

A Predator-Prey Perspective on the Tumor Microbiome

Michael Lam



Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2023-94

<http://www2.eecs.berkeley.edu/Pubs/TechRpts/2023/EECS-2023-94.html>

May 11, 2023

Copyright © 2023, by the author(s).
All rights reserved.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission.

A Predator-Prey Perspective on the Tumor Microbiome

Michael Lam

Research Project

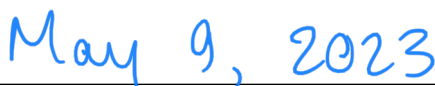
Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee



Claire Tomlin
Research Advisor



(Date)

★ ★ ★ ★ ★ ★ ★



Murat Arcak
Second Reader

May 10, 2023

(Date)

A Predator-Prey Perspective on the Tumor Microbiome

by

Michael Lam

A thesis submitted in partial satisfaction of the
requirements for the degree of

Master of Science

in

Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Claire Tomlin, Chair
Professor Murat Arcak

Spring 2023

A Predator-Prey Perspective on the Tumor Microbiome

Copyright 2023
by
Michael Lam

Abstract

A Predator-Prey Perspective on the Tumor Microbiome

by

Michael Lam

Master of Science in Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Claire Tomlin, Chair

It is difficult to model systems on the cellular level because of the large numbers of variables involved in these environments. The tumor microbiome is no exception. Linear models are too simplistic and unstructured to model long-term time dependencies while neural network models are so complex that it becomes difficult to gain any biological insight from the parameters. We propose using the generalized Lotka-Volterra equations, a predator-prey model that imposes an ecological structure onto the problem of inferring population-to-population interactions within the tumor microbiome. Inference was performed via model optimization on experimental data obtained from three-dimensional bioprinted tumor spheroids. We found that this deterministic method can often produce results comparable to those produced by stochastic algorithms in the past. This optimization procedure has the added benefit of being robust to experimental noise.

To my father

Who taught me that there is always something to learn from every person in the room

Contents

Contents	ii
List of Figures	iv
List of Tables	vi
1 Introduction	1
2 Background	3
2.1 Bioprinted Tumor Tissues	3
2.2 Cell Slice Image Processing	4
2.3 Differential Equation-Based Biological Modeling	5
2.4 The Lotka-Volterra Equations	7
2.5 Lotka-Volterra Representation of Tumor Environments	8
2.6 Genetic Algorithmic Approach	9
2.7 Reference Formulas	10
3 Parameter Inference	12
3.1 Problem Formulation	12
3.2 Baseline: Genetic Algorithm	13
3.3 Deterministic Approach	14
4 Experimental Evaluation	23
4.1 Two-Population Analysis	23
4.2 Three-Population Analysis	28
5 Model Robustness	38
5.1 Integral Formulation	38
5.2 Logarithmic Integral Formulation	39
6 Conclusions and Future Work	42
6.1 Summary of Results	42
6.2 Future Directions	43

Bibliography	44
A Training Curves	46
B Parameter Robustness Charts	49
B.1 Two Populations	49
B.2 Three Populations	53

List of Figures

2.1	A bioprinted tumor spheroid.	4
2.2	CycIFAAP exclusiveness plot	5
2.3	Differentiation state finite state machine.	7
2.4	Generalized Lotka-Volterra (gLV) equations for n populations.	10
2.5	Generalized Lotka-Volterra (gLV) equations for 2 populations.	10
2.6	Generalized Lotka-Volterra (gLV) equations for 3 populations.	10
2.7	Integral formulation of 2 population generalized Lotka-Volterra equations.	11
2.8	Logarithmic integral formulation of 2 population generalized Lotka-Volterra equations.	11
3.1	Error Function	13
3.2	Raw data points	17
3.3	Polynomial interpolations	18
3.4	Derivative from Interpolant	19
4.1	Two-Population Error Comparison	24
4.2	2-population Network Interaction Graphs	26
4.3	2-population model vs. experimental Data, integral formulation	27
4.4	2-population model vs. experimental Data, logarithmic integral formulation	28
4.5	Three-Population Error Comparison	30
4.6	3-population Network Interaction Graphs	32
4.7	SKBR3 Data vs. Model, 3-population, integral formulation	33
4.8	MDAMB231 Data vs. Model, 3-population, integral formulation	33
4.9	MCF7 Data vs. Model, 3-population, integral formulation	34
4.10	HCC1143 Data vs. Model, 3-population, integral formulation	34
4.11	SKBR3 Data vs. Model, 3-population, logarithmic integral formulation	35
4.12	MDAMB231 Data vs. Model, 3-population, logarithmic integral formulation	36
4.13	MCF7 Data vs. ModelM, 3-population, logarithmic integral formulation	37
4.14	HCC1143 Data vs. Model, 3-population, logarithmic integral formulation	37
A.1	Two-Population SKBR3 and MDAMB231 Training Errors	46
A.2	Two-Population MCF7 and HCC1143 Training Errors	47
A.3	Three-Population SKBR3 and MDAMB231 Training Errors	47

