Abstract

CHATTOPADHYAY, SOMRITA. Model Discrimination Using Bounded Error Estimation Approach. (Under the direction of Stephen Campbell.)

Understanding the dynamic behavior of biological systems is one of the major challenges in biological research and the success of systems biology lies in the ability to use mathematical models for formulating hypotheses about the interaction of components that can then be tested experimentally. A correct model helps us to understand the underlying mechanisms and on the other hand, one can use that model to predict the future behavior of the biological system under various circumstances. A commonly encountered problem in such scenarios is multiple models presenting different or competing hypotheses i.e. structurally different models explaining the same set of data equally well. Discrimination among these models is crucial to understand the fundamental behavior of the examined process. This distinction can be particularly difficult when training/evaluating models based on data that contains uncertainty. In this project, we have addressed the problem of model ambiguity by introducing a novel approach for model discrimination that employs an interval analysis approach, based on the techniques developed for parameter estimation in a bounded error context using Extended Mean Value (EMV) and Set Inversion via Interval Analysis (SIVIA) algorithms. This approach allows us to choose the most suitable model for a particular condition in presence of uncertainty. Thus, in the long run this work will help the researchers to represent various biological processes in the most accurate, predictable and tractable manner.
Model Discrimination Using Bounded Error Estimation Approach

by
Somrita Chattopadhyay

A thesis submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Master of Science

Mathematics

Raleigh, North Carolina

2015

APPROVED BY:

Mansoor Haider
Negash Medhin

Stephen Campbell
Chair of Advisory Committee
Dedication

To my parents.

Prasanta Kumar Chatterjee and Ratna Chattopadhyay
Biography

Somrita Chattopadhyay was born on May 23, 1987 in Kolkata, (known as the 'City of Joy’), the cultural capital of India. She completed her schooling at Patha Bhavan, one of the top educational institutes in India. After graduating from Heritage Institute of Technology in August 2009 with a Bachelor of Technology degree in Electronics and Communication Engineering, she joined the Electrical and Computer Engineering Department of North Carolina State University in Fall 2010 for graduate studies. High school onwards the rationale working behind Mathematics and the logical solutions of day-to-day problem with its help always intrigued her. She was always passionate about Mathematics and decided to receive a Master of Science in Mathematics as well.
Acknowledgements

I would like to thank my advisor Dr. Campbell for his excellent guidance and endless patience during the course of my Masters research. His valuable suggestions have not only helped me to achieve success in my research work, but have also contributed immensely towards my professional and personal growth.

I also acknowledge Dr. Haider and Dr. Medhin; their valuable inputs and constant support towards my degree completion kept me calm as well as motivated throughout my graduate studies.

I would also like to thank my two best friends, Sagar and Subharup. They were always there cheering me up and stood by me through the good times and bad. Last but not the least, I would extend earnest thanks and love to my parents for being the constant source of support and encouragement at every stage of my career and life.
# Table of Contents

List of Tables .......................................................................... vii

List of Figures ........................................................................ viii

Chapter 1 Introduction ................................................................. 1
  1.1 Overview ........................................................................ 1
  1.2 Motivation ...................................................................... 2
  1.3 Thesis Contribution .......................................................... 4
    1.3.1 Intellectual Merit ...................................................... 4
    1.3.2 Broader Impacts ...................................................... 4
  1.4 Thesis Organization ............................................................ 5

Chapter 2 Background ................................................................. 6
  2.1 Overview ........................................................................ 6
  2.2 State of Art ..................................................................... 8
  2.3 Our Approach .................................................................. 14
  2.4 Discussion ...................................................................... 14

Chapter 3 Methodology ............................................................... 16
  3.1 Overview ........................................................................ 16
  3.2 Biological Modeling ........................................................ 17
  3.3 Problem Statement .......................................................... 18
  3.4 Interval Analysis .............................................................. 19
    3.4.1 Basic Tools ................................................................ 19
    3.4.2 Inverse Image Evaluation (Set Inversion) .................... 22
    3.4.3 Direct Image Evaluation .......................................... 23
  3.5 Extended Mean Value Theorem and Set Inversion via Interval Analysis ........................................................................ 24
  3.6 Parameter and State Estimation ........................................ 26
  3.7 Model Discrimination ........................................................ 26
  3.8 Discussion ...................................................................... 29

Chapter 4 Results .................................................................... 30
  4.1 Case Study I ................................................................... 31
    4.1.1 Model Discrimination Between Two Predator - Prey Population Models ......................................................... 31
    4.1.2 Experimental Setup ................................................... 32
    4.1.3 Results and Analysis ............................................... 34
  4.2 Case Study II ................................................................... 37
Chapter 4

4.2.1 Model Discrimination Between Henry and Michaelis-Menten Mechanisms .................................. 37
4.2.2 Experimental Setup .................................................. 39
4.2.3 Results and Analysis .................................................. 41
4.3 Discussion ............................................................... 44

Chapter 5 Conclusion and Future Works .................................................. 47

5.1 Conclusion ............................................................... 47
5.2 Limitations of the Proposed Method and Future Scope .................................................. 48

References ............................................................... 50
# List of Tables

<table>
<thead>
<tr>
<th>Table 4.1</th>
<th>Results of PE for Model 1</th>
<th>35</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 4.2</td>
<td>Results of PE for Model 2</td>
<td>36</td>
</tr>
<tr>
<td>Table 4.3</td>
<td>Results of PE for the Michaelis-Menten Model</td>
<td>43</td>
</tr>
</tbody>
</table>
## List of Figures

| Figure 3.1 | Mapping from state space $R^n$ to data space $R^m$ using the function $y(t) = g(x(t), p)$[19] | 21 |
| Figure 3.2 | Feasibility of boxes using SIVIA | 25 |
| Figure 3.3 | Parameter Estimation Algorithm[19] | 27 |
| Figure 3.4 | Block Diagram for Model Discrimination method | 28 |
| Figure 4.1 | State Estimation of the Reference Model for 6 different initial conditions. X-axes represent time, Y-axes represent bounded state outputs. The red and black colored lines represent the lower and upper bounds of the state intervals. | 33 |
| Figure 4.2 | State Estimation of the Reference Model for 10 different initial conditions. X-axes represent time, Y-axes represent bounded state outputs. The red and black colored lines represent the lower and upper bounds of the state intervals. | 40 |
| Figure 4.3 | Parameter Estimation of the Michaelis-Menten Model for 10 different initial conditions. X-axes represent the parameter space of $c$ and Y-axes represent parameter space of $f$. | 42 |
| Figure 4.4 | Parameter Bound for Model 1 | 45 |
| Figure 4.5 | Parameter Bound for Model 2 | 46 |
Chapter 1

Introduction

1.1 Overview

Systems biology is an emerging research field that deals with quantitative representation of dynamic biological systems. Systems biologists mainly focus on understanding biological processes at system level by studying the interactions between individual components of the system and representing them in a mathematical framework. In recent years, it has become a groundbreaking scientific approach towards developing accurately predictive biological models, which have the potential to bring advancement in the fields of medicine, agriculture, ecosystems and food science[8]. Like engineers, systems biologists first develop an overall understanding of a particular system and then apply that knowledge to control it. Thus, systems biology is not only a scientific field, but also an engineering discipline[28].

Mathematical modeling of biological systems is not only confined to the handling of a large amount of data from multiple experimental sources, but also extends to integrating these data into a comprehensive analytical description of biological function with
predictive power [13]. Analyzing, representing, manipulating and integrating such data requires application of ideas from diverse areas of electrical and computer engineering like artificial intelligence, signal processing, control theory and many more. Thus systems biology is highly reliant on computational and analytical tools developed by electrical and computer engineers.

From medicine and environmental science to alternative fuel technology and agriculture, systems biologists are bringing revolutionary changes [28]. The most eminent impact will be in the health-care sector by means of new drug discovery [6] and innovative treatment of complex diseases like cancer, dementia, aging[8] and neurological problems [38]. It is also leading to the development of bio-fuels and artificial algae design to process carbon-dioxide to reduce power-plant emission [28]. Systems Biology also provides a platform for the foundations of Synthetic Biology, another emerging research field, which deals with design and construction of new biological systems that are not available naturally.

