the assumption of a physically incorrect model, is a more complex issue begetting a range of additional techniques such as discrepancy terms (72) and posterior predictive assessment (73). Future work should investigate this more general sense of robustness using alternative flow models, network structures, and defects in more localized arterial subdomains.

Appendix: Convergence of TMCMC

Figure 3.6 shows eight generations of TMCMC convergence from parameter estimation in Section 3.3.1. Samples are initialized from the prior, i.e., sampled uniformly

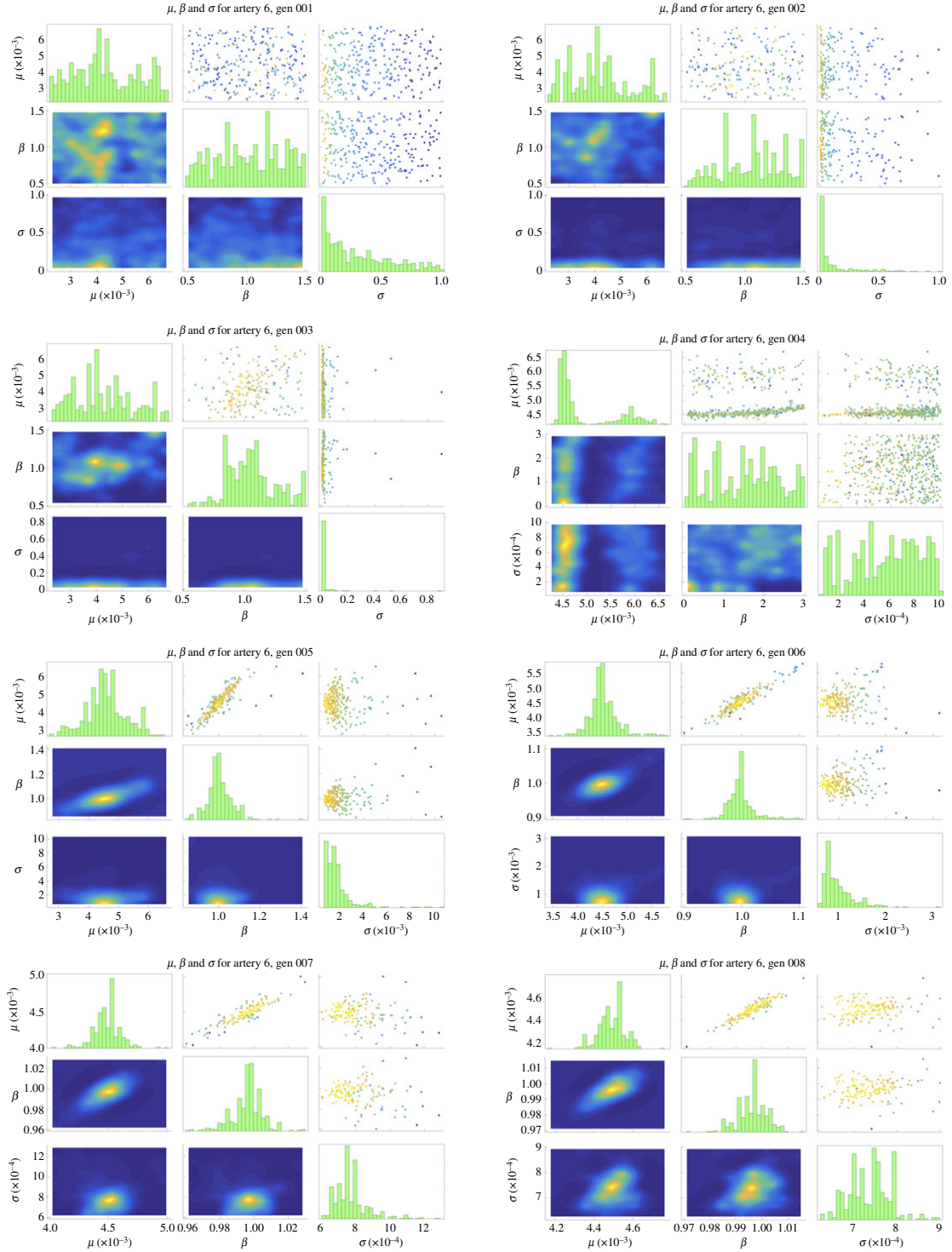


Figure 3.6: Samples from eight generations of TMCMC in Section 3.3.1. Plot limits are not constant (the sampled region of high probability monotonically decreases in size).

from all feasible parameter values. The transitional target distribution yields gradual coalescence on regions of high likelihood, eventually resampling directly from the posterior. Each generation displays the 500 samples taken on that stage of TMCMC, yielding a total of 4500 samples for all 9 generations. We do not include a visualization for generation 0, as it is simply a uniform sampling over the space of the prior distribution. In these figures, we can observe how TMCMC intelligently samples the parameter space: the first generation (not displayed) takes takes samples from the prior distribution, allowing the method to find regions of higher probability. The next few generations of samples all mostly focus on finding a better range of σ values, as the prior for σ was $[0, 1]$, and the nominal parameter value was ~ 0.00074 . Once σ 's value is in the appropriate range of values, the algorithm is able to identify the proper values for μ and β , and the intermediate distributions begin to center and narrow around the nominal parameter values.

CHAPTER FOUR

Data-driven prediction and origin identification of epidemics in population networks

4.1 Introduction

Robust prediction for the spread of an epidemic is critical to monitoring and halting its progress. The reliability of these predictions, which have high clinical and societal significance, hinges on the underlying mathematical models which quantify the spread and virulence of diseases. Several models have been proposed for predicting the spread of epidemics in real-world populations, allowing for the development of strategies for effectively managing disease spread via organized intervention. Perhaps the most well-known approach is Kermack and McKendrick’s compartmental SIR model (and its extensions, such as SIRS and SEIR), a differential equation model which divides populations into groups corresponding to their relation with the disease (e.g., susceptible or removed), which is widely studied due to its simplicity and predictiveness for several common diseases (4, 74). More recent work has also incorporated the topological aspect of network structure by modeling explicit (or random) population networks through which diseases propagate (6, 75, 76, 77), working toward a more holistic view of disease modeling which can include aspects such as demography, land use, and climate change (78). The predictions made by these mathematical models have been used to study a diverse set of historical and modern epidemics, including HIV (79, 80), malaria (81, 82, 83), polio (84), and tuberculosis (85), by using a wide range of data assimilation techniques (86, 87).

Many aspects of these models have also been analyzed from a more abstract mathematical perspective. Local bifurcation analysis has been performed on the man-environment-man and SIR models (5, 88), while other work has used Lyapunov functions to determine endemic equilibria for SIRS and SEIR models (89, 90) or has considered a mean-field approach (91, 77, 92, 93). Recent work has addressed how the network structure influences the spread of disease via the initial conditions

and network topologies (75, 76) and the ways in which epidemics spread on random networks (6). Analytical results have also been obtained for the case of two competing (or promoting) diseases on a network (88, 94). Moreover, many of the models in question have been used to design intervention policies or allocate vaccines via optimal control (90, 95) or randomized interventions (96).

In what follows, we apply the Bayesian uncertainty quantification framework $\Pi 4U$ (17) from Chapter 2 to an extension of the SIR model to graphs. Using noisy, simulated data, we show that our method is able to efficiently estimate values of the model parameters and their underlying uncertainties. The $\Pi 4U$ framework is also convenient for Bayesian model selection, which we use to identify the origin of the epidemic by considering each possible starting location as a separate model. Bayesian approaches thus show significant potential to aid in real-world epidemic modeling and mitigation.

4.2 SIR model

The SIR epidemic model decomposes a population into three eponymous groups: hosts who are *susceptible* to the disease, hosts who are infected and contagious (the *infective* group), and hosts who are neither susceptible nor infected, either via gained immunity from recovery or due to a vaccine, quarantine policies, or disease-related death (the *removed* group).

4.2.1 Single population model

Let $S(t)$, $I(t)$, and $R(t)$ denote the size of the susceptible, infective, and removed groups, respectively, as functions of a continuous time t . The SIR model is based on three main assumptions: first, since the timescale on which the disease evolves is assumed to be much shorter than the timescale on which the population may evolve via, e.g., births or natural deaths, the population Y is assumed constant, and so

$$S(t) + I(t) + R(t) = Y \quad (4.2.1)$$

for all t (note that individuals killed by the disease are considered part of the removed group). Second, members of the population are assumed to come into contact uniformly at random and at a constant rate β – this parameter governs the rate at which an infection can spread. Finally, the infective population recovers (or is otherwise removed from the infectives via, e.g., death) at a constant rate γ . These assumptions can thus be visualized as



yielding the following set of ordinary differential equations:

$$\frac{dS(t)}{dt} = -\beta IS, \quad \frac{dI(t)}{dt} = \beta IS - \gamma I, \quad \frac{dR(t)}{dt} = \gamma I. \quad (4.2.3)$$

Namely, at a particular time t , $S(t)$ susceptibles and $I(t)$ infectives come into contact at a rate β , yielding βIS transitions from susceptible to infective (implicitly assuming that contact with an infective immediately infects a susceptible – if this assumption is not desired, the chance of disease transfer can be incorporated in β). Meanwhile,

$I(t)$ infectives are removed at a rate γ , yielding γI transitions from infective to removed.

4.2.2 Epidemic model on graphs

The SIR model is readily generalized to a directed graph with N vertices, a mathematical construct which can be thought of as modeling a collection of N distinct communities. Namely, let each node (community) be a distinct population whose dynamics evolve according to Eq. 4.2.3; the directed edges (connections between communities) are a convenient framework to dictate transfer between populations. Since each population itself has three groups (susceptible, infective, removed), three quantities are needed to describe movement. Here, we use $\lambda_{i,j}$, $\eta_{i,j}$, and $g_{i,j}$ to describe the rate of movement from node i to node j on the susceptible, infective, and removed groups, respectively; identifying each transition rate as the weight of the edge connecting i to j , these rates are naturally written as weighted adjacency matrices, here denoted $\mathbf{\Lambda}$, $\mathbf{H}$, and $\mathbf{G}$. The SIR model on a network, now a system of N models corresponding to each population i , can then be written as

$$\begin{aligned}
 \frac{dS_i(t)}{dt} &= -\beta I_i S_i + \sum_{j=1}^N \lambda_{j,i} S_j - \sum_{j=1}^N \lambda_{i,j} S_i, \\
 \frac{dI_i(t)}{dt} &= \beta I_i S_i - \gamma I_i + \sum_{j=1}^N \eta_{j,i} I_j - \sum_{j=1}^N \eta_{i,j} I_i, \\
 \frac{dR_i(t)}{dt} &= \gamma I_i + \sum_{j=1}^N g_{j,i} R_j - \sum_{j=1}^N g_{i,j} R_i,
 \end{aligned} \tag{4.2.4}$$

or more succinctly in matrix form:

$$\begin{aligned}
\frac{d\mathbf{S}}{dt} &= -\beta\mathbf{I} \cdot \mathbf{S} + \mathbf{\Lambda}^T \mathbf{S} - (\mathbf{\Lambda}\mathbf{F}) \cdot \mathbf{S}, \\
\frac{d\mathbf{I}}{dt} &= \beta\mathbf{I} \cdot \mathbf{S} - \gamma\mathbf{I} + \mathbf{H}^T \mathbf{I} - (\mathbf{H}\mathbf{F}) \cdot \mathbf{I}, \\
\frac{d\mathbf{R}}{dt} &= \gamma\mathbf{I} + \mathbf{G}^T \mathbf{R} - (\mathbf{G}\mathbf{F}) \cdot \mathbf{R}.
\end{aligned} \tag{4.2.5}$$

(Here, $\mathbf{F} = (1, 1, \dots, 1)^T$ is a vector of ones which simplifies the notation.) It should be emphasized that $\mathbf{S}$, $\mathbf{I}$, and $\mathbf{R}$ are $N \times 1$ vectors whose i^{th} element corresponds to the i^{th} population. Note that if $\lambda_{i,j} = \eta_{i,j} = g_{i,j} = 0$ for all i, j , i.e., there is no movement between populations, each model reduces to the single population model (Eq. 4.2.3).

4.3 Results

In the following results, we apply our high-performance implementation of Bayesian uncertainty quantification $\Pi 4U$ to a case study with two example network structures. In each case, we fix particular values of the system parameters (infectivity β , rate of removal γ , and time of observation T) and use Eq. 4.2.5 to evolve the network forward in time via a 4th-order Runge-Kutta method. At the single observation time T (unless otherwise noted), the infective populations (and sometimes also the removed populations) from a selected subset of communities, corrupted by additive Gaussian noise with noise level σ , are output as the noisy observed data. Namely, observed data D_k are generated as

$$D_k = q_k + \sigma \epsilon_k,$$

where each deterministic population datum q_k generated from the reference model is added to a zero-mean, unit-variance Gaussian ϵ_k scaled by the noise level σ to yield the observed noisy datum D_k . In order that the signal-to-noise ratio be high enough for meaningful estimation, we choose σ to be a fraction $\sigma = 0.01\alpha$ (or sometimes $\sigma = 0.05\alpha$) of the standard deviation α of all model outputs p_k .

We then use our method to approximately solve the inference problem by generating 10^4 samples from the posterior distribution of the model parameters given these noisy outputs, checking the validity of our approach by comparing the resulting distributions to the known reference values. While we use synthetic, model-generated noisy data, we note that the framework is readily extended to the incorporation of real-world data.

For ease of comparison, we present numerical results in terms of the rescaled parameters $(\theta_\beta, \theta_\gamma, \theta_T, \theta_\sigma)$, given by $\theta_\beta = \beta/\beta_0$, i.e., the ratio between the estimated value and the true reference value. Accurate estimation thus results in scaled parameters close to 1. As the $\Pi4U$ approach does not rely on the choice of a particular prior, we use a simple uniform distribution on $[0.01, 2] \times [0.5, 2] \times [0.02, 5] \times [0.5, 10]$ in this scaled parameter space.

4.3.1 Network 1: the 20-barbell graph

We first consider a network with two distinct populations, each with many highly interacting sub-communities. The two populations mix via a single route, modeled by a single connecting edge. This “barbell graph” – two complete 20-node graphs connected by a single edge – is illustrated in Fig 4.1. In this case, we impose uniform transition rates between adjacent vertices of 0.02, 0.3, and 0.05 for the susceptible,

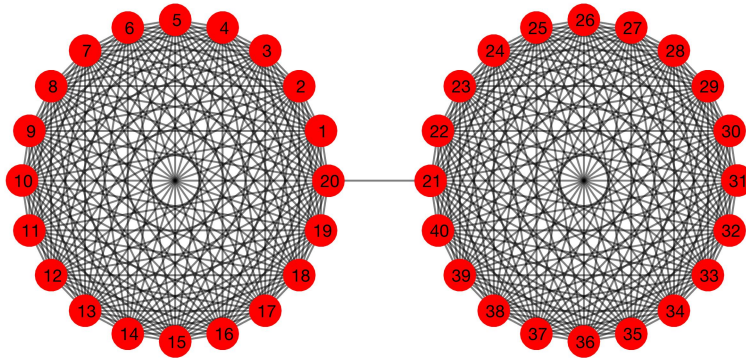


Figure 4.1: The 20-barbell graph. Two complete 20-node graphs are connected by a single edge.

infective, and removed populations, respectively. The infection begins at node 1 with the configuration $S_1(0) = 5$, $I_1(0) = 95$, $R_1(0) = 0$, and all other nodes are fully susceptible with configuration $S_i(0) = 100$, $I_i(0) = R_i(0) = 0$, $i \neq 1$.

We consider in particular the case of having information only from a limited subset of nodes; by placing the “sensors” at different locations (i.e., observing different subsets of nodes), we can test how the sensor configuration influences the parameter estimation procedure and the corresponding uncertainties. In particular, we assume observations of both the infective and removed populations at the sensor locations and at a single time point.

We consider three sensor configurations: in the first experiment, we place two sensors at nodes 3 and 12, which are members of the same complete subgraph as node 1, the origin of the epidemic (see Fig 4.1). In the second experiment, we gather data from both complete subgraphs by placing sensors at nodes 3 and 24. Finally, in the third experiment, we focus on nodes 24 and 27, which are part of the initially healthy complete subgraph. The disease evolves according to reference values $\beta = 0.02$ and $\gamma = 0.3$, while the observation time T is chosen as $T = 3$, $T = 5$, and $T = 7$

Data	Node Pair	θ_β	u_β (%)	θ_γ	u_γ (%)	θ_T	u_T (%)	θ_σ	u_σ (%)
$T = 3,$ $\sigma = 0.01\alpha$	3 and 12	1.1102	28.15	1.1152	30.62	0.9924	34.38	3.8534	51.61
	3 and 24	1.0057	11.73	0.9992	10.79	1.0059	8.47	3.0804	69.60
	24 and 27	1.1303	23.25	0.9624	24.29	0.9282	19.11	5.5323	37.56
$T = 5,$ $\sigma = 0.01\alpha$	3 and 12	0.9462	20.05	0.7806	31.86	1.2366	15.38	2.2167	102.2
	3 and 24	1.0452	15.01	1.0314	12.18	0.9788	10.86	3.9955	62.45
	24 and 27	0.9584	29.40	0.9526	26.75	1.0962	21.93	2.8993	67.70
$T = 7,$ $\sigma = 0.01\alpha$	3 and 12	1.1997	25.04	1.0390	29.32	0.9947	19.43	2.1569	81.12
	3 and 24	1.0549	17.93	1.0340	11.19	0.9758	11.53	2.8422	73.70
	24 and 27	1.3086	25.23	1.0503	35.14	0.9907	23.08	2.9131	66.75
$T = 5,$ $\sigma = 0.05\alpha$	3 and 12	1.1113	23.69	0.9226	46.58	1.1355	27.14	5.4781	41.96
	3 and 24	1.0316	11.67	1.0101	9.58	0.9814	8.95	4.1174	52.97
	24 and 27	1.1436	26.45	1.1129	24.19	0.9361	22.49	3.6762	60.00

Table 4.1: Numerical results for estimation of β , γ , T and σ for the 20-barbell graph. Blocks of the table are different times of observation or noise levels. Parameters are reported in terms of scaled values θ ; accurate estimation thus results in values close to 1. Uncertainties, e.g., u_β , are the ratio of a parameter’s standard deviation to its mean.

in three separate cases, yielding nine total experiments (three sensor configurations each observed at three different times).

Results at the intermediate time $T = 5$ are summarized in the second block of Table 4.1 and displayed in Fig 4.2a, Fig 4.2b, and Fig 4.2c. All reference values for β , γ , and T are within one standard deviation of estimated means with the exception of T when observing nodes 3 and 12 (~ 1.24 standard deviations). β and γ are positively correlated, i.e., similar outputs can be achieved by simultaneously raising both the infection rate and the rate of removal. Intuitively, a faster-spreading disease must be counteracted by quicker recovery in order for the dynamics to remain consistent. Similarly, both β and γ are negatively correlated with T ; a more infectious disease or quicker recovery would increase the speed of the system dynamics, meaning similar outputs would be observed earlier.

The recovered noise standard deviation σ has significantly larger uncertainty

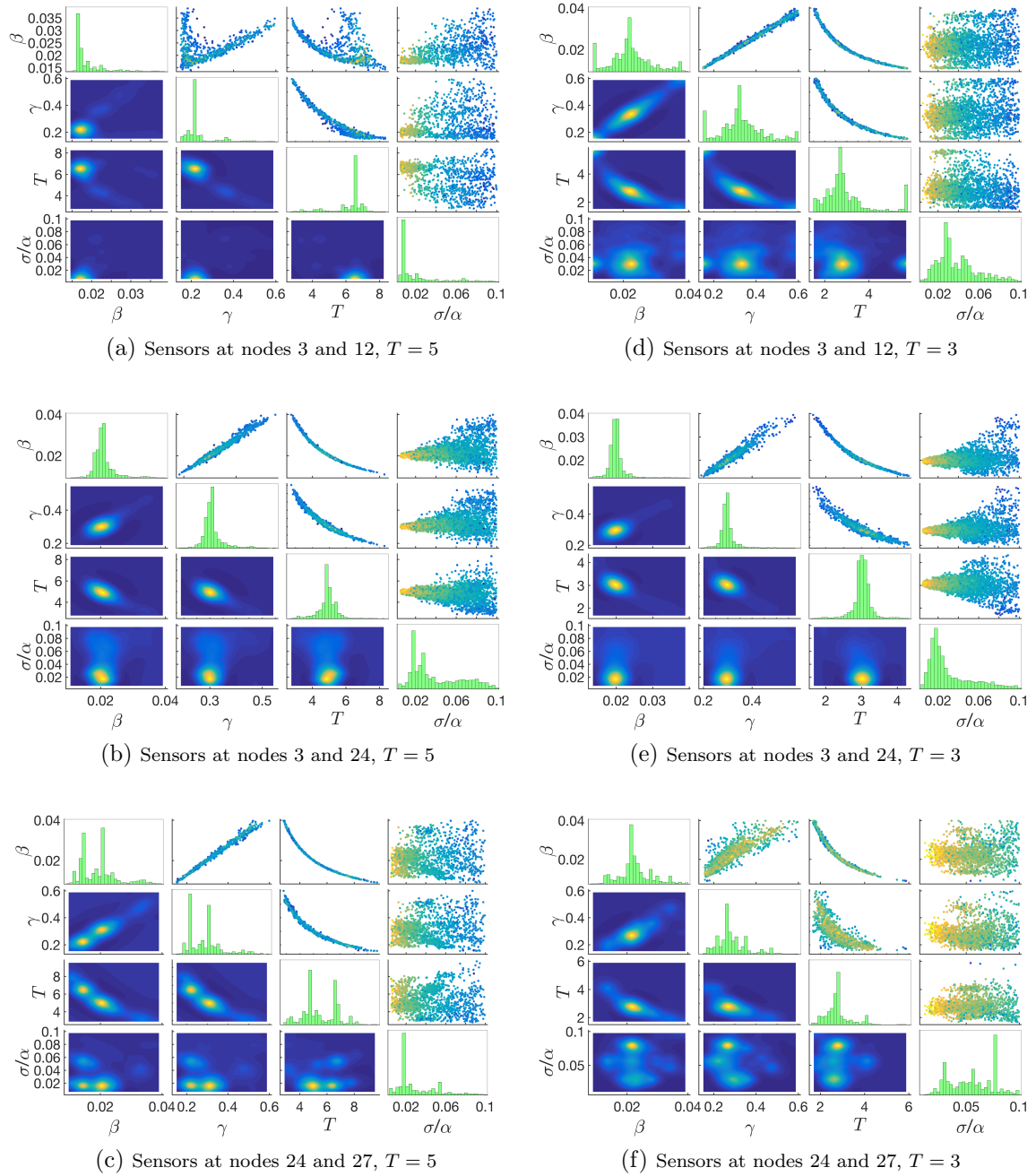


Figure 4.2: Parameter estimation results for the 20-barbell graph with infection rate $\beta = 0.02$, rate of removal $\gamma = 0.3$, noise level $\sigma = 0.01\alpha$, and time step $\Delta t = 0.005$. In each experiment, noisy data from two nodes were used to track the epidemic. For each pair of nodes, the experiment was run for time of observation $T = 5$ ((a), (b), (c)), $T = 3$ ((d), (e), (f)), and $T = 7$ (shown in Fig 4.4). Histograms for each parameter are displayed along the main diagonal of the figure. Subfigures below the diagonal show the marginal joint density functions for each pair of parameters, while subfigures above the diagonal show the samples used in the final stage of TMCMC. Colors correspond to probabilities, with yellow likely and blue unlikely.

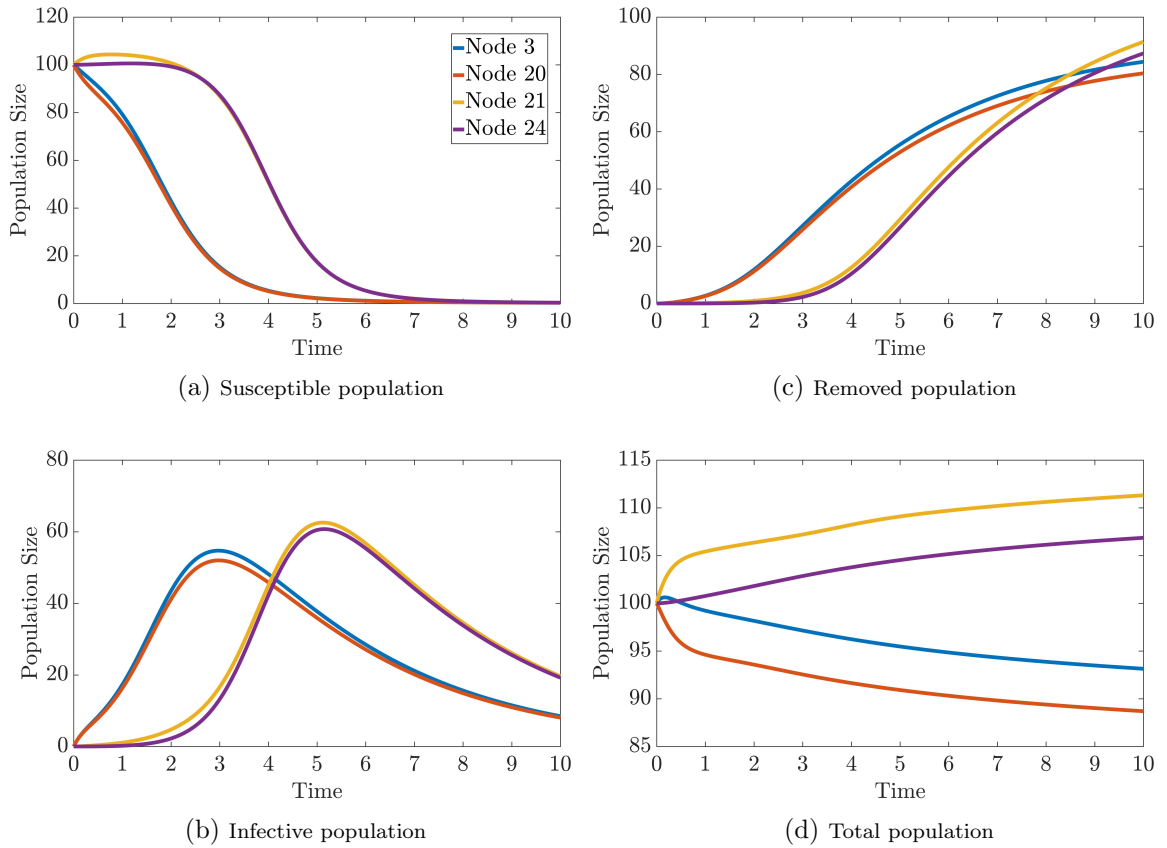


Figure 4.3: Time evolution of the susceptible, infective, removed, and total populations at nodes 3, 20, 21, and 27 of the 20-barbell graph. Nodes from different complete subgraphs have different trends and peak times.

than do the system parameters β , γ , and T . Despite this comparatively large uncertainty, the reference noise value σ is recovered to within two standard deviations for all three sensor configurations, though the posterior means θ_σ consistently overestimate the true magnitude. Uncertainties in other parameters depend strongly on sensor location; observing nodes on opposite sides of the graph yields much smaller uncertainties in β , γ , and T than observing nodes on the same side of the graph.

Fig 4.3 shows the deterministic populations of the susceptible, infected, and removed groups as a function of time for a selection of nodes involved in the experiments (since, e.g., node 12 is identical to node 3, only one is shown). Nodes 21 and 27,

both contained in the initially susceptible subgraph, reach peak infective population around time $t \approx 5$. Nodes on the side of the infection origin, conversely, achieve peak infective population at $t \approx 3$. The results of parameter estimation at the reference observation time $T = 5$ thus suggest that observing nodes around the time when the infective population peaks improves the accuracy of the recovered parameters. This effect can also be seen in Fig 4.2 in the joint marginals of the system parameters β , γ , and T : nodes 3 and 12, which peak at $t \approx 3$, give much more smeared marginals in Fig 4.2a, obtained using observations at time $T = 5$, than in Fig 4.2d, corresponding to observations at $T = 3$; meanwhile nodes 24 and 27, which peak at $t \approx 5$, have sharper distributions in Fig 4.2c, with observations taken at $T = 5$, than in Fig 4.2f, which has observations at $T = 3$. The parameter distributions obtained for data from nodes 3 and 24, shown in Fig 4.2b and Fig 4.2e, are very similar in both cases, suggesting that placing one sensor in each subgraph leverages information from both timescales.

The parameter estimation results at $T = 7$, shown in the left column of Fig 4.4, corroborate this conclusion. Though all nodes in the graph are well past peak infective population at this time, using information on two different timescales (the two subgraphs) yields much sharper joint marginals (see, e.g., the joint distribution of β and T in Fig 4.4b as compared to Fig 4.4a and Fig 4.4c).

Numerical values for $T = 3$ and $T = 7$ appear in the first and third block of Table 4.1. For both observation times, all sensor configurations recover β , γ , and T to within one standard deviation. The experiment with sensors at nodes 3 and 24 continues to be both the most accurate (in terms of posterior means) and precise (in terms of uncertainties), highlighting the importance of sensor placement in leveraging information from different timescales.