A.4 Three-Population MCF7 and HCC1143 Training Errors	48
---	----

List of Tables

4.1	Genetic algorithm parameters, 2-population	23
4.2	2-population interactions, genetic algorithm	25
4.3	2-population interactions, integral formulation	26
4.4	3-population genetic algorithm parameters	29
4.5	3-population interactions, genetic algorithm	31
4.6	Inter-population interactions for three populations as predicted by the deterministic algorithm.	31
5.1	HCC1143 robustness results for two populations.	39
5.2	HCC1143 robustness results for 3 populations.	40
B.1	SKBR3 Robustness, Integral Formulation, 2-population	49
B.2	MDAMB231 Robustness, Integral Formulation, 2-population	50
B.3	MCF7 Robustness, Integral Formulation, 2-population	50
B.4	HCC1143 Robustness, Integral Formulation, 2-population	50
B.5	SKBR3 Robustness, Logarithmic Integral Formulation, 2-population	51
B.6	MDAMB231 Robustness, Logarithmic Integral Formulation, 2-population	51
B.7	MCF7 Robustness, Logarithmic Integral Formulation, 2-population	51
B.8	HCC1143 Robustness, Logarithmic Integral Formulation, 2-population	52
B.9	SKBR3 Robustness, Integral Formulation, 3-population	53
B.10	MDAMB231 Robustness, Integral Formulation, 3-population	54
B.11	MCF7 Robustness, Integral Formulation, 3-population	55
B.12	HCC1143 Robustness, Integral Formulation, 3-population	56
B.13	SKBR3 Robustness, Logarithmic Integral Formulation, 3-population	57
B.14	MDAMB231 Robustness, Logarithmic Integral Formulation, 3-population	58
B.15	MCF7 Robustness, Logarithmic Integral Formulation, 3-population	59
B.16	HCC1143 Robustness, Logarithmic Integral Formulation, 3-population	60

Acknowledgments

I would first like to thank my research advisor, Professor Claire Tomlin, for her constant support throughout the entirety of this past year. She has been so much more than an academic mentor to me—the encouraging kindness and instructive clarity of thought so clearly apparent in every one of her interactions is a clear testament to the strength of her character not just as an educator, but as a role model of what a true leader can be. I could not be more grateful for the opportunity to have worked with her this past year.

I would next like to thank Professor Murat Arcaak for the meticulous feedback he gave on the early drafts of this report. When I first emailed him during the hustle and bustle that marked the beginning of this semester, asking whether he'd be willing to be the second reader for this paper, he promptly responded with just three words:

“Yes, happy to.”

I am most grateful for those first three words and the ensuing critiques that would come to be so instrumental to the report that I now have before me. Professor Arcaak, thank you.

Last, but not least, I would like to express my thanks to my family, without whose unceasing prayers and endless encouragement I never would have even embarked on this journey toward my graduate degree. They have shown me far more loving support than I could ever hope to repay. I was taught to keep my eyes on the stars and my feet on the ground, but my family has shown me that this delicate balance between courageous ambition and belonging is something that cannot be achieved alone—to my family, thank you for lifting my eyes when I feel down and reminding me I have a home when I feel lost. I thank God for placing you in my life.

Chapter 1

Introduction

Accurately describing biological systems at the cellular level is a challenge due largely to the enormous numbers of variables at play, each operating on hundreds, perhaps even thousands of cells. This plethora of data complicates the all-important task of developing models in order to gain insight into the systems that govern the properties of these cells as well as the interactions between them, insights that can be essential to designing drugs and other treatments that deal intimately with environments and their ensuing impacts on the entire organism. The tumor environment is one such system that poses these difficulties. Not only do even the tiniest tumor bodies include thousands of cells, but each individual cell can be one of many cell types (fibroblast, cancerous epithelial, endothelial, etc.) [1]. The way these cells interact with each other depends on parameters like type, the relative population value, and even the three-dimensional spatial orientation among many other measures—the system’s complexity is astounding.