The work described in this thesis focuses on finding the most appropriate model among many dynamic models used to represent the same biological process. A set based approach has been developed for this purpose. The algorithm has been applied on two different test cases and the results obtained show that even in presence of non-deterministic uncertainty, our method can outperform the traditional approaches designed for model discrimination.

1.2 Motivation

Finding suitable models of dynamic biological systems is the most important area of systems biology. There are several processes in nature that can not be described ex-
perimentally or observed. So, the need arises for representing biological processes in a framework of mathematical equations. For any given system, there can be a number of equation sets with unique characteristics that distinguishes one model from the other. This gives rise to situations where several models can explain the available set of data equally well. As these mathematical models are mainly used to predict and control the future behavior of the given biological process, a sufficiently accurate model with less complexity is required to understand the actual underlying phenomenon. The question now becomes how to find the best suited model out of different equally likely hypotheses. This gives rise to different model discrimination theories.

In this paper, we have aimed at developing a set based model discrimination approach. Usually, most of the classical methods treat uncertainty associated with data sets as statistically defined noise or require a detailed priori knowledge about the system parameters. However, this type of assumption is not appropriate for biological systems. Uncertainty in these non-linear systems can arise from a variety of factors, such as - measurement error, non-stringent experimental setup, etc. Moreover in most cases, sufficient experimental data is not available at the beginning of modeling. Thus, heuristic approaches, based on experience or intuition, are also not applicable. Our method characterizes errors in the data by simple sets such as interval boxes, ellipsoids, zonotopes, etc [16]. Uncertainty associated with biological systems are more aptly represented by a continuous range of values. The method is very appealing, because our proposed algorithm analyzes entire sets, instead of just checking some finite amount of distinct points. This gives us a better understanding of the true dynamics of the system.

Some other researchers have also proposed set-based approaches [25], [31] to address the model discrimination problem. However, they either face huge computational burden while analyzing entire data sets or need some reliable prior knowledge where only few
measurements with high uncertainty are available. The work presented in this document is based on an improved parameter estimation algorithm [20] developed over the works of Raissi [30] and Jaulin [11]. Thus our approach for model discrimination handles the problem of increased computational time and availability of only sparse, discrete data in a better and concrete manner by employing techniques like parallelization and preprocessing.

1.3 Thesis Contribution

1.3.1 Intellectual Merit

The intellectual merit of this research work is described below:

A technique has been developed to distinguish competing model hypotheses, all representing the same biological activity. This is achieved without imposing any unrealistic assumption or probabilistic characterization on the data sets. The uncertainties associated with the data are supposed to be bounded without any other assumption on their distribution.

1.3.2 Broader Impacts

Development of model discrimination tools will have huge impact in the areas like synthetic biology, gene regulation, plant system engineering, personalized health-care and assessment and treatment selection. As an example, traditionally new drugs are discovered with an assumption that the patients belong to homogeneous groups within disease categories. But evolution of model discrimination theories will be the key in innovating individualized medicines [8]. To identify a drug target or develop the best treatment
strategy, one needs to choose the best representative model of the particular situation. Finding the best representative model is not a trivial task. The contribution of this research lies in solving such problems in a more concrete and realistic manner, without taking into consideration any probabilistic assumption. Thus, the work presented in this thesis will prove to be an important tool for system biologists in the future. This work will not only be beneficial for modeling of biological systems, but can also be applicable for other control systems.

1.4 Thesis Organization

The entire thesis is divided into five chapters.

Chapter 1 gives us an overview of the field ‘Systems Biology’ and its application in today’s world. It also describes in brief the contribution of this research.

Chapter 2 explains the foundation, terminologies and concepts used for this research work. It also discusses some of the other work conducted in the area of model discrimination and points out how the work presented here is different from the traditional approaches.

Chapter 3 gives a detailed description about our approach towards developing a model discrimination algorithm for biological systems.

Chapter 4 mainly illustrates the results of our proposed method. Outcome of our algorithm is demonstrated by its application on two different case studies.

Chapter 5 summarizes the work presented in this thesis and discusses some future topics for development in this field.
Chapter 2

Background

This chapter discusses the recent state of art in the field of model discrimination for biological systems and advantages of our method over others.

2.1 Overview

Mathematical modeling of biological systems has gained a lot of attention in recent times due to its immense impact in the fields of medicine, agriculture, food science, bio-materials and bio-mechanics. Here are several processes in nature that cannot be described experimentally or observed. This motivates the need for representing biological processes in a framework of mathematical equations. Mathematical models can be used to predict dynamic behavior of biological systems and establish hypotheses for the interaction of components that can be tested experimentally. Finding suitable models of biological systems, thus, has become one of the most important areas of systems biology.

For any given system, there can be different hypotheses with unique characteristics that distinguish one model from the others. This gives rise to situations where several
models can explain the available set of data equally well. As these mathematical models are mainly used to predict and control the future behavior of the given biological process, a sufficiently accurate model with less complexity is required to understand the actual underlying phenomenon. In the early phase of mathematical modeling, while establishing the interactions and reactions in the pathway, mathematical descriptions of the competing hypotheses can be used to verify whether or not these hypotheses are viable by comparing with the experimental data that has been obtained. The question now becomes how to find the best model out of different equally likely hypotheses. This becomes a very compelling application for model discrimination, particularly since the experimental data that has been obtained contains some sort of uncertainty.

Accurate models are required for understanding, predicting and finally controlling biological systems. Adding new features to these models are, thus, often necessary. But this gives rise to increased complexity in terms of increased number of parameters. As these parameters play an important role in influencing the model’s dynamics, efficient parameter estimation algorithms are crucial to discriminate between different candidate models.

Parameter estimation and model discrimination are more complicated in biological systems due to many computational, analytical and experimental issues. Their non-linear and uncertain dynamics result from issues like measurement error, non-stringent experimental set-up, unpredictable environmental conditions, used stimuli, etc. Moreover, often sufficient information about the data sets are not available at the beginning of the experiment. Thus, in this type of mathematical modeling, one of the most important and complicated tasks is to characterize the uncertainty associated with the biological processes. Most of the classical methods treat uncertainty associated with the data sets as statistically defined noise i.e. the errors produced due to noise are assumed to be
independent and identically distributed, usually Gaussian with zero mean and variance \( \sigma^2 \) or require detailed a-priori knowledge about the noise characteristics. Thus, probabilistic and heuristic approaches, based on experience or intuition, are not appropriate for biological systems. Moreover, the classical approaches are deficient as they evaluate a finite number of points in the parameter space. But, uncertainty in biological systems is more aptly represented by a finite range of continuous values where true dynamics of the system lie [30].

In this chapter, we discuss the common methods of model discrimination and some of the drawbacks associated with them. We also explain why our discrimination approach is aptly suited for biological systems.

2.2 State of Art

A majority of the classical approaches in parameter estimation and model discrimination usually focus on finding the difference between the experimental data and the predicted model output. This is usually achieved by employing weighted least square methods or maximum likelihood functions. These methods usually employ Fisher Information Matrix and associated A-, D- or E- optimality [9],[3],[39]. These approaches usually give rise to solving non-convex optimization problems, which are very difficult to solve. Thus, locally optimal solutions with some desired global properties are searched for and this, at times, may lead to unsatisfying parameter estimates due to non-convergence to a global optimum in finite time [21],[23],[17].

Another work on model discrimination is presented in [2], where an input signal is designed to drive model output through selected trajectory and a feedback controller is used to measure dynamic stimulus using which models with mechanistic differences can
be distinguished. However, this method depends on the development of a controller in laboratory, which may turn out to be a difficult task.

Dominik et al. in [34] explains a KL- optimality criterion for model discrimination. It calculates the number of optimal measurement location and time points instead of continuous design. It also gives the optimal initial condition and optimal perturbation to the system. This method employs a fitting technique and uncertainty associated with experimental data is assumed to have a fixed probability distribution.