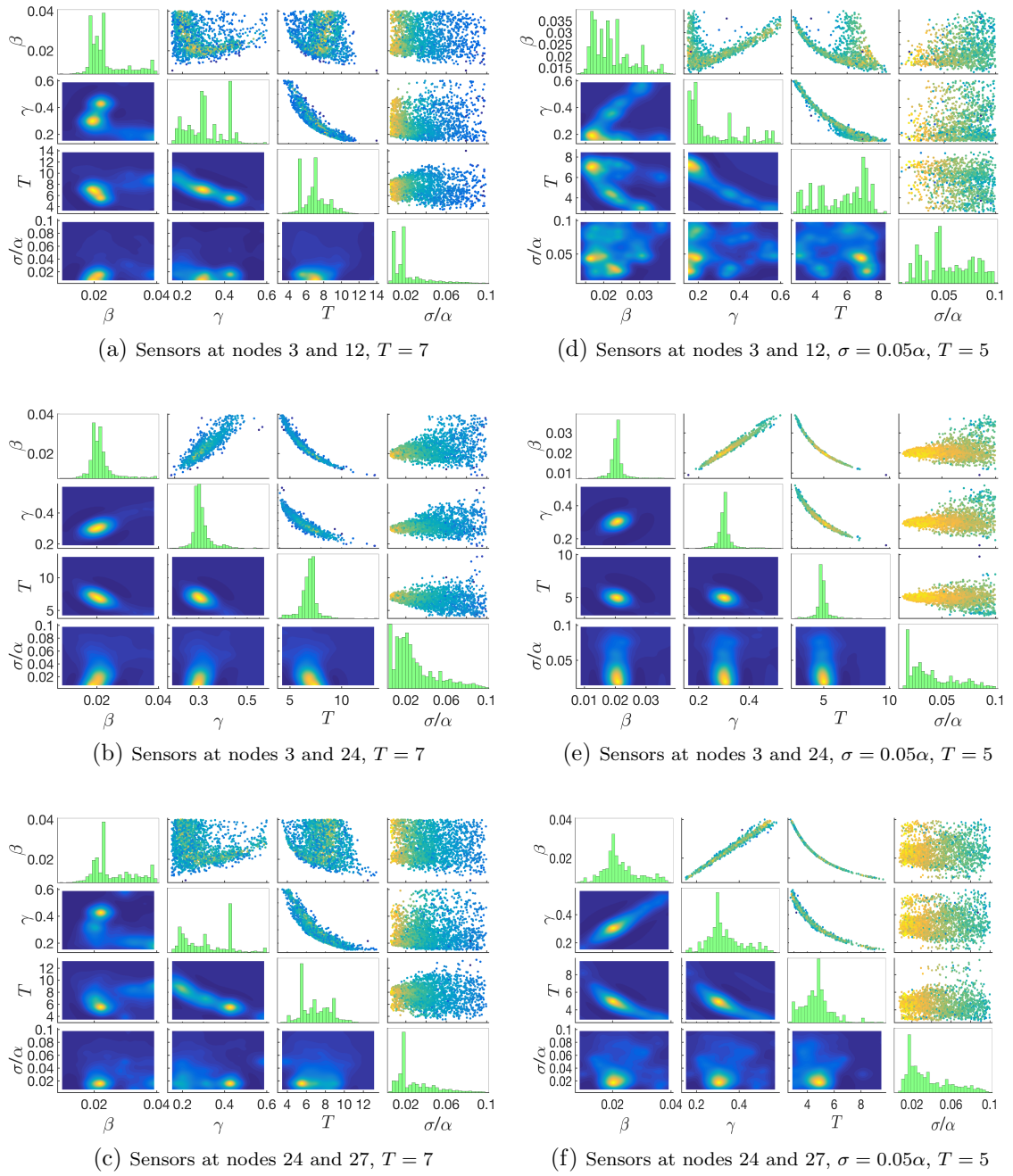


Figure 4.4: Parameter estimation results for the 20-barbell graph with infection rate $\beta = 0.02$, rate of removal $\gamma = 0.3$, time step $\Delta t = 0.005$, and noise level $\sigma = 0.01\alpha$ at time $T = 7$ ((a), (b), (c)) and with increased noise level $\sigma = 0.05\alpha$ at time $T = 5$ ((d), (e), (f)). See Fig 4.2 for description of subfigures.

To test the robustness of the parameter estimation to increased noise, we next reconsider the time $T = 5$ case with noise level $\sigma = 0.05\alpha$, five times the previously-used $\sigma = 0.01\alpha$. Results appear in the right column of Fig 4.4 and in the fourth block of Table 4.1. Again, the experiment using simulated data from nodes 3 and 24 (Fig 4.4e) recovers the parameters with comparatively lower uncertainty and greater accuracy. For all three sensor configurations, parameters are recovered to within one standard deviation, and so we conclude that the Bayesian uncertainty quantification approach to SIR models on the 20-barbell graph has significant robustness to observational noise.

A final experiment tests the approach in the context of model misspecification. We perturb the network by introducing an extra edge between nodes 2 and 40; while observations are generated from a network which includes this edge, we perform parameter estimation using the original model (without the extra edge). Results for observing nodes 3 and 24 for the $T = 3$, $\sigma = 0.01\alpha$ case with two different perturbation strengths appear in Figure 4.5 and Table 4.2. As compared to the original case (second row of Table 4.1 and Fig 4.2e), the most notable change is the significant increase in the estimated noise level – disagreements between the original model (now misspecified) and the observed data are reconciled by assuming a much higher magnitude of noise. Nonetheless, estimation of system parameters (β, γ, T) for both perturbations is reasonably successful, with all reference values recovered to within one standard deviation, though uncertainties in estimation are 3 – 4 times larger than without the perturbation.

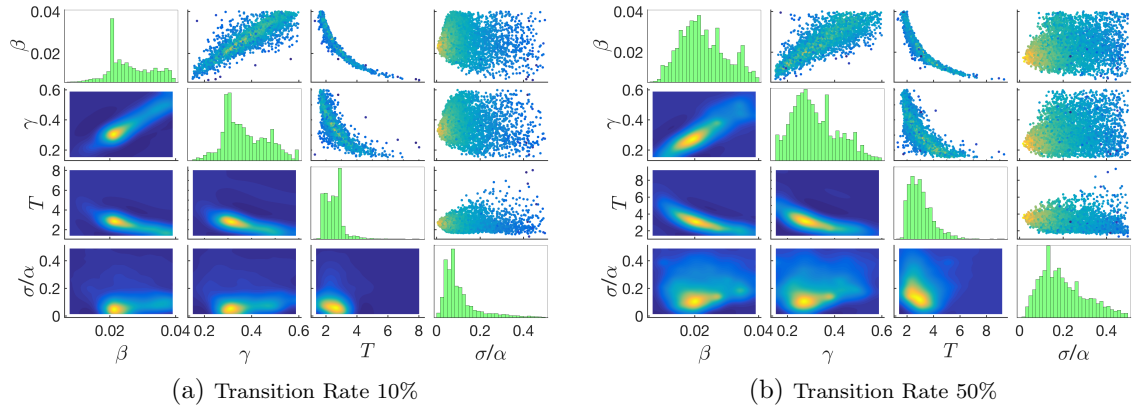


Figure 4.5: Parameter estimation results for the perturbed 20-barbell graph with infection rate $\beta = 0.02$, rate of removal $\gamma = 0.3$, time step $\Delta t = 0.005$, and noise level $\sigma = 0.01\alpha$ at time $T = 3$ using observations from nodes 3 and 24. Data are generated with an additional edge between nodes 2 and 40, which uses (a) 10% or (b) 50% of the transition rate of other edges; parameter estimation uses the original model. See Fig 4.2 for description of subfigures.

Transition Rate	θ_β	u_{θ_β} (%)	θ_γ	u_{θ_γ} (%)	θ_T	u_{θ_T} (%)	σ	u_σ (%)
10%	1.3141	24.88	1.2408	24.36	0.8371	23.39	10.4449	78.56
50%	1.1396	31.16	1.0657	30.21	1.0046	29.37	20.1990	51.89

Table 4.2: Numerical results for estimation of β , γ , T and σ for the perturbed 20-barbell graph. Data are generated with an additional edge between nodes 2 and 40, which uses the specified fraction of the transition rate of other edges; parameter estimation uses the original model. See Table 4.1 for description of values.

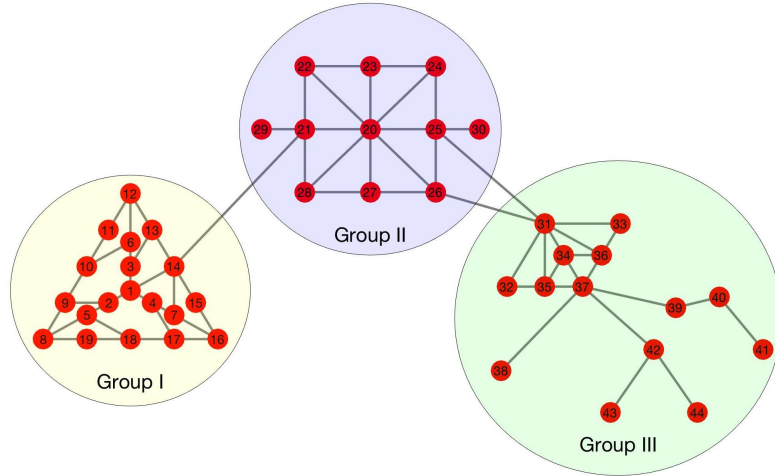


Figure 4.6: The three-group network. Group I (yellow) comprises nodes 1-19, Group II (purple) 21-30, and Group III (green) 31-44. Groups are sparsely connected.

4.3.2 Network 2: the three-group network

The second network considered is a 44-node graph comprising three large sub-networks with limited interaction (Fig 4.6). Each sub-network has a distinct topological structure and set of nonuniform transition rates (explicit values appear in Appendix 4.4). We again consider three sensor configurations: a 7-node set (nodes 1–4, 20, 31, and 34), a 23-node set (nodes 1–7, 20–28, and 31–37), and a 35-node set (nodes 1–28 and 31–37), each in the presence of observational noise $\sigma = 0.01\alpha$. Parameter estimation results for this network use only observations of the infective populations (in contrast with results for the barbell graph, which additionally used observations of the removed populations).

4.3.2.1 Parameter estimation

First, we attempt to recover β , γ , and T (reference values 0.02, 0.3, 5, respectively) for a disease which starts at node 34 with $S_{34}(0) = 5$, $I_{34}(0) = 90$, $R_{34}(0) = 5$. All

Experiment	Observed Set	θ_β	u_β (%)	θ_γ	u_γ (%)	θ_T	u_T (%)	θ_σ	u_σ (%)
1 Sample	7-node	1.0282	8.05	1.0149	6.10	0.9844	5.23	5.4603	33.95
	23-node	1.0138	0.58	1.0127	0.49	0.9910	0.39	1.0651	13.69
	35-node	0.9865	0.54	0.9903	0.47	1.0095	0.37	0.9401	10.50
2 Samples	7-node	1.0059	0.35	0.9987	0.36	0.9956	0.26	0.8247	23.18
	23-node	1.0009	0.22	1.0006	0.18	0.9992	0.17	0.9744	8.35
	35-node	0.9956	0.24	0.9972	0.20	1.0034	0.18	0.9665	6.74
3 Samples	7-node	1.0062	0.37	1.0030	0.30	0.9956	0.29	1.0531	14.32
	23-node	1.0013	0.15	1.0011	0.11	0.9989	0.12	0.9851	6.81
	35-node	0.9990	0.17	0.9992	0.13	1.0009	0.13	0.9700	5.68
4 Samples	7-node	1.0059	0.26	1.0034	0.18	0.9959	0.21	0.9451	11.98
	23-node	1.0008	0.11	1.0005	0.08	0.9993	0.09	0.9387	5.26
	35-node	0.9997	0.11	0.9998	0.09	1.0004	0.10	0.9998	4.55
Perturbation	35-node	0.9712	0.73	0.9796	0.66	1.0167	0.49	1.4183	9.57

Table 4.3: Numerical results for estimation of β , γ , and T for the three group network with $\sigma = 0.01\alpha$. 7-node set is nodes 1–4, 20, 31, and 34; 23-node set is nodes 1–7, 20–28, and 31–37; and 35-node set is nodes 1–28 and 31–37. In the case of multiple samples (blocks 2 – 4), observations are evenly spaced in time with known interval 1. Perturbation experiment generated data with an additional edge between nodes 16 and 44. See Table 4.1 for description of parameter values.

other nodes are fully susceptible, i.e., $S_i(0) = 100$, $I_i(0) = R_i(0) = 0$, for all $i \neq 34$.

Results are shown in Fig 4.7a, Fig 4.7b, Fig 4.7c, and in the first block of Table 4.3. Sensors at the 7-node subset recover β , γ , and T to within one standard deviation, while larger subsets recover all parameters with comparable accuracy and much greater precision: the 7-node 95% confidence interval for the scaled infection rate θ_β is $[0.8627, 1.1937]$, which narrows significantly to $[1.0020, 1.0256]$ in the 23-node case. Increasing the number of observed nodes has the additional effect of accurately recovering the scaled noise level θ_σ , which is significantly overestimated when observing only the 7-node set.

Many real epidemic data sets include multiple observations over time; it is worth verifying that our approach can reasonably incorporate such observations. T (time from infection to first sample) remains the only unknown time in this setting, as the

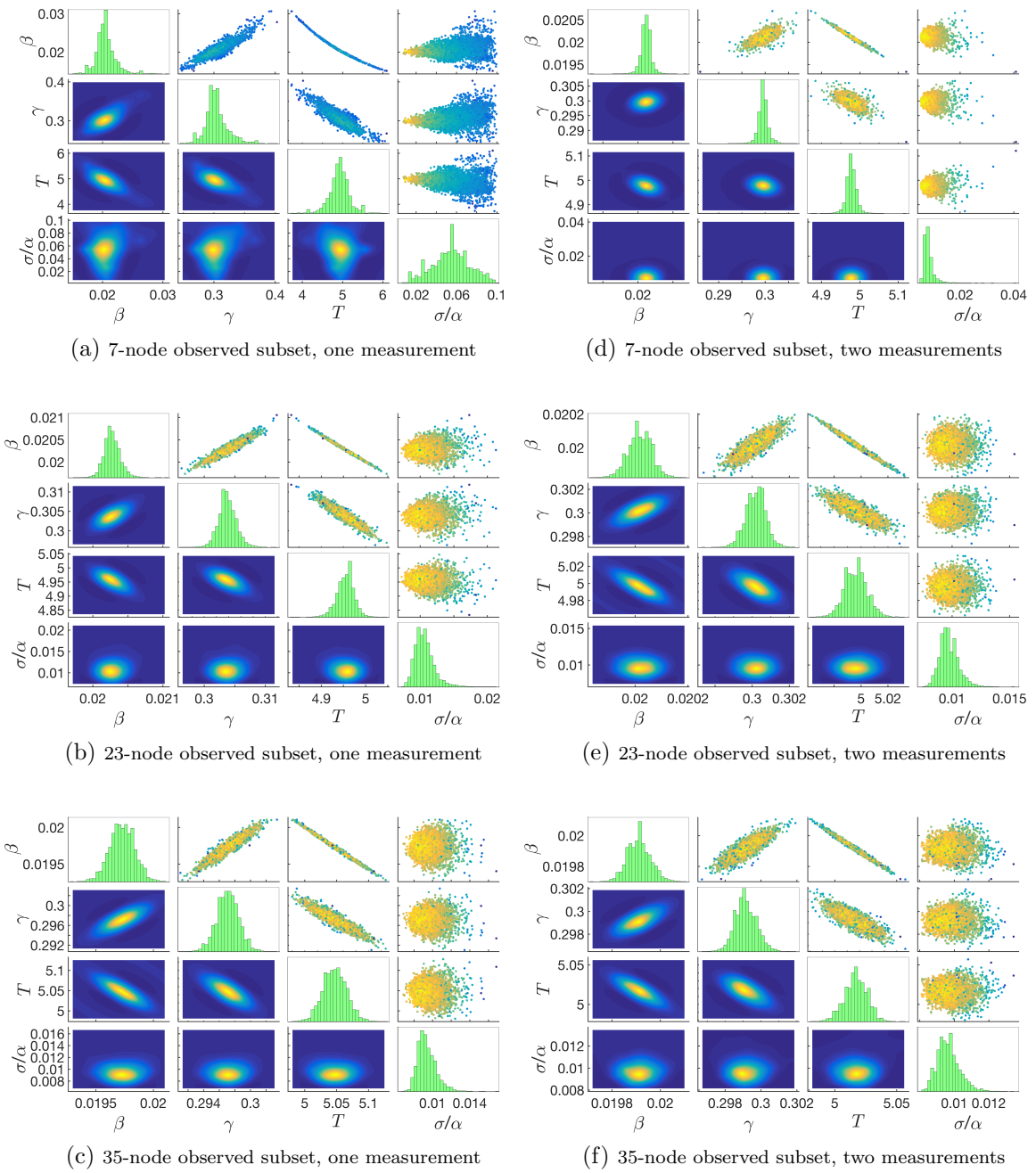


Figure 4.7: Parameter estimation results for the three group network. ((a), (b), (c)) Parameter estimation for β , γ , T based on information from 7-node, 23-node and 35-node subsets, respectively. Reference values are $\beta = 0.02$, $\gamma = 0.3$, and $T = 5$, with time step $\Delta t = 0.0005$. ((d), (e), (f)) Same experiment using two observations from each sensor separated by an interval $t = 1$. See Fig 4.2 for description of subfigures.

relative timing between samples is known. Blocks 2–4 of Table 4.3, the right column of Fig 4.7, and the left column of Fig 4.8 present results when considering additional samples taken at evenly spaced intervals of length 1. Including a second observation yields a significant reduction in estimated uncertainties, especially when observing only 7 nodes (uncertainties for β , γ , T , and σ are reduced by factors of 23.51, 17.22, 19.89, and 9.70, respectively). Comparatively, additional samples beyond the second provide only marginal benefit (when observing the 7-node set, moving from 2 to 4 samples reduces uncertainties for β , γ , T , and σ by factors of 1.35, 1.99, 1.24, and 1.69, respectively). We additionally note that observing multiple samples over time can reduce the correlation between parameter uncertainties (see, e.g., the highly-correlated posterior of β and γ in Fig 4.7a as compared to the nearly uncorrelated ellipse of Fig 4.8a).

As with the 20-barbell graph, we also consider model misspecification via perturbation of the network. The final row of Table 4.3 presents parameter estimation results when observations at the 35-node subset are generated using an additional edge connecting nodes 16 and 44 (and thus Groups I and III); transition rates along the additional edge are chosen to match those of the edges connecting Groups II and III. As before, uncertainties for system parameter estimates (β , γ , and T) are larger than without the perturbation (by factors of 1.33, 1.39, and 1.33, respectively), though recovered parameter values are nonetheless accurate (scaled values off by less than 0.03), and so parameter estimation in this setting remains successful. σ is again overestimated, as it must additionally account for disagreement between the model used to generate the data and the model used to perform parameter estimation.

We next augment the parameter set θ with the initial population vector S_0 , I_0 , and R_0 of the initially infected node, but take as known the observation time $T = 5$. The initial population vector at node 34 is altered to $S_{34}(0) = 5$, $I_{34}(0) = 90$,

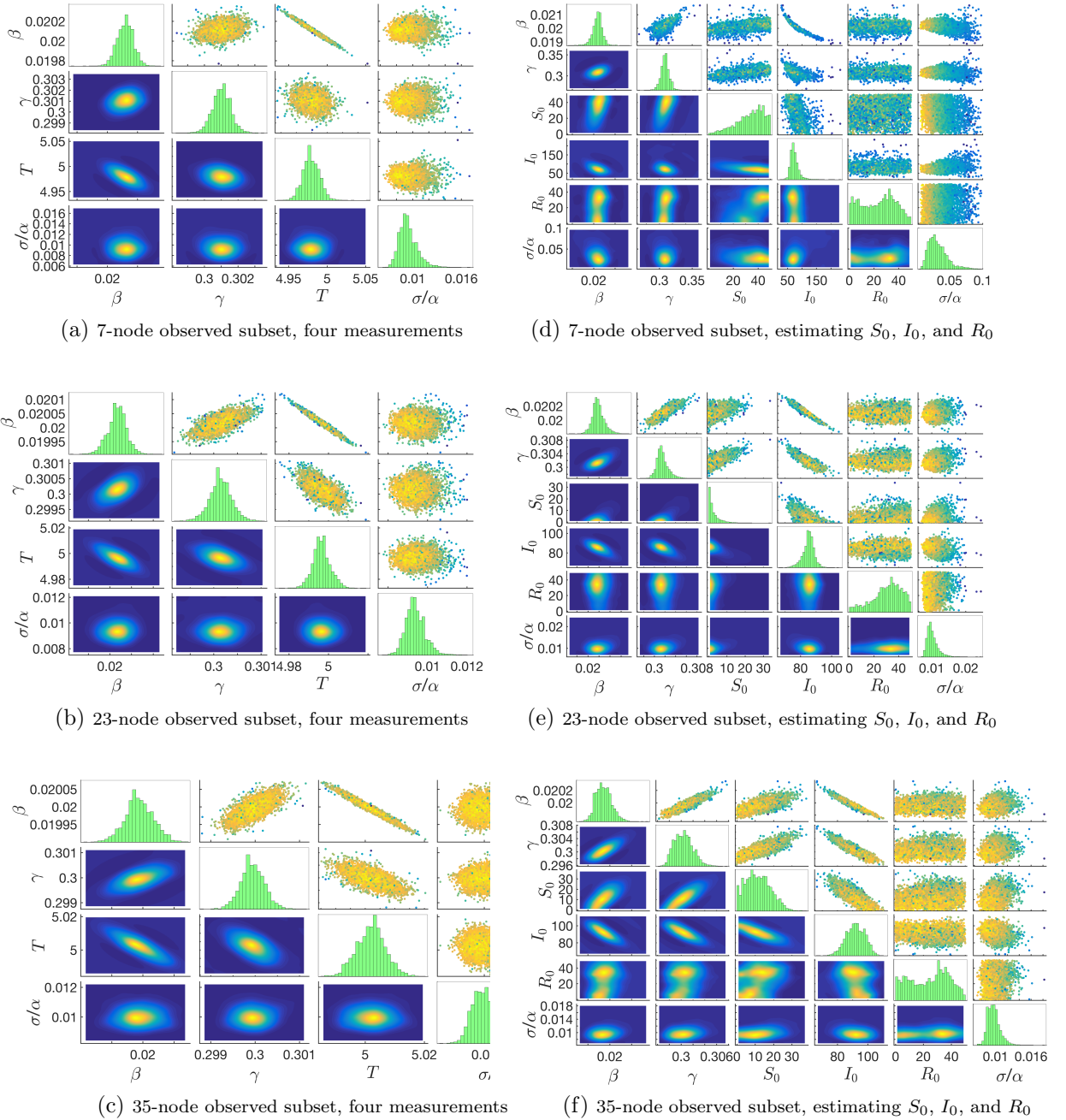


Figure 4.8: Additional parameter estimation results for the three group network. ((a), (b), (c)) Estimating β , γ , T using four measurements at intervals $t = 1$. Reference values are again $\beta = 0.02$, $\gamma = 0.3$, and $T = 5$, with time step $\Delta t = 0.0005$. ((d), (e), (f)) Estimating β , γ , S_0 , I_0 , and R_0 with one measurement at known time $T = 5$. Reference values are $\beta = 0.02$, $\gamma = 0.3$, $S_0 = 5$, $I_0 = 90$, and $R_0 = 5$, with time step $\Delta t = 0.02$. See Fig 4.2 for description of subfigures.

Observed Set	θ_β	u_β (%)	θ_γ	u_γ (%)	θ_σ	u_σ (%)
7-node	1.0092	1.16	1.0289	2.12	3.1965	45.40
23-node	1.0047	0.27	1.0051	0.36	1.0214	15.73
35-node	0.9973	0.36	1.0014	0.52	0.9678	10.87
Observed Set	θ_{S_0}	u_{S_0} (%)	θ_{I_0}	u_{I_0} (%)	θ_{R_0}	u_{R_0} (%)
7-node	6.6752	33.64	0.8206	20.85	4.7458	57.01
23-node	0.5463	114.80	0.9455	4.28	6.1614	37.56
35-node	2.3354	57.14	1.0212	6.64	4.9126	55.53

Table 4.4: Numerical results for estimation of β , γ , S_0 , I_0 , and R_0 for the three group network with $\sigma = 0.01\alpha$. See Table 4.1 for description of values.

$R_{34}(0) = 5$. The scaled parameter set $(\theta_\beta, \theta_\gamma, \theta_{S_0}, \theta_{I_0}, \theta_{R_0}, \theta_\sigma)$ uses a uniform prior on $[0.02, 2] \times [0.02, 2] \times [0, 10] \times [0, 10] \times [0, 10] \times [0.01, 10]$.

Results are shown in Fig 4.8d, Fig 4.8e, and Fig 4.8f, and appear numerically in Table 4.4. Compared to estimation of β , γ , and T , correlations between parameters are generally weaker in this context, though there do exist clear relationships (e.g., larger I_0 necessitates smaller β for the infection to spread at the same absolute rate). All sensor configurations recover the disease parameters β and γ accurately, with scaled values off by less than 10^{-2} in all but one case. Conversely, uncertainty in the recovered distributions for S_0 and R_0 is significantly higher than for other parameters, owing to the small ground-truth values for these parameters relative to the total population at the origin ($S_0 = R_0 = 5$ out of 100 individuals at time $t = 0$) and their correspondingly minor effect on the behavior of the initial outbreak.

4.3.2.2 Origin of disease identification

Finally, we introduce a method for probabilistically identifying the origin of the epidemic using the Bayesian model selection framework described in the Bayesian methodology section; recall that all observations are at the future time T , and so the

origin may not be identified with certainty even when included in the set of observed nodes. We initialize the disease at node 1 with the standard initial configuration $S_1(0) = 5$, $I_1(0) = 95$ and $R_1(0) = 0$, with all other nodes fully susceptible with $S_i(0) = 100$, $I_i(0) = R_i(0) = 0$, and use corrupted observations of infective and removed populations from the 35-node subset. Other parameter values are identical to those used in the previous section; in this case, $T = 5$ is assumed to be known, with β and γ estimated from observations. Defining the model class M_j as the model under which the disease originated from node j with the given initial vector, the log evidence for each model is generated from the model selection framework (see Equation 2.1.2).

A representative selection of results appear in Table 4.5. Model M_1 , corresponding to the correct origin of the disease at node 1, has a significantly larger log evidence than all other models considered. Models which place the origin at increasingly distant points generate increasingly less accurate and more uncertain results; models M_{32} and M_{43} , which originate the disease in Group III, find the estimated noise σ to be two orders of magnitude larger than the reference value. Out of the models shown, if the correct model M_1 is not considered, the probabilities shift to 0.003 for M_8 , 0.997 for M_{14} , and all other probabilities $\approx 10^{-13}$ or smaller, suggesting that topographic proximity to the true origin is the dominant factor in the evidence.

The effect of topographic proximity on the model evidence appears visually in Fig 4.9. Models M_i , $i = 1, \dots, 44$ correspond to the epidemic beginning at node i ; each node in the graph is colored by the estimated log evidence ($\log \rho(\mathbf{D}|M_i)$), where $\mathbf{D}$ are the noisy data obtained from a reference simulation beginning at node 1. In order that the fine detail be more visible, the color mapping is nonuniform such that node 1 (log evidence -8.15) is the only node in its color bin; other bins from log evidence -360 to -250 capture the range of behavior in the remaining nodes. Evidence

Model	Log Evidence	$Pr(M_j D)$	θ_β	u_β (%)	θ_γ	u_γ (%)	θ_σ	u_σ (%)
M_1	-8.1478	1.00	0.9998	0.03	1.0000	0.07	0.9628	9.33
M_8	-296.0472	~ 0.00	1.7900	2.86	1.1537	7.08	108.4015	11.32
M_{14}	-290.3494	~ 0.00	0.9397	2.20	0.9414	5.45	81.8066	9.31
M_{21}	-319.3832	~ 0.00	0.9233	5.56	0.9574	10.74	117.0764	11.07
M_{32}	-340.3117	~ 0.00	1.9765	1.40	0.9497	10.92	182.1469	6.32
M_{43}	-359.4945	~ 0.00	1.7543	10.10	1.4097	22.35	184.1593	5.92

Table 4.5: Subset of model selection results for the three-group network. Recovered scaled parameters θ and uncertainties u appear with the estimated log evidence for the model.

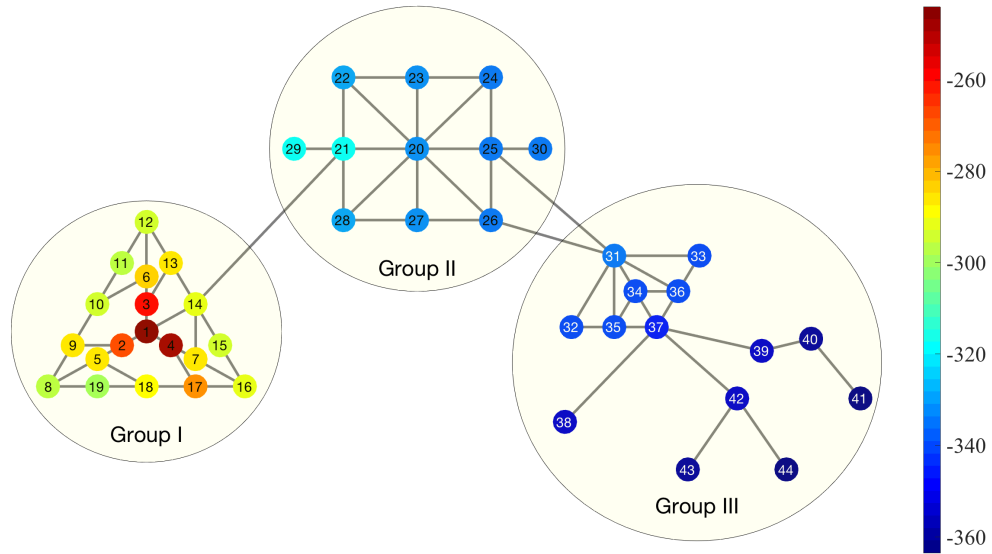


Figure 4.9: Origin identification model evidence by node, three-group network. Color map of model selection log evidence for models initializing the epidemic at each node in the three-group network using noisy data from an epidemic simulation beginning at node 1. Nonuniform color mapping (right label) emphasizes differences among incorrect models (log evidence -360 to -250); log evidence for the correct origin (node 1) is notably larger (log evidence -8.15).

Model	Log Evidence	$Pr(M_j D)$	θ_β	u_β (%)	θ_γ	u_γ (%)	θ_σ	u_σ (%)
M_1	-193.3884	1.00	1.0115	4.89	1.0408	4.35	12.7405	6.97
M_8	-299.4659	~ 0	0.9353	10.70	0.6534	9.57	48.6215	2.13
M_{14}	-291.7504	~ 0	0.6533	26.55	0.7343	19.09	47.9977	3.47
M_{21}	-320.8051	~ 0	0.3025	8.38	0.5136	3.33	49.3168	0.88
M_{32}	-396.6718	~ 0	1.9126	2.61	0.5041	1.05	49.8335	0.38
M_{43}	-390.3285	~ 0	1.2828	29.34	0.6798	54.33	49.6430	0.29

Table 4.6: Subset of model selection results for the perturbed three-group network. Data are generated with an additional edge between nodes 16 and 44; parameter estimation uses the original model. Recovered scaled parameters θ and uncertainties u appear with the estimated log evidence for the model.

decays with topographic distance within the first group and becomes negligible in the second and third groups, highlighting the improbability of the epidemic starting at these distant nodes and producing the noisy observations $\mathbf{D}$ corresponding to an epidemic outbreak at node 1.

Lastly, we attempt origin identification via the same model selection approach for the perturbed model which generates data using an additional edge connecting nodes 16 and 44; results appear in Table 4.6. Despite the misspecification, the correct model M_1 is selected with near certainty, though its log evidence is many orders of magnitude smaller than in the no-perturbation experiment (-193.3884 vs. -8.1478) owing to the high level of noise required to explain disagreements with the observed data. Models M_{32} and M_{43} , both located in Group III, are also subject to significant drops in log evidence. A second notable difference is the reduced level of estimated noise for incorrect models M_j – the additional edge allows the disease to quickly reach nodes which are distant under the original model, thereby requiring less noise to explain discrepancies with the observed data. Overall, selection among origin models appears robust to perturbation.