Linear models, though easy to understand and optimize, only provide rudimentary insights into the biological contexts to which they are applied and can be poor predictors of variables within time-dependent systems [9, 8]. On the opposite end of the spectrum are deep neural network architectures that can often be tuned to provide extremely accurate cell-to-cell interaction predictions but whose enormous model complexity (perhaps even including as many parameters as the experimental data itself, if not more) obscures any biological insights due to its lack of interpretability. We focus on an ecologically-inspired differential model that is more complex than the linear model but is simple enough to be interpretable unlike neural networks: the generalized Lotka-Volterra equations, a system of differential equations describing multi-population predator-prey interactions [12].

In this report, we infer the generalized Lotka-Volterra model’s defining parameters (self-reproduction, self-interaction, and inter-population impact) from experimental data gathered from bioprinted tumor environments [1]. This is done by first formulating the inference as a constrained least-squares problem, then optimizing over the generalized Lotka-Volterra parameters such that the resulting model describes the data with minimal error according to a predefined error metric. The model’s robustness to experimental noise is also evaluated. The results obtained through this deterministic process is compared to baseline parameter

sets obtained via a stochastic genetic algorithm.

We found that, at least for the two-population model, the types of interactions inferred by the deterministic model generally agreed with those predicted by the genetic algorithm. Results were more unclear in three-population interactions. For both two-population and three-population interactions, the deterministic algorithm predicts similar interactions even in the presence of large volumes of experimental noise, implying model robustness.

Chapter 2

Background

This chapter provides a brief overview of the past work that contextualizes and inspires the approaches explained later in this report. It also introduces the specific formulations of the generalized Lotka-Volterra model that will be used.

2.1 Bioprinted Tumor Tissues

A major challenge to understanding the inner workings of biological systems is the replication of massively complex cellular environments in a controlled manner, systems that are governed by a myriad of variables, including cell types, spatio-temporal interactions, physical and chemical environments, etc. Moreover, gathering data about these environments *in vitro* often involves killing the entities involved, which can deter the use of live subjects for these kinds of studies. Bioprinting is one alternative to performing experiments on live subjects. To study intercellular interactions within the tumor microbiome, the model inference in this technical report is performed on population data obtained from three-dimensional bioprinted tumor structures in the shape of a spheroid.

Each of these spheroids contains identically-sized cancerous cores surrounded by similar numbers of human mammary fibroblasts and human umbilical vein endothelial cells (Figure 2.1). Depending on the specific spheroid, the cancerous core will be one of several subtypes: luminal (MCF-7), HER2 amplified (SKBR3), basal-like (HCC1143), or claudin low (MDA-MB-231). These spheroids were bioprinted using Organovo's Novogen MMX Bioprinter Platform, which generates complex, multicellular tissues free of scaffolding according to a predefined architecture. On days 4, 7, and 10, a subset of each type of spheroid is fixed and stained using Vimentin (VIM), Cytokeratin 8/18 (KRT8/18), and CD31 antibodies to identify stromal fibroblasts, epithelial cancer cells, and endothelial cells, respectively. For several of the spheroid slices, only cancer cell and fibroblast staining were performed [1]. This will be reflected later in our Lotka-Volterra data inference, which is accordingly split into two parts: a two-population and a three-population analysis.

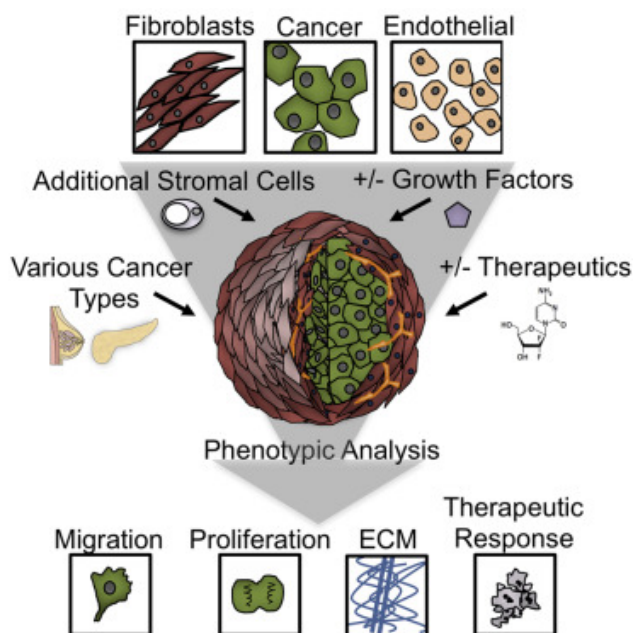


Figure 2.1: An illustration of a spheroid from “Modeling Tumor Phenotypes *In Vitro* with Three-Dimensional Bioprinting” by Langer et al.