Literature shows that model selection is also approached using various types of Kalman filters. These methods basically tackle the above stated problem from a control theory point of view by using state observers. One of the recent works is presented in [15] where the authors have developed a constrained hybrid extended Kalman filtering algorithm that can handle sparse and noisy data, applied to large parameter spaces. However, the Kalman filter is not generally an optimal estimator in a non-linear setting. Also, if the initial conditions are not close-by, the filter may end in an estimate whose mean is different from the true mean. In such cases, these methods produce unreliable results.

[22] describes discrimination method of multiple model candidates based on Akaike’s Information Criterion. Problem of model lumping while applying RSOS methods is overcome using the AWDC criterion which reduces the number of experimental designs in the early stage of model identification and thus, reduces the overall cost of the system. AIC minimizes KL-distance between model and truth. The best model is the one with lowest AIC. This is a relative likelihood approach.

Several methods have also applied semi-definite programming to solve the problem of model discrimination in a convex optimization framework [25],[27].

[27] focuses on determining the best initial condition that can maximize the upper bound between the outputs of the possible network models. Model is selected based on the
trend of the state or state variables behavior. However, models chosen are deterministic and thus, maximizing the upper bound can not guarantee discrimination.

[25] proposes three different approaches to maximize the output between different candidate models. Those are (a) best initial condition is determined which discriminates the models the most. (b) Best input distribution is chosen to maximize the difference between model outputs. (c) Best initial condition is modified with structural modifications to achieve maximum difference in model outputs. Here the SOS technique has been used and a Bayesian approach is followed. Time series data having maximum information (how measurements are spaced in time, how many are needed, etc.) is collected. In these methods, same input or initial conditions are taken and outputs of different patterns are generated. Then the same conditions are applied to the actual physical system and the model whose pattern differs hugely from the actual system is discarded. But, only deterministic models that do not consider noise measurements directly are used here. The authors have tried to make the outputs of the two models as distant as possible so that even a noisy measurement has a good chance of discriminating between them.

Simulation based approaches cited in [36], [37] aim at developing model selection framework based on approximate Bayesian computation and employing Monte Carlo sampling. Bayesian model selection strikes a balance between the complexity of the simulation models and their ability to describe observed data. Some other works applying Bayesian conditioning are described in [10], [12], [4]. These kind of approaches assign some prior probabilities to each candidate model and chooses the model with maximum likelihood. However, success of these methods relies largely on the quality of prior information. Recent developments in Bayesian methods are Markov Chain models [5], ensemble methods [14],[7] and sequential Monte Carlo methods[37],[33]. Though these approaches do not employ likelihood approaches, they are usually applied in lower
dimensional problems or where large data samples are available.

In this type of mathematical modeling, one of the most important and complicated tasks is to characterize the uncertainty associated with the corresponding biological processes. In most of the works described above, variation associated with data measurements are modeled as statistically defined noise i.e. the errors produced due to noise are assumed to be independent and identically distributed, usually Gaussian with zero mean and variance $\sigma^2$. However, non-linear and poorly understood system dynamics make this characterization inappropriate for biological processes. In these cases, it is preferable to bound the error associated with the models by simple range of values than any statistical assumption. Moreover, the classical approaches are deficient as they evaluate at a finite number of points in the parameter space. Thus, uncertainty in biological systems is more aptly represented by a a continuous range of values where the true dynamics of the system lie [30].

Recent works in this field have focused on characterizing noise present in biological systems in a bounded error context. Set-based approaches for parameter estimation and model discrimination have started receiving immense attention since the last decade. The only assumption, considered in these methods, is that we have prior knowledge about the upper and lower bounds of the parameter sets and measurement errors, but no hypothesis regarding the error distribution in these sets is known. Sophisticated approaches, nowadays, employ interval analysis and semi-definite programming to address the same problem.

Literature has addressed several model discrimination approaches based on semi-definite programming. SDP is basically an optimization approach designed for convex problems. However, for non-convex cases, SDP applies some relaxation criteria by constructing barrier functions or Lagrange multipliers. The basic idea of these approaches
is to determine an outer approximation of the parameters which are consistent with the available data (called SCP - set of consistent parameters). To efficiently outer bound the set of consistent parameters, infeasibility certificates are obtained by SDP. Model falsification is done by showing SCP to be empty. This method aims at overcoming the problems of noisy data and system’s non-convexity with the help of bounded measurement error.

Our method also characterizes errors in the data by simple sets such as interval boxes, ellipsoids, zonotopes, etc [16]. In our work, we bound the uncertainty associated with the system by a continuous range of values. The method is very appealing, because our proposed algorithm analyzes entire sets, instead of just checking a finite number of distinct points. This gives us a better understanding of the true dynamics of the system.

The main concept used by both the Interval Analysis and SDP methods is that the parameter space is bounded and not defined by any distribution. However, we have chosen Interval Estimation method over SDP because of the following reasons -

- Firstly, other set-based approaches, including SDP, mainly aimed at model invalidation rather than model discrimination. These approaches discarded the models having empty SCP. However, no solution has been provided for cases where more than one model has feasible parameter sets. Our method addresses this issue by designing a metric which can discriminate between models having non-empty SCPs. To our knowledge, this problem has not been investigated previously.

- Secondly, interval estimation method is based on the use of validated methods for solving the initial value problem (IVP) for ODE [30]. Higher-order Taylor expansion methods are employed to find intervals which are guaranteed to contain the solution of the IVP for the ODE. Thus, if a solution is obtained, then it is guaranteed that
the problem has a unique solution and thus, a guaranteed enclosure of the solution can be derived. However, the SDP approaches employ some relaxation criteria to solve the non-convex problems. This relaxed version makes sure that there are no false negatives in the solution i.e. no feasible parameterization is lost. However, the result may contain false positives [31]. This may end up producing some inaccurate solutions. This means that there could be cases in which the actual solving problem would allow us to invalidate a model, while solving the relaxed version does not. However, if the relaxed version is infeasible, this guarantees that the actual non-convex problem is infeasible as well, and hence that the model is inconsistent with the experimental data. Interval method may also produce false positives. However, as stated previously, our aim is not to invalidate a model; we discriminate between the models. Thus, presence of false positives do not affect our solution vigorously. Moreover, we recursively divide the interval boxes until the user defined tolerance is reached. As the boxes get smaller and smaller, the occurrence of false positives also decrease. This makes our approach more robust.

In practice, both the SDP and Interval Analysis approaches suffer from disadvantages like large computational burden and time consumption. Our method, however, handles these issues better, as it is based on a Parallelized Parameter Estimation algorithm developed by Marvel & Williams [18]. This speeds up the estimation process by using multi-threading technique and the performance also improves by almost the number of cores used for processing.
2.3 Our Approach

In this paper, we have utilized the interval analysis method[24] to solve the model discrimination problem in a set based framework. Interval boxes are used to bound error associated with the data. We have represented the models in terms of ordinary differential equations. Our novel approach is based on the works in [11], [30] and [20]. If two or more model variants describe the available experimental data, a new experiment must be designed to discriminate between the hypothetical models. We basically want to see how well the old model describes the new data set. The key idea is to find an input profile that maximizes the difference of the outputs of the competing models. Once we have designed the new data set, the states and parameters of the models are first estimated using Extended mean Value (EMV) theorem and Set Inversion Via Interval Analysis (SIVIA) approach and, then the uncertainty associated with those parameters are checked. Model discrimination can then be achieved by comparing the parameter bounds of the competing models. The model with most consistent parameter set is chosen as the best model.