4.4 Discussion

We found that Bayesian uncertainty quantification via TMCMC effectively recovered SIR network model parameters, such as the infection rate β and rate of removal γ , using only noisy observations from a limited set of nodes. The approach was tested on two example networks with distinct topologies, with two possible sets of observed data (infective populations vs. both infective and removed populations), using different sets of free parameters, and in a number of additional contexts (in particular, with model perturbations and with multiple observations at each node). Given its explicit estimation of parameter uncertainty, the framework permits comparison of precision in distinct scenarios – for example, uncertainties in recovered parameter values were found to be inversely related to the number of sensors (for the three-group network, increasing the number of sensors from 7 to 23 decreased parameter uncertainties by a factor of 12 – 14).

The 20-barbell graph, wherein only pairs of nodes were observed, provided additional insight into the effect of sensor placement on the uncertainty of parameter estimates. Sensors which were close together, e.g., nodes 3 and 12 (whose connectivity is identical), yielded similar noisy data, thereby affording less information about the underlying dynamics. In contrast, placing sensors on opposite sides of the network to gain information about dynamics on different time scales yielded significantly less uncertainty. Given limited resources for monitoring an epidemic, it may thus be beneficial to track a set of communities which are at varying stages of outbreak rather than allocating resources directly to those communities nearest to which the disease was initially observed.

Our approach proved robust both to perturbations in the model and to increased

observational noise. Results for the 20-barbell graph with increased noise level $\sigma = 0.05\alpha$, five times the original $\sigma = 0.01\alpha$, recovered reference parameters with comparable accuracy. For the three-group network, we were also able to identify the disease origin with near certainty via selection among models corresponding to potential starting points, even when data were generated from a distinct model with an additional edge; model selection thus has the potential to locate real outbreaks even when observations begin well after the time of infection and the network structure is not known exactly. We remark that the evidence calculation required for this procedure is an intermediate step of the Bayesian uncertainty quantification framework, and so it does not incur any additional computational cost.

Appendix: Transition matrices for three-group network

Transition rates were chosen to vary among groups in the three-group network in order to test the robustness of our approach to nonuniform rates. Populations moved between Group I and Group II (via the edge connecting node 14 to node 21) at a rate of 0.4, while the transition rate between Group II and Group III (via the edges connecting node 31 to nodes 25 and 26) was 0.2. Rates $(\lambda_{i,j}, \eta_{i,j}, g_{i,j})$ within Group I were selected randomly to be either (0.2, 0.05, 0.1) or (0.1, 0.1, 0.2), those for Group II were selected randomly to be either (0.15, 0.2, 0.1) or (0.3, 0.1, 0.1), and Group III had uniform rates of 0.05.

CHAPTER FIVE

**Data-driven selection and
parameter estimation for DNA
methylation mathematical models**

5.1 Introduction

DNA methylation has a pivotal epigenetic role to play during embryonic development (97). The covalent bonding of methyl groups to DNA serves to regulate gene expression during this period and is a process which culminates with the formation of tissue specific methylation patterns. However, during aging, mammalian methylation patterns change. Aging is synonymous with genome wide hypomethylation, whilst, paradoxically, it is associated with regional increases in DNA methylation, most notably at the promoter region of a diverse array of genes (7).

Intriguingly, several disease processes display similar characteristics. Specifically, cancers invariably show global hypomethylation and gene specific hypermethylation, while autoimmune diseases routinely exhibit hypomethylation both globally and on specific genes (8). Moreover, changes to genomic methylation patterns with age have a burgeoning role to play in cardiovascular disease (98), Alzheimer's disease (9), and osteoporosis/osteoarthritis (99). Thus, it is clear the dysregulation of this fundamental epigenetic process is vital to a variety of age related pathologies and potentially aging. In order to identify why an increase in age results in aberrant DNA methylation, it is necessary to understand the molecular mechanisms which govern this biochemical system (100, 101). Additionally, it is imperative to appreciate how the dynamics of this molecular system change with age.

DNA methylation occurs in mammals primarily at CpG dyads; more specifically, the methyl group is attached to the fifth carbon of the cytosine at the CpG site (Cytosine - Guanine dinucleotide sequence separated by a phosphate group). Within the vertebrate genome, global methylation can be defined by CpG islands (CGIs). These are genomic regions which comprise 1000 base pairs and consist of high levels of G+C

base levels. In addition, they are characterized by a deficiency in DNA methylation (102). Although CGI are scantily decorated throughout the genome, their biological imperative has been getting more notice in recent years. Chiefly, CGIs are sites of transcriptional initiation and thus act as promoters in mammalian genomes. Consequently, any change in the methylation status of a CGI will potentially effect gene transcription, and this is exactly what happens as hypermethylation of CGIs are routinely correlated with the transcriptional silencing of gene promoters, a phenomenon which is often a feature of diseases such as cancer (103). Besides, increasing age has been correlated with the hypermethylation of a wide variety of gene promoters belonging to genes which have been associated with aging (104). Consequently, it is clear from the above discussion aging has a profound effect on the dynamics of DNA methylation and age related changes to processes which control the reactions which govern DNA methylation ultimately drive the formation of aberrant DNA methylation.

This biological system remains to be fully delineated; what is known is that it is characterized by the activities of several enzymes (105). The enzymes operate as follows: post replicatively, new CpG dinucleotides are attached to the complementary strand of the daughter cells, which are unmethylated. DNA methyltransferase (DNMT1) then uses S-Adenosyl methionine as a substrate to transfer methyl groups to the DNA molecule (106). As DNMT1 preferentially acts on hemimethylated DNA it is thought to be chiefly a maintenance enzyme (107). Therefore, other enzymes are a necessity for *de novo* DNA methylation. Current thinking suggests DNMT3a and DNMT3b are the enzymes which perform this task. Enzymatic maintenance and *de novo* methylation reactions are in turn counterbalanced by passive and active demethylation (108). Passive demethylation usually occurs as a result of replication and DNA methylation levels can decrease after several rounds of this process

(109). On the other hand it is suggested active methylation requires Ten-Eleven Translocation (TET) dioxygenases, which oxidize the methyl groups of cytosine; a process which eventually results in the reincorporation of an unmethylated cytosine into DNA (110). As a result the maintenance of DNA methylation levels can be viewed as a subtle balancing act between maintenance/*de novo* methylation and passive/active demethylation.

In recent years mathematical models have been used as effective tools to help advance our understanding of the dynamics which underpin DNA methylation – reviewed in (111). In this chapter, we present linear and nonlinear mathematical models which encapsulate the molecular mechanisms which define DNA methylation (112). In addition, we use our recently developed Bayesian algorithm for estimating the parameters of a model and, furthermore, select the model that best fits given DNA methylation data. The Bayesian parameter estimation allows us to leverage prior knowledge of the methylation rates to guide the search in the parameter space. To test the viability of using parallel transitional Markov chain Monte Carlo (TMCMC) for the DNA methylation problem, parameters are estimated using noisy model observations generated from two potential models. Furthermore, the sampling algorithm used allows for model selection without any additional computational resources. With this in mind, we will use the algorithm to identify which model best matches noisy model observations and will test which model is most biologically feasible. The model selected as “best” is the one that most closely fits the biological system and discovers correlations that match with the underlying biological mechanisms.

5.2 Models and Methods

5.2.1 DNA methylation models

Based on the approach in (113), three types of population are considered; unmethylated CpG dyads, the total number of which is denoted as $x_1(t)$, hemimethylated CpG dyads, $x_2(t)$, and methylated CpG dyads, as $x_3(t)$, see Fig. 5.1. An unmethylated CpG dyad is a CpG dyad with neither of the two CpG sites methylated. Similarly, a hemimethylated CpG dyad has only one methylated CpG site and the opposing unmethylated, and a methylated CpG dyad has both opposing sites methylated. The methylation enzymes DNMT1, DNMT3a, and DNMT3b, TET family of demethylation enzymes, and DNA replication are responsible for the transitions between the possible states of CpG dyads. The methylation rates of unmethylated CpG dyads and hemimethylated CpG dyads are k_1 and k_2 , respectively, and the demethylation rates of hemimethylated and methylated CpG dyads are k_3 and k_4 , respectively. D denotes the rate of cell division, see Fig. 5.2.

Inspired by (114) the mechanism behind DNA division is described as follows. Unmethylated DNA strands bond with the parental strands during DNA replication. Therefore, all parental methylated CpG dyads form hemimethylated CpG dyads in the daughter cells (115). The hemimethylated CpG dyads in the parental cell either become unmethylated or remain hemimethylated in the daughter cells. Thus, we assume that half of the parental hemimethylated CpG dyads become unmethylated in the daughter cells and the other half remain hemimethylated. Unmethylated dyads remain unmethylated.

The above biological mechanisms can be translated in a set of ordinary differential

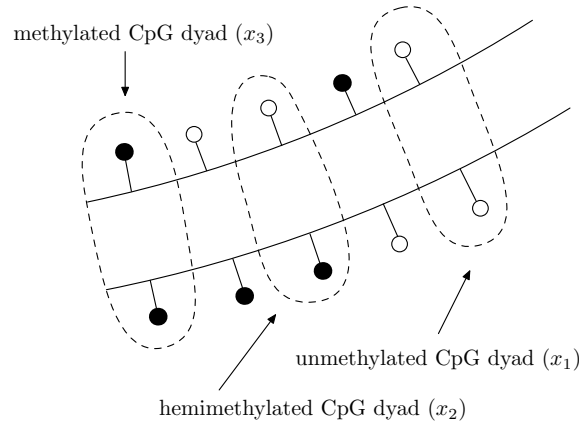


Figure 5.1: The three different states of a CpG dyad; unmethylated (x_1), hemimethylated (x_2) and methylated (x_3) CpG dyads. A white circle denotes an unmethylated CpG site whereas a black circle represents a methylated CpG site. An unmethylated (x_1) CpG dyad consists of two unmethylated opposite CpG sites, a hemimethylated dyad (x_2) has only one of the two sites methylated and a methylated dyad (x_3) has both opposing sites methylated.

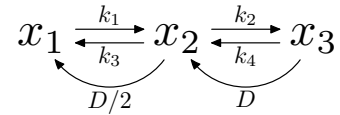


Figure 5.2: The diagram for the methylation rates between x_1 , x_2 and x_3 .

equations (ODEs) as follows, see (112),

$$\frac{dx_1(t)}{dt} = -k_1x_1(t) + \left(k_3 + \frac{1}{2}D\right)x_2(t) \quad (5.2.1)$$

$$\frac{dx_2(t)}{dt} = k_1x_1(t) - \left(k_2 + k_3 + \frac{1}{2}D\right)x_2(t) + \left(k_4 + D\right)x_3(t) \quad (5.2.2)$$

$$\frac{dx_3(t)}{dt} = k_2x_2(t) - \left(k_4 + D\right)x_3(t). \quad (5.2.3)$$

It is important to note the difference between D and our various rates k_i . D corresponds to the cell division rate, while the model parameters correspond to the methylation rates between the various states of the CpG dyads. Since these have

separate biological functions, we include both in the model and do not collapse them to a single parameter. Gene promoters are regions of interest in terms of DNA methylation levels. Monitoring the evolution of the populations of CpG sites in gene promoters shows that the total number of CpG dyads has to be constant. We assume that $x_1(t) + x_2(t) + x_3(t) = C$, with $C > 0$. Substituting the above equation into the set of equations (5.2.1)-(5.2.3), we deduce the equivalent non-homogeneous system

$$\begin{aligned}\frac{dx_1(t)}{dt} &= -\left(k_1 + k_3 + \frac{1}{2}D\right)x_1(t) - \left(k_3 + \frac{1}{2}D\right)x_3(t) + C\left(k_3 + \frac{1}{2}D\right) \\ \frac{dx_3(t)}{dt} &= -k_2x_1(t) - \left(k_2 + k_4 + D\right)x_3(t) + Ck_2 \\ x_2(t) &= C - x_1(t) - x_3(t).\end{aligned}$$

Depending on the specific type of tissue, methylation levels in gene promoters may significantly vary (116). There are two different patterns observed in the DNA methylation levels in gene promoters; hypomethylated and hypermethylated (117). In a hypomethylated location, the vast majority of the CpG sites are unmethylated. On the contrary, a hypermethylated region consists mainly of methylated CpG sites. Thus, it is necessary to introduce some nonlinear terms to obtain the observed behaviour, namely, the bistable state of gene promoters. A key question is developing the most appropriate nonlinear model that will give the expected behaviour with respect to the data collected.

From studying the biological mechanisms of the system, there are a few ways to account for a potential transition between two states. If there is an abundance of unmethylated CpG dyads (x_1) and the gene promoter is hypomethylated, it is reasonable to assume that with time unmethylated CpG dyads will become methy-

lated. Biologically this could happen as a result of fluctuating levels of DNMT3a and DNMT3b (denoted in the model by an increase in k_1 rate). As the number of unmethylated CpG dyads (x_1) drops, the methylation rate k_1 increases. While the level of unmethylated dyads decreases, the number of hemimethylated dyads (x_2) increases. This can be interpreted as k_1 being a decreasing function of $x_1(t)$ or an increasing function of $x_2(t)$. It remains unknown if the transition between the two different states is due to an increase in the *de novo* methylation enzymes (DNMT3a and DNMT3b) or whether it is due to a decrease in the demethylation enzymes (the TET protein family). To account for the latter, it can be assumed that as the number of unmethylated CpG dyads (x_1) decreases, the demethylation rate k_3 drops, due to a change in TET enzyme activity and consequently the number of hemimethylated CpG dyads increases. This can be described by denoting the rate k_3 as an increasing function of $x_1(t)$ or a decreasing function of $x_2(t)$. A similar premise can be suggested for the transition rates k_2 and k_4 . A significant rise in methylated CpG dyads can be a result of an increase in DNMT1 maintenance levels or due to a decrease in the TET enzymes, namely either a k_2 increase or a k_4 drop. Thus, it is logical that k_2 can be a decreasing function of x_2 or an increasing function of x_3 and, k_4 can be an increasing function of x_2 or a decreasing function of x_3 .

Following the biological assumptions made above, there are two plausible approaches of representing the transition rates as functions of the CpG dyads. These expressions can be considered either in terms of x_2 or in terms of x_1 and x_3 , as follows. In the equations below, $\nearrow$ represents that the function that increases with respect to x_1 , x_2 , or x_3 , and $\searrow$ represents a that the function decreases with respect to x_1 , x_2 , or x_3 .

$$k_1(x_2) = k_{11} + k_{12}x_2^{\gamma_1}(\nearrow),$$

$$k_2(x_2) = k_{21} + k_{22}x_2^{\gamma_2}(\searrow),$$

$$k_3(x_2) = k_{31} + k_{32}x_2^{\gamma_3}(\searrow),$$

$$k_4(x_2) = k_{41} + k_{42}x_2^{\gamma_4}(\nearrow),$$

or

$$k_1(x_1) = k_{11} + k_{12}x_1^{\gamma_1}(\searrow),$$

$$k_2(x_3) = k_{21} + k_{22}x_3^{\gamma_2}(\nearrow),$$

$$k_3(x_1) = k_{31} + k_{32}x_1^{\gamma_3}(\nearrow),$$

$$k_4(x_3) = k_{41} + k_{42}x_3^{\gamma_4}(\searrow),$$

where $k_j(x_i) > 0$ and $\gamma_i \in \mathbb{R}$. We assume here that $\gamma_i = 2$, since these reactions are akin to second order kinetics which are common in biochemical systems. The arrow next to each formula denotes an increasing or decreasing transition function of the populations x_i , as biology dictates. In our previous work (112), we selected methylation rates as functions of x_1 and x_3 . Therefore, the following system is obtained

$$\begin{aligned} \frac{dx_1(t)}{dt} &= -A_1(x_1(t))x_1(t) - A_2(x_1(t))x_3(t) + A_3(x_1(t))C \\ \frac{dx_3(t)}{dt} &= -B_1(x_3(t))x_1(t) - B_2(x_3(t))x_3(t) + B_3(x_3(t))C \\ x_2(t) &= C - x_1(t) - x_3(t), \end{aligned}$$

where

$$\begin{aligned}
A_1(x_1(t)) &= k_{11} - k_{12}x_1^2(t) + k_{31} + k_{32}x_1^2(t) + \frac{1}{2}D, \\
A_2(x_1(t)) &= k_{31} + k_{32}x_1^2(t) + \frac{1}{2}D, \\
A_3(x_1(t)) &= k_{31} + k_{32}x_1^2(t) + \frac{1}{2}D, \\
B_1(x_3(t)) &= k_{21} + k_{22}x_3^2(t), \\
B_2(x_3(t)) &= k_{21} + k_{22}x_3^2(t) + k_{41} - k_{42}x_3^2(t) + D, \\
B_3(x_3(t)) &= k_{21} + k_{22}x_3^2(t),
\end{aligned}$$

see (112). The results of our model corroborate experimental work which has investigated the epigenetic nature of gene promoters (118). Moreover, sensitivity analysis was able to suggest which parameters were vulnerable to small perturbations. However, given the uncertainty which surrounds these parameter values generally and the lack of quantitative biological information, it is necessary to consolidate our findings. One approach to these problems is through utilizing Bayesian inference.

5.3 Results

We apply $\Pi 4U$ to the nonlinear DNA methylation model described earlier. The four-stage Runge-Kutta method was used to generate the model outputs $\mathbf{g}$ in the model prediction equation. The model outputs considered are the time-series evolution for x_1 , x_2 , and x_3 from $t = 0$ to $t = 10$, with step size $\Delta t = 0.0001$ and every hundredth step recorded, resulting in 300 sample points. Measured data are simulated by computing all outputs using a reference model and corrupting each output by Gaussian noise as

$$D_k = \xi_k + \sigma \epsilon_k$$

where D_k is the observation data from the k^{th} position of the vector, ξ_k is the k^{th} model output, ϵ_k is a zero-mean, unit-variance Gaussian variable, and σ is the level of the noise. The reference model is selected to correspond to some nominal values of the model parameters. In order for the signal-to-noise ratio to be high enough for meaningful estimation, we choose σ to be a fraction $\sigma = 0.01\alpha$ of the standard deviation α of all model outputs. The model prediction error covariance Σ is assumed to be a diagonal matrix $\Sigma = \sigma \mathbf{I}$ whose nonzero entries all have the same magnitude σ .

In the following results, we estimate parameters via the generation of 10^4 samples from the posterior for the DNA methylation model. For ease of comparison, numerical results are computed in terms of the rescaled parameters $(\theta_{k_{31}}, \theta_{k_{32}}, \theta_\gamma, \sigma/\alpha)$, given by $\theta_{k_{31}} = k_{31}/k_{31_0}$, i.e., the ratio between the estimated value and the nominal value. The prior is assumed uniform on $[-4, 4] \times [-4, 4] \times [-4, 4] \times [0, 0.05]$ in the scaled parameter space. We consider two formats for the non-linear term k_3 : $k_3(x_1) = k_{31} + k_{32}x_1^\gamma$, an increasing function of x_1 , and $k_3(x_2) = k_{31} + k_{32}x_2^\gamma$, a decreasing function of x_2 . We take the following parameter values for the nominal parameter values $k_{31_0} = 1$, $k_{32_0} = 0.01$, $\gamma_0 = 2$, and $\sigma = 0.01\alpha$ for $k_3(x_1)$ and $k_{31_0} = 100$, $k_{32_0} = -0.01$, $\gamma_0 = 2$, and $\sigma = 0.01\alpha$ for $k_3(x_2)$.

5.3.1 Parameter Estimation

In this case, we have two sets of reference data. The first, denoted O_1 , comes from the time history generated by the model M_1 where the expression k_3 depends on x_1 and

the second, denoted O_2 , comes from the model M_2 where the expression k_3 depends on x_2 . In both cases, $k_1 = 0.012$, $k_2 = 99$, $k_4 = 0.08$ and the other parameters are as described above.

The results using $k_3(x_1)$ are displayed in Figure 5.3, which show a strong negative correlation with k_{31} and k_{32} , which determine the hemimethylation rate of CpG dyads (x_2 to x_1). This makes intuitive sense, as an increase in k_{31} should correspond with a decrease in k_{32} to match the dynamics. Correspondingly, both k_{32} and γ have a strong negative correlation.

The recovered scaled mean parameter values are in Table 5.1, recovering values $(k_{31}, k_{32}, \gamma, \sigma/\alpha) = (0.961, 0.0105, 1.998, 0.010)$. To quantify the degree of uncertainty for each parameter's posterior distribution, we compute the coefficient of variation, defined as the ratio of its standard deviation to its mean (denoting the results $u_{k_{31}}, u_{k_{32}}, u_{\gamma}, u_{\sigma/\alpha}$). In all cases, the parameter values are recovered within one standard deviation of the nominal values.

The results using $k_3(x_2)$ are displayed in Figure 5.4, which displays different correlations than the $k_3(x_1)$ case. Here, the k_{31} parameter is found almost exactly, while k_{32} and γ have a range of negative and positive values that are found. This is likely due to small changes in k_{32} and γ not greatly effecting the dynamics of $k_3(x_2)$.

The recovered scaled mean parameter values are in Table 5.1, recovering values $(k_{31}, k_{32}, \gamma, \sigma/\alpha) = (99.1, -0.014, 0.576, 0.010)$. In this case, the parameters are all recovered within two standard deviations of the mean, due to the wide smear in the recovered parameter values.

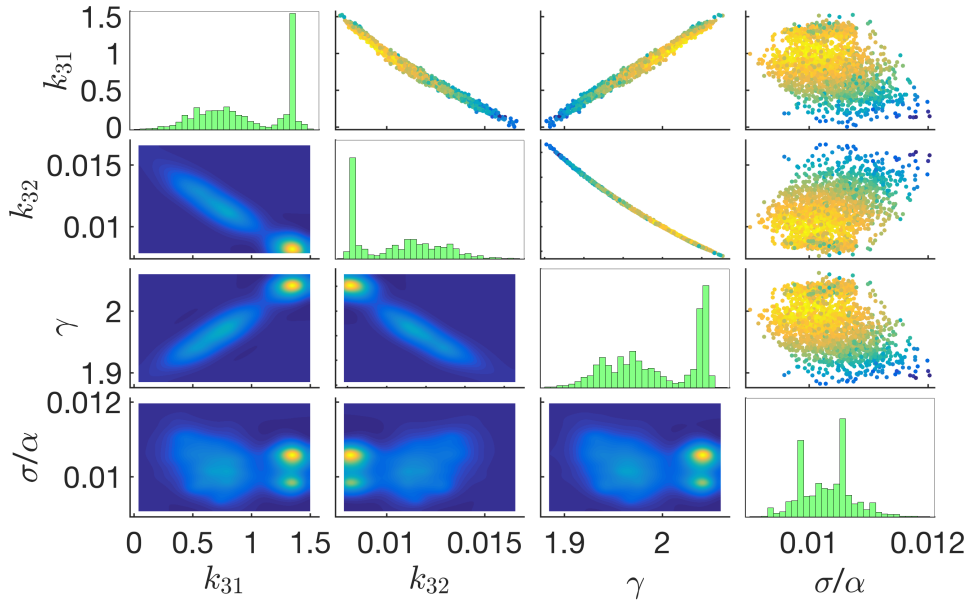


Figure 5.3: Parameter estimation results using reference data from the model where $k_3 = k_{31} + k_{32}x_1^\gamma$ with a time history from $T = 0$ to $T = 10$. The nominal parameter values used were $k_{31} = 1$, $k_{32} = 0.01$, $\gamma = 2$, and noise level $\sigma = 0.01\alpha$. The model used in parameter estimation is $k_3 = k_{31} + k_{32}x_1^\gamma$. Histograms for each parameter are displayed along the main diagonal of the figure. Sub-figures below the diagonal show the marginal joint density functions for each pair of parameters, while sub-figures above the diagonal show the samples used in the final stage of TMCMC. Colors correspond to probabilities, with yellow likely and blue unlikely.

5.3.2 Model Selection

Next, our aim was to identify which model generated which reference data set. To do this, we used the reference data set for $k_3(x_2)$ and use the $k_3(x_1)$ model to recover parameters; correspondingly the $k_3(x_2)$ model also uses the $k_3(x_1)$ reference data to perform parameter estimation. The results for these two experiments are displayed in Figures 5.5a and 5.5b, respectively. Since $k_3(x_2)$ takes on parameter values in a much larger range and $k_3(x_1)$ is increasing, the prior distributions were expanded to $[0, 400] \times [0, 400] \times [0, 4] \times [0, 0.20]$ and $[-4, 4] \times [-4, 4] \times [-4, 4] \times [0, 0.20]$ for models $k_3(x_1)$ and $k_3(x_2)$, respectively.

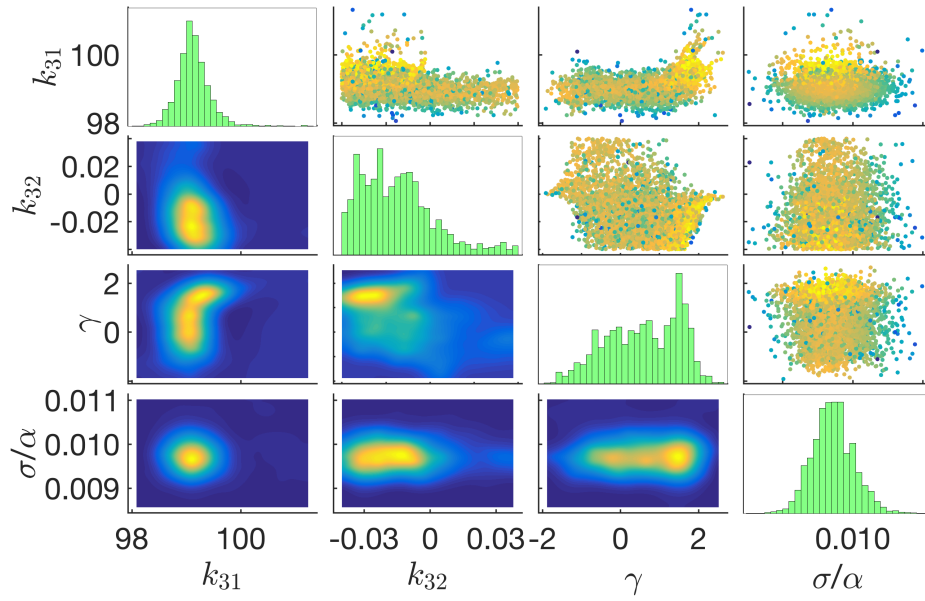
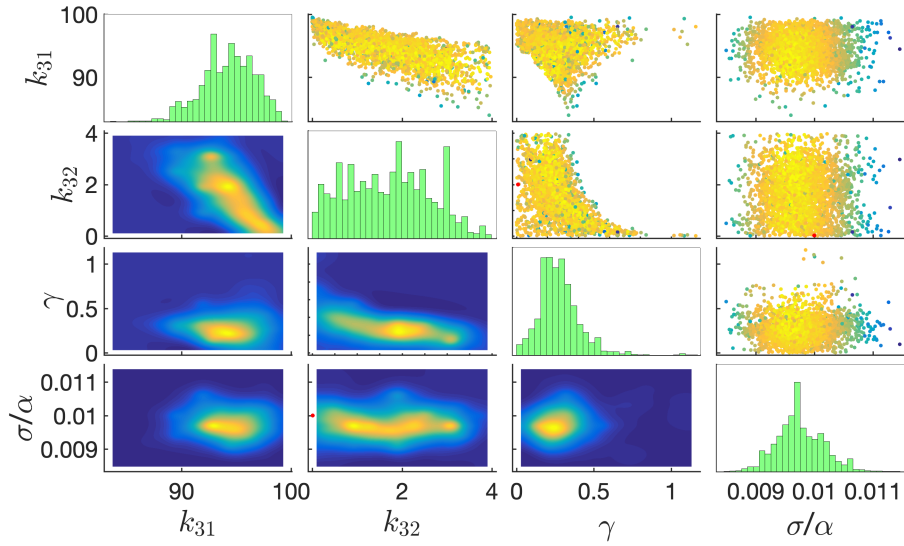
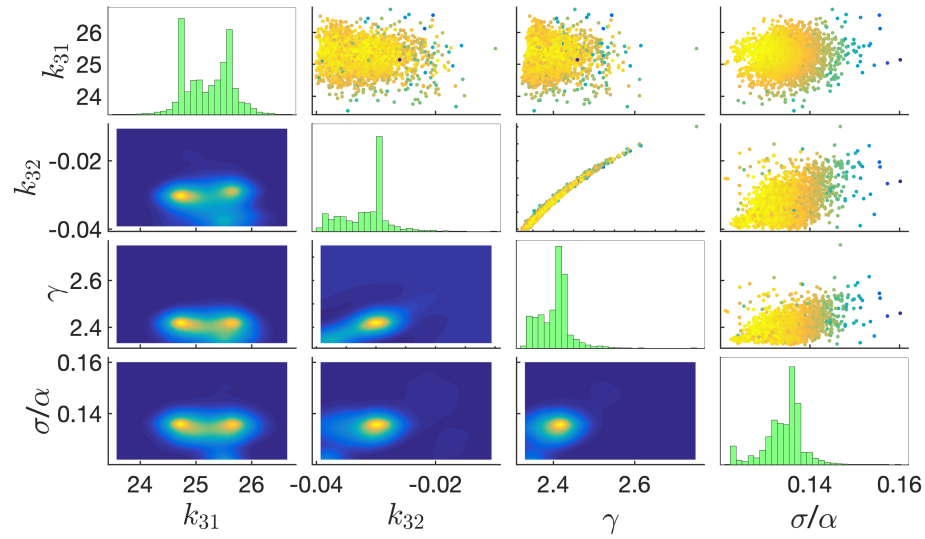


Figure 5.4: Parameter estimation results using reference data from model where $k_3 = k_{31} + k_{32}x_2^\gamma$ with a time history from $T = 0$ to $T = 10$. The nominal parameter values used were $k_{31} = 100$, $k_{32} = -0.01$, $\gamma = 2$, and noise level $\sigma = 0.01\alpha$. The model used in parameter estimation is $k_3 = k_{31} + k_{32}x_2^\gamma$. Descriptions of the sub-figures can be found in Figure 5.3.

To analyze the various models, the results from the previous four experiments are displayed in Table 5.1. By comparing the probabilities for both models that used reference data from $k_3(x_1)$, we find that with probability one that data came from model $k_3(x_1)$, regardless of the range of parameter values that are explored. Correspondingly, model $k_3(x_2)$ recovered parameters that are very different from the nominal parameter values, further demonstrating the inability of that model to recover dynamics similar to those of the observed data. When comparing the model probabilities for the observation data from $k_3(x_2)$, we find with probability ~ 0.91 that the data came from model $k_3(x_2)$ when we use a sufficiently large prior distribution for model $k_3(x_1)$. However, to achieve this result requires an immense parameter domain that seems infeasible in general as the recovered parameters are 200 times the expected value. Thus, we are able to confidently resolve a model misspecification problem; i.e., we are able to correctly recover the model that generated each of the



(a) Results for using model $k_3(x_1)$ on reference data from $k_3(x_2)$.