2.2 Cell Slice Image Processing

Cell type population abundance data are extracted from images of these cell slices using the Cyclic ImmunoFluorescence Automatic Analysis Pipeline (cycIFAAP) developed by Guillaume Thibault and Young Hwan Chang at the Oregon Health and Science University (OHSU). This pipeline is composed of two main parts: registration/segmentation and feature extraction [17].

The first segment performs registration on the raw images, aligning them to common axes. It then uses a deep learning-based mask R-CNN to segment the individual cell bodies and nuclei. Background subtraction is applied to stabilize the feature counts between different images. After these pre-processing stages, the pipeline tabulates the individual marker levels across four separate locations within each cell (nucleus, rim, ring, and cell body) for all cells segmented earlier. Marker exclusiveness plots are derived from these data (Figure 2.2). Although each marker only has an affinity for one cell type, in practice, many of the cells are double-marked (or even triple-marked in rare cases). As such, the thresholds determined by these exclusiveness are of utmost importance in determining which cell belongs to which type when multiple markers may be present in the same cell.

Part two of the pipeline uses the marker levels and the thresholds calculated in the first part to extract the cell counts for each type. Any cell that exhibits marker levels above the threshold is categorized into that respective marker’s cell type.

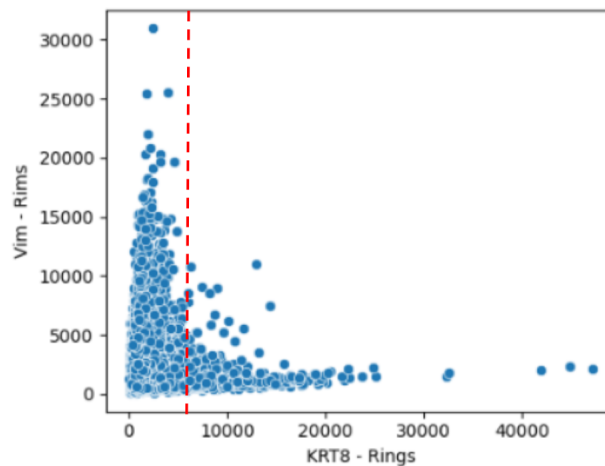


Figure 2.2: A cycIFAAP exclusiveness plot generated using the cycIFAAP software between the Vim marker levels found at the cell rims and the KRT8 marker levels at the cell rings [17]. Here, the KRT8 threshold (red dotted line) is set to 5500. Any cell that exhibits KRT8 levels above 5500 would be classified as a cancer cell.

The resulting pipeline output is a set of cell type counts for individual cell slice images taken on different days of the experiment—time series population data on which the generalized Lotka-Volterra model can be applied.

2.3 Differential Equation-Based Biological Modeling

Data is not particularly useful without a framework to organize and understand the patterns expressed. Mathematical modeling has long been an important way of explaining natural phenomena due largely to its ability to condense a massive collection of experimental data into an abstract framework. These formulaic principles can then, in theory, be interpreted in ways that elucidate aspects of the natural phenomenon that would otherwise have been obscured by the sheer amount of irrelevant details and noise present in the raw data. This type of modeling is thus particularly useful in biological contexts, where hundreds, perhaps even thousands of cells can be abstracted into just a few distinct types based on common properties, and interactions between these cells can similarly be described in terms of relations between these abstract types.

To further capture the temporal aspect of these phenomena, a time-dependent model is needed. Differential equation-based models, by definition, fall into this class. In cases where the general physics of the phenomena is known with respect to the time domain—whether that be planetary motion, predator-prey interactions within an ecosystem, or even economic trends—these equations are particularly useful because they compactly and elegantly

capture the changes of the relevant quantities with respect to time in terms of closed-form mathematical expressions.

Notably, the parameters involved in these expressions often carry true, physical meanings. In equations relating to predator-prey interactions, these parameters may represent carrying capacity, predation interaction, or reproductive rates; in a mechanical system, perhaps radius, spring force, or gravitational force. Importantly, this is not generally true of the variables parameterizing the weight matrices in neural networks. While these neural models often yield extremely accurate predictions, these models are often trained end-to-end as ready-made “black boxes”, meaning the individual weights within the network aren’t specific to the phenomenon and thus offer little insight into the inner workings of the physical phenomenon itself. It is largely because of this property that differential equations-based models are so popular. Qualitative analyses of these differential models yield further phenomenological understanding than traditional experimentation alone and can be thought of as an extension of the statistical analyses usually performed on the experimental data [9].