2.4 Discussion

As we have discussed previously, the classical approaches which modeled variation by statistically defined noise are not very apt for capturing the non-linear, uncertain dynamics of the biological processes. Our method overcomes that problem by representing the noise associated with a system by range of values where true dynamics of the system lies. The set based approaches employing SDP also suffer from drawbacks like high computation time and lack of scalability. One significant drawback of SDP approaches
is that only model falsification is possible because if no parameter values are found for a given data set, it can not be guaranteed that no such parameter values exist. Thus, if there are several models for which parameter values are obtained, it is not possible to discriminate them. In our approach, we have shown that even if parameters sets are found for different models, we can discriminate between them based on their consistencies with the available data. Also, the SDP approaches employ some relaxation criteria to solve the non-convex problems. This relaxed version makes sure that there are no false negatives in the solution. However, there may be false positives [31], leading to inaccurate solutions at times. Moreover, our method is based on a parallel parameter estimation algorithm which efficiently uses multi-core chips of a computer and thus, our algorithm is much faster than other conventional approaches. Our algorithm can also produce reliable parameter estimates, even when only sparse and discrete data measurements are available. Detailed description of this approach has been discussed in Chapter 3.
Chapter 3

Methodology

This chapter introduces all the important concepts used in the paper and gives a detailed description of our approach to find the best model among a set of candidate models.

3.1 Overview

We start with explaining how biological systems are modeled using mathematical equations and then formulate the problem we aim to solve. As mentioned previously, our approach is based on the parameter estimation and state estimated works presented in [11], [30], [20]. The relevant interval arithmetic algorithms are briefly presented here. We explain why interval analysis is used and how it helps us to characterize the uncertainty of our system in a better way. Important algorithms like Extended Mean Value theorem (EMV) and Set Inversion via Interval Analysis (SIVIA), which have helped us to design our discrimination algorithm, are also discussed. And finally we present our own method to select the most suitable mathematical model for a certain biological procedure.
3.2 Biological Modeling

Many biological systems can be modeled by using a set of ordinary differential equations. The general form of a model can be represented as -

\[
\begin{align*}
\dot{x}(t) &= f(x(t), p) \quad (3.1) \\
y(t) &= g(x(t), p), \quad x(t_0) \in X_0 \quad (3.2)
\end{align*}
\]

where \( x \in \mathbb{R}^n \) denote the n-dimensional state vector, \( y \in \mathbb{R}^m \) denote the m-dimensional measured output vector and \( p \in \mathbb{R}^p \) are the p model parameters. The initial conditions \( x(t_0) \) are supposed to belong to an initial “box” \([X_0]\), the notion of ”box” being described in the following.

An interval vector (or box) \([X]\) is a vector with interval components and may equivalently be seen as a Cartesian product of scalar intervals: \([X] = [x_1] \times [x_2] \times \ldots \times [x_n]\), where \([x_i]\) represents the interval set \( \underline{x}_i \leq x_i \leq \overline{x}_i \). The set of \( n \)-dimensional real interval vectors is denoted by \( IR^n \).

Time is assumed to belong to \([0, t_{\text{max}}]\). The functions \( f \) and \( g \) are real and finite on \( M \), where \( M \) is the open set of \( \mathbb{R}^n \) such that \( x(t) \in M \) for every \( t \in [0, t_{\text{max}}] \). Moreover the function \( f \) is assumed to be at least \( k \)-times continuously differentiable in the domain \( M \). The structures of functions \( f \) and \( g \) depend on the modeling frameworks required to describe specific biological systems.

The output error is assumed to be given by:

\[
v(t_i) = y(t_i) - y_m(t_i), \quad i = 1, 2, ..., N. \quad (3.3)
\]

The model output \( y_m \) may be computed by a function or finite algorithm or as the
solution of a differential equation. \( v(t) \) and \( \overline{v(t)} \) are known lower and upper bounds for the acceptable output errors. Such bounds may, for instance, correspond to a bounded measurement noise. The integer \( N \) is the total number of sample times.

Interval arithmetic is used to compute guaranteed bounds for the solution of (3.1) and (3.2) at the sampling times \( \{t_1, t_2, ..., t_N\} \).

### 3.3 Problem Statement

In this work, our goal is to discriminate between two different mechanisms that produce similar biological phenotypes. We develop an approach that investigates transitions to the steady states and uncertainty to discriminate between competing mechanistic hypotheses. We do this by exploring the consistency of estimated uncertain parameters over uncertain experiments that explore model dynamics.

If we have two or more models which have same steady states and fit an already available experimental data set equally well, our task is to see how well these models explain a new set of data and based on our observation, we try to discriminate the models. For this purpose, we generate new sets of parameters when we have the same models, but new data sets. If the difference between the previous and the current parameter sets for any model comes out to be very large, then this indicates that model is not accurate. And if the difference is not large, then we can take that model as valid and refine the model parameters appropriately by concatenating the two data sets. The main goal of this approach is to see how well the old models describe the new data sets. Later on, in the section 3.7, we have discussed this approach in detail.
3.4 Interval Analysis

Interval arithmetic is a method of bounding measurements and rounding off errors in mathematical computations. Interval analysis was initially developed to assess the numerical errors resulting from the use of finite-precision arithmetic. It now makes it possible to obtain guaranteed solutions to standard engineering problems such as computing solution of non-linear system of equations or all global minimizers of a cost function, or computing sets guaranteed to contain all solutions of sets of nonlinear inequalities.

This approach can characterize the uncertainty associated with biological systems better than other approaches which are based on statistical assumptions. Thus, due to limited accuracy of available experimental data, interval analysis [24] is becoming an essential tool for biological modeling.

In the following section, we present some basics tools of interval arithmetic and two popular algorithms to compute the outer approximation of sets of arbitrary shape.

3.4.1 Basic Tools

A real interval \([x] = [x, x]\) is a closed and connected subset of \(\mathbb{R}\) where \(x\) represents the lower bound of \([x]\) and \(x\) represents the upper bound of \([x]\). The width of an interval \([x]\) is given by \(w(x) = x - x\), and its midpoint by \(m(x) = (x + x)/2\).

Two intervals \([x]\) and \([y]\) are considered equal if and only if \(x = y\) and \(x = y\). We can also extend real arithmetic operations to intervals [24].

Arithmetic operations on two intervals \([x]\) and \([y]\) can be defined by:

\[
\circ \in \{+, -, *, /\}, \ [x] \circ [y] = \{x \circ y \mid x \in [x], y \in [y]\}.
\]
An interval corresponding to $[u] \circ [v]$ for the first three operators can be easily obtained by:

$$[x] + [y] = [x + y, \overline{x} + \overline{y}], \quad [x] - [y] = [x - \overline{y}, \overline{x} - y],$$

$$[x] \ast [y] = [\min(xy, x\overline{y}, \overline{x}y, \overline{x}\overline{y}), \max(xy, x\overline{y}, \overline{x}y, \overline{x}\overline{y})].$$

However, for division, when $0 \not\in [y]$,

$$[x]/[y] = [\min(x/y, x/\overline{y}, \overline{x}/y, \overline{x}/\overline{y}), \max(x/y, x/\overline{y}, \overline{x}/y, \overline{x}/\overline{y})].$$

and extended intervals have to be introduced, when $0 \in [y]$.

Thus, the interval counterpart of a real-valued function is an interval-valued function defined as:

$$f([x]) = \{f(x) \mid x \in [x]\},$$

where $[S]$ is the interval hull of $S$, i.e., the smallest interval containing it.

Interval counterparts to continuous elementary functions can also be obtained easily. For monotonic functions, only computations on bounds are required:

$$\exp([x]) = [\exp(\underline{x}), \exp(\overline{x})],$$

$$\log([x]) = [\log(\underline{x}), \log(\overline{x})], \quad \text{if } \underline{x} > 0$$

For non-monotonic elementary functions, like the trigonometric functions, algorithmic definitions are still easily obtained. For example, the interval square function is computed
as:

\[
[x]^2 = \begin{cases} 
0, \max(x^2, x^2), & \text{if } 0 \in [x] \\
\min(x^2, x^2), \max(x^2, x^2), & \text{else}
\end{cases}
\]

For more complicated functions, it is no longer possible to evaluate the interval counterparts easily and, thus the inclusion functions were introduced. An inclusion function \([f(.)]\) for a function \(f(.)\) defined over a domain \(D \subset R\) is such that the image of an interval by this function is an interval, guaranteed to contain the image of the same interval by the original function:

\[
\forall [x] \subset D, f([x]) \subset [f([x])]. \tag{3.4}
\]

When the inclusion in (3.4) becomes an equality, the inclusion function is minimal.