(b) Results for using model $k_3(x_2)$ on reference data from $k_3(x_1)$.

Figure 5.5: Parameter estimation results using the “incorrect” model on reference data, i.e. the model $k_3(x_1)$ on noisy data generated from $k_3(x_2)$ and similarly for model $k_3(x_2)$. Descriptions of the sub-figures can be found in Figure 5.3, and the parameter values used can be found in Figures 5.3 and 5.4.

data sets used.

Model	$p(M_j D)$	$\theta_{k_{31}}$	$u_{k_{31}}$ (%)	$\theta_{k_{32}}$	$u_{k_{32}}$ (%)
M_1, O_1	~ 1.0	0.961	37.96	1.045	19.79
M_2, O_1	~ 0.0	0.252	1.64	-3.161	11.89
M_1, O_2	0.0936	94.242	2.52	170.437	5.18
M_2, O_2	0.9064	0.991	0.32	-1.445	-116.22

Model	θ_γ	u_γ (%)	σ/α	$u_{\sigma/\alpha}$ (%)
M_1, O_1	0.999	2.223	0.010	4.22
M_2, O_1	1.202	1.50	0.134	3.24
M_1, O_2	0.138	5.73	0.010	3.94
M_2, O_2	0.288	162.73	0.010	3.08

Table 5.1: Subset of model selection results for DNA methylation models.

5.3.3 Two Standard Deviation Added-Noise and Nonlinear Model

Another method to add noise is to use a different standard deviation for each model output. Previously, we had assumed that the noise was added with a constant factor that was the same for x_1 , x_2 , and x_3 . However, the spread of these three populations may be different, so we instead considered the case where we had a different standard deviation, σ_1 and σ_3 , for x_1 and x_3 . The covariance matrix was still assumed to be diagonal, but now $\Sigma_{j,j} = \sigma_1$ for $j = 1, 3, \dots, m-1$ and $\Sigma_{j,j} = \sigma_3$ for $j = 2, 4, \dots, m$. The noise was added as before, but where we used σ_1 when the output corresponds to an x_1 observation, and σ_3 otherwise, i.e.

$$D_k = \begin{cases} \xi_k + \sigma_1 \epsilon_k, & k = 1, 3, \dots, m-1, \\ \xi_k + \sigma_3 \epsilon_k, & k = 2, 4, \dots, m. \end{cases}$$

In the following experiments, we worked with the nonlinear coefficient cases and considered two potential models: $M_{1,3}$, where the coefficients for k_1 and k_3 depend on x_1 and k_2 and k_4 depend on x_3 and M_2 , in which all coefficients depend on x_2 .

For each of these models, we generated reference data as described above using noisy observations from model $M_{1,3}$, denoted $O_{1,3}$, or noisy observations from model M_2 , denoted O_2 . We first estimated parameters for each model using the noisy data that was generated from that model. Secondly, we tried to identify which model generated what data. Three experiments were performed: first we attempted to recover the parameters for models $M_{1,3}$ and M_2 . Second, model selection was performed to see if the $\Pi 4U$ framework was able to recover which model generated which data set. Finally, we tried to recover all eight model parameters, as well as the corresponding noise levels.

Again, for the ease of comparison, numerical results were computed in terms of the rescaled parameters $(\theta_{k_{11}}, \theta_{k_{21}}, \theta_{k_{32}}, \theta_{k_{41}}, \theta_\gamma, \sigma_1/\alpha_1, \sigma_3/\alpha_3)$, given by $\theta_{k_{11}} = k_{11}/k_{11_0}$, i.e. the ratio between the estimated value and the nominal value. The prior was assumed uniform on $[0, 4] \times [0, 4] \times [0, 4] \times [0, 4] \times [0, 0.05] \times [0, 0.05]$ in the scaled parameter space. We took the following parameter values for the nominal parameter values $k_{11_0} = 2.1$, $k_{21_0} = 10$, $k_{32_0} = 0.0099$, and $k_{41_0} = 4$ for $M_{1,3}$ and $k_{11_0} = 1.9$, $k_{21_0} = 110$, $k_{32_0} = -0.0099$, and $k_{41_0} = 2$ for M_2 .

5.3.3.1 The Four Most Sensitive Parameters

In this experiment, we used model $M_{1,3}$ with noisy observations $O_{1,3}$ or M_2 with noisy observations O_2 to perform parameter estimation for the four model parameters k_{11} , k_{21} , k_{32} , and k_{41} , which were found to be the four parameters that caused the largest changes in x_1 , x_2 , and x_3 (112). In addition, we also tried to recover the amount of added noise. In these experiments, 5% noise was added to the model outputs of x_1 and x_3 to generate observations $O_{1,3}$, where $\sigma_i = 0.05\alpha_i$, for $i = 1, 3$ and α_i is the standard deviation for model outputs x_i .

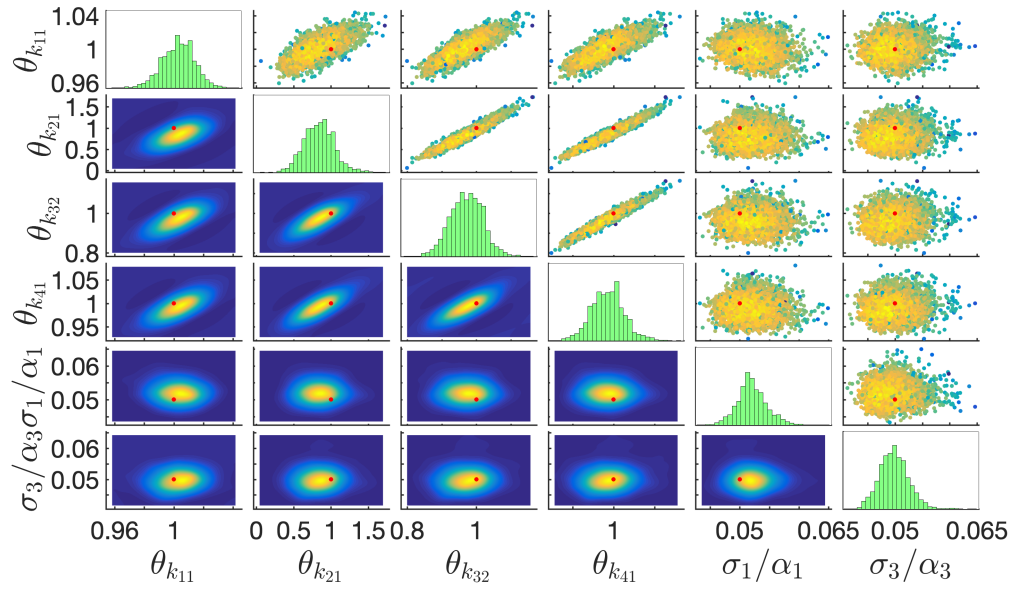


Figure 5.6: Parameter estimation results for model $M_{1,3}$ using reference data $O_{1,3}$ with 5% added noise. Histograms for each parameter are displayed along the main diagonal of the figure. Sub-figures below the diagonal show the marginal joint density functions for each pair of parameters, while sub-figures above the diagonal show the samples used in the final stage of TCMCMC. Colors correspond to probabilities, with yellow likely and blue unlikely. Red dots indicate the parameter values used for generating the noisy data set used for performing parameter estimation and model selection.

The case where observations were created using model $M_{1,3}$ has results displayed in Figure 5.6, where the parameters used for generating the reference data are marked with red dots. We see clear correlations between the parameters: k_{11} and k_{32} have a positive correlation, which makes intuitive sense as those two coefficients have opposite effects: one controls the methylation rate, while the other influences the demethylation rate. Correspondingly, we see positive correlations with k_{21} and k_{41} . In addition to the marginal posterior distributions matching the expected correlations, they also include the nominal parameter values used to create the reference data: for all four model parameters, the “true” parameter value is recovered within one standard deviation; the level of the noise is also recovered well for this case. The marginal distributions from σ_1/α_1 and σ_3/α_3 are both centered around 0.05,

the noise level used to create the noisy observations.

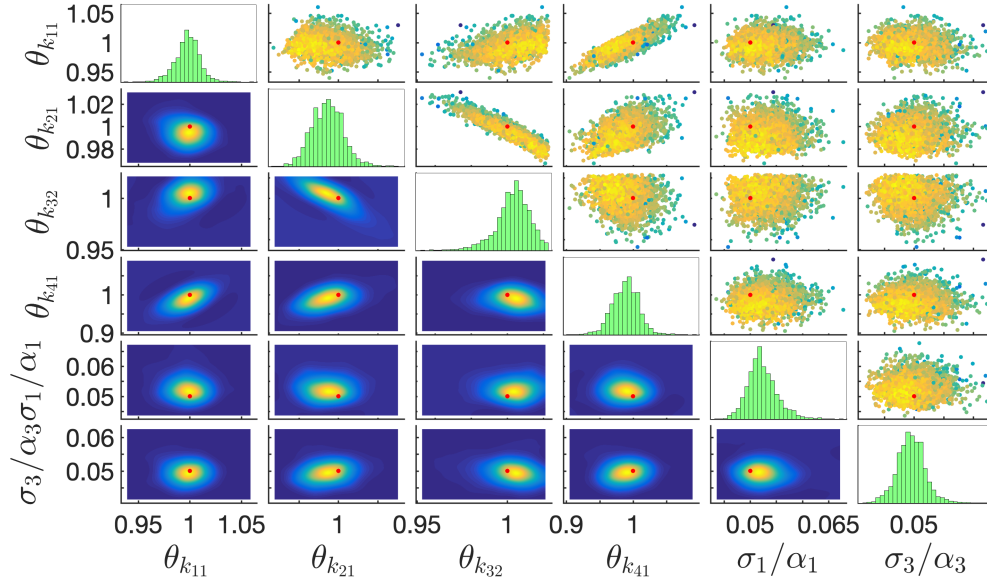
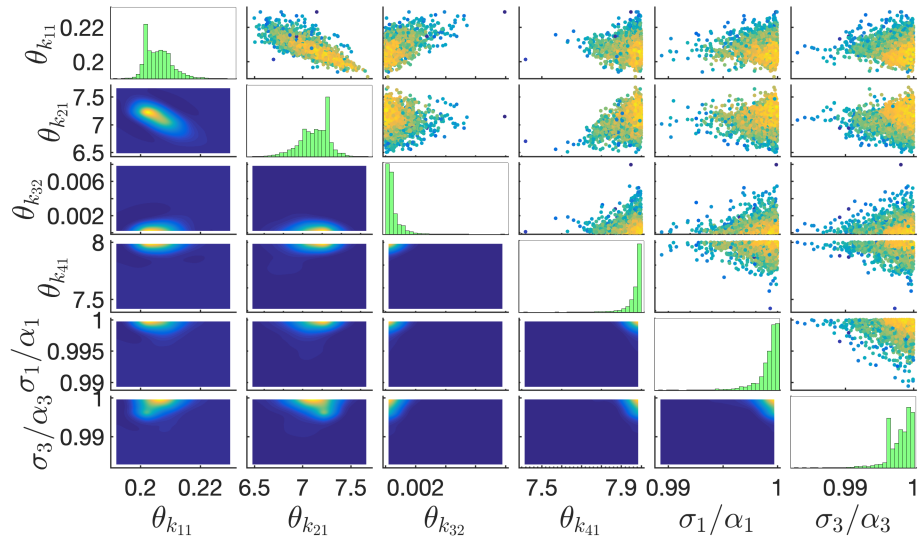


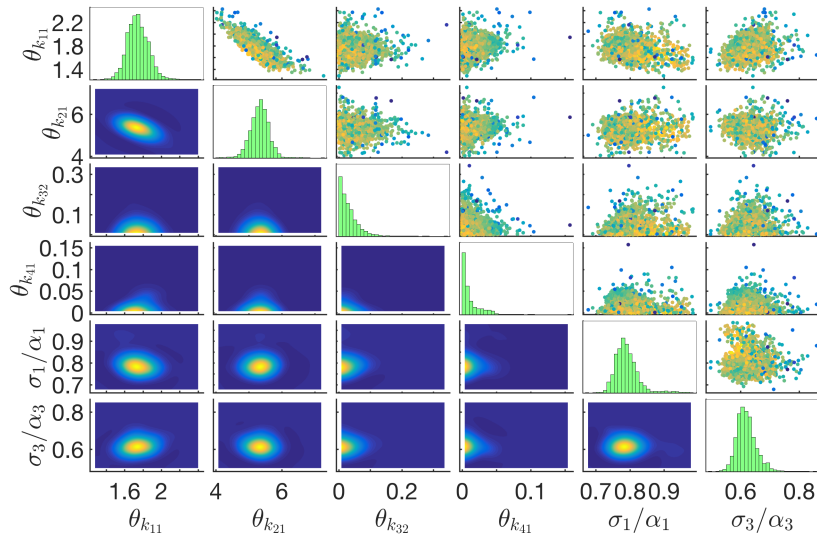
Figure 5.7: Parameter estimation results for model M_2 using reference data O_2 with 5% added noise.

Our second case, where model M_2 was used to create noisy observations O_2 , also performs well: as with the previous case, all of the nominal parameter values are recovered within one standard deviation of the estimated means. Furthermore, there are similar correlations to the previous case. Now the correlation between k_{21} and k_{32} is negative due to k_{32} having a negative coefficient: for larger values of k_{21} , a more negative coefficient is needed to have similar dynamics as the observations. In both cases, the relative uncertainty of all parameters are on the same scale. The exception is k_{21} for $M_{1,3}$, which has a relatively bigger uncertainty than the other parameters. This large uncertainty in k_{21} could be due to the relatively larger parameter values that $\theta_{k_{21}}$ explores during the TMCMC algorithm.

Next, we performed model selection. To do this, we used model M_2 on $O_{1,3}$ and model $M_{1,3}$ on O_2 to see whether it could still produce the same dynamic behavior as the other model. We compared both models' results on each set of the noisy



(a) Results for using model $M_{1,3}$ on reference data from O_2 .



(b) Results for using model M_2 on reference data from $O_{1,3}$.

Figure 5.8: Parameter estimation results using a different model than the one that generated the reference data. Descriptions of the sub-figures can be found in Figure 5.3, and the parameter values used can be found in Figures 5.6 and 5.7.

observations. In both cases, as seen in Table 5.2, the other model is unable to recover the dynamics using the uniform prior $[0, 8] \times [0, 8] \times [0, 8] \times [0, 8] \times [0, 1] \times [0, 1]$. When the probabilities for the two models are compared, the likelihood of M_2 on $O_{1,3}$ is negligible, and similarly for M_1 on O_2 . These results can be intuitively seen by also

looking at how the incorrect models are inadequate at estimating the parameters, as they are not able to match the dynamics of the problem. Demonstrated in Figure 5.8, the “best” parameters in these cases are along the border of the prior distribution and to better match the observed data, portraying that parameter values much larger than those expected biologically are required to recover dynamics close to the noisy observations.

Model	$p(M_i D)$	$\theta_{k_{11}}$	$u_{k_{11}}$ (%)	$\theta_{k_{21}}$	$u_{k_{21}}$ (%)	$\theta_{k_{32}}$	
$M_{1,3}, O_{1,3}$	~ 1.00	1.0030	1.09	0.8374	23.93	0.9710	
$M_2, O_{1,3}$	~ 0.00	1.7466	6.66	5.3280	5.21	0.0316	
$M_{1,3}, O_2$	~ 0.00	0.2060	2.14	7.1298	2.09	0.0004	
M_2, O_2	~ 1.00	0.9975	1.20	0.9945	0.80	1.0032	
Model	$u_{k_{32}}$ (%)	$\theta_{k_{41}}$	$u_{k_{41}}$ (%)	σ_1/α_1	u_{σ_1/α_1}	σ_3/α_3	u_{σ_3/α_3}
$M_{1,3}, O_{1,3}$	4.82	0.9926	2.01	0.0521	4.78	0.0498	5.36
$M_2, O_{1,3}$	92.75	0.0112	112.73	0.7903	4.37	0.6196	5.34
$M_{1,3}, O_2$	115.23	7.9701	0.47	0.9992	0.10	0.9980	0.17
M_2, O_2	0.92	0.9884	1.99	0.0524	5.39	0.0495	4.77

Table 5.2: Model selection results for estimating four most sensitive transition rates for two model scenarios and two observed data sets for 5% added noise.

5.3.4 All Parameters

Finally, we attempted to recover the eight coefficients for all four parameters k_1 , k_2 , k_3 , and k_4 using noisy observations from model $M_{1,3}$ and M_2 . We did this for two cases of added noise: 1% and 5% noise.

For the 1% added noise case using model $M_{1,3}$ displayed in Figure 5.9, all parameters are recovered within two standard deviations of the means. As in the four parameter case, the distribution for k_{21} has a rather wide standard deviation. Furthermore, we note that the parameters k_{12} , k_{22} , k_{42} all have wider distributions than the more sensitive parameters k_{11} , k_{32} , and k_{41} . This matches intuition, as the pa-

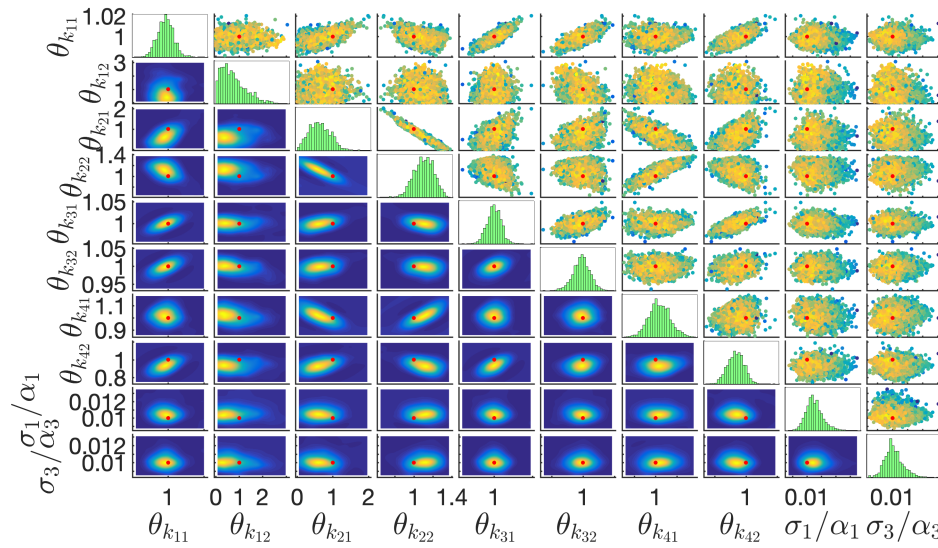


Figure 5.9: Parameter estimation results for model $M_{1,3}$ using reference data $O_{1,3}$ with 1% added noise.

rameters that cause larger changes in the model outputs are more easily recovered since we want to minimize the difference between model outputs and observed data. In addition, the model correlations match what is expected from the system dynamics: for example, k_{21} and k_{22} have a negative correlation. Since the coefficient $k_2(x_3)$ is an increasing function, a larger value of k_{21} would require a smaller value of k_{22} to result in similar values for $k_2(x_3)$.

In addition, we again considered the 1% added noise case for M_2 , shown in Figure 5.10. For the model where coefficients depend on x_2 all nominal parameter values are found within two standard deviations of the recovered means, except k_{12} . It was found that, model M_2 is highly insensitive to k_{12} : increasing or decreasing k_{12} by 50% change the model outputs on the order of 10^{-5} ; however, the added noise adjusts the parameter values on the order of 10^{-1} . Due to this discrepancy in the model sensitivity and noise level, it seems that the parameter value for k_{12} is difficult to recover and, as a result of model robustness to this parameter, is not as important to recover accurately. Similar to the $M_{1,3}$ case, the recovered marginal

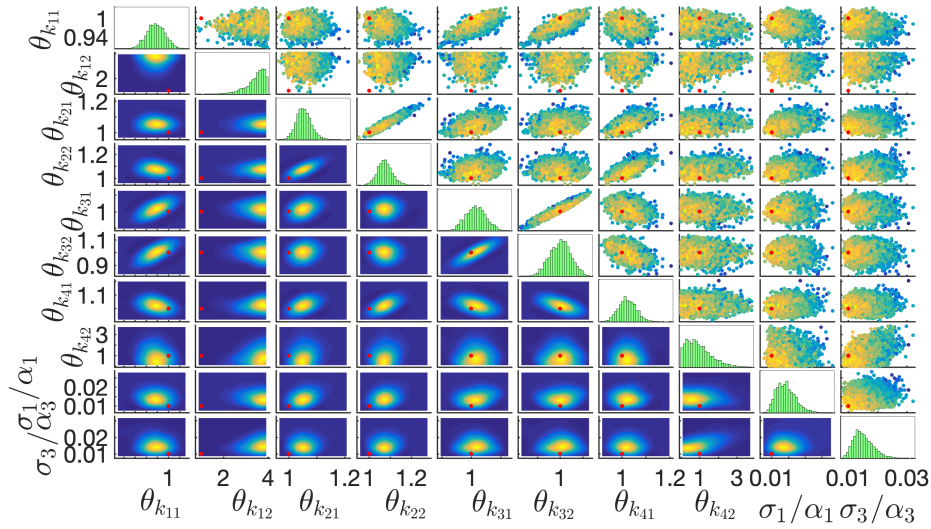


Figure 5.10: Parameter estimation results for model M_2 using reference data O_2 with 1% added noise.

distributions have correlations that match intuition. For example, k_{21} and k_{22} have a positive correlation. Since $k_2(x_2)$ is a decreasing function, smaller values of k_{21} correspondingly need smaller values of k_{22} to keep k_2 close to the same values as the nominal parameter values.

Model	$p(M_j D)$	$\theta_{k_{11}}$	$u_{k_{11}}$ (%)	$\theta_{k_{12}}$	$u_{k_{12}}$ (%)	$\theta_{k_{21}}$	$u_{k_{21}}$ (%)
$M_{1,3}, O_{1,3}$	~ 1.0	0.999	0.43	0.743	71.00	0.636	46.81
$M_2, O_{1,3}$	~ 0.0	3.899	0.85	0.650	59.11	4.000	0.01
$M_{1,3}, O_2$	~ 0.0	0.326	1.14	0.011	118.97	2.943	6.43
M_2, O_2	~ 1.0	0.992	0.52	3.151	16.66	1.010	0.88
		$\theta_{k_{22}}$	$u_{k_{22}}$	$\theta_{k_{31}}$	$u_{k_{31}}$	$\theta_{k_{32}}$	$u_{k_{32}}$
		1.106	9.36	1.000	0.90	0.998	1.34
		0.006	115.00	1.740	0.84	0.001	99.19
		0.001	102.08	4.000	0.05	0.001	114.44
		1.017	1.13	1.000	0.80	0.996	1.16
		$\theta_{k_{41}}$	$u_{k_{41}}$	σ_1/α_1	u_{σ_1/α_1} (%)	σ_3/α_3	u_{σ_3/α_3} (%)
		1.018	4.88	0.011	5.57	0.010	5.90
		3.696	6.05	0.200	0.0003	0.200	0.01
		0.005	94.86	0.200	0.03	0.200	0.02
		0.536	72.47	0.01	6.18	0.01	6.37

Table 5.3: Model selection results for estimating all transition rates added noise for two model scenarios and two observed data sets for 1% added noise.

Model selection was also performed on these parameter values. The scaled model parameters were given a prior of $[0, 4]$ and the standard deviations had a uniform prior on $[0, 0.20]$. In these cases, we find that $M_{1,3}$ is the model that best matches $O_{1,3}$ with probability one, and similarly M_2 best matches O_2 with probability one. In addition, the incorrect models are unable to recover the proper parameter values. As seen in Table 5.3, many of the recovered parameter values are on the boundary of the prior distribution, demonstrating that the usual domain for the parameters of each model are not able to recover the dynamics of the other model. In both of our cases, we are able to recover that $O_{1,3}$ came from $M_{1,3}$ and O_2 is a noisy version of model M_2 .

The same experiment was repeated using 5% added noise. Using model $M_{1,3}$ (seen in Figure 5.11), the 5% noise case recovers all parameters within two standard deviations, although the distributions are comparatively wider than those found in previous experiments: most of the recovered standard deviations are 2–5 times as large for 5% added noise as the 1% added noise case. Again, this matches intuition as noisier data should result in more uncertainty in the estimated distributions.

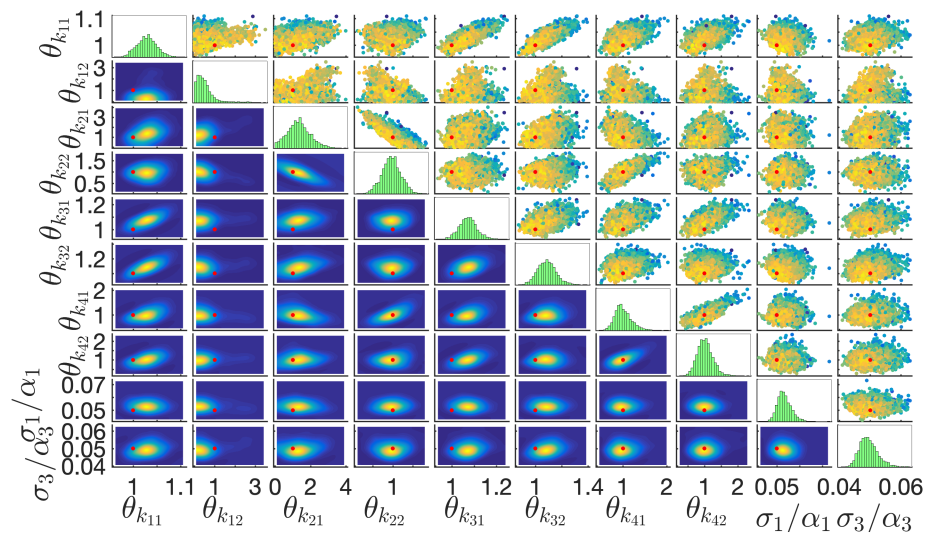


Figure 5.11: Parameter estimation results $M_{1,3}$ using reference data $O_{1,3}$ with 5% added noise.

For model M_2 , more model parameters are not estimated well: k_{12} , k_{31} , k_{32} , and k_{42} are not recovered within two standard deviations, and most parameters are only recovered due to having very large standard deviations. For M_2 , all standard deviations for the 5% noise case are 2–5 times as large as those for 1% noise.

Model	$p(M_j D)$	$\theta_{k_{11}}$	$u_{k_{11}}$ (%)	$\theta_{k_{12}}$	$u_{k_{12}}$ (%)	$\theta_{k_{21}}$	$u_{k_{21}}$ (%)
$M_{1,3}, O_{1,3}$	~ 1.0	1.029	2.01	0.521	88.73	1.34	47.35
$M_2, O_{1,3}$	~ 0.0	3.429	9.10	1.269	68.52	3.907	3.08
$M_{1,3}, O_2$	~ 0.0	0.340	3.36	0.078	100.10	3.687	4.93
M_2, O_2	~ 1.0	0.925	2.23	0.667	69.54	1.173	4.78
	$\theta_{k_{22}}$	$u_{k_{22}}$	$\theta_{k_{31}}$	$u_{k_{31}}$	$\theta_{k_{32}}$	$u_{k_{32}}$	$\theta_{k_{41}}$
	0.967	23.81	1.073	3.75	1.087	6.50	1.008
	0.334	90.12	1.559	10.10	0.126	102.09	0.007
	0.004	81.31	3.986	0.34	0.004	109.10	1.869
	1.244	5.53	0.995	5.46	0.917	7.37	1.139
	$\theta_{k_{42}}$	$u_{k_{42}}$	σ_1/α_1	u_{σ_1/α_1} (%)	σ_3/α_3	u_{σ_3/α_3} (%)	
	1.017	23.21	0.053	5.61	0.050	6.10	
	1.765	59.56	0.837	5.05	0.665	5.80	
	0.031	102.78	0.499	0.18	0.499	0.12	
	1.061	81.26	0.55	6.91	0.051	6.87	

Table 5.4: Model selection results for estimating all transition rates added noise for two model scenarios and two observed data sets for 5% added noise.

Finally, the model selection was again performed and similar results to the 1% noise case occurred: again, the model probabilities accurately conclude which observation data came from which model. Secondly, the distributions again move to the boundaries of the domain, demonstrating that for the typical parameter values, M_2 cannot give the dynamics of $M_{1,3}$ and vice versa.

5.4 Discussion

In this paper, we presented an uncertainty quantification framework that is applicable to a wide array of parameter estimation and model selection problems. For

experimental biologists with wet lab data, the methodology can be used to test and improve various models, as well as guide future research. In order to use the method described in the paper, three components are necessary: first, experimental data are needed as the reference data. Second, model or models for the biological phenomena are needed to estimate parameters or test which model best describes the experimental data. Finally, a “connector” code is required to compare the model to the experimental data. Our Bayesian framework propagates the forward model through sampled sets of parameters and uses the goodness of the fit to guide the next generation of samples. In addition, this approach is highly parallelizable, lending itself easily to applications that were previously too computationally intensive to be feasible (31). Due to these simple three requirements and its parallelizability, this method is useable for parameter estimation and model selection problems for a wide array of fields and backgrounds.

We find that the Bayesian uncertainty quantification framework allows us to recover robust predictions for parameters of the DNA methylation model, for example the demethylation rate k_3 , despite only using a limited amount of noisy data. When examining the model with unknown exponent γ , the method was able to accurately recover all parameters within two standard deviations of the nominal parameter values for observed data that was generated from the corresponding model. Furthermore, the method was able to discern which model the observed data were generated from, as long as the biological constraints were maintained. Our methodology prefers that demethylation is dependent upon the total number of unmethylated CpG dyads for the noisy data generated from that assumption. If the parameter space searched allows the model to no longer obey the observed monotonicity of the demethylation parameters, the observed data generated from the demethylation parameter depending on hemimethylated CpG dyads x_2 , i.e. $k_3(x_2)$, can be matched

to both models.

In our second scenario, we considered various noise levels and estimated more model parameters for the demethylation rates, although we assumed that the exponent was fixed $\gamma = 2$ for all populations. For the four most sensitive parameter scenario, all parameters were recovered within a reasonable deviation from the nominal values. Furthermore, correlations that match the intuition of the biological model are observed in the various recovered parameter values. Finally, we were also able to identify whether the parameters depended on the unmethylated and methylated or the hemimethylated CpG dyads, showing that our method may be able to help better understand the biological process if given real data.