We will focus on the usage of differential equations to describe physical changes that occur only with respect to time. As such, our models will be governed specifically by *ordinary differential equations*, equations with only one independent variable (time). These can be written in the following form:

$$\frac{d\mathbf{x}(t)}{dt} = \mathbf{f}(\mathbf{x}(t), \mathbf{u}(t), t) \quad \mathbf{x}(t) \in \mathbb{R}^n, \mathbf{u}(t) \in \mathbb{R}^m \quad (2.1)$$

$\mathbf{x}(t)$ is the state vector containing n states; $\mathbf{u}(t)$, a vector containing m distinct inputs. The function \mathbf{f} governs how the state vector $\mathbf{x}(t)$ changes with respect to the current state, the current input, and the time. Note that this function may be nonlinear, making parameters difficult to infer algebraically. However, in the special case where \mathbf{f} is indeed linear with respect to the function parameters, linear algebra techniques like least-squares regression can conveniently be applied to solve for the optimal parameters. When the parameters must be constrained within a certain range, as is often the case with parameters that must respect biological feasibility, solving for these parameters can be reduced to a convex optimization problem with a norm objective and affine constraints.

Linear time-invariant positive dynamical differential equation models have been effectively applied to describing the differentiation-state transitions that allow triple-negative breast cancer cells to evade certain drug treatments (Figure 2.3) [2]. This specific model imposes the Markovian assumption that all transitions within a system can be described as a function of the relative populations of each cell type in the previous time step and the transition probabilities between these states; moreover, the current model, given the state in the previous time step, is completely independent of all other time steps. Clearly, a certain amount of information regarding long-term effects extending past the previous time step is consequently lost, but even this model’s estimation of the dynamical parameters between each of the cell types was enough to reveal that small-molecule targeted therapy strongly affected the differentiation-state transition rates, allowing cancerous environments to resist therapies [8]. A combination of empirical evidence and ensembling was used to derive the

rate parameters in this linear dynamical model [2]. It is worth noting that the Markovian assumption on which this approach is based is not generally true in biological systems, whose states are likely to depend on past states well beyond the time step immediately prior. To capture these long-term dependencies, different classes of differential equation models are necessary.

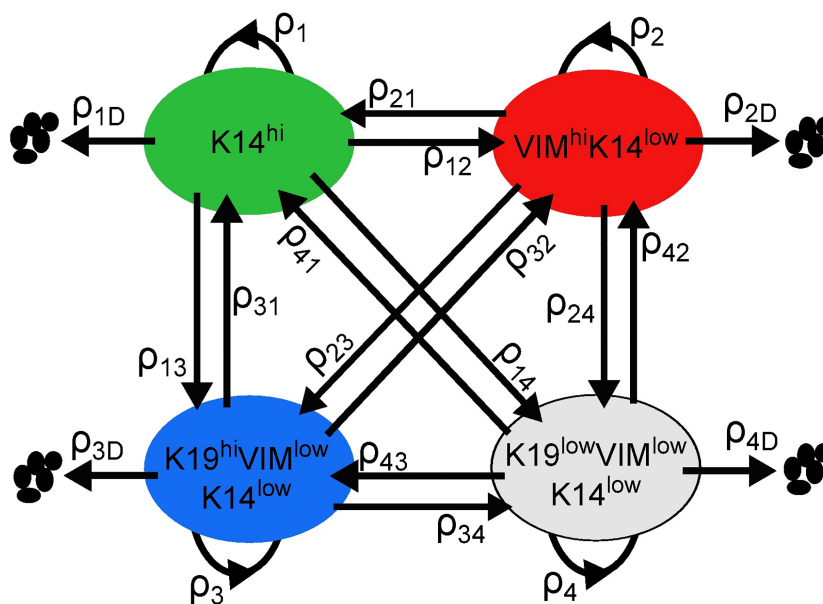


Figure 2.3: A diagram of a finite state machine from Chapman’s “Modeling differentiation-state transitions linked to therapeutic escape in triple-negative breast cancer”. Each of 4 nodes represents a differentiation state. A cell can either transition to a different state, divide, or die. The value ρ represents the probability with which a cell in each of the differentiation states takes the respective action [2].