This inclusion function is convergent if \(\lim_{w([x]) \to 0} w([f([x])]) = 0\), where \(w\) is the width of the interval \([x]\), and inclusion monotonic if \([x] \subset [y]\) implies \([f([x])] \subset [f([y])].\)

Figure 3.1: Mapping from state space \(R^n\) to data space \(R^m\) using the function \(y(t) = g(x(t), p)\)[19]

Interval box enclosures and inclusion functions are used extensively nowadays to
model linear as well as non-linear systems in a bounded error context.[20],[19]. In Figure 3.2, it is shown that the model function \( g \) maps a state interval box \([x]\) to another image, \( g([x]) \), in the output data space, \( R^m \). Here, \([x]\) represents the Cartesian product of \( n \) scalar intervals i.e. \([x] = [x_1] \times [x_2] \times ... \times [x_n] \), where \([x_i]\) represents the interval set \( x_i \leq x_i \leq \overline{x}_i \). \([g]\) is a unique minimum inclusion function that maps \( x \) to the smallest box that encases the image, \( g([x]) \). As in practical scenarios, these minimum inclusion functions are not always attainable, an approximation is made such that \([g](x) \subset G([x])\).

### 3.4.2 Inverse Image Evaluation (Set Inversion)

The SIVIA algorithm (for Set Inverter Via Interval Analysis) [11] is dedicated to the characterization of sets defined by :

\[
S = \{ u \in U | \Phi(u) \in [y] \} = \Phi^{-1}([y]) \cap U,
\]

where \([y]\) is known a priori, \( U \) is an a priori search set for \( u \) and \( \Phi \) a nonlinear function not necessarily invertible. (3.5) involves computing the reciprocal image of \( \Phi \) and is known as a set inversion problem which can be solved using the algorithm Set Inversion Via Interval Analysis (denoted SIVIA). The algorithm SIVIA [11] is a recursive algorithm which explores the entire search space without losing any solution and ensures a guaranteed enclosure of the solution set \( S \). The algorithm recursively builds two unions of non-overlapping boxes (subpavings) \( \underline{S} \) and \( \overline{S} \) such that :

\[
\underline{S} \subseteq S \subseteq \overline{S}
\]
The inner enclosure $\mathcal{S}$ is composed of the feasible boxes. To show that a box $[u]$ is feasible it is sufficient to prove that $\Phi([u]) \subseteq [y]$. Reversely, if $\Phi([u]) \cap [y] = \emptyset$, the box $[u]$ is unfeasible. Otherwise, no conclusion can be reached and the box $[u]$ is indeterminate. The latter is then bisected and checked again until its size reaches a user-specified precision threshold $\epsilon > 0$. This ensures that the SIVIA algorithm terminates after a finite number of iterations.

### 3.4.3 Direct Image Evaluation

Another popular algorithm using the interval analysis concept is Direct image evaluation. This is the characterization of

$$Y = f(X)$$

when $X$ is a known set and $f$ a known function. This task is usually quite complex. The algorithm IMAGESP, again based on interval analysis, builds a sub-paving $Y$ such that $Y \subset \overline{Y}$ when $\overline{Y}$ is an outer approximation of the set $Y$, $X$ is itself a sub-paving and an inclusion function $[f]$ is available for $f$. A subpaving of a box $[x] \in \mathbb{R}^n$ is a union of non-overlapping subboxes of $[x]$ with non-zero width. Two boxes in the same subpaving may have a non-empty intersection if they have a boundary in common, but their interiors must have empty intersection. IMAGESP consists of three steps:

1. Firstly, all boxes in $X$ are bisected in order to obtain a sub-paving $X'$ consisting only of boxes of width less than a pre-specified precision parameter $\epsilon$.

2. Then, the images of all boxes of $X'$ are evaluated using $[f]$ and stored into a list of image boxes $\mathcal{Y}$. 
3. Finally, all boxes in $\mathcal{Y}$ are merged into a new sub-paving $\bar{Y}$.

$\bar{Y}$ is guaranteed to contain $Y$. The precision of the approximation is controlled by an user-defined threshold $\epsilon > 0$.

### 3.5 Extended Mean Value Theorem and Set Inversion via Interval Analysis

The Extended Mean Value theorem (EMV) is a method to find a valid solution of eqs. (3.1) and (3.2) at an initial condition $x_0 = x(t_0)$ using bounded error context. This algorithm produces a enclosure of the state estimate at each time step, ensuring that all possible solutions are retained. This is done by using Taylor expansion method [24, 26]. The Taylor expansion is evaluated using the method explained in [29] with the help of mean value theorems [1] and matrix preconditioning.

EMV uses interval analysis and generates state intervals. Whenever a state value is generated at a time where previous measurements are available, SIVIA is employed to compare those two values. If we have a function $h$, output $y$ and an initial condition $x_0$, SIVIA is used to find a solution of $x = h^{-1}(y)$. The state space $x$ is iteratively searched for guaranteed enclosure of the solution space. In the solution space, the feasible boxes are surrounded by the set of indeterminate boxes. The parameter boxes are classified as follows -

- If the state intervals are completely contained within the data measurements, the parameter subset is feasible.
- If the state intervals span outside but still contain part of the data measurement, the parameter subset is indeterminate.
• If the data measurement is completely outside the state interval the parameter subset is unfeasible.

This basic concept of the set inversion algorithm can be illustrated using a simple example:

Let us consider, the set \( X = f^{-1}([4, 9]) \) where \( f(x_1, x_2) = x_1^2 + x_2^2 \). As explained above, there can be the following three scenarios:

• Since \([2, 1]^2 + [4, 5]^2 = [4, 1] + [16, 25] = [20, 26]\) does not intersect the interval \([4, 9]\), the box \([-2, 1] \times [4, 5]\) is outside \( X \) and considered unfeasible.

• Since \([0, 2]^2 + [2, 1]^2 = [0, 4] + [4, 1] = [4, 5]\) is completely inside the interval \([4, 9]\), the box \([0, 2] \times [2, 1]\) is considered feasible.

• Since \([1, 1]^2 + [0, 2]^2 = [1, 1] + [0, 4] = [1, 5]\) is partially inside the interval \([4, 9]\), the box \([-2, 1] \times [4, 5]\) is considered indeterminate.

Below, we give a pictorial representation of the SIVIA algorithm:

![Diagram of SIVIA algorithm](image)

Figure 3.2: Feasibility of boxes using SIVIA
3.6 Parameter and State Estimation

The parameter and state estimation approach used in this work are originally developed by Raissi [30] and Jaulin [11] and later on parallelized by Skylar Marvel[19]. This estimation method efficiently combines the EMV and SIVIA algorithms for evaluating the parameter boxes and checks whether each box produces trajectories consistent with the available data or not.

The algorithm starts with an unclassified parameter box and Extended Mean Value algorithm is applied on it. Whenever a data measurement is available during the execution of the EMV algorithm, the corresponding state output is compared to the bounded measurement using SIVIA. The unfeasible sets are immediately discarded. For all other cases the EMV algorithm will run until the final time and the parameter box is then classified. Feasible boxes are retained and other boxes are regarded as indeterminate. When a box is found to be indeterminate, the longest edge is compared to a user defined width, epsilon. If the longest box edge is larger than epsilon, the indeterminate box is bisected and each resulting box is reevaluated. The PE algorithm runs until every parameter box has been classified as feasible, unfeasible, or is small enough to be classified as indeterminate. Block diagram of this algorithm is shown in Figure 3.3.