Finally, we estimate all eight parameters related to demethylation rates for a few added noise scenarios to demonstrate the power of the Bayesian framework. Again, when the parameters depend on unmethylated and methylated CpG dyads, all values are recovered within two standard deviations of the nominal parameter values. The scenario where the parameters depend on the hemimethylated CpG dyads has a bit more difficulty recovering the nominal parameter values for 5% added noise, but this is likely due to some model insensitivity to those parameter values. In all cases, however, we are able to recover which model generated what observed data. In addition, by comparing the recovered levels of noise, we observe that the $\sigma = 0.05\alpha$ scenarios correspondingly have larger uncertainties in the recovered parameter values than the $\sigma = 0.01\alpha$ cases.

It is important to put the results of our work within a biological context. For example, a finding of the work is the ability of our technique to determine if model parameter values are dependent on different methylation states. The issue of whether or not rates of methylation/demethylation are inexorably linked to promoter topol-

ogy is an ongoing question the experimental community has been attempting to unravel. For instance, it is generally regarded that during replication DNMTs are responsible for the remethylation of hemimethylated DNA. However, due to a number of experimental findings it has been suggested by (118) that this is unfeasible and that the methylation of a CpG site is affected by the methylation levels of the nearby CpG sites. This idea couples DNMTs activity with CpG level. Therefore, the fact that our theoretical analysis can explore this issue is a significant feature of our work which can be explored further by the experimental community. Another area of biological significance is that our models are based on the assumption of enzymatic processivity (i.e. how DNMT1, DNMT3a, and DNMT3b behave with various levels of methylated, demethylated, and hemimethylated dyads) (119). A way to fully test the findings from our model would be determine in the laboratory whether our findings hold in case it turns out that DNMTs are not processive enzymes and that rather other factors are at play, as for example, RNA directed DNA methylation.

CHAPTER SIX

Bayesian uncertainty quantification for micro-swimmers with fully resolved hydrodynamics

6.1 Introduction

Mathematical models to investigate the swimming speeds and efficiency of microswimmers date back to the foundational work of Taylor (11, 120, 10, 121). This pioneering work focused on studying a swimmer represented as either a planar sheet in 2-dimensions (2D) or a waving cylindrical tail in 3-dimensions (3D), both of infinite extent and with a prescribed beat form. State of the art now involves highly computational 2D or 3D models of swimmers with a finite length flagellum (tail) and accurate geometries where the beat form is either prescribed or an emergent property of the coupled system where the fluid dynamics are fully resolved (122, 123, 124, 125, 126, 127, 128, 129, 130, 131). These models are able to incorporate material properties of the actual flagellum and the fluid stress may be linear or nonlinear with respect to the strain (132, 12, 133, 134).

Currently, there is great interest in understanding the coupling between fluids and the swimming structure. In the case of sperm motility, understanding motility patterns could lead to the development of treatments to either aid or hinder sperm progression (135). In experiments, CASA (computer aided sperm analysis) is a widely used software that is able to track the head or cell body of sperm to determine swimming velocities and can also be used to determine beat frequency of the flagellum (tail). However, these data are often noisy and one might want to understand the fluid properties or swimmer properties (geometry or constitutive parameters) that would allow for these types of trajectories. Previous estimation of constitutive parameters such as flexural rigidity and shear stiffness has usually been completed using experiments that either analyze the beat form of the flagellum (with or without manipulation) and modifications of elastic beam theory (136, 137, 138, 139, 140). Since these approaches do not have fully resolved hydrodynamics, nor do they ac-

count for all of the active force generation by dynein motors, this of course is very restrictive. These parameters give insight into the constitutive parameters but do not necessarily correlate with those that would be used in accurate fluid-structure interaction simulations.

In addition, an active area is in the development of artificial micro-swimmers for applications such as drug delivery and microsurgery (141, 142, 143, 144). In the case of artificial micro-swimmers, most are bio- inspired and have flagella. The swimming is achieved through physical mechanisms and could be driven by a magnetic field, electric field, or a chemical reaction. In the context of artificial micro-swimmers, one might want to determine parameters that lead to an optimal swimming gate for a particular application. Optimal could be defined as the fastest swimming speed, maximizing hydrodynamic efficiency, or leading to a preferred path or trajectory.

Little prior work has been completed in terms of parameter estimation or uncertainty quantification for micro-swimmers. Plouraboue et al. (145) used a simplified bead model to represent a swimmer with linear elasticity. Identifiability conditions determined that parameter identification was possible in the presence of noise. However, even in this simplified model, forcing and passive parameters could only be identified separately, not simultaneously. The assumed linear relationship between the observed data and the parameters to be solved for allowed estimations of the parameters, but models of micro-swimmers are generally non-linear due to the complex coupling of the emergent flagellar bending with the fluid dynamics (11). Recently, Tsang et al.(146) explored the idea of self-learning to optimize locomotion of synthetic micro-swimmers in the different fluid environments that might be encountered in the body for drug delivery applications. These artificial micro-swimmers were self-learning and adaptive.

In this work, we utilize a Bayesian framework for uncertainty quantification to understand the emergent trajectories and swimming speeds of micro-swimmers. We assume the micro-swimmers are immersed in a fluid with a sparse network of fibers, which is captured via a flow dependent term with a resistance parameter, and the hydrodynamics are fully resolved utilizing the method of regularized fundamental solutions (147, 148, 149). The resistance parameter allows exploration of a Stokes fluid (in the limit of no resistance) to a range of applicable *in vivo* fluid environments. In the swimmer model, we include non-linearities in the elasticity portion since the elastic energy depends on curvature— a nonlinear function of shape, making previous analytical methods on simplified models (as in (145)) not feasible. In this context, Bayesian estimation methods allow us to leverage our prior knowledge of micro-swimmers and the viscous medium to inform the search of the parameter space. We examine the validity of using the parallel transitional Markov chain Monte Carlo (TMCMC) algorithm for micro-swimmer questions. This is done by using the high performance computing framework $\Pi 4U$ (17, 31, 27) to estimate parameters for the fluid resistance and material parameters of the flagella using noisy *in silico* observations generated from the biological model, which allows us to garner correlations between parameters and gain insights into the underlying biological mechanisms. Through the marginal joint densities for each pair of estimated parameters, we gain further insight into the fluid dynamics of micro-swimmers by understanding which sets of parameters give similar swimmer trajectories.

6.2 Micro-swimmer Model

6.2.1 Fluid Equations

Micro-swimmers navigate in the regime where viscous forces dominate and acceleration is negligible (120, 125). A common approach is to model the fluid that the swimmer is immersed in via the steady Stokes equations. This of course will be representative of experimental data in a lab where sperm are in water. However, *in vivo*, many micro-swimmers encounter non-homogeneous fluids where there is a protein network or other obstacles. For example, mammalian spermatozoa navigate in vaginal and cervical fluids with a network of mucin fibers (150, 151) and bacteria navigate through a mucus layer or in biofilms, both consisting of a network of polymeric substances (152, 153). The appropriate governing equation for the fluid will depend on several factors relating to the volume fraction, material properties, and movement of the immersed fibers or proteins.

In the case of a stationary and sparse network of fibers or obstacles in the fluid, the Brinkman equation can be used (154, 155, 156, 157, 158). This is a fluid model that does not model the exact architecture and location of the fibers, but captures the homogenized fluid flow, and has been successfully used to model different types of micro-swimmers (159, 160, 161, 162, 149). The incompressible Brinkman equation is

$$\mu_f \Delta \mathbf{v}^* - \frac{\mu_f}{K_D} \mathbf{v}^* = \nabla p^* - \mathbf{F}^*, \quad \nabla \cdot \mathbf{v}^* = 0 \quad (6.2.1)$$

where $\mathbf{v}^*$ is the fluid flow, p^* is the pressure, μ_f is the viscosity of the fluid, and the fibers or particles are represented via the Darcy permeability K_D (units of length squared). Here, $\mathbf{F}^*$ is a force density that a micro-swimmer exerts on the surround-

ing fluid (163, 149, 164, 148, 165). We use an immersed boundary approach where the swimmer is assumed to be a neutrally buoyant structure immersed in the incompressible fluid (166). We note that for micro-swimmers, this governing equation is only valid when the fibers are sparse enough such that there is room for the swimmer to navigate without pushing on any of the fibers. Spielman (158) derived a relationship between the darcy permeability and the volume fraction of fibers and we have previously shown that a volume fraction of less than 2% fibers is valid and in a biologically relevant range for micro-swimmers such as sperm (149, 163, 148).

Defining L as the characteristic length scale (length of the micro-swimmer) and $\alpha = L/\sqrt{K_D}$ as the resistance parameter, we arrive at the nondimensional Brinkman equation,

$$\Delta \mathbf{v} - \alpha^2 \mathbf{v} = \nabla p - \mathbf{F}, \quad \nabla \cdot \mathbf{v} = 0. \quad (6.2.2)$$

This equation governs the fluid flow $\mathbf{v}$ at any point $\mathbf{x}$ in the domain. The utility in using this equation is that as $\alpha \rightarrow 0$, we recover a Stokes flow and by varying this single parameter, we can explore different *in vivo* fluid environments. Since the force density corresponding to the micro-swimmers will be a singular force layer, we write $\mathbf{F}(\mathbf{x}) = \mathcal{F}\delta(\mathbf{x} - \mathbf{X})$ where the micro-swimmer location(s) are given by $\mathbf{X}$ and δ is the delta distribution. Since the Brinkman equation is linear, we can utilize fundamental solutions to solve for the resulting flow. Although we note that due to the coupled fluid-structure interaction, the elastic response of the swimmer to drag or flow is non-linear.

The force density corresponding to the micro-swimmers in Eq. (6.2.2) is a singular force layer. We write $\mathbf{F}$ as $\mathbf{F}(\mathbf{x}) = \mathcal{F}\delta(\mathbf{x} - \mathbf{X})$, for micro-swimmer location(s) $\mathbf{X}$ and delta distribution δ . However, swimmers represented as curves will lead to singular integrals. To resolve this issue, we can regularize the forces using a mollifier

$\phi_\varepsilon(\mathbf{x} - \mathbf{X})$ and then solve for the resulting regularized fundamental solution for the Brinkman equation, known as the regularized Brinkmanlet (147, 148). In this case, $\mathbf{F}$ in (6.2.2) is replaced with $\mathcal{F}\phi_\varepsilon(\mathbf{x} - \mathbf{X})$, where ε is the regularization parameter corresponding to the region where most of the force is spread to the fluid. In the case of a micro-swimmer, we can choose this width to correspond to the radius of the flagellum (tail); the width we choose has emergent swimming speeds for swimmers that match previous asymptotic results (163, 149, 160, 148).

We will study micro-swimmers such as sperm in a 2-dimensional, infinite fluid. Each swimmer is represented as a discretized set of points, $\mathbf{X}_k$ for $k = 1, \dots, N$, where $r = \|\mathbf{x} - \mathbf{X}_k\|$ (utilizing the standard Euclidean norm). The solution approach involves taking the divergence of the regularized version of (6.2.2) to solve for the pressure, which is then plugged back into (6.2.2) to solve for the resulting Brinkman flow. Since the Brinkman equation is linear, we have a solution that is a superposition of all of the point forces. As described in (148), the resulting fluid flow and pressure at a point $\mathbf{x}$ due to $\mathcal{N}$ regularized point forces on the structure of the swimmer is given as

$$\mathbf{p}(\mathbf{x}) = \sum_{i=1}^{\mathcal{N}} \mathcal{F}_i \nabla G_\varepsilon(r), \quad (6.2.3a)$$

$$\mathbf{v}(\mathbf{x}) = \sum_{i=1}^{\mathcal{N}} \left[-\mathcal{F}_i B_\varepsilon''(r) + (\mathcal{F}_i \cdot (\mathbf{x} - \mathbf{X}_i))(\mathbf{x} - \mathbf{X}_i) \frac{r B_\varepsilon''(r) - B_\varepsilon'(r)}{r^3} \right], \quad (6.2.3b)$$

where $\phi_\varepsilon = \nabla^2 G_\varepsilon$ and B_ε is implicitly defined as $(\Delta - \alpha^2)B_\varepsilon = G_\varepsilon$. Note that $\int_0^\infty r \phi_\varepsilon dr = \frac{1}{2\pi}$ and as $\varepsilon \rightarrow 0$, the dirac delta distribution is recovered. Details of the numerical method are given in Leiderman et al. (148, 167). For the swimmer, at each point in time, it will have a given configuration $\mathbf{X}$ that will lead to forces $\mathcal{F}$ along the structure (details in the Appendix). Thus, at each moment in time, we are solving for the instantaneous fluid flow. To march forward in time, we assume

that the swimmer is moving with the resulting fluid flow, i.e. $d\mathbf{X}/dt = \mathbf{v}(\mathbf{X})$. In the case of $\mathcal{M}_S$ swimmers each with $\mathcal{N}_T$ points, the summation in (6.2.3a)–(6.2.3b) is then taken over the $\mathcal{M}_S\mathcal{N}_T$ points. This becomes computationally intensive when solving for a large number of time steps or for a large number of micro-swimmers since there are approximately $\mathcal{O}(\mathcal{M}_S\mathcal{N}_T)$ calculations at each time step.

The micro-swimmer we model is a simplified representation of a sperm with a head (cell body) attached to a flagellum (tail). Since we are at zero-Reynolds number, this will correspond to force-free and torque-free swimming (120). (Note that in the case when the sum of forces in the system is non-zero, extra terms must be accounted for in (6.2.3b) (168, 167).) The active force generation along the elastic flagellum propels the swimmer forward whereas the head is a passive and fairly rigid structure (containing the genetic material) (11). The flagellum is discretized into $\mathcal{N}_F$ points ($\mathbf{X}_F^j$) and the head is discretized into $\mathcal{N}_H$ points ($\mathbf{X}_H^j$), for a total of $\mathcal{N}_T = \mathcal{N}_F + \mathcal{N}_H$ point forces on each micro-swimmer ($j = 1, \dots, \mathcal{M}_S$). We detail the specifics of the force models in the Appendix. Briefly, using a model similar to Fauci and McDonald (169), we derive an energy $E(\mathbf{X}^j)$ for each swimmer j , where a variational derivative of this energy results in the forces,

$$\begin{aligned} \mathcal{F}^j &= -\frac{\partial E^j}{\partial \mathbf{X}^j} \\ &= -\frac{\partial}{\partial \mathbf{X}^j} (E_{F,bend}^j(\mathbf{X}_F^j) + E_{F,tens}^j(\mathbf{X}_F^j) + E_{H,bend}^j(\mathbf{X}_H^j) + E_{H,tens}^j(\mathbf{X}_H^j) + E_N^j(\mathbf{X}_F^j, \mathbf{X}_H^j)), \end{aligned} \quad (6.2.4)$$

and this is the force density that is put into (6.2.2) as a regularized force, $\mathbf{F} = \mathcal{F}\phi_\epsilon$. The total energy $E(\mathbf{X})$, will be a sum of different energy components corresponding to a bending and tensile energy in the flagellum ($E_{F,bend}$ and $E_{F,tens}$) and head ($E_{H,bend}$ and $E_{H,tens}$), as well as an energy component corresponding to the connection of the head and flagellum in the neck region (E_N). The bending energy will drive

the dynamics and resulting motion of the flagellum and is detailed in the Appendix.

6.3 Results

We now apply the Bayesian framework of Chapter 2 to the cases of a single micro-swimmer and two interacting micro-swimmers. Previous studies have shown complex relationships between stiffness of the flagellum and emergent swimming speeds and trajectories (122, 126, 165, 170); fluid resistance, characteristic of *in vivo* environments, also increases this complexity (163, 149, 164, 148). The mapping between parameter space (material and/or fluid parameters) and swimmer trajectories is not fully understood due to the non-linear elastic responses of the flagellum. To ensure we fully explored the parameter space, as well as ensuring posterior distributions were continuous, we used 2000 samples for each generation of TMCMC.

The forward model for the micro-swimmers uses a time-step of Δt (refer to Table 6.6) to propagate the solution forward in time up to nondimensional time $T = 20$ (corresponding to 40 beat cycles of the tail and 2 seconds of real time). Although we are interested in both the *in vivo* (fluids with $\alpha \neq 0$) and the *in vitro* case with $\alpha \rightarrow 0$, we choose our *in silico* reference data as either a swimming speed based on the center of the head of the swimmer or the distance between the heads and tails of swimmers (in the multi-swimmer case). This is motivated by the *in vitro* setup where images are taken in the range of 20 to 60 Hz and resolution often only allows tracking of the center of the head (135). To estimate quantities of interest, we utilize observations at ~ 25 Hz. As a proof of concept, we use simulated observation data

by perturbing model outputs with Gaussian noise as

$$D_k = \xi_k + \sigma \epsilon_k$$

where D_k is the observation from the k^{th} position of the vector, ξ_k is the k^{th} model output, ϵ_k is a zero-mean, unit-variance Gaussian variable, and σ is the level of the noise. In order to have a meaningful noise-level, we choose σ to be a fraction $\sigma = \gamma\beta$, where γ is the percent of noise considered, and β is the mean of the absolute value for all model outputs. The model prediction error covariance Σ from Equation 2.1.5 is assumed to be diagonal, such that $\Sigma = \sigma \mathbf{I}$ whose non-zero entries all have the same magnitude σ .

6.3.1 Single Micro-Swimmer

As illustrated in Fig. 6.1, even though each simulation has the same initial conditions and beat form parameters (Table 6.6), the emergent swimming speed and trajectory of the deterministic swimmer model is a function of both the resistance parameter α (corresponding to the volume fraction of stationary fibers) as well as the parameters relating to the stiffness of the flagellum that is actively bending to propel the swimmer forward. Asymptotic analysis of infinite length swimmers with a prescribed beat form in a Brinkman fluid provides insight into these relationships; increased swimming speeds can be obtained with increased amplitude and/or increased resistance (160, 163). However, the work required to achieve larger amplitude bending at a higher volume fraction of fibers in the fluid (corresponding to a larger α) becomes too large of a hurdle for micro-swimmers to overcome (163, 149). That is why in models with emergent beat forms, there is often an enhancement in swimming speeds for a moderate range of α and then a decreased swimming speed due to a decreased

achieved amplitude at larger α (148, 149). In addition, the parameters related to the tail stiffness will lead to a range of emergent beat forms, with different achieved amplitudes, which then results in different swimming speeds. As can be seen from Fig. 6.1(A) to (B), halving only the tail curvature stiffness parameter K_C results in a slower swimmer with smaller achieved amplitude (seen in tail traces). From Eq. (6.4.1), we can observe that the magnitude of K_C controls how closely the swimmer will try to maintain the preferred beat form. In (C), the x -location of the center of the head can be seen for different α and K_C . At this range of parameter values, we observe that the higher α results in a small increase in forward progression and that the larger K_C results in a larger increase in forward progression. The small oscillations in the graphs correspond to the beat form where the swimmer may actually move backwards for a small portion of the beat but in the overall beat, forward progression is made.

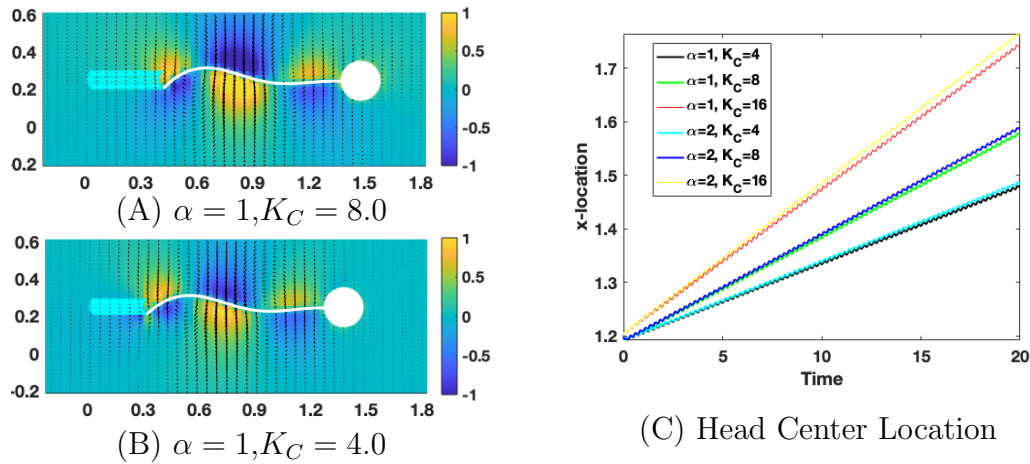


Figure 6.1: Representative computational model results for the micro-swimmer. A swimmer at $T = 20$ using resistance parameter $\alpha = 1$ and tail curvature stiffness $K_C = 8.0$ in (A) and $\alpha = 1$ and $K_C = 4.0$ in (B). Flow field (arrows) and pressure (colorbar) in (A)-(B) are normalized by the maximum values in (A). The trace of the end of the tail for $T = 0 - 20$ is also shown. A comparison of the x -location of the center of the swimmer head is in (C) for different values of α and K_C .

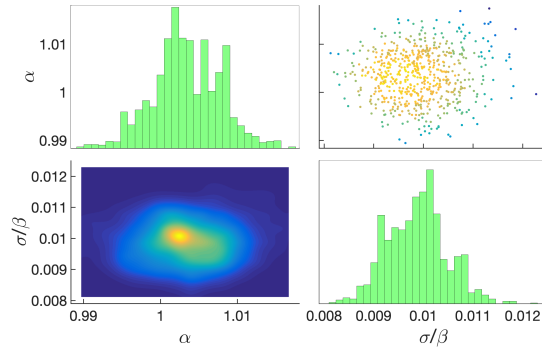
For the single micro-swimmer case, the model outputs $\mathbf{g}$ are generated from our

deterministic model. Using the center of the head of the micro-swimmer, the x - and y - velocities are estimated, resulting in 50 x -velocity and 50 y -velocity observations when solving up to $T = 20$ (data at 25 Hz). This test can be summarized by the following question: Which swimmer stiffness and/or fluid resistance parameter would be the likely parameters for this data and with what certainty are we sure of these parameter values? Due to the non-linear complexities of the model, we do not know *a priori* whether a single mode is likely or which combinations of values are likely in the case of more than one unknown parameters. To illustrate the feasibility of this approach, we consider two scenarios for parameter estimation: first, estimating only the fluid resistance α and the noise level σ/β and second, estimating the fluid resistance α , the tail curvature stiffness parameter K_C , and the noise level σ/β . The nominal parameter values taken are $\alpha = 1.0$ and $K_C = 8.0$. In both of these cases, we also demonstrate the strength of the method by perturbing the observation data by 1%, 10%, and 20% noise.

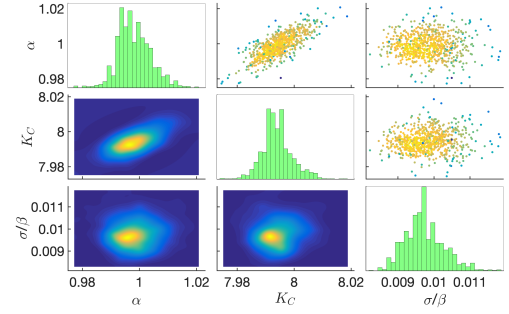
Noise Level	α	u_α (%)	σ/β	$u_{\sigma/\beta}$ (%)
1%	1.0035	0.436	0.00987	6.006
10%	1.0297	3.354	0.0968	5.376
20%	1.0421	7.780	0.196	6.006

Table 6.1: Posterior means and uncertainties for parameter estimation of resistance parameter α and noise level σ/β for the single swimmer using velocity data with varying noise levels (given in left hand column). The degree of uncertainty for the parameters is characterized by the coefficient of variation, u_α and $u_{\sigma/\beta}$, for resistance α and noise σ/β , respectively.

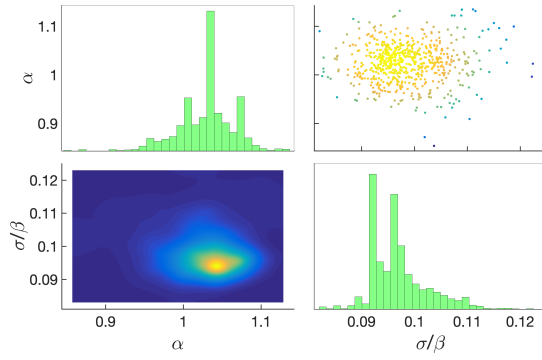
For the first case, we assume a uniform prior for the parameters $(\alpha, \sigma/\beta)$ on the space $[0.0, 2.0] \times [0.0, 0.5]$. Oviductal fluids often have a very low volume fraction of fibers, corresponding to α in this range (148, 151), and the nominal value is $\alpha = 1$. Since we are not sure which value is more likely *in vivo*, a uniform prior is appropriate. The results are summarized in Table 6.1, and the 1% noise case is displayed in Fig. 6.2(A). In this case, the joint marginal distribution is an ellipsoid,



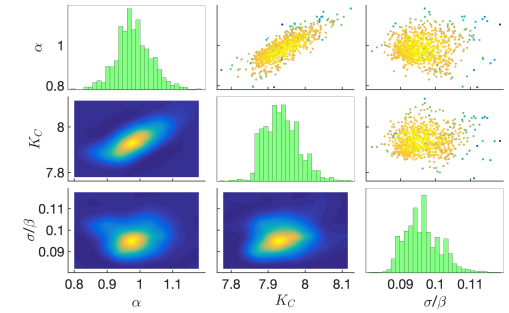
(A) One Parameter, 1% noise



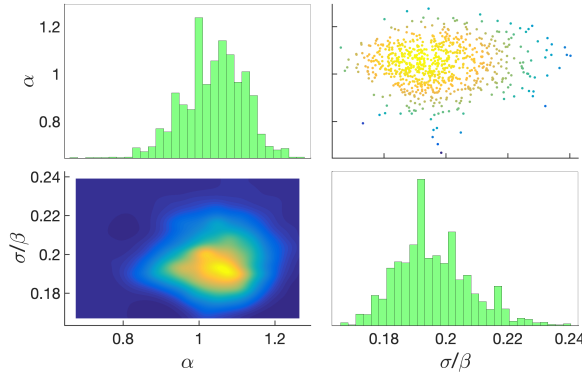
(D) Two Parameters, 1% noise



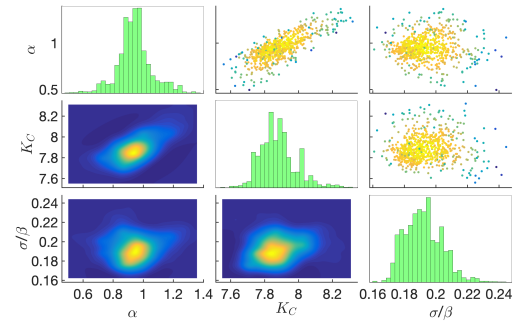
(B) One Parameter, 10% noise



(E) Two Parameters, 10% noise



(C) One Parameter, 20% noise



(F) Two Parameters, 20% noise

Figure 6.2: Parameter estimation results for a single swimmer with fluid resistance $\alpha = 1.0$ and tail curvature stiffness $K_C = 8.0$. In each experiment, noisy velocity data from the center of the head were used to track the movement. For each noise level, the experiment was run until $T = 20$. Histograms for each parameter are displayed along the main diagonal of the figure. Subfigures below the diagonal show the marginal joint density functions for each pair of parameters, while subfigures above the diagonal show the samples used in the final stage of TMCMC. Colors correspond to probabilities, with yellow likely and blue unlikely.

matching intuition that the noise σ/β and resistance parameter α are uncorrelated. Considering the recovered parameter values in row one of Table 6.1, the parameters found are 1.0035 for α and 0.00987 for σ/β , matching closely to the nominal values of 1 and 0.01, respectively. Beyond recovering the nominal parameters from the data, the joint marginal distribution shows how different combinations of noise and resistance parameter can lead to the synthetic data used. We observe a very small region in the middle that is yellow, corresponding to a small region with very high likelihood.

To quantify the degree of uncertainty for each parameter's posterior distribution, we compute the coefficient of variation, defined as the ratio between a parameter's standard deviation to its mean (denoted by u_α and $u_{\sigma/\beta}$ here). In this case, the coefficient of variation for α is 0.436%, showing that we are fairly certain in the recovered mean. In addition, the nominal parameter values are included within two-standard deviations of the recovered means for both parameters. Hence, for only a noisy set of head locations, we can discover the likely fluid parameters when the noise is at 1%.

The same experiment is run using 10% and 20% additive Gaussian noise, displayed in Figs. 6.2(B) and (C) and Table 6.1. For both of these cases, we see that the nominal parameter values are within two-standard deviations of the recovered means. However, the distributions are much more spread out than the 1% noise case. For example, the 1% additive noise joint posterior distribution for α ranges in value from 0.99 to 1.01, while the 10% additive noise ranges from 0.9 to 1.1 and the 20% noise ranges from 0.8 to 1.2. Even though there is a large amount of noise and a wider distribution, the methodology is still be able to recover likely parameter values. The wider distributions led to a coefficient of variation for the α parameter with the following trends: u_α for 10% additive noise is about 10 times as large as

that for 1% noise and the 20% noise is approximately double that of the 10% noise, matching very closely to the increase in the noise level. In addition, we can see that the coefficient of variation for the noise level stays the same between the various noise levels, matching intuition that the certainty in the noise should not change as more noise is added to the system.

Noise Level	α	u_α (%)	K_C	u_{K_C} (%)	σ/β	$u_{\sigma/\beta}$ (%)
1%	0.9982	0.580	7.9944	0.0647	0.009728	5.47
10%	0.9820	5.47	7.9368	0.58	0.0964	5.34
20%	0.9426	11.93	7.8800	1.35	0.1922	6.11

Table 6.2: Posterior means and uncertainties for parameter estimation on resistance α and tail curvature stiffness K_C for the single swimmer using velocity data for varying noise levels. The degree of uncertainty for the parameters is characterized by the coefficient of variation, $u_\star$ for parameter $\star$ where the estimated noise is σ/β .

Next, we estimate both α and tail curvature stiffness parameter K_C , assuming a uniform prior for the parameters $(\alpha, K_C, \sigma/\beta)$ on $[0, 2] \times [0, 16] \times [0, 0.5]$. Again, we also perturb the data using 1%, 10%, and 20% Gaussian noise. In the results displayed in Table 6.2 and Fig. 6.2(D) for the 1% noise level, we can see that there is a positive correlation between the fluid resistance parameter α and tail curvature stiffness K_C with a Pearson correlation coefficient of 0.7554: as the volume fraction of fibers in the fluid increases (larger α), a stiffer tail is required to achieve similar swimming speeds. Like the one-parameter case, the joint marginal distributions including the noise have a single mode and are uncorrelated, i.e. the parameters are independent of the noise. Since both α and K_C lead to different emergent trajectories, we can use the joint marginal distributions to further understand the complex relationships between parameter space and model outputs.

We repeated these experiments using 10% and 20% additive noise, displayed in Figs. 6.2(E) and (F) and Table 6.2. From the table, we can see that all nominal parameter values are recovered within two-standard deviations of the estimated means.

Similar to the single parameter case, the noise for α and K_C also scales with the overall level of additive noise: the 10% noise case has about 10 times more uncertainty than the 1% case, and the 20% noise case has standard deviations twice as large as the 10% noise case. As seen in the 1% noise case, all the figures also show a strong positive correlation between α and K_C with Pearson correlation coefficients of 0.8065 and 0.7362 for 10% and 20%, respectively.

6.3.2 Single Swimmer and Stokes Flow

In order to show that the constitutive parameters can be identified for cases of Stokes flow, we consider the case where $\alpha = 0$ and estimate parameters for the properties of micro-swimmer tails. In this case, the observed data again are the x - and y -velocities when solving up to $T = 20$ (data at 25 Hz). We consider two scenarios: first, estimating the tail curvature stiffness parameter K_C and the noise level σ/β and second, estimating the tail curvature stiffness parameter K_C , tail spring stiffness parameter K_T , and the noise level σ/β . The nominal parameter values taken are $K_C = 8.0$ and $K_T = 10000$. In both cases, we perturb the observation data by 1% noise.

We first consider a uniform prior for the parameters $(K_C, \sigma/\beta)$ on the space $[0.0, 16.0] \times [0.0, 0.1]$. As with the previous case of only estimating α , the recovered joint posterior distribution is ellipsoid, matching intuition that the noise σ/β and tail curvature stiffness K_C are uncorrelated, seen in Figure 6.3. The recovered parameters found are 7.996 for K_C and 0.00973 for σ/β with coefficients of variation of 0.052% and 5.76%, respectively. In other words, the nominal parameter values are recovered within one-standard deviation of the recovered means, and the recovered posterior distributions are centered tightly around the nominal values.

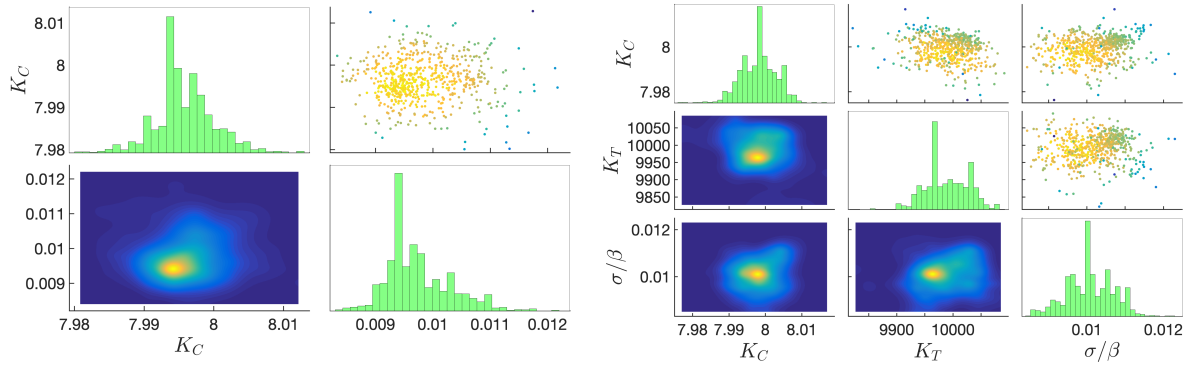
(A) K_C , 1% noise(B) K_C and K_T , 1% noise

Figure 6.3: Parameter estimation results for a single swimmer with fluid resistance $\alpha = 0.0$ (corresponding to Stokes flow), tail curvature stiffness $K_C = 8.0$, and tail spring stiffness $K_T = 10000$. In each experiment, noisy velocity data from the center of the head were used to track the movement. For each noise level, the experiment was run until $T = 20$.

Secondly, we estimate both K_C and the tail spring stiffness parameter K_T , assuming a uniform prior for the parameters $(K_C, K_T, \sigma/\beta)$ on $[0, 16] \times [0, 20000] \times [0, 0.1]$ and perturb the data using 1% Gaussian noise. The results are shown in Fig. 6.3(B) where we see that all of the parameters seem relatively uncorrelated, with the posterior distributions having an ellipsoidal shape. We recovered means of 7.9984, 9989.57, 0.0101 for K_C , K_T , and σ/β , respectively. The corresponding coefficients of variation are 0.058%, 0.38%, and 5.93%, i.e. in all three cases, the nominal parameter values are recovered within one-standard deviation of the estimated means. Due to the coefficient of variations being so small, the distributions are fairly tight around the nominal parameter values.

6.3.3 Two Micro-Swimmers

Representative results of the deterministic model for the two micro-swimmer case are shown in Fig. 6.4. Two swimmers are initialized in the same exact initial condition,

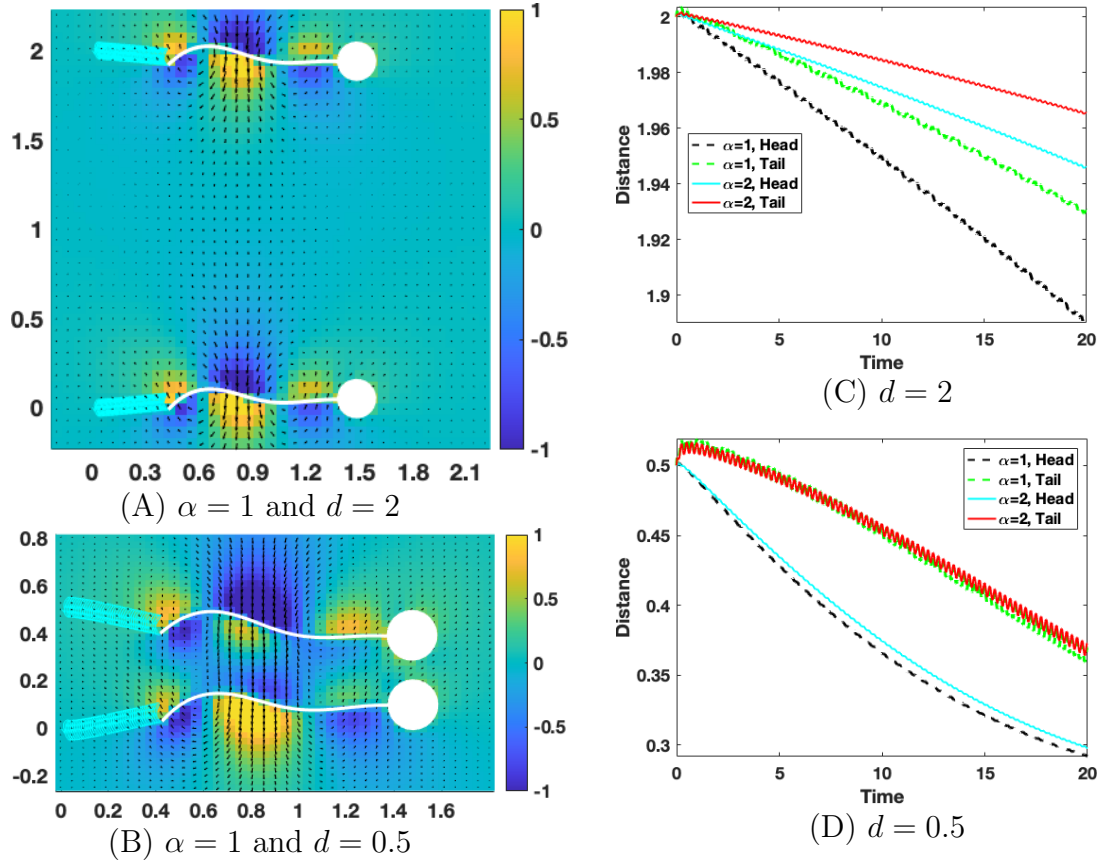


Figure 6.4: Representative simulation results for two micro-swimmers that are initially a distance of 2 units apart in (A) and 0.25 units apart in (B). These are both for the resistance parameter $\alpha = 1$ at $T = 20$; the time course of the tail undulations of each swimmer are highlighted to the left of the swimmer and the normalized flow field is depicted with arrows and the pressure is visualized with the colorbar. For each of the cases, the distance between the center of the heads of the swimmers as well as the distance between the last point of each of the swimmers is given in (C) for 2 units apart initially and 0.25 units apart in (D). These are all for tail curvature stiffness $K_C = 8.0$.

separated by a vertical distance of $d = 2$ in (A) and $d = 0.5$ in (B). The flow field and swimmer locations are at $T = 20$ and the entire tail trajectory from $T = 0 - 20$ is shown for each swimmer. Each swimmer feels the other swimmer via the fluid and for $\alpha = 1$, attraction is observed in both cases, although it is much stronger when initialized at $d = 0.5$ in (B). As previously observed, the time scale of attraction will be a function of α as well as the stiffness parameters of the tail (165, 148). In (C)-(D), the distance between the head centers and the last points on the tail are

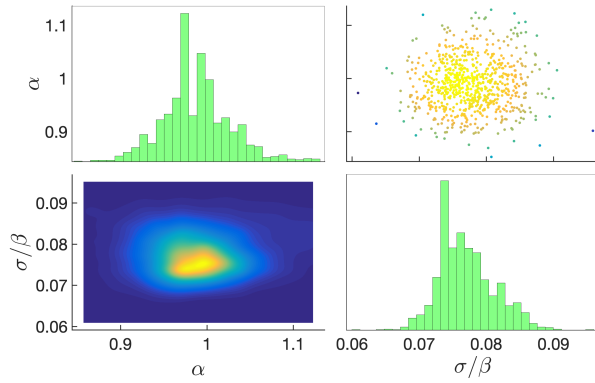
shown for initial separations of $d = 2$ and $d = 0.5$, respectively.

To investigate parameter estimation and uncertainty quantification in the two micro-swimmer case, the observations at 25 Hz will correspond to two data points: the distance between the center of the two swimmers' heads and the distance between the last points on the tail of the swimmers. The swimmers initially start parallel to each other with the center of their heads $d = 2$ units apart and we solve up to $T = 20$. We use the same amount of noise on both sets of observations.

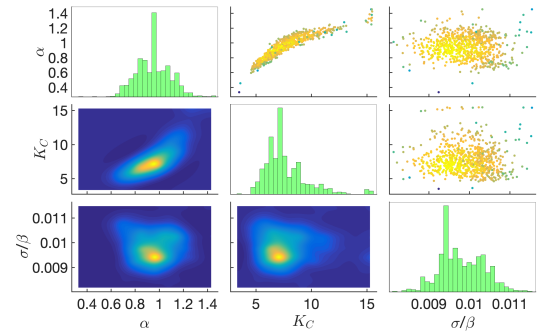
The 1% noise case using 2000 samples for each generation of TMCMC for estimating α and σ/β is displayed in Fig. 6.5(A). The joint marginal distribution for α and σ/β is unimodal and uncorrelated. In addition, it is centered closely around the nominal parameter value for α and includes the nominal parameter value for the level of the noise. The distributions for the 10% noise case are displayed in Fig. 6.5(B) and the results are summarized in Table 6.3. Similar to the 1% noise case, the joint marginal distribution for α and σ/β is ellipsoidal, displaying that the two parameters are uncorrelated. The recovered means include the nominal parameter values within one-standard deviation. Furthermore, the coefficient of variation for the parameters increases in proportion to the increase in the noise level: for 1% noise, the coefficient of variation $u_\alpha = 3.94\%$, while for the 10% noise case, the coefficient of variation $u_\alpha = 34.67\%$. This shows as the noise in the parameters increase, the certainty in our parameter values correspondingly decreases.

Noise Level	α	u_α (%)	σ/β	$u_{\sigma/\beta}$ (%)
1%	0.9878	3.94	0.00962	5.35
10%	1.0763	34.67	0.09646	5.14

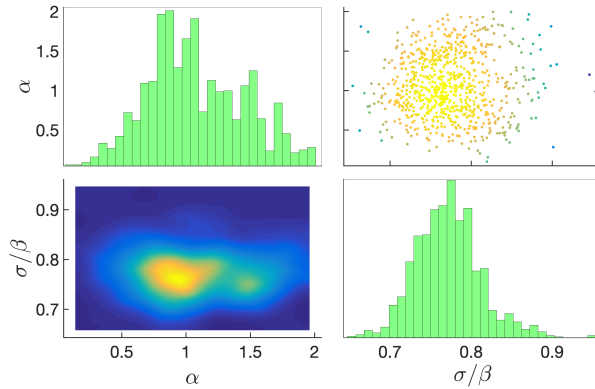
Table 6.3: Posterior means and uncertainties for parameter estimation of resistance α for two swimmers using distance between the heads and tails as noisy observation data. The degree of uncertainty for the parameters is characterized by the coefficient of variation, $u_\star$ for parameter $\star$ where the estimated noise is σ/β .



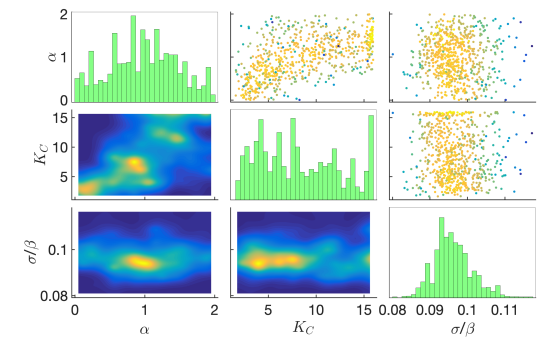
(A) One Parameter, 1% noise



(C) Two Parameters, 1% noise



(B) One Parameter, 10% noise



(D) Two Parameters, 10% noise

Figure 6.5: Parameter estimation results for two swimmers with fluid resistance $\alpha = 1.0$ and tail curvature stiffness $K_C = 8.0$. In each experiment, noisy distance data between the center of the heads and tips of the tails were used as reference data. For each noise level, the experiment was run until $T = 20$. Histograms for each parameter are displayed along the main diagonal of the figure. Subfigures below the diagonal show the marginal joint density functions for each pair of parameters, while subfigures above the diagonal show the samples used in the final stage of TMCMC. Colors correspond to probabilities, with yellow likely and blue unlikely.

As with the single swimmer case, we also estimated both the fluid resistance α and the tail curvature stiffness K_C of the micro-swimmers through observing the distance between the head and tails of two swimmers. The results are summarized in Fig. 6.5(C)-(D) and Table 6.4. As in the single swimmer case, there is a positive correlation between the fluid resistance and the curvature stiffness for the tails of the micro-swimmers with Pearson correlation coefficients of 0.9062 and 0.6123 for the 1% and 10% noise cases. In comparison to the single swimmer case in Fig. 6.2(D)-(F), we observe more complex relationships between noisy trajectories with parameters K_C and α . In Fig. 6.5(D), there are distinct combinations of K_C and α that have similar trajectories to the data. This highlights how different parameter combinations may lead to similar trajectories, which may not have been discovered with previous methods. More details on this can be seen in Table 6.7 in the Appendix.

However, we note that the uncertainty in the parameters are much larger than the single-swimmer case, and grows as the noise in the observations increase. One reason for this is that the dynamics of attraction may not yet be achieved by time $T = 20$ because of the distance apart that the swimmers began (refer to Fig. 6.4(A) and (C)). Due to this and the large amount of noise in the system, the resulting distributions are very spread out in parameter space. In order to better approach the parameter estimation question in this case, data need to be recorded for a longer amount of time so that the swimmers attract and attain their beat formation.

Noise Level	α	u_α (%)	K_C	u_{K_C} (%)	σ/β	$u_{\sigma/\beta}$ (%)
1%	0.9533	14.35	7.6632	24.56	0.00980	5.46
10%	0.9768	49.46	8.2374	50.27	0.09616	5.27

Table 6.4: Posterior means and uncertainties for parameter estimation on resistance α and curvature stiffness K_C for two swimmers using distance between the heads and last point on tails as noisy observation data.

6.3.4 Two Micro-Swimmers, Two Standard Deviations

The distance between the swimmers' tails and the swimmers' head distances can vary in time and can be on different scales. In Fig. 6.4(D), the head's and tail's are $d = 0.5$ apart but by $t = 20$, the head's are less than 0.3 apart and the tail's are greater than 0.36 apart. If the same amount of noise is added for the entire simulation to both head and tail data, the noisy trajectories could cause large uncertainty in likely parameters. Since we generally do not know the exact time course of the change in distances, but we do know that in general the cell body will attract before the tails (165), we choose to use different standard deviations σ_t and σ_h for the tail data x_t and head data x_h , respectively. The goal is to have a different, and possibly more accurate, representation of how noise is added to the system. Here, the covariance matrix is still assumed to be diagonal, but now $\Sigma_{j,j} = \sigma_t$ for $j = 1, 3, \dots, m-1$ and $\Sigma_{j,j} = \sigma_h$ for $j = 2, 4, \dots, m$. The data are corrupted according to

$$D_k = \begin{cases} \xi_k + \sigma_t \epsilon_k, & k = 1, 3, \dots, m-1, \\ \xi_k + \sigma_h \epsilon_k, & k = 2, 4, \dots, m. \end{cases}$$

Since the previous case had issues resolving the position of the swimmers in part because the swimmers began too far apart, we initially set the center of the heads to be $d = 0.5$ units apart and measure the distance between the heads and tails at 25 Hz. We take 2000 samples at each generation of TMCMC and again compare the results for two different noise levels.

For the 1% noise level case, results are shown in Fig. 6.6(A) and Table 6.5. In the figure, a clear correlation between the fluid resistance α and the tail curvature stiffness K_C , as seen in the previous experiments. We observe that parameter combinations that could be representative of the data include pairs of α and K_C where

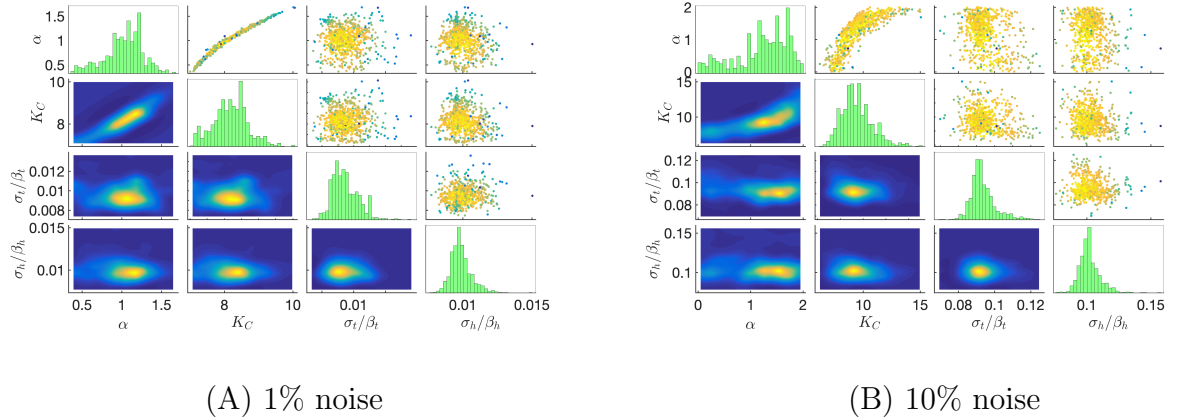


Figure 6.6: Parameter estimation results for two swimmers initially 0.5 apart with fluid resistance $\alpha = 1.0$ and curvature stiffness of tail $K_C = 8.0$ using two different noise levels for the tails and the heads. In each experiment, noisy distance data between the center of the heads and tips of the tails were used as reference data. The experiment was run until $T = 20$. Histograms for each parameter are displayed along the main diagonal of the figure. Subfigures below the diagonal show the marginal joint density functions for each pair of parameters, while subfigures above the diagonal show the samples used in the final stage of TMCMC. Colors correspond to probabilities, with yellow likely and blue unlikely.

higher α corresponds to higher K_C . In addition, we see that the marginal distributions between the parameters and the noise appear to be uncorrelated. Finally, all nominal parameter values are recovered within two-standard deviations of the recovered means, indicating the ability of the method to recover parameters from noisy data.

Noise Level	α	u_α (%)	K_C	u_{K_C} (%)	σ_t/β_t	u_{σ_t/β_t} (%)	σ_h/β_h	u_{σ_h/β_h} (%)
1%	1.0597	27.27	1.0353	7.56	0.00959	11.18	0.0100	10.51
10%	1.2275	40.70	1.1833	14.22	0.09343	7.25	0.1020	7.68

Table 6.5: Posterior means and uncertainties for parameter estimation of resistance α and curvature stiffness K_C for two swimmers using distance between the heads and tails as noisy observation data and looking for two different noise levels. The degree of uncertainty for the parameters is characterized by the coefficient of variation, $u_\star$ for parameter $\star$ where the estimated noise is σ/β for either the head or tail data (h and t subscripts, respectively).

The 10% noise case in Fig. 6.6(B) and Table 6.5 shows similar results to the 1% noise case. Again, a strong positive correlation is seen between the α and K_C

parameters with Pearson correlation coefficient of 0.7888 and the joint marginal distributions are unimodal and ellipsoidal, indicating low correlation between the noise and the parameter values with Pearson correlation coefficients on the order of 0.01 to 0.25. Furthermore, all nominal parameter values are recovered within two-standard deviations of the recovered means. We note that the distributions of the parameters are more spread out than the 1% noise case, but this is due to the added noise from the observations. Additionally, even with the additional noise, these distributions give a clear understanding of likely combinations of α and K_C that lead to the given noisy trajectory.

6.4 Discussion and Conclusions

Overall, the results highlight that the Bayesian framework can estimate both fluid and swimmer material parameters while quantifying uncertainty in the estimates. Since there are complex dependencies on emergent swimming speeds and trajectories, it is not surprising that the potential parameter distributions from which the data could come from grew as the noise in the data increased. Through several test cases we have highlighted that the amount of data used as well as the type of data (along with noise) must be considered carefully as it will have an effect on the ability to estimate parameters. In micro-swimmer applications, a long enough time interval must be chosen to capture the full dynamics in order to estimate the parameter with high certainty. In addition, in the case of multiple swimmers, if data is of slightly different magnitudes, using two different noise levels led to tighter distributions of parameter values.

In terms of *in vitro* data, CASA data motivates the use of only head or tail trajec-

tories in the 20-60 Hz (135). We were able to successfully recover likely parameters given noisy data representative of CASA data. In our computational framework, we utilized the Brinkman equation for the fluid governing equation. This has a resistance parameter where water is representative of fluids with $\alpha \rightarrow 0$ where as gels or oviductal fluids would correspond to $\alpha > 0$. We have shown through simulations, that parameter estimation can be done in both of these cases. Additionally, we note that uniform prior distributions were used for parameters, but this is sensible since there is generally a known range *a priori*. For example, oviductal fluid has a known range for the volume fraction of fibers (148, 151), corresponding to resistance $\alpha \in [0, 2]$ for this model. Additionally, experiments are often completed to determine elastic moduli of the tail (137, 138). One interesting thing to note is that measured quantities for elastic moduli are based on the particular experimental protocol and may not be a direct match for the term in the model. In these cases, the parameter range could be specified and this parameter estimation framework could be utilized to determine the likely parameters for the model that fit the data.

Due to the non-linearity in the curvature portion of the elastic force of the swimmer, the fluid-structure interaction problem leads to complex relationships between emergent trajectories of the swimmer as a function of fluid resistance and/or elastic moduli of the swimmer. Prior to running the simulations, we did not know how swimmer head trajectories would correlate with these parameters; intuition gives general relationships that for a fixed tail stiffness parameter, increasing resistance would most likely lead to a trajectory that makes less forward progress. However, it was not known whether there could be multiple pairs of values, or a range of values that could lead to similar trajectories. The joint marginal distributions highlight the new knowledge about the fluid dynamics of swimming. For example in Fig. 6.2(D)-(F), if the data was a noisy trajectory of the solo swimmer, there is a range of highly

likely parameter values. Through these simulations, we learn that as the noise increases in the trajectory, there is a larger range of likely parameters that is generally in a compact region that is not disjoint. Based on this current setup, there are not not disconnected modes in the posterior distributions, which is something that we predicted but have now verified. These joint marginal distributions from this methodology will continue to provide further insight into the complex relationships between parameter space and model outputs.

The ability to use this Bayesian framework to estimate parameters will aid in further understanding emergent properties of micro-swimmers since computational resources limit running models throughout the entire parameter space to understand emergent wave forms and trajectories of swimmers in different fluid environments (when using fully resolved fluid-structure interaction models). Experiments show synchronization of bull sperm beat forms (171), alignment of mouse sperm (head to tail) (172), and sea urchin sperm swimming in vortices (173). Utilizing the proposed framework with noisy data of sperm trajectories, there is the potential to identify what parameter ranges or distributions that could lead to these emergent patterns.

Although we have focused on micro-swimmers in a Brinkman fluid, we emphasize that it is possible to use this methodology in a variety of contexts. Given noisy data of an elastic micro-swimmer, parameter identification could be used to identify fluid properties e.g. visocisty in a Stokesian fluid or viscosity and/or relaxation time of elastic polymers in the fluid if viscoelastic and assuming governed by an Oldroyd B type model. Additionally, we focused on estimating the tail curvature stiffness parameter but in general, the algorithms have the potential to identify any parameters, including internal properties of swimmer, e.g. constitutive parameters (e.g. stiffness parameter or elastic moduli).

In this chapter, our chief concern was to offset the computational cost of the micro-swimmer model through minimizing the number of evaluations needed for the likelihood computation. Our TMCMC-based approach allowed us to thoroughly explore the parameter space and resulted in reliable estimates for the posterior distributions. We remark that if observed data are not fully known from the onset or if there is a prohibitively large amount of observations, filtering-based approaches (174, 175, 20) can aid in re-sampling either to incorporate new data as they become available or to split the data for computational efficiency.

Although we did not explore model selection, this framework can be utilized to not only estimate parameters but decide the likelihood that the data is described by a particular model. We have demonstrated these methods previously in identifying arterial wall abnormalities in Chapter 3, discovering epidemic origin in Chapter 4, and choosing the best model for DNA methylation in Chapter 5. In the future, we could use these methods to test different constitutive laws and models for micro-swimmers.

Appendix

Swimmer Model

Details of the force derivation and numerical algorithm are given here. We assume that the flagellum will beat with a curvature wave corresponding to a low amplitude tapered sine wave, as observed in experiments with human sperm (176). This

corresponds to a preferred flagellum configuration

$$\hat{X}_F^j(s, t) = [\hat{x}(s, t), \hat{y}^j(s, t)] = [s, a^j(1 - s) \sin(\eta^j s - \omega^j t)],$$

where the j -th micro-swimmer has amplitude a , wavelength $2\pi/\eta$, and beat frequency $\omega/2\pi$. Here, s is a parameter initialized as arc length where $0 \leq s \leq 1$ (the nondimensional length of the flagellum is 1). The flagellum will attempt to reach this preferred configuration and we will define the bending energy as

$$E_{F,bend}^j = K_C^j \int_{\Gamma_F^j} \left(\zeta^j(s, t) - \hat{\zeta}^j(s, t) \right)^2 ds, \quad (6.4.1)$$

where Γ_F^j is the centerline curve corresponding to the j -th flagellum and K_C is a stiffness coefficient enforcing the curvature or bending constraint. The preferred curvature $\hat{\zeta}^j$ and actual curvature $\zeta^j(s, t)$ of the flagellum are given as

$$\hat{\zeta}^j = \frac{\partial^2 \hat{y}^j}{\partial s^2}, \quad \zeta^j(s, t) = \frac{\frac{\partial^2 y^j}{\partial s^2} \frac{\partial x^j}{\partial s} - \frac{\partial^2 x^j}{\partial s^2} \frac{\partial y^j}{\partial s}}{\left(\left(\frac{\partial x^j}{\partial s} \right)^2 + \left(\frac{\partial y^j}{\partial s} \right)^2 \right)^{3/2}}, \quad (6.4.2)$$

where $\mathbf{X}_F^j = [x^j, y^j]$.

In addition to the bending component, we will account for an additional energy component that will tend to maintain the inextensibility of the flagellum. This results in

$$E_{F,tens}^j = \int_{\Gamma_F^j} K_T^j \left(\left\| \frac{\partial^2 \mathbf{X}_F^j}{\partial s^2} \right\| - 1 \right)^2 ds, \quad (6.4.3)$$

which in a discretized form, corresponds to Hookean springs between points on the flagellum with stiffness coefficient K_T .

Similar to the flagellum, we assume a preferred shape or curvature of the head.

In this simple model, we will assume a head shape with radius H_r and preferred curvature $\hat{\kappa} = 1/H_r$. The corresponding energy is

$$E_{H,bend}^j = \int_{\Gamma_H^j} K_{H,C}^j (\kappa^j(s,t) - \hat{\kappa}(s,t))^2 ds,$$

where Γ_H^j corresponds to the circular head. Here, the actual curvature $\kappa^j(s,t)$ is calculated using the same equation as ζ in (6.4.2), but now $\mathbf{X}_H^j = [x^j, y^j]$. In addition, we also have an energy to maintain inextensibility in the head, the same as (6.4.3) using Γ_H^j , $\mathbf{X}_H^j$, and $H_{C,tens}^j$, where we envision Hookean springs between points on the membrane of the head as well as springs connecting points on the circular head that are π apart (we choose $\mathcal{N}_H$ to be even to ensure points and springs exactly π apart).

The swimmer is initialized (left to right) to have the center of the circular head be placed with a y-coordinate the same as the rightmost point on the flagellum and an x-coordinate that is shifted to the right of the rightmost point by H_r and an additional small distance apart, dN . To ensure that the passive head remains attached to the actively bending flagellum, and to represent the stiff neck region of a sperm, we connect the head and flagellum with 5 springs. These springs connect the rightmost point (the $\mathcal{N}_F$ -th point) of the flagellum to the points on the circle with $\theta = (\pi - d\theta), \pi, (\pi + d\theta)$ where $d\theta$ is the angular spacing between the $\mathcal{N}_H$ points on the head. Additionally, there are two springs connecting the second rightmost point on the flagellum ($\mathcal{N}_F - 1$) to the points on the circle with $\theta = \pi \pm d\theta$. These springs will have a stiffness coefficient $K_{N,tens}$. There is also an energy based on the desired angle between the flagellum and the head. Let z_1 be the vector connecting the $\mathcal{N}_F$ -th point on the flagellum and the point on the head with $\theta = \pi$ and let z_2 be the vector connecting the points on the head with $\theta = \pi \pm d\theta$. In general, we wish for these vectors to be approximately orthogonal, and we can derive an energy

and hence forces that penalize this deviation, tending to maintain $\mathbf{z}_1 \cdot \mathbf{z}_2 = 0$ with stiffness coefficient $K_{N,ang}$ (169).

Given a configuration for each of the $\mathcal{M}_S$ swimmers at the initial time point, we determine the forces on the $\mathcal{M}_S \mathcal{N}_T$ discretized points using (6.2.4), where each of the components are calculated using (6.4.1)–(6.4.3). Second order finite difference approximations are utilized in the calculation of all derivatives in the energy components and a trapezoidal rule is used to calculate the integrals. The forces are then used to calculate the resulting velocity at points along the discretized swimmer, (6.2.3b). The location of the swimmer is updated using the no-slip condition, numerically implemented with a forward Euler method. The next time step is reached, where this calculation is repeated.