3.7 Model Discrimination

The work in this paper is based on the parallelized parameter estimation (PPE) algorithm developed by Marvel & Williams [18]. The discrimination problem is designed using one reference model and several candidate models, representing competing hypotheses. As determining model parameters directly through experimentation is difficult, the parame-
Parameters are often obtained indirectly by estimating values based on the model equations and a collection of data measurements. Thus in our research, for the given reference model, we use EMV algorithm to generate state outputs (which can correspond to component concentrations or quantities) over time for different initial states of the biological system and as measurement uncertainties are better characterized by bounded sets, set based parameter estimation techniques are used to obtain feasible parameter sets of each candidate model, which generate model outputs consistent with the uncertain data measurements. The process is repeated for all the candidate models using the same initial conditions and the corresponding data sets of the reference model. Candidate models having empty feasible parameter sets are directly discarded, as biologically, this implies those models do not explain the biological system under every circumstance. Models having feasible parameter sets for all initial states are again evaluated to check the consistency of those models’ estimated parameter sets over the uncertain experimental data with the help of some biologically meaningful metric. In this paper, we have used parameter bounds [19] as our consistency metric. Bound for a parameter is computed by taking union of all the feasible parameter sets corresponding to different initial conditions. This metric can help us to predict the uncertainty associated with each parameter of the candidate
models. Single parameter values are compared using the width of the uncertainty for the parameter of interest, e.g. $P_p = \bar{p}_i - \underline{p}_i$, where $\bar{p}_i$ denotes the upper bound, $\underline{p}_i$ the lower bound of the selected parameter over all the feasible parameter sets for all the initial conditions. This parameter bound represents the maximum amount of uncertainty associated with the selected parameter of a particular model. More the width of the parameter bound, more is the noise associated with that parameter. So, smaller parameter bound of a model implies that the parameter of that particular model is more consistent. The model with minimum parameter bound, thus, represents the reference model most accurately and is chosen as the most suitable model i.e. the best among the competing hypotheses.

A high level block diagram of our model discrimination approach is shown in Fig 3.4.

![Block Diagram for Model Discrimination method](image-url)
3.8 Discussion

This chapter has focused on a detailed description of our model discrimination approach. We have also provided brief overviews of the relevant algorithms. Our framework is entirely based on the works done by Skylar et al. [19]. Thus, we have used a parallelized parameter estimation algorithm which proved to be faster than traditional PE algorithms. Previous model discrimination algorithms, rather than discrimination, aimed at invalidating the models with empty feasible parameter sets. However, our approach can not only invalidate the models with empty parameter sets, but can also make discrimination among the candidate models with feasible parameter sets.

In the next chapter, we will discuss the usefulness of our approach by illustrating its implementation on two different test cases.
Chapter 4

Results

In the previous chapter, we have given a detailed description of our model discrimination methodology. In this chapter, we illustrate its application by providing two case studies. In the first case, we discriminate between two different predator-prey population models and in the second case study, we apply our method to discriminate between two simple alternative reaction schemes, the Michaelis-Menten and the Henry mechanisms. In both the cases, we aim at model discrimination when we are already provided with uncertain data measurement sets corresponding to different initial conditions. All the numerical experiments were performed on a MacBook Pro with an Intel Core 2 Duo 3.06 GHz processor. The code is written in C++ language using Boost interval arithmetic library and compiled in Xcode 3.2.6. The phase portraits are generated in PPLANE and all other graphs are plotted using MATLAB.
4.1 Case Study I

The first case studies two different Lotka-Volterra predator-prey population models. This kind of model describes the population dynamics of two species. We assume that both the species are competing for the same food and the food supply is limited. Each species grow according to its own capacity in absence of the other. This can be modeled assuming logistic growth of each species[35]. However, the problem starts when they confront each other. We assume that the growth rate for each species reduces due to this conflict. In our work, we have taken two specific models that incorporates these assumptions.

4.1.1 Model Discrimination Between Two Predator - Prey Population Models

In both the models, the prey and the predator populations are denoted by $x_1$ and $x_2$. The general form of the predator-prey model is -

\[
\dot{x}_1 = p_1 x_1 - p_2 x_1^2 - p_3 x_1 x_2 \tag{4.1}
\]

\[
\dot{x}_2 = p_4 x_2 - p_5 x_2^2 - p_6 x_1 x_2 \tag{4.2}
\]

where $p_i$ are the parameters reflecting the rates of population change of the predators and the prey.

The parameters for our reference model are taken as: $p_1 = 3, p_2 = 2, p_3 = 2, p_4 = 2, p_5 = 1, p_6 = 1$. Data is generated by the EMV algorithm from this model and uncertainty is added to obtain bounded interval measurements.

For the candidate models, we assume that the parameters - $p_1, p_2, p_4$ and $p_5$ are known. Our aim is to estimate the feasible parameter sets corresponding to $p_3$ and $p_6$ using the
data generated from the reference model.

For the candidate models, the known parameters are taken as -

- Model 1: \( p_1 = 3, p_2 = 2, p_4 = 2, p_5 = 1 \)
- Model 2: \( p_1 = 3, p_2 = 1, p_4 = 2, p_5 = 1 \)

4.1.2 Experimental Setup

The first step of our algorithm deals with generation of data measurements from the reference model. The measurements are simulated for 6 different initial conditions, \((x_1(0), x_2(0)) = \{(1, 2), (1, 3), (2, 2), (2, 3), (3, 2), (3, 3)\}\) and for each set of initial condition, a 4th order Taylor expansion is used to obtained 30 data points. We run this experiment with a constant time step of 0.005 and \(\alpha = 0.005\). The bounded interval measurements (shown in Figure 4.1) are then obtained by adding ±0.01 uncertainty to all the generated state values.

The second step aims at finding the feasible parameter values of \(p_3\) and \(p_6\) for the candidate models corresponding to the data generated from the reference model. Here, we only use the bounded error measurements for the prey population i.e. \(y = x_1\). The initial intervals assumed for the parameters are as follows -

- Model 1: \(p_3 = [1, 3]\) and \(p_6 = [0, 2]\)
- Model 2: \(p_3 = [1, 3]\) and \(p_6 = [0, 2]\)

For both the models, the indeterminate boxes are bisected until the width is less than \(\epsilon = 0.00001\).

Based on the consistency of the feasible parameter sets, the candidate models are discriminated.
Figure 4.1: State Estimation of the Reference Model for 6 different initial conditions. X-axes represent time, Y-axes represent bounded state outputs. The red and black colored lines represent the lower and upper bounds of the state intervals.
The state and parameter estimation stage of our experiment is most time consuming. Thus, we ran the algorithm with either 1, 2, 4 or 8 threads. The average computation times with respect to each initial condition were 33.35, 15.48, 7.92 and 4.24 seconds for 1, 2, 4 and 8 threads, respectively. Observed decreases in computation time were factors of 2.15, 4.21 and 7.86 when increasing the number of threads from 1 to 2, 4 and 8, respectively.

4.1.3 Results and Analysis

The results for parameter estimation of candidate model 1 and 2 are displayed in the tables 4.1 and 4.2.

It is observed that model 1 has feasible parameter sets consistent with the data generated from the reference model for all the available initial conditions. However, in case of model 2, there are some empty feasible parameter sets corresponding to the initial conditions (3,2) and (3,3). Thus, we can say that for those initial conditions, the candidate model 2 is unable to explain the data of the reference model. As a consequence, we can invalidate model 2 and take model 1 as the more appropriate model for the given reference model.
Table 4.1: Results of PE for Model 1

<table>
<thead>
<tr>
<th>Initial Conditions</th>
<th>Indeterminate set of $p_3$</th>
<th>Feasible Set of $p_3$</th>
<th>Indeterminate set of $p_6$</th>
<th>Feasible Set of $p_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,2)</td>
<td>[1.936256, 2.063865]</td>
<td>[1.936294, 2.063819]</td>
<td>[0.941009, 1.060043]</td>
<td>[0.941047, 1.060005]</td>
</tr>
<tr>
<td>(1,3)</td>
<td>[1.946128, 2.054740]</td>
<td>[1.946174, 2.054687]</td>
<td>[0.945365, 1.055259]</td>
<td>[0.945419, 1.055198]</td>
</tr>
<tr>
<td>(2,2)</td>
<td>[1.953315, 2.046562]</td>
<td>[1.953392, 2.046485]</td>
<td>[0.9642486, 1.035720]</td>
<td>[0.964302, 1.035667]</td>
</tr>
<tr>
<td>(2,3)</td>
<td>[1.961991, 2.038322]</td>
<td>[1.962089, 2.038223]</td>
<td>[0.9679641, 1.032402]</td>
<td>[0.968048, 1.03231]</td>
</tr>
<tr>
<td>(3,2)</td>
<td>[1.957725, 2.042152]</td>
<td>[1.957847, 2.042015]</td>
<td>[0.9721679, 1.02787]</td>
<td>[0.9722366, 1.027793]</td>
</tr>
<tr>
<td>(3,3)</td>
<td>[1.966384, 2.033607]</td>
<td>[1.966552, 2.033439]</td>
<td>[0.975433, 1.024749]</td>
<td>[0.975554, 1.02462]</td>
</tr>
</tbody>
</table>