In these simulations, when there is more than one swimmer, we assume that the beat form parameters such as the amplitude and beat frequency are the same for each swimmer. In addition, we assume that all stiffness parameters are the same. All parameters are given in Table 6.6.

Table 6.6: Parameters for swimmer model.

L , characteristic length scale	100 μm
$\mathcal{N}_F$, points on flagellum	50
$\mathcal{N}_H$, points on head	24
ε , regularization parameter	0.045
$K_{N,tens}$, tensile stiffness of neck	500
$K_{H,tens}$, tensile stiffness of head	10000
K_{tens} , tensile stiffness of flagellum	10000
K_C , curvature stiffness of flagellum	8
$K_{H,C}$, curvature stiffness of head	0.8
$K_{N,ang}$, stiffness for flagellum-body connecting angle	10000
Δt , time step	1×10^{-4}
η , wavenumber	2π
a , amplitude	0.1
ω , frequency	4π

In Region			Out of Region		
α	K_C	L_2 error	α	K_C	L_2 error
0.70	5.00	0.042125	0.70	7.50	0.052375
0.90	7.50	0.041838	0.80	10.00	0.058950
1.00	8.00	0.041432	0.90	10.00	0.049307
1.10	9.00	0.041947	1.00	5.00	0.050663
1.20	15.00	0.044655	1.10	14.00	0.048324

Table 6.7: L_2 error between the observations and model outputs for a sample of α and K_C pairs.

Spread of Posterior Distributions

We remark that sometimes the coefficient of variation is not the best measure of performance for the sampling method as the distributions can be multimodal. However, what TMCMC recovers is a region of possible parameter values and discovers relationships between the various model parameters that affect the dynamics of the swimmers. In Fig. 6.5(C)-(D), a positive correlation is seen between the parameters for K_C and α , indicating that an increase in one correspondingly requires an increase in the other to maintain the same system dynamics. Table 6.7 contains the L_2 error between the observations and model outputs for a sample of α and K_C pairs to show that those contained within the recovered posterior distribution result in small changes from the original parameter pairs. For example, the L_2 error for the true parameter values for $\alpha = 1.0$ and $K_C = 8.00$ is approximately 0.0414. Other pairs of parameters along the curve, such as (1.1, 9.0) and (0.70, 5.0), have similar L_2 error values of 0.420 and 0.0421, respectively. If we look at parameters not included in the distribution, we find that the L_2 error is much larger, e.g. for the pair (0.80, 10.00), we recover an error of 0.589. One of the key benefits of TMCMC is its ability to uncover these multiple modes in posterior distributions, which is demonstrated in this case.

CHAPTER SEVEN

Conclusion

7.1 Summary of Contributions

In this work, we tailored and used a Bayesian uncertainty quantification framework to study parameter estimation and model selection for a wide array of biological systems. We demonstrated that the framework serves as a tool to discover underlying properties of the mechanics of the models, through discovering correlations between various parameters. In addition, the methodology is easily used to compare various models and find the one that best fits the observed data. Below I summarize the key insights we learned about these systems.

In Chapter 2, we presented a method for parameter estimation and model selection. The key benefit of the TMCMC algorithm is that it fully explores the parameter space; traditional Markov chain Monte Carlo methods can get “stuck” in one specific region of parameters and, as a result, take a long time to converge. If the forward problem is computationally expensive, some MCMC algorithms may require running the algorithm longer than feasible. TMCMC combines simulated annealing with traditional MCMC methods in order to slowly transition between intermediate distributions, allowing it to better explore the parameter space. The $\Pi 4U$ implementation of TMCMC is highly parallelizable, making the problems studied here more tractable.

We first explored arterial blood flow models in Chapter 3 and identified where aneurysms occurred in a network using noisy *in silico* velocity data from a 19-artery network. Studying the parameter estimation problem allowed us to see correlations in the parameters, e.g. that a stiffer artery requires higher blood viscosity in order to achieve a similar velocity profile. We demonstrated that the method recovers the system parameters and identifies where abnormalities occur in the network. Further-

more, we demonstrated the robustness of the approach by perturbing the underlying model, so no combination of free parameters is capable of reproducing the observed data. The method was still able to recover which arteries were damaged, but over-compensated the magnitude of the damage due to all parameters being perturbed.

Chapter 4 introduced a model for epidemic spread on a network of populations, and we used the uncertainty quantification framework to estimate the infectiousness of a disease and where the epidemic originated. Again, visualizations of the parameter estimation framework allowed us to see how the parameters interact in order to maintain the system dynamics – a higher infection rate correspondingly has a higher recovery rate. The method allowed us to look at some basic sensor placement strategies: by studying the 20-Barbell graph, we found that taking observations from each side of the graph led to less noisy parameter estimation. The method was robust to perturbations of the network: missing edges in the graph did not effect the performance of the parameter estimation or origin identification, but just led to higher overall noise tolerances. Finally, we could identify where an epidemic originated.

In Chapter 5, we compared two different DNA methylation models and determined whether we could differentiate between the two models using noisy observations. Again, we could see correlations in the system parameters, which measured the various methylation rates that describe the system dynamics. A key finding of our work was whether model parameters were dependent upon various methylation rates. The issue of whether or not rates of methylation/demethylation are inexorably linked to promoter topology is an ongoing question the experimental community has been attempting to unravel. Therefore, the fact that our theoretical analysis can explore this issue is a significant feature of our work which can be explored further by the experimental community.

Due to the computational complexity of micro-swimmer models with fully resolved hydrodynamics, parameter estimation has been prohibitively expensive; however, due to its parallel nature, the Bayesian framework described above was utilized to understand the emergent trajectories and swimming speeds of micro-swimmers in Chapter 6. Each swimmer feels the other via the fluid and the time scale of attraction is a function of α (the volume fraction of fibers in the fluid) as well as the stiffness parameters of the tail K_c . Taking sparse observations, the distance between the center of the two micro-swimmers' heads and the distance between the last two points on the tails were used. The distributions resulted in a strong positive correlation between the viscous parameter α and the tail stiffness parameter K_c : as the volume fraction of the fibers in the fluid increases (larger α), a stiffer tail is required to achieve similar swimmer distances. The noise parameters for the observations appeared to be uncorrelated with the other parameters, i.e. the parameters are independent of the noise. Also, all nominal parameter values are included within two-standard deviations of the means of the recovered posterior distributions, demonstrating the method's effectiveness in recovering system parameters. Next steps in this project would be to use model selection to test different constitutive laws and models for micro-swimmers to discover which best match observed data.

In addition to the work detailed in this thesis, Clark Bowman, Alexander Roitershtein, Derek Stein, Anastasios Matzavinos, and I also developed a dissipative particle dynamics (DPD) model for a lipid bilayer membrane to study its biological properties (31). Due to the high computational complexity of DPD simulations, parameter estimation has been prohibitively expensive. Using the membrane curvature, we estimated the system parameters (the lipid tail ϕ_t and ϕ_b of the top and bottom lipid layers, respectively). The joint posterior distributions recovered were multimodal, demonstrating multiple parameters could feasibly generate the reference

curvature.

7.2 Future Directions

Now that we have demonstrated the robustness of the methodology to a wide array of biological systems, we are delving more deeply into better understanding several of the applications. Below I give a brief description of some of our future projects and the first steps that we have taken in approaching these problems.

7.2.1 Classification of Patients

In Chapter 3, we introduced a method for estimating the stiffness in arteries in order to identify abnormalities in the network. Recently, we have been working with Dr. Roy Aaron from the Department of Orthopedic Surgery at Brown University to apply this methodology to study patients with certain autoimmune diseases. Our goal is to observe distinct posterior distributions for the arterial stiffness, indicating that the diseases impact the arterial elasticity. This could ultimately create a protocol to help identify potential patients and better understand how to treat the disease using non-invasive procedures.

We consider four arteries in the leg: the common femoral (CF), deep femoral (DF), superficial femoral (SFA), and popliteal. The common femoral artery bifurcates into the deep femoral and superficial femoral, and the superficial femoral transitions into the popliteal artery. There is a belief that abnormal arterial flow coexist with osteoarthritis of the knee to a significant degree than in control subjects, suggesting

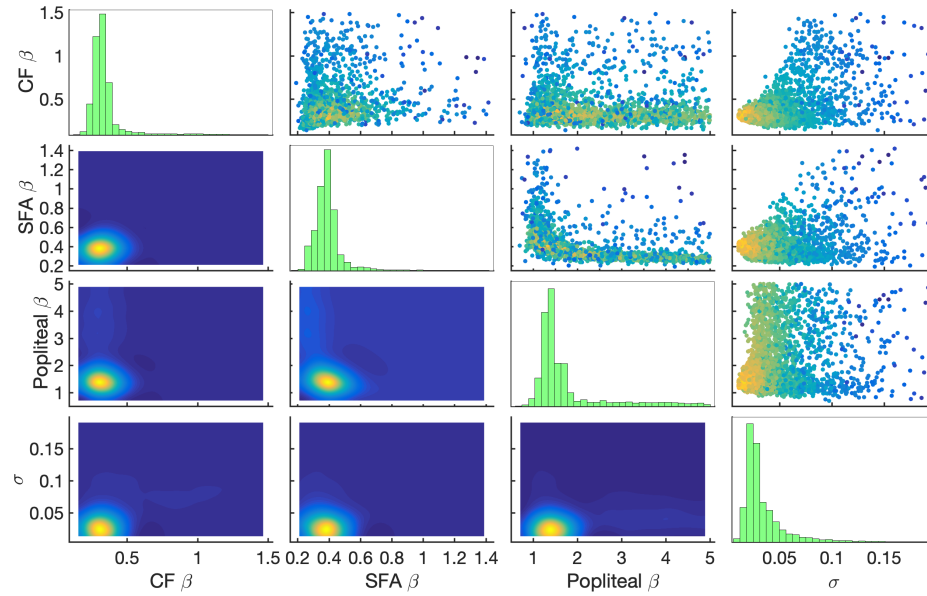


Figure 7.1: Parameter estimation results for the three stiffnesses β for the common femoral (CF), superficial femoral (SFA), and popliteal arteries and noise level σ using minimum and maximum flow velocity data from a patient. Figures on the diagonal show histograms for each parameter. Subfigures below the diagonal show the marginal joint densities for each pair of parameters, while subfigures above the diagonal show the samples used in the final (convergent) stage of TCMCMC. Colors correspond to likelihoods, with yellow likely and blue unlikely. Red dots indicate the nominal parameter values.

a connection between osteoarthritis and cardiovascular disease.

To take first steps to study this question, we use the one-dimensional blood flow model described in Chapter 3 to approximate the flow in this four-artery network. Using minimum and maximum flow velocity for the common femoral, superficial femoral, and popliteal arteries, we try to recover parameters related to the stiffness of those three arteries. The arterial lengths and initial cross-sectional artery parameters are taken from medical measurements.

The results for our first experiment is displayed in Figure 7.1. For all of the arterial stiffness parameters, we recover tight distributions surrounding the param-

eters. The noise parameters are a little more noisy than the other samples, but this is likely due to using limited data (two velocity values for each artery) and the simplified model. We cannot validate the recovered parameter values yet, but we are interested in comparing the parameters recovered for healthy patients versus those with osteoarthritis.

Next steps of this project involve gathering more detailed information about the flow velocity through these four arteries, either a time-series or more flow data, and performing model selection for a healthy patient and a sick patient to see if the recovered distributions differ.

7.2.2 Graphical Model Estimation

A natural question that arises from these problems is “How are the networks of populations generated?” In part, this can be done from using real population migration behavior to construct where edges should appear on the graphs and the corresponding weights. However, this sort of data can be difficult to record or obtain, so we started a basic process to construct the population network based upon epidemic data from each of the communities.

Most simply, we considered the case of three interacting populations with the SIR epidemic on a graph and considered three scenarios visualized below: [7.2a](#) all three populations mix, [7.2b](#) the first and third populations do not directly mix, but both interact with the second population and [7.2c](#) only two populations mix and the third is separate. We denote the transition rate as $r_{i,j}$ for the transfer between nodes i and j . For simplicity, unlike in Chapter [4](#), we use the same transition rates for susceptible, infective, and removed individuals.

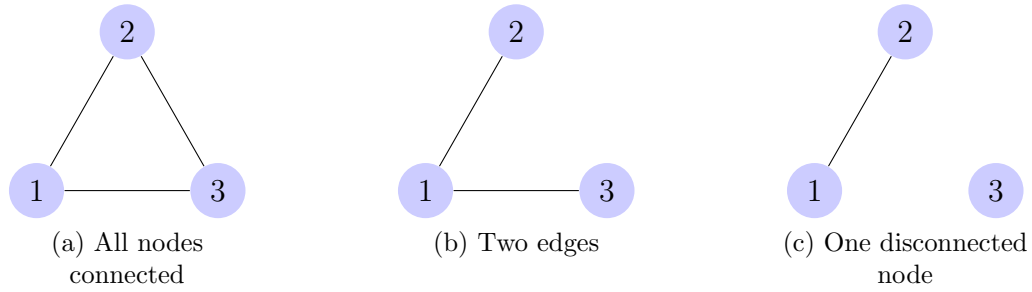


Figure 7.2: Three possible configurations for a three node graph.

One way to approximate the topology of the graph would be to model the graph as having all edges connected, but to estimate the transportation weights on each of the edges. To demonstrate that we are able to recover the structure of the network, we consider three different edge and weight scenarios. In all three cases, the populations for nodes 2 and 3 are purely susceptible with 100 individuals, and the first node has 5 susceptible and 95 infective individuals.

The first case considered is represented in Figure 7.2a where all three nodes are directly connected with transportation weights of 0.1 from node 1 to node 2, 0.05 between nodes 1 and 3, and 0.025 between nodes 2 and 3. We use noisy, *in silico* observation data of the infective and removed populations from all three nodes at times $t = 5, 6, 7$, and 8 to estimate the transition rates $r_{i,j}$ between the each of the nodes. In all three cases, the parameters are sampled from the prior $[0.00, 1.00] \times [0.00, 1.00] \times [0.00, 1.00] \times [0.00, 0.50]$.

Model	$\hat{r}_{1,2}$	$u_{\hat{r}_{1,2}}$ (%)	$\hat{r}_{1,3}$	$u_{\hat{r}_{1,3}}$ (%)	$\hat{r}_{2,3}$	$u_{\hat{r}_{2,3}}$ (%)	$\hat{\sigma}/\alpha$	$u_{\hat{\sigma}/\alpha}$ (%)
3 Edges	0.1005	2.37	0.0503	2.31	0.0254	1.37	0.0119	21.29
2 Edges	0.0991	1.83	0.0503	1.57	0.000341	60.78	0.0110	15.91
1 Edge	0.1067	20.25	0.0000000508	86.87	0.000000360	93.32	0.0958	0.97

Table 7.1: Posterior means and uncertainties for parameter estimation on the three-node graph for three cases for the number of edges (noise level $\sigma = 0.01\alpha$).

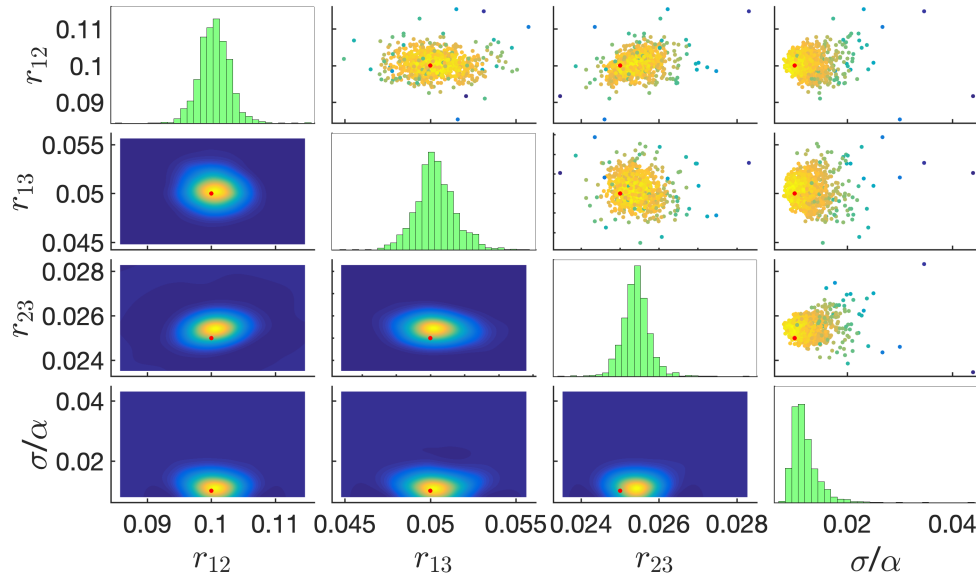


Figure 7.3: Parameter estimation results for the three transition rates between the three nodes on the network $r_{1,2}$, $r_{1,3}$, $r_{2,3}$ and noise level σ using corrupted infective and susceptible data from times $t = 5, 6, 7$, and 8 from a graph with three edges (Figure 7.2a). Figures on the diagonal show histograms for each parameter. Subfigures below the diagonal show the marginal joint densities for each pair of parameters, while subfigures above the diagonal show the samples used in the final (convergent) stage of TMCMC. Colors correspond to likelihoods, with yellow likely and blue unlikely. Red dots indicate the nominal parameter values.

The first row of Table 7.1 and Figure 7.3 display the results for this parameter estimation. Looking at the recovered joint marginal distributions for each pair of parameters, we find minimal correlation between the various parameter values. In addition, all recovered parameter means are very closely centered around the nominal parameter values with the recovered values found as $(0.1005, 0.0503, 0.0254, 0.0119)$ and the nominal values used were $(0.100, 0.050, 0.025, 0.10)$. The coefficient of variation is small for the transition parameters, about 2% of the recovered means, but a bit more uncertain for the recovered noise.

The second case considered is the two edge scenario where node 1 is directly connected to both nodes 2 and node 3, but node 2 is only indirectly connected to node 3 (see Figure 7.2b). Here, the transition rate between nodes 1 and 2 ($r_{1,2}$)

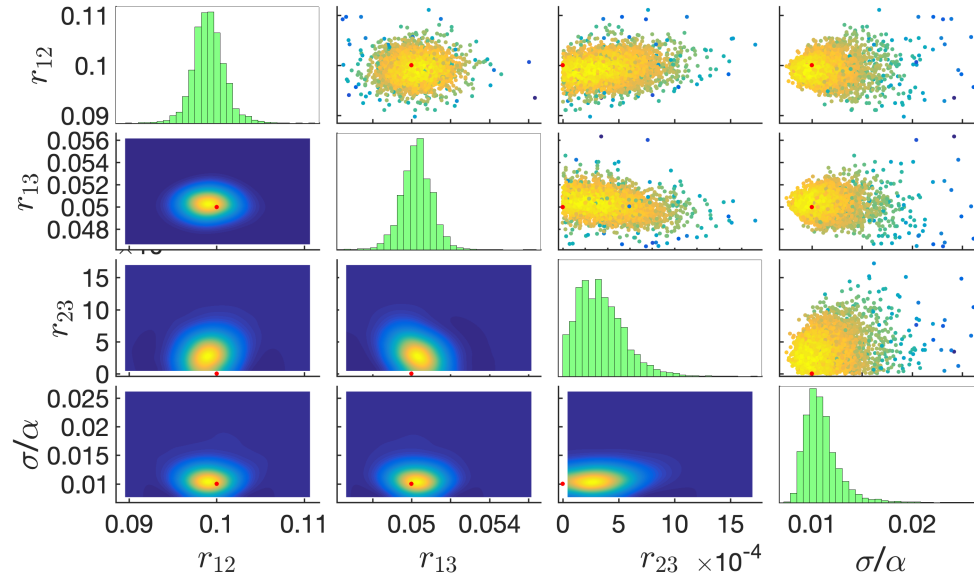


Figure 7.4: Parameter estimation results for the three transition rates between the three nodes on the network $r_{1,2}$, $r_{1,3}$, $r_{2,3}$ and noise level σ using corrupted infective and susceptible data from times $t = 5, 6, 7$, and 8 from a graph with two edges (Figure 7.2b).

is 0.100 and the transition rate between nodes 1 and 3 ($r_{1,3}$) is 0.050. As in the previous case, we use noisy, *in silico* observation data of the infective and removed populations from all three nodes at times $t = 5, 6, 7$, and 8 to estimate the transition rates $r_{i,j}$ between the each of the nodes.

The second row of Table 7.1 and Figure 7.4 display the corresponding results for this experiment. In this case, we see that the estimated values for $r_{1,2}$, $r_{1,3}$, and σ are very close to the nominal parameter values of 0.100, 0.050, and 0.01, recovering means of 0.0991, 0.0503, and 0.0110, respectively. In addition, all three of these parameters have relatively low coefficients of variation. However, the recovered mean for $r_{2,3}$ is 0.000341 with relatively high uncertainty. The value is a bit perturbed from 0, likely due to the indirect connection that node 3 has with node 2. However, it should be noted that the recovered value is quite small.

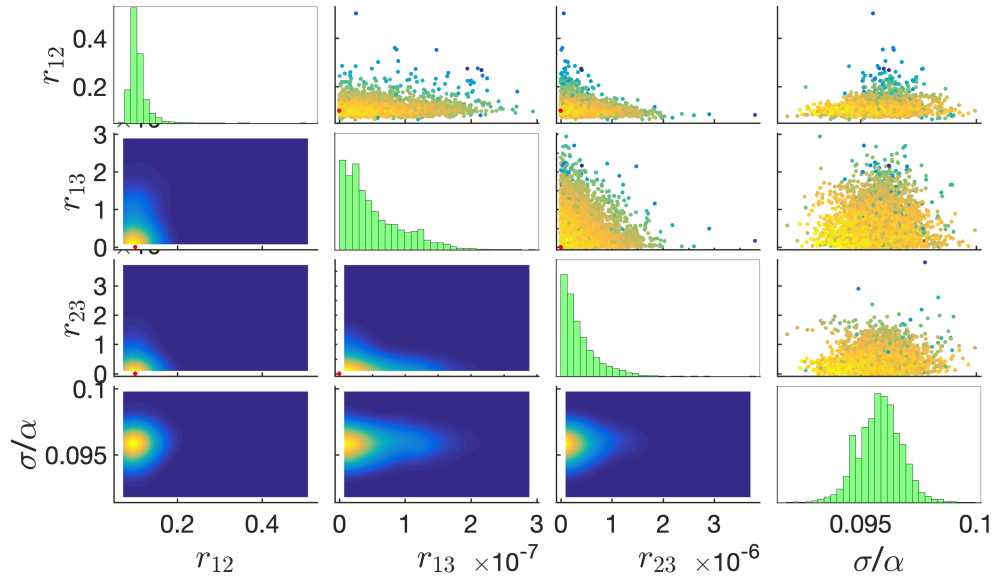


Figure 7.5: Parameter estimation results for the three transition rates between the three nodes on the network $r_{1,2}$, $r_{1,3}$, $r_{2,3}$ and noise level σ using corrupted infective and susceptible data from times $t = 5, 6, 7$, and 8 from a graph with one edge (Figure 7.2c).

Finally, we consider the one edge scenario, where only nodes 1 and nodes 2 are connected, and node 3 is isolated (see Figure 7.2c). In this case, the nominal parameter values are taken to be $(0, 0.10, 0.000, 0.000, 0.0100)$ for $(r_{1,2}, r_{1,3}, r_{2,3}, \sigma)$.

The results are shown in the final row of Table 7.1 and Figure 7.5 and are definitive. The recovered parameter means for the transition rates are $(0.1067, 5.08 \times 10^{-8}, 3.60 \times 10^{-7}, 0.0958)$ for $(r_{1,2}, r_{1,3}, r_{2,3}, \sigma)$. The coefficient of variation for $r_{1,3}$ and $r_{2,3}$ seem large, however this is due to the smallness of the recovered parameter values. The noise parameter is relatively large in this instance, but that is likely due to method being unable to exactly match the noisy observation data for node 3.

In this, we have demonstrated that the methodology is able to determine the structure of a network of interacting populations using noisy *in silico* epidemic data. In other words, we are able to determine both where there are connections between

communities as well as the strength of those connections. Natural extensions to this work would be considering more complicated graphs or using different sorts of population dynamics to estimate the structure of a community.

Bibliography

- [1] Christopher JW Breward, Helen M Byrne, and Claire E Lewis. A multiphase model describing vascular tumour growth. *Bulletin of Mathematical Biology*, 65(4):609–640, 2003.
- [2] Anastasios Matzavinos, Chiu-Yen Kao, J Edward F Green, Alok Sutradhar, Michael Miller, and Avner Friedman. Modeling oxygen transport in surgical tissue transfer. *Proceedings of the National Academy of Sciences*, 106(29):12091–12096, 2009.
- [3] Yue Yu, Paris Perdikaris, and George Em Karniadakis. Fractional modeling of viscoelasticity in 3D cerebral arteries and aneurysms. *Journal of Computational Physics*, 323:219–242, 2016.
- [4] W.O. Kermack and A.G. McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 115(772):700–721, 1927.
- [5] V. Capasso and R.E. Wilson. Analysis of a reaction-diffusion system modeling man–environment–man epidemics. *SIAM J. Appl. Math.*, 57(2):327–346, 1997.
- [6] M.J Keeling and K.T.D. Eames. Networks and epidemic models. *J. R. Soc. Interface*, 2(4):295–307., 2005.
- [7] Meaghan J Jones, Sarah J Goodman, and Michael S Kobor. DNA methylation and healthy human aging. *Aging cell*, 14(6):924–932, 2015.
- [8] Michael Klutstein, Deborah Nejman, Razi Greenfield, and Howard Cedar. DNA methylation in cancer and aging. *Cancer research*, 76(12):3446–3450, 2016.
- [9] Philip L De Jager, Gyan Srivastava, Katie Lunnon, Jeremy Burgess, Leonard C Schalkwyk, Lei Yu, Matthew L Eaton, Brendan T Keenan, Jason Ernst, Cristin McCabe, et al. Alzheimer’s disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. *Nature neuroscience*, 17(9):1156, 2014.
- [10] GI Taylor. Analysis of the swimming of microscopic organisms. *Proc Roy Soc Lond Ser A*, 209:447–461, 1951.

- [11] E. A. Gaffney, H. Gadêlha, D. J. Smith, J. R. Blake, and J. C. Kirkman-Brown. Mammalian sperm motility: observation and theory. *Annu Rev Fluid Mech*, 43:501–528, 2011.
- [12] SD Olson, SS Suarez, and LJ Fauci. Coupling biochemistry and hydrodynamics captures hyperactivated sperm motility in a simple flagellar model. *J Theor Bio*, 283:203–216, 2011.
- [13] Panagiotis Angelikopoulos, Costas Papadimitriou, and Petros Koumoutsakos. Bayesian uncertainty quantification and propagation in molecular dynamics simulations: a high performance computing framework. *The Journal of chemical physics*, 137(14):144103, 2012.
- [14] Shuoxing Wu, P Angelikopoulos, C Papadimitriou, R Moser, and Petros Koumoutsakos. A hierarchical Bayesian framework for force field selection in molecular dynamics simulations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2060):20150032, 2016.
- [15] David F Anderson and Thomas G Kurtz. *Stochastic analysis of biochemical systems*, volume 1. Springer, 2015.
- [16] Dongbin Xiu. *Numerical methods for stochastic computations: a spectral method approach*. Princeton university press, 2010.
- [17] Panagiotis E Hadjidoukas, Panagiotis Angelikopoulos, Costas Papadimitriou, and Petros Koumoutsakos. Π4U: A high performance computing framework for Bayesian uncertainty quantification of complex models. *Journal of Computational Physics*, 284:1–21, 2015.
- [18] Olivier Le Maître and Omar M Knio. *Spectral methods for uncertainty quantification: with applications to computational fluid dynamics*. Springer Science & Business Media, 2010.
- [19] Ralph C Smith. *Uncertainty Quantification: Theory, Implementation, and Applications*, volume 12. SIAM, 2013.
- [20] Albert Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*, volume 89. SIAM, 2005.
- [21] David Barber. *Bayesian reasoning and machine learning*. Cambridge University Press, 2012.
- [22] Bradley Efron and Trevor Hastie. *Computer age statistical inference*, volume 5. Cambridge University Press, 2016.
- [23] Panagiotis Angelikopoulos, Costas Papadimitriou, and Petros Koumoutsakos. X-TMCMC: Adaptive kriging for bayesian inverse modeling. *Computer Methods in Applied Mechanics and Engineering*, 289:409–428, 2015.
- [24] Karen R Larson, Clark Bowman, Zhizhong Chen, Panagiotis Hadjidoukas, Costas Papadimitriou, Petros Koumoutsakos, and Anastasios Matzavinos. Data-driven prediction and origin identification of epidemics in population networks. *Submitted*, 2020.

- [25] Jianye Ching and Yi-Chu Chen. Transitional Markov chain Monte Carlo method for Bayesian model updating, model class selection, and model averaging. *Journal of engineering mechanics*, 133(7):816–832, 2007.
- [26] K Larson, C Bowman, C Papadimitriou, P Koumoutsakos, and A Matzavinos. Detection of arterial wall abnormalities via Bayesian model selection. *Royal Society Open Science*, 6:182229, 2019.
- [27] Karen Larson, Loukas Zagkos, Mark Mc Auley, Jason Roberts, Nikos I. Kaval-laris, and Anastasios Matzavinos. Data-driven selection and parameter esti-mation for DNA methylation mathematical models. *J Theor Biol*, 467:87 – 99, 2019.
- [28] Karen R Larson, Sarah D Olson, and Anastasios Matzavinos. Bayesian uncer-tainty quantification for micro-swimmers with fully resolved hydrodynamics. *Submitted*, 2020.
- [29] James L Beck and Ka-Veng Yuen. Model selection using response measure-ments: Bayesian probabilistic approach. *Journal of Engineering Mechanics*, 130(2):192–203, 2004.
- [30] Ka-Veng Yuen. *Bayesian methods for structural dynamics and civil engineer-ing*. John Wiley & Sons, 2010.
- [31] Clark Bowman, Karen Larson, Alexander Roitershtein, Derek Stein, and Anas-tasios Matzavinos. Bayesian uncertainty quantification for particle-based sim-ulation of lipid bilayer membranes. In Magda Stolarska and Nicoleta Tarfulea, editors, *Cell Movement: Modeling and Applications*. Springer, 2018.
- [32] Michael W Vanik, James L Beck, and Siu-Kui Au. Bayesian probabilistic approach to structural health monitoring. *Journal of Engineering Mechanics*, 126(7):738–745, 2000.
- [33] Daniel W Stroock. *An introduction to Markov processes*, volume 230. Springer Science & Business Media, 2013.
- [34] Bruce Hajek. Cooling schedules for optimal annealing. *Mathematics of opera-tions research*, 13(2):311–329, 1988.
- [35] Richard Holley and Daniel Stroock. Simulated annealing via Sobolev inequal-ities. *Communications in Mathematical Physics*, 115(4):553–569, 1988.
- [36] Peter Salamon, Paolo Sibani, and Richard Frost. *Facts, Conjectures, and Improvements for Simulated Annealing*, volume 7. SIAM, 2002.
- [37] Heikki Haario, Marko Laine, Antonietta Mira, and Eero Saksman. DRAM: Efficient adaptive MCMC. *Statistics and computing*, 16(4):339–354, 2006.
- [38] Jasper A Vrugt, CJF Ter Braak, CGH Diks, Bruce A Robinson, James M Hy-man, and Dave Higdon. Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *Inter-national Journal of Nonlinear Sciences and Numerical Simulation*, 10(3):273–290, 2009.

- [39] Majid K. Vakilzadeh, James L. Beck, and Thomas Abrahamsson. Approximate Bayesian Computation by Subset Simulation for model selection in dynamical systems. *Procedia Engineering*, 199(Supplement C):1056 – 1061, 2017. X International Conference on Structural Dynamics, EURODYN 2017.
- [40] Leonhard Euler. Principia pro motu sanguinis per arterias determinando. *Opera posthuma mathematica et physica anno*, pages 814–823, 1844.
- [41] SJ Sherwin, V Franke, J Peiró, and K Parker. One-dimensional modelling of a vascular network in space-time variables. *Journal of Engineering Mathematics*, 47(3-4):217–250, 2003.
- [42] T. Alarcon, H.M. Byrne, and P.K. Maini. A design principle for vascular beds: The effects of complex blood rheology. *Microvasc. Res.*, 69:156–172, 2005.
- [43] J. Alastruey, K.H. Parker, J. Peiró, and S.J. Sherwin. Can the modified Allen’s test always detect sufficient collateral flow in the hand? A computational study. *Comput Methods Biomech Biomed Engin*, 9:353–361, 2006.
- [44] L. Formaggia, D. Lamponi, and A. Quarteroni. One-dimensional models for blood flow in arteries. *Journal of Engineering Mathematics*, 47:251–276, 2003.
- [45] A.D. Marquis, A. Arnold, C. Dean-Bernhoft, B.E. Carlson, and M. S. Olufsen. Practical identifiability and uncertainty quantification of a pulsatile cardiovascular model. *Mathematical Biosciences*, 304:9–24, 2018.
- [46] Jordi Alastruey, Ashraf W Khir, Koen S Matthys, Patrick Segers, Spencer J Sherwin, Pascal R Verdonck, Kim H Parker, and Joaquim Peiró. Pulse wave propagation in a model human arterial network: assessment of 1-d viscoelastic simulations against in vitro measurements. *Journal of biomechanics*, 44(12):2250–2258, 2011.
- [47] Paris Perdikaris, Leopold Grinberg, and George Em Karniadakis. An effective fractal-tree closure model for simulating blood flow in large arterial networks. *Annals of Biomedical Engineering*, 43(6):1432–1442, 2015.
- [48] Paris Perdikaris and George Em Karniadakis. Fractional-order viscoelasticity in one-dimensional blood flow models. *Annals of Biomedical Engineering*, 42(5):1012–1023, 2014.
- [49] M. Wu, H.B. Frieboes, M.A.J. Chaplain, S.R. McDougall, V. Cristini, and J.S. Lowengrub. The effect of interstitial pressure on therapeutic agent transport: Coupling with the tumor blood and lymphatic vascular systems. *J. Theor. Biol.*, 355:194–207, 2014.
- [50] Marco Fedele, Elena Faggiano, Luca Dedè, and Alfio Quarteroni. A patient-specific aortic valve model based on moving resistive immersed implicit surfaces. *Biomechanics and Modeling in Mechanobiology*, 16(5):1779–1803, Oct 2017.
- [51] Luca Formaggia, Alfio Quarteroni, and Allesandro Veneziani. *Cardiovascular Mathematics: Modeling and simulation of the circulatory system*. Springer, 2009.

- [52] A. Quarteroni, A. Manzoni, and C. Vergara. The cardiovascular system: Mathematical modelling, numerical algorithms and clinical applications. *Acta Numerica*, 26:365–590, 2017.
- [53] Hiroshi Suito, Kenji Takizawa, Viet QH Huynh, Daniel Sze, Takuya Ueda, and Tayfun E Tezduyar. A geometrical-characteristics study in patient-specific FSI analysis of blood flow in the thoracic aorta. In *Advances in Computational Fluid-Structure Interaction and Flow Simulation*, pages 379–386. Springer, 2016.
- [54] Gabor Zavodszky, Britt van Rooij, Victor Azizi, and Alfons Hoekstra. Cellular Level *in-silico* Modeling of Blood Rheology with an Improved Material Model for Red Blood Cells. *Frontiers in Physiology*, 8:563, 2017.
- [55] Pavel S. Zun, Tatiana Anikina, Andrew Svitenkov, and Alfons G. Hoekstra. A Comparison of Fully-Coupled 3D In-Stent Restenosis Simulations to *in-vivo* Data. *Frontiers in Physiology*, 8:284, 2017.
- [56] Rocco Michele Lancellotti, Christian Vergara, Lorenzo Valdetaro, Sanjeeb Bose, and Alfio Quarteroni. Large eddy simulations for blood dynamics in realistic stenotic carotids. *International Journal for Numerical Methods in Biomedical Engineering*, 33(11):e2868, 2017. e2868 cnm.2868.
- [57] Donald A Drew and Stephen L Passman. *Theory of multicomponent fluids*, volume 135. Springer Science & Business Media, 2006.
- [58] Anastasios Matzavinos and Mariya Ptashnyk. Homogenization of oxygen transport in biological tissues. *Applicable Analysis*, 95(5):1013–1049, 2016.
- [59] Emilie Marchandise, Marie Willemet, and Valérie Lacroix. A numerical hemodynamic tool for predictive vascular surgery. *Medical Engineering and Physics*, 31(1):131–144, 2009.
- [60] Vincent Martin, François Clément, Astrid Decoene, and Jean-Frédéric Gerbeau. Parameter identification for a one-dimensional blood flow model. In *ESAIM: Proceedings*, volume 14, pages 174–200. EDP Sciences, 2005.
- [61] Peng Chen, Alfio Quarteroni, and Gianluigi Rozza. Simulation-based uncertainty quantification of human arterial network hemodynamics. *International Journal for Numerical Methods in Biomedical Engineering*, 29(6):698–721, 2013.
- [62] Paris Perdikaris and George Em Karniadakis. Model inversion via multi-fidelity Bayesian optimization: a new paradigm for parameter estimation in haemodynamics, and beyond. *Journal of The Royal Society Interface*, 13(118):20151107, 2016.
- [63] W. Betz, I. Papaioannou, and D. Straub. Transitional Markov chain Monte Carlo: Observations and improvements. *Journal of Engineering Mechanics*, 142(5):04016016, 2016.

- [64] Peter Vennemann, Ralph Lindken, and Jerry Westerweel. In vivo whole-field blood velocity measurement techniques. *Experiments in Fluids*, 42(4):495–511, Apr 2007.
- [65] Hanwook Park, Yeom Eunseop, and Sang Joon Lee. X-ray PIV measurement of blood flow in deep vessels of a rat: An in vivo feasibility study. *Scientific Reports*, 6:19194, 01 2016.
- [66] Kosuke Tsukada, Haruyuki Minamitani, Eiichi Sekizuka, and Chikara Oshio. Image correlation method for measuring blood flow velocity in microcirculation: correlation ‘window’ simulation and *in vivo* image analysis. *Physiological measurement*, 21(4):459, 2000.
- [67] Kweon-Ho Nam, Eunseop Yeom, Hojin Ha, and Sang Joon Lee. Velocity field measurements of valvular blood flow in a human superficial vein using high-frequency ultrasound speckle image velocimetry. *The International Journal of Cardiovascular Imaging*, 28:69–77, 2010.
- [68] Johnny T Ottesen, Mette S Olufsen, and Jesper K Larsen. *Applied mathematical models in human physiology*. SIAM, 2004.
- [69] Bernardo Cockburn. Discontinuous Galerkin Methods. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 83(11):731–754, 2003.
- [70] Jan S Hesthaven, Sigal Gottlieb, and David Gottlieb. *Spectral methods for time-dependent problems*, volume 21. Cambridge University Press, 2007.
- [71] Jan S Hesthaven and Tim Warburton. *Nodal discontinuous Galerkin methods: algorithms, analysis, and applications*. Springer, 2007.
- [72] Marc C. Kennedy and Anthony O’Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.
- [73] Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, pages 733–760, 1996.
- [74] Herbert W. Hethcote. The Mathematics of Infectious Diseases. *SIAM Review*, 42(4):599–653, 2000.
- [75] M. Draief, A. Ganesh, and L. Massoulié. Thresholds for virus spread on networks. *Ann. Appl. Probab.*, 18(2):359–378, 2008.
- [76] Moez Draief and Laurent Massouli. *Epidemics and Rumours in Complex Networks*. Cambridge University Press, New York, NY, USA, 1st edition, 2010.
- [77] S.C. Ferreira, C. Castellano, and R. Pastor-Satorras. Epidemic thresholds of the susceptible-infected-susceptible model on networks: A comparison of numerical and theoretical results. *Phys. Rev. E*, 86(4):041125, 2012.

- [78] M. Woolhouse. How to make predictions about future infectious disease risks. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 366(1573):2045–2054, 2011.
- [79] Diego Cuadros, Jingjing Li, Adam Branscum, Adam Akullian, Peng Jia, Elizabeth N. Mziray, and Frank Tanser. Mapping the spatial variability of HIV infection in Sub-Saharan Africa: Effective information for localized HIV prevention and control. *Scientific Reports*, 7(9093), 12 2017.
- [80] Mingwang Shen, Yanni Xiao, Libin Rong, Lauren Ancel Meyers, and Steven E. Bellan. Early antiretroviral therapy and potent second-line drugs could decrease HIV incidence of drug resistance. *Proceedings of the Royal Society of London B: Biological Sciences*, 284(1857), 2017.
- [81] Oliver J Brady, Hannah C Slater, Peter Pemberton-Ross, Edward Wenger, Richard J Maude, Azra C Ghani, Melissa A Penny, Jaline Gerardin, Lisa J White, Nakul Chitnis, Ricardo Aguas, Simon I Hay, David L Smith, Erin M Stuckey, Emelda A Okiro, Thomas A Smith, and Lucy C Okell. Role of mass drug administration in elimination of *Plasmodium falciparum* malaria: a consensus modelling study. *The Lancet Global Health*, 5(7):e680 – e687, 2017.
- [82] Philip A. Eckhoff, Edward A. Wenger, H. Charles J. Godfray, and Austin Burt. Impact of mosquito gene drive on malaria elimination in a computational model with explicit spatial and temporal dynamics. *Proceedings of the National Academy of Sciences*, 114(2):E255–E264, 2017.
- [83] J. Challenger, K. Bruxvoort, A. Ghani, and L. Okell. Assessing the impact of imperfect adherence to artemether-lumefantrine on malaria treatment outcomes using within-host modelling. *Nature Communications*, 8(1373), 2017.
- [84] Steve J. Kroiss, Michael Famulare, Hil Lyons, Kevin A. McCarthy, Laina D. Mercer, and Guillaume Chabot-Couture. Evaluating cessation of the type 2 oral polio vaccine by modeling pre- and post-cessation detection rates. *Vaccine*, 35(42):5674 – 5681, 2017.
- [85] Nicolas A Menzies, Gabriela B Gomez, Fiammetta Bozzani, Susmita Chatterjee, Nicola Foster, Ines Garcia Baena, Yoko V Laurence, Sun Qiang, Andrew Siroka, Sedona Sweeney, Stéphane Verguet, Nimalan Arinaminpathy, Andrew S Azman, Eran Bendavid, Stewart T Chang, Ted Cohen, Justin T Denholm, David W Dowdy, Philip A Eckhoff, Jeremy D Goldhaber-Fiebert, Andreas Handel, Grace H Huynh, Marek Lalli, Hsien-Ho Lin, Sandip Mandal, Emma S McBryde, Surabhi Pandey, Joshua A Salomon, Sze chuan Suen, Tom Sumner, James M Trauer, Bradley G Wagner, Christopher C Whalen, Chieh-Yin Wu, Delia Boccia, Vineet K Chadha, Salome Charalambous, Daniel P Chin, Gavin Churchyard, Colleen Daniels, Puneet Dewan, Lucica Ditiu, Jeffrey W Eaton, Alison D Grant, Piotr Hippner, Mehran Hosseini, David Mametja, Carel Pretorius, Yogan Pillay, Kiran Rade, Suvanand Sahu, Lixia Wang, Rein M G J Houben, Michael E Kimerling, Richard G White, and Anna Vassall. Cost-effectiveness and resource implications of aggressive action on tuberculosis in China, India, and South Africa: a combined analysis of nine models. *The Lancet Global Health*, 4(11):e816 – e826, 2016.

- [86] Gerardo Chowell, Hiroshi Nishiura, and Luís M.A Bettencourt. Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. *Journal of The Royal Society Interface*, 4(12):155–166, 2007.
- [87] J. Nathan Kutz, Steven L. Brunton, Bingni W. Brunton, and Joshua L. Proctor. *Dynamic Mode Decomposition: Data-Driven Modeling of Complex Systems*. SIAM, Philadelphia, PA, USA, 2016.
- [88] B. Karrer and M.E.J. Newman. Competing epidemics on complex networks. *Phys. Rev. E*, 84:036106, 2011.
- [89] A. Kumar, P.K. Srivastava, and Y. Takeuchi. Modeling the role of information and limited optimal treatment on disease prevalence. *Journal of Theoretical Biology*, 414:103–119, 2017.
- [90] H. Guo, M.Y. Li, and Z. Shuai. A graph-theoretic approach to the method of global Lyapunov functions. *Proc. Amer. Math. Soc.*, 136:2793–2802, 2008.
- [91] B. Armbruster and E. Beck. Elementary proof of convergence to the mean-field model for the SIR process. *J. Math. Biol.*, 75:327–339, 2016.
- [92] P. Shu, W. Wang, M. Tang, and Y. Do. Numerical identification of epidemic thresholds for susceptible-infected-recovered model on finite-size networks. *Chaos*, 25(6):063104, 2015.
- [93] J. Truscott and N.M. Ferguson. Evaluating the adequacy of gravity models as a description of human mobility for epidemic modelling. *PLoS Comput Biol*, 8(10):e1002699, 2012.
- [94] M.E.J. Newman and C.R. Ferrario. Interacting epidemics and coinfection on contact networks. *PLoS ONE*, 8(9):e71321, 2013.
- [95] E. Duijzer, W. van Jaarsveld, J. Wallinga, and R. Dekker. The most efficient critical vaccination coverage and its equivalence with maximizing the herd effect. *Math Biosci.*, 282:68–81, 2016.
- [96] G. Tanaka, C. Urabe, and K. Aihara. Random and targeted interventions for epidemic control in metapopulation models. *Scientific Reports*, 4:5522, 2014.
- [97] Zachary D Smith and Alexander Meissner. DNA methylation: roles in mammalian development. *Nature Reviews Genetics*, 14(3):204, 2013.
- [98] Jia Zhong, Golareh Agha, and Andrea A Baccarelli. The role of DNA methylation in cardiovascular risk and disease: methodological aspects, study design, and data analysis for epidemiological studies. *Circulation research*, 118(1):119–131, 2016.
- [99] Jesús Delgado-Calle, Agustín F Fernández, Jesús Sainz, María T Zarrabeitia, Carolina Sañudo, Raúl García-Renedo, María I Pérez-Núñez, Carmen García-Ibarbia, Mario F Fraga, and José A Riancho. Genome-wide profiling of bone reveals differentially methylated regions in osteoporosis and osteoarthritis. *Arthritis & Rheumatism*, 65(1):197–205, 2013.

- [100] Steve Horvath. DNA methylation age of human tissues and cell types. *Genome biology*, 14(10):3156, 2013.
- [101] Steve Horvath, Michael Gurven, Morgan E Levine, Benjamin C Trumble, Hillard Kaplan, Hooman Allayee, Beate R Ritz, Brian Chen, Ake T Lu, Tammy M Rickabaugh, et al. An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. *Genome biology*, 17(1):171, 2016.
- [102] Peter A Jones and Gangning Liang. Rethinking how DNA methylation patterns are maintained. *Nature Reviews Genetics*, 10(11):805, 2009.
- [103] Manel Esteller, Jose Maria Silva, Gemma Dominguez, Felix Bonilla, Xavier Matias-Guiu, Enrique Lerma, Elena Bussaglia, Jaime Prat, I Clara Harkes, Elizabeth A Repasky, et al. Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors. *JNCI: Journal of the National Cancer Institute*, 92(7):564–569, 2000.
- [104] Allan Hunter, Paul A Spechler, Alyssa Cwanger, Ying Song, Zhe Zhang, Guishuang Ying, Anna K Hunter, Edwin Dezoeten, and Joshua L Dunaief. DNA methylation is associated with altered gene expression in AMD. *Investigative ophthalmology & visual science*, 53(4):2089–2105, 2012.
- [105] Keith D Robertson. DNA methylation, methyltransferases, and cancer. *Oncogene*, 20(24):3139, 2001.
- [106] Krista S. Crider, Thomas P. Yang, Robert J Berry, and Lynn B. Bailey. Folate and DNA Methylation: A Review of Molecular Mechanisms and the Evidence for Folate’s Role. *Advances in Nutrition*, 3(1):21–38, 01 2012.
- [107] Keith D Robertson, Eva Uzvolgyi, Gangning Liang, Cathy Talmadge, Janos Sumegi, Felicidad A Gonzales, and Peter A Jones. The human DNA methyltransferases (DNMTs) 1, 3a and 3b: coordinate mRNA expression in normal tissues and overexpression in tumors. *Nucleic acids research*, 27(11):2291–2298, 1999.
- [108] Zhao-xia Chen and Arthur D Riggs. DNA methylation and demethylation in mammals. *Journal of Biological Chemistry*, 286(21):18347–18353, 2011.
- [109] Aharon Razin and Arthur D Riggs. Dna methylation and gene function. *Science*, 210(4470):604–610, 1980.
- [110] Laurianne Scourzic, Enguerran Mouly, and Olivier A Bernard. TET proteins and the control of cytosine demethylation in cancer. *Genome medicine*, 7(1):9, 2015.
- [111] Mark T Mc Auley, Kathleen M Mooney, and J Enrique Salcedo-Sora. Computational modelling folate metabolism and DNA methylation: implications for understanding health and ageing. *Briefings in bioinformatics*, 19(2):303–317, 2016.
- [112] Loukas Zagkos, Mark Mc Auley, Jason Roberts, and Nikos I Kavallaris. Mathematical models of DNA methylation dynamics: Implications for health and ageing. *Journal of theoretical biology*, 462:184–193, 2019.

- [113] Andrew P McGovern, Benjamin E Powell, and Timothy JT Chevassut. A dynamic multi-compartmental model of DNA methylation with demonstrable predictive value in hematological malignancies. *Journal of theoretical biology*, 310:14–20, 2012.
- [114] Albert Jeltsch and Renata Z Jurkowska. New concepts in DNA methylation. *Trends in biochemical sciences*, 39(7):310–318, 2014.
- [115] Robert A Waterland and Karin B Michels. Epigenetic epidemiology of the developmental origins hypothesis. *Annu. Rev. Nutr.*, 27:363–388, 2007.
- [116] Kaie Lokk, Vijayachitra Modhukur, Balaji Rajashekar, Kaspar Märtens, Reedik Mägi, Raivo Kolde, Marina Koltšina, Torbjörn K Nilsson, Jaak Vilo, Andres Salumets, et al. DNA methylome profiling of human tissues identifies global and tissue-specific methylation patterns. *Genome biology*, 15(4):3248, 2014.
- [117] Jan O Haerter, Cecilia Lövkvist, Ian B Dodd, and Kim Sneppen. Collaboration between CpG sites is needed for stable somatic inheritance of DNA methylation states. *Nucleic acids research*, 42(4):2235–2244, 2013.
- [118] Cecilia Lövkvist, Ian B Dodd, Kim Sneppen, and Jan O Haerter. DNA methylation in human epigenomes depends on local topology of cpG sites. *Nucleic acids research*, 44(11):5123–5132, 2016.
- [119] Eric Hervouet, Paul Peixoto, Régis Delage-Mourroux, Michaël Boyer-Guittaut, and Pierre-François Cartron. Specific or not specific recruitment of DNMTs for DNA methylation, an epigenetic dilemma. *Clinical Epigenetics*, 10(1):17, Feb 2018.
- [120] E Lauga and TR Powers. The hydrodynamics of swimming microorganisms. *Rep Prog Phys*, 72:096601, 2009.
- [121] GI Taylor. The action of waving cylindrical tails in propelling microscopic organisms. *Proc Roy Soc Lond Ser A*, 211:225–239, 1952.
- [122] L Carichino and SD Olson. Emergent three-dimensional sperm motility: coupling calcium dynamics and preferred curvature in a kirchhoff rod model. *J Math Med Biol*, 2018. doi:10.1093/imammb/dqy015.
- [123] RH Dillon, LJ Fauci, and X Yang. Sperm motility and multiciliary beating: an integrative mechanical model. *Comput Math Appl*, 52:749–758, 2006.
- [124] J Elgeti, UB Kaupp, and G Gompper. Hydrodynamics of sperm cells near surfaces. *Biophys J*, 99(4):1018–1026, 2010.
- [125] J Elgeti, RG Winkler, and G Gompper. Physics of Microswimmers- Single Particle Motion and Collective Behavior: a review. *Rep Prog Phys*, 78:056601, 2015.
- [126] J Huang, L Carichino, and SD Olson. Hydrodynamic interactions of actuated elastic filaments near a planar wall with applications to sperm motility. *J Coupled Syst Multiscale Dyn*, 6:163–175, 2018.

- [127] K Ishimoto and EA Gaffney. An elastohydrodynamical simulation study of filament and spermatozoan swimming driven by internal couples. *IMA J Applied Math*, 83:655–679, 2018.
- [128] SF Schoeller and EE Keaveny. Flagellar undulations to collective motion: predicting the dynamics of sperm suspensions. *J Roy Soc Interface*, 15:20170834, 2018.
- [129] DJ Smith, EA Gaffney, JR Blake, and JC Kirkman-Brown. Human sperm accumulation near surfaces: a simulation study. *J Fluid Mech*, 621:289–320, 2009.
- [130] T Omori and T Ishikawa. Swimming of Spermatozoa in a Maxwell Fluid. *Micromach*, 10:78, 2019.
- [131] Y Yang, J Elgeti, and G Gompper. Cooperation of sperm in two dimensions: synchronization, attraction, and aggregation through hydrodynamic interactions. *Phys Rev E*, 78:061903–1–9, 2008.
- [132] K Ishimoto and EA Gaffney. Hydrodynamic Clustering of Human Sperm in Viscoelastic Fluids. *Sci Rep*, 8:15600, 2018.
- [133] J Teran, L Fauci, and M Shelley. Viscoelastic fluid response can increase the speed of a free swimmer. *Phys Rev Lett*, 104:038101–4, 2010.
- [134] B Thomases and RD Guy. Mechanisms of elastic enhancement and hindrance for finite-length undulatory swimmers in viscoelastic fluids. *Phys Rev Lett*, 113:098102, Aug 2014.
- [135] ST Mortimer. CASA—practical aspects. *J Androl*, 21:515–524, 2000.
- [136] H Gadelha, EA Gaffney, and A Goriely. The counterbend phenomenon in flagellar axonemes and cross-linked filament bundles. *Proc Natl Acad Sci USA*, 110:12180–12195, 2013.
- [137] C.B. Lindemann, L.J. Macauley, and K.A. Lesich. The counterbend phenomenon in dynein-disabled rat sperm flagella and what it reveals about the interdoubtlet elasticity. *Biophys. J.*, 89:1165–1174, 2005.
- [138] D.W. Pelle, C.J. Brokaw, and C.B. Lindemann. Mechanical properties of the passive sea urchin sperm flagellum. *Cell Motil. Cytoskeleton*, 66:721–735, 2009.
- [139] IH Riedel-Kruse, A Hilfinger, J Howard, and F Julicher. How molecular motors shape the flagellar beat. *HFSP J*, 1:192–208, 2007.
- [140] G Xu, KS Wilson, RJ Okamoto, JY Shao, SK Dutcher, and PV Bayly. Flexural Rigidity and Shear Stiffness of Flagella Estimated from Induced Bends and Counterbends. *Biophys J*, 110:2750–2768, 2016.
- [141] W Gao and J Wang. Synthetic micro/nanomotors in drug delivery. *Nanoscale*, 6:10486–10494, 2014.
- [142] BJ1 Nelson, IK Kaliakatsos, and JJ Abbott. Microrobots for minimally invasive medicine. *Annu Rev Biomed Eng*, 12:55–85, 2010.

- [143] L Sanders. Microswimmers make a splash: tiny travelers take on a viscous world. *Science News*, 176:22–25, 2009.
- [144] P Tierno, R Golestanian, I Pagonabarraga, and F Sagues. Magnetically actuated colloidal micro swimmers. *J Phys Chem B*, 112:16525–16528, 2008.
- [145] F Plouraboue, Thiam EI, B Delmotte, and E Climent. Identification of internal properties of fibres and micro-swimmers. *Proc Roy Soc A*, 473:20160517, 2017.
- [146] Alan Cheng Hou Tsang, Pun Wai Tong, Shreyes Nallan, and On Shun Pak. Self-learning how to swim at low Reynolds number. *arXiv:1808.07639*, 2018.
- [147] R Cortez, B Cummins, K Leiderman, and D Varela. Computation of three-dimensional Brinkman flows using regularized methods. *J Comput Phys*, 229:7609–7624, 2010.
- [148] K Leiderman and SD Olson. Swimming in a two-dimensional Brinkman fluid: Computational modeling and regularized solutions. *Phys Fluids*, 28(2):021902, 2016.
- [149] N. Ho, K. Leiderman, and S.D. Olson. A 3-dimensional model of flagellar swimming in a Brinkman fluid. *J Fluid Mech*, 864:1088–1124, 2019.
- [150] J. Rutllant, M. Lopez-Bejar, and F. Lopez-Gatius. Ultrastructural and rheological properties of bovine vaginal fluid and its relation to sperm motility and fertilization: a review. *Reprod Dom Anim*, 40:79–86, 2005.
- [151] W. M. Saltzman, M. L. Radomsky, K. J. Whaley, and R. A. Cone. Antibody diffusion in human cervical mucus. *Biophys J*, 66:508, 1994.
- [152] HC Flemming and Wingender J. The biofilm matrix. *Nat Rev Microbiol*, 8:623–633, 2010.
- [153] SA Miradbagheri and HC Fu. Helicobacter pylori couples motility and diffusion to actively create a heterogeneous complex medium in gastric diseases. *Phys Rev Lett*, 116, 2016.
- [154] J. L. Auriault. On the domain of validity of Brinkman’s equation. *Trans. Porous Media*, 79:215–223, 2009.
- [155] H. C. Brinkman. A calculation of the viscous force exerted by a flowing fluid on a dense swarm of particles. *Appl Sci Res*, 1:27–34, 1947.
- [156] L. Durlofsky and J. F. Brady. Analysis of the Brinkman equation as a model for flow in porous media. *Phys Fluids*, 30(11):3329–3341, 1987.
- [157] I. D. Howells. Drag due to the motion of a Newtonian fluid through a sparse random array of small fixed rigid objects. *J Fluid Mech*, 64:449–475, 1974.
- [158] L. Spielman and S. L. Goren. Model for predicting pressure drop and filtration efficiency in fibrous media. *Env Science Tech*, 1(4):279–287, 1968.
- [159] H Fu, VB Shenoy, and TR Powers. Low Reynolds number swimming in gels. *Europhys Lett*, 91, 2010.

- [160] AM Leshansky. Enhanced low-Reynolds-number propulsion in heterogeneous viscous environments. *Phys Rev E*, 80:051911, 2009.
- [161] M Morandotti. Self-propelled micro-swimmers in a Brinkman fluid. *J Biol Dynamics*, 6:88–103, 2012.
- [162] H Nganguia and OS Pak. Squirring motion in a Brinkman medium. *J Fluid Mech*, 855:554–573, 2018.
- [163] N Ho, K Leiderman, and SD Olson. Swimming speeds of filaments in viscous fluids with resistance. *Phys Rev E*, 93(4):043108, 2016.
- [164] C Jeznach and SD Olson. Dynamics of Swimmers in Fluids with Resistance. *Fluids*, 5:14:1–20, 2020.
- [165] SD Olson and K Leiderman. Effect of fluid resistance on symmetric and asymmetric flagellar waveforms. *J Aero Aqua Bio-mech*, 4(1):12–17, 2015.
- [166] C.S. Peskin. The immersed boundary method. *Acta Numer*, 11:459–517, 2002.
- [167] K. Leiderman and S.D. Olson. Erratum: “Swimming in a two-dimensional Brinkman fluid: Computational modeling and regularized solutions” [Phys. Fluids 28, 021902 (2016)]. *Phys Fluids*, 29:029901, 2017.
- [168] E Ahmadi, R Cortez, and H Fujioka. Boundary integral formulation for flows containing an interface between two porous media. *J Fluid Mech*, 816:71–93, 2017.
- [169] LJ Fauci and A McDonald. Sperm motility in the presence of boundaries. *Bull Math Biol*, 57:679–699, 1995.
- [170] J Simons, L Fauci, and R Cortez. A fully three-dimensional model of the interaction of driven elastic filaments in a stokes flow with applications to sperm motility. *J Biomech*, 48:1639–1651, 2015.
- [171] DM Woolley, RF Crockett, WDI Groom, and SG Revell. A study of synchronisation between the flagella of bull spermatozoa, with related observations. *J Exp Biol*, 212:2215–2223, 2009.
- [172] H Moore, K Dvorakova, N Jenkins, and W Breed. Exceptional sperm cooperation in the wood mouse. *Nature*, 418:174–177, 2002.
- [173] IH Riedel, K Kruse, and J Howard. A self-organized vortex array of hydrodynamically entrained sperm cells. *Science*, 309:300–303, 2005.
- [174] A. M. Stuart. Inverse problems: A Bayesian perspective. *Acta Numerica*, 19:451–559, 2010.
- [175] Jari P. Kaipio and Erkki Somersalo. *Statistical and Computational Inverse Problems*. Springer, 2005.
- [176] DJ Smith, EA Gaffney, H Gadêlha, N Kapur, and JC Kirkman-Brown. Bend propagation in the flagella of migrating human sperm, and its modulation by viscosity. *Cell Motil Cytoskel*, 66(4):220–236, 2009.