Overall range of parameter variation:
- \([1.936256, 2.063865]\)
- \([1.936294, 2.063819]\)
- \([0.941009, 1.060043]\)
- \([0.941047, 1.060005]\)
<table>
<thead>
<tr>
<th>Initial Conditions</th>
<th>Indeterminate set of $p_3$</th>
<th>Feasible Set of $p_3$</th>
<th>Indeterminate set of $p_6$</th>
<th>Feasible Set of $p_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,2)</td>
<td>[2.380584, 2.500030]</td>
<td>[2.380607, 2.499984]</td>
<td>[0.939163, 1.057167]</td>
<td>[0.9391937, 1.057128]</td>
</tr>
<tr>
<td>(1,3)</td>
<td>[2.231399, 2.330024]</td>
<td>[2.231422, 2.329971]</td>
<td>[0.9434814, 1.051651]</td>
<td>[0.9435348, 1.051589]</td>
</tr>
<tr>
<td>(2,2)</td>
<td>[2.864250, 2.878952]</td>
<td>[2.864318, 2.878852]</td>
<td>[0.9831924, 1.02911]</td>
<td>[0.9836578, 1.029052]</td>
</tr>
<tr>
<td>(2,3)</td>
<td>[2.561820, 2.567054]</td>
<td>[2.561897, 2.566932]</td>
<td>[1.003822, 1.02456]</td>
<td>[1.004516, 1.024482]</td>
</tr>
<tr>
<td>(3,2)</td>
<td>[0, 0]</td>
<td>[0, 0]</td>
<td>[0, 0]</td>
<td>[0, 0]</td>
</tr>
<tr>
<td>(3,3)</td>
<td>[0, 0]</td>
<td>[0, 0]</td>
<td>[0, 0]</td>
<td>[0, 0]</td>
</tr>
</tbody>
</table>

Overall range of parameter variation
(empty sets not considered)  
[2.231399, 2.878952]  
[2.231422, 2.878852]  
[0.939163, 1.057167]  
[0.9391937, 1.057128]
4.2 Case Study II

Our second case study deals with two alternative reaction schemes - Henry mechanism and Michaelis-Menten mechanism. Both the schemes [31] represent the same chemical reaction in which an enzyme (E) reacts with a substrate (S) to form a complex (C) and gives the final product (P). The reactions are represented as -

- Michaelis-Menten Model: \( E + S \xrightleftharpoons[k_2]{k_1} C \xrightarrow{k_3} E + P \)

- Henry Model: \( C \xrightleftharpoons[k_2]{k_1} E + S \xrightarrow{k_3} E + P \)

where \( k_i \) are the rate constants in both models. [32] gives detailed description about the relevance of these models. In our work, we have discriminated these two schemes by capturing their transient phase behavior, as they are analytically indistinguishable during the steady state conditions.

4.2.1 Model Discrimination Between Henry and Michaelis-Menten Mechanisms

The above reaction schemes can be modeled as 4-state models according to the law of mass action.

- Michaelis-Menten Model
  
  \[
  \begin{align*}
  \dot{S} &= -k_1 E S + k_2 C \\
  \dot{C} &= k_1 E S - k_2 C - k_3 C \\
  \dot{E} &= -k_1 E S + k_2 C + k_3 C \\
  \dot{P} &= k_3 C
  \end{align*}
  \]
• Henry Model
\[\dot{S} = k_1 C - k_2 ES\]
\[\dot{C} = -k_1 C + k_2 ES\]
\[\dot{E} = k_1 C - k_2 ES + k_3 ES\]
\[\dot{P} = k_3 ES\]

These 4-state models can again be converted into 2-state models without any loss of generality by utilizing a conservation relationship satisfied by both the schemes. First we assume that the systems are dependent only on the concentrations of S and C. And we fix the total enzyme concentration \(E + C\) to a constant value 1 [31].

The above equations transform into the following two state forms -

• Michaelis-Menten Model
\[\dot{S} = -k_1 S + k_2 C + k_1 ES\]
\[\dot{C} = k_1 S - (k_2 + k_3)C - k_1 ES\]

• Henry Model
\[\dot{S} = -(k_2 + k_3)S + k_1 C + (k_2 + k_3)ES\]
\[\dot{C} = k_2 S - k_1 C - k_2 ES\]

For simplicit, we take the substrate and the complex concentrations as \(x_1\) and \(x_2\) respectively. We then rewrite the above schemes into a general form as follows -

\[\dot{x}_1 = -ax_1 + bx_2 + cx_1x_2\]  \hspace{1cm} (4.3)
\[\dot{x}_2 = dx_1 - ex_2 - fx_1x_2\]  \hspace{1cm} (4.4)

where \(a, b, c, d, e\) and \(f\) denote the parameter constants.
We aim at invalidating the Michaelis-Menten model based on the uncertain measurements generated from the Henry model which we have considered to be the reference model in this case. The parameters for our reference model are taken as: \(a = k_2 + k_3 = 2, b = k_1 = 1, c = k_2 + k_3 = 2, d = k_2 = 1, e = k_1 = 1, f = k_2 = 1\). Data is generated by the EMV algorithm from this model and uncertainty is added to obtain bounded interval measurements.

For the Michaelis-Menten model, we assume that the parameters - \(a, b, d, e\) are known. Our aim is to estimate the feasible parameter sets corresponding to \(c\) and \(f\) using the data generated from the Henry model. The known parameters for this model are taken as - \(a = 1.5, b = 1.3, d = 1.5\) and \(e = 1.7\).

### 4.2.2 Experimental Setup

Data is generated from the Henry model for 10 different initial conditions: \((x_1(0), x_2(0)) = \{(0, 3), (1, 1), (1, 2), (1, 3), (2, 1), (2, 2), (2, 3), (3, 1), (3, 2), (3, 3)\}\) and a 4th order Taylor expansion is employed to obtain 30 data points for each initial condition. As both models are analytically indistinguishable during the steady state conditions, the initial transient dynamics are taken into consideration. The experiment is performed with a constant time step \(h = 0.005\) and \(\alpha = 0.005\). The bounded interval measurements (shown in Figure 4.2) are then obtained by adding \(\pm 0.01\) uncertainty to all the generated state values. Once the data is generated from the reference model, we aim at finding the feasible parameter values of \(c\) and \(f\) for the Michaelis-Menten model consistent with the generated data. Here, we only use the bounded error measurements of the substrate concentration i.e. \(y = x_1\). Initial parameter intervals of \(c = [0.5, 2.5]\) and \(f = [0.5, 2.5]\) are used and the indeterminate boxes are bisected until the width is less than \(\epsilon = 0.00001\).
Figure 4.2: State Estimation of the Reference Model for 10 different initial conditions. X-axes represent time, Y-axes represent bounded state outputs. The red and black colored lines represent the lower and upper bounds of the state intervals.
ran the algorithm with either 1, 2, 4 or 8 threads. The average computation times with respect to each initial condition were 43.15, 21.54, 11.27 and 5.44 seconds for 1, 2, 4 and 8 threads, respectively. Observed decreases in computation time were factors of 2.01, 3.83 and 7.97 when increasing the number of threads from 1 to 2, 4 and 8, respectively.

4.2.3 Results and Analysis

The output of parameter estimation of the Michaelis-Menten model is displayed in Figure 4.3 as well in table 4.3.

The simulation results in several empty parameter sets which implies that the Michaelis Menten model can not describe the data of the Henry model for all the considered initial conditions. Thus, the Michaelis-Menten model can be invalidated.
Figure 4.3: Parameter Estimation of the Michaelis-Menten Model for 10 different initial conditions. X-axes represent the parameter space of $c$ and Y-axes represent parameter space of $f$. 
Table 4.3: Results of PE for the Michaelis-Menten Model

<table>
<thead>
<tr>
<th>Initial Conditions</th>
<th>Indeterminate set of $c$</th>
<th>Feasible Set of $c$</th>
<th>Indeterminate set of $f$</th>
<th>Feasible Set of $f$</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0,3)</td>
<td>[0, 0]</td>
<td>[0, 0]</td>
<td>[0, 0]</td>
<td>[0, 0]</td>
</tr>
<tr>
<td>(1,1)</td>
<td>[1.101234, 1.269065]</td>
<td>[1.101249, 1.269050]</td>
<td>[0.7969512, 0.9675521]</td>
<td>[0.7969665, 0.9675445]</td>
</tr>
<tr>
<td>(1,2)</td>
<td>[1.460266, 1.516708]</td>
<td>[1.460281, 1.516693]</td>
<td>[0.7094268, 0.7210769]</td>
<td>[0.7094345, 0.7210693]</td>
</tr>
<tr>
<td>(1,3)</td>
<td>[0, 0]</td>
<td>[0, 0]</td>
<td>[0, 0]</td>
<td>[0, 0]</td>
</tr>
<tr>
<td>(2,1)</td>
<td>[1.276832, 1.370780]</td>
<td>[1.276847, 1.370765]</td>
<td>[1.150283, 1.245109]</td>
<td>[1.150291, 1.245101]</td>
</tr>
<tr>
<td>(2,2)</td>
<td>[1.577049, 1.604072]</td>
<td>[1.577064, 1.604049]</td>
<td>[0.9864959, 0.9906387]</td>
<td>[0.9865036, 0.9906311]</td>
</tr>
<tr>
<td>(2,3)</td>
<td>[0, 0]</td>
<td>[0, 0]</td>
<td>[0, 0]</td>
<td>[0, 0]</td>
</tr>
<tr>
<td>(3,1)</td>
<td>[1.341156, 1.407348]</td>
<td>[1.341178, 1.407333]</td>
<td>[1.272621, 1.340484]</td>
<td>[1.272628, 1.34047]</td>
</tr>
<tr>
<td>(3,2)</td>
<td>[0, 0]</td>
<td>[0, 0]</td>
<td>[0, 0]</td>
<td>[0, 0]</td>
</tr>
<tr>
<td>(3,3)</td>
<td>[0, 0]</td>
<td>[0, 0]</td>
<td>[0, 0]</td>
<td>[0, 0]</td>
</tr>
</tbody>
</table>
4.3 Discussion

In both the examples of this chapter, we discarded the model which had empty feasible parameter sets for certain initial conditions and as a consequence, the other model is considered as the best model. However, our algorithm can also be applied for cases where the competing models have only non-empty feasible parameter sets. In such scenarios, the discrimination is done by evaluating the overall parameter bounds of each model.

As an example, we can go to back case study I and refer to tables 4.1 and 4.2. We can see the range of feasible parameter set corresponding to $p_5$ is almost the same for both the models. However, a sufficient dissimilarity can be observed in the parameter range corresponding to $p_3$. For model, the feasible parameter bound comes out to be 0.127525329 while for model 2, it is 0.64743042. Thus, $p_3$ varies over a larger range in model 2 which implies amount of uncertainty associated with the parameters of model 2 is larger than that of model 1. Also, Figure 4.4 and Figure 4.5 show that the parameters of model 1 are more consistent than that of model 1. Thus, we can say that model 1 fits the data of the reference model better than model 2. In this chapter, we have shown how our algorithm can be applied to solve model discrimination problems. Using this approach we have captured the steady as well as the transient behavior of the given biological models. However, it is observed that during steady states, the models either bear very similar dynamics and thus can not be distinguished analytically, or give strange results due to cumulative error. As a result, we have preferred to analyze the models based on their transient phase dynamics. Also, the algorithm accumulates error, as time goes on. Thus, we have generated very few data points from the reference model and restricted the simulation time of the EMV algorithm to prevent the system from exploding.

Also, we realize computational tractability is the major concern with interval analy-
Figure 4.4: Parameter Bound for Model 1
sis method. Combining EMV and SIVIA for bounded parameter estimation in biological systems is limited due to the increase in computation time for larger models and number of unknown parameters. Execution times were reduced by utilizing multi-threading techniques as proposed by Skylar et al. in [19] and were able to increase performance of our approach by nearly the number of cores used for processing. For further improvement, the algorithm can potentially be scaled up for use on supercomputers and may be modified to use the many processors of graphics processing units.
Chapter 5

Conclusion and Future Works

5.1 Conclusion

This thesis addressed the issue of model discrimination in biological systems. We proposed a solution method which is capable of providing promising results even with uncertain measurements and model parameters. A set-based approach has been developed where the uncertainty associated with the system is characterized by a continuous range of values and thus, gives more reliable results than methods assuming statistically defined noise or probabilistic uncertainty.

Chapters 1 and 2 gives an overview of our model discrimination approach and also presents the motivation and current state of art associated with it. In chapter 3, we explained all the relevant concepts and interval arithmetic algorithms needed for our approach and a detailed description of our algorithm is provided. Then, we demonstrated our approach, in chapter 4, by applying the algorithm to two different categories of biological systems - (i) predator-prey models and (ii) enzyme-catalyzed biochemical models. The illustrations showed that our bounded error discrimination method is not only ap-
plicable for invalidating models based on empty feasible parameter sets, but also for scenarios, where all the candidate models are having feasible parameter sets consistent with the generated data.

The key to our approach is in designing our algorithm in a set-based feasibility framework. This allows us to discard the regions which are not consistent with the experimental data and bounding the measurements in interval sets guarantees that no valid solution is lost.

Moreover, our algorithm is based on the improved parallelized parameter estimation algorithm [19] developed by Skylar et al. which efficiently uses multicore-chips of a computer to speed up the entire process. Based on the works presented in [20], this method is capable of producing reliable parameter estimates, even when uncertain and incomplete measurements are available.

Thus, we conclude that our proposed method is more effective than the other classical and contemporary model discrimination methods.

5.2 Limitations of the Proposed Method and Future Scope

There is further scope of improvement in the work presented here.

- Computational burden resulting from the complexity of biological systems is still a big issue. Our approach is mainly suitable for lower dimensional problems. Also, the code takes long time to run when using large error bounds. The computational requirements of the bounded-error methods increase exponentially with the number of model parameters and states. And for our case, while handling more than one
model, this problem becomes more severe. Though, our algorithm is based on a parallelized parameter estimation algorithm, future developments can be made to run it on supercomputers using multi-processor GPUs. We can also reduce the computational burden by retaining larger feasible boxes, but that will result in inaccurate parameter estimates. An alternative solution may be to perform the sensitivity analysis of the model parameters and focus on estimating only those parameters upon which our system is sufficiently sensitive.

- Moreover, the EMV and SIVIA algorithms seem to be very model sensitive. Adding error bounds on the state estimates result in negative lower state bounds for some model. This is not valid in practice as we are dealing with population models and chemical reaction models. Thus, our method works better in systems where relatively high number of data samples are available.

- One of the other drawbacks associated with set based approaches is the wrapping effect. This can lead to very unreliable model predictions due to encasement of the predicted state output in an enclosure (e.g. interval boxes in our case) at each time step. Proper choice of enclosure shape is very crucial for this reason. Several methods have been proposed in the literature to resolve this issue. Incorporating those methods in our algorithm will lead to better results.

Thus in the future works, we will focus on reducing the wrapping effect and computational burden associated with this approach. Moreover, applications of this method can be extended to model discrimination of other control system based models (for example, mechanical systems).
References