2.4 The Lotka-Volterra Equations

The Lotka-Volterra equation is one such system of ordinary differential equations that better captures these dependencies. This model has the added advantage that the derivatives are linear with respect to the model parameters, meaning parameter inference can be performed via constrained least-squares algorithms. Originally formulated by Alfred Lotka and Vito Volterra to describe the oscillatory behavior of animal populations, these equations are now best known as a “predator-prey” model that captures the predatory interactions between species within an ecosystem [12]. The generalized form of these relations can be expressed as

$$\frac{dx_i(t)}{dt} = x_i(t) \left(r_i + \sum_{j=1}^n \alpha_{ij} x_j(t) \right) \quad i = 1, \dots, n \quad (2.2)$$

where $x_i(t)$ is the population of species i at time t , r_i is the population's innate growth rate, and α_{ij} is the inter-species ‘‘interaction’’ coefficient. Note that a positive value of α_{ij} indicates that population j positively enhances population i 's growth rate while a negative value similarly signifies a reduction of i 's derivative.

Though this model was primarily used to explain inter-species interactions in an ecosystem on a macroscopic scale, recent research has shown that such these equations can be strong descriptors of cell interactions in the microbiome, an important field of study that connects with key areas of disease, ranging from immune and neurological responses to cancer. In this vein, a study led by Brunner and Chia has found that a generalized Lotka-Volterra model can accurately predict whether certain probiotics can successfully engraft themselves into an individual's microbiome [5]. The motivation for the somewhat unconventional use of this model was the observation that cell type-to-cell type interactions within a microbiome are ecological in nature and function not so differently from a macroscopic ecosystem of different animal species.

As aforementioned, a strength of these differential equation models is their interpretability. In this case, a species' growth rate is completely parameterized by its innate growth rate (a quantity that can actually be measured independent of other populations) and its interaction coefficient with each of the other species in the microbiome. Conversely, and perhaps more importantly, given just population time series data and estimates of their time derivatives, we can use constrained least-squares algorithms to solve for these parameters to gain insight into how different cell types affect each other.

2.5 Lotka-Volterra Representation of Tumor Environments

A particularly interesting and important microbiome is the tumor environment, a heterogeneous mixture of multiple cancer types, fibroblasts, endothelial cells, and epithelial cells among many other cell types. Given this rich variety of cell types, multiple studies have been conducted to analyze the effects of other cell types on the cancer growth rate in hopes of understanding the factors that promote or suppress cancer growth. Much groundwork has been laid in the analysis of competing cancer cell lines in a simulated tumor spheroid environment and the viability of the Lotka-Volterra model as a tool to infer the type of interaction between the lines (competitive, mutualistic, or antagonistic) [10] [14].

It has been shown using simulated data sets that the Lotka-Volterra model is *structurally identifiable*, meaning that a unique set of Lotka-Volterra parameters can be inferred from the data. Perturbing a model parameter will not result in the same model output via compensations from other parameters. The model's further *practical identifiability* implies

that these parameters can be uniquely estimated even in the presence of noise if proper calibration is applied [10] [15].

These properties are encouraging because they imply that any set of parameters optimally inferred from a time-series data set describes the changes unambiguously. In practice, however, inferring such a parameter set from *in vitro* spheroid time series data is extremely difficult. Unlike in simulated data sets, there are no guarantees that the noise in experimental data will follow a certain distribution. The steep cost of running these spheroid experiments *in vitro* leads to a paucity in data, complicating the estimation of derivatives and rendering any noise estimates rudimentary at best. The bulk of this technical report will be dedicated toward developing deterministic generalized Lotka-Volterra parameter inference methods and verifying its robustness in the presence of these difficulties.

2.6 Genetic Algorithmic Approach

William Sharpless from the Lawrence Berkeley National Laboratory (LBNL), now at the University of California, Santa Cruz, has applied an ensembled genetic algorithm to infer the generalized Lotka-Volterra parameters from the bioprinted spheroid data [16]. A genetic algorithm is a stochastic iterative optimization algorithm inspired by Darwinian evolution and biological reproduction. The algorithm is initialized with a set (“population”) of random parameter sets. At each iteration (“generation”), the error function is run on each of the members of the “population”. The top k performing parameter sets are then used to seed the creation of the next generation’s members, forming a new population of parameter sets similar to the best performing k parameter sets from the previous generation. This iterative process continues until convergence or some other predetermined criterion is reached.

For this specific application, the genetic algorithm was run 5 times on each of the cell line experimental data sets in ensemble fashion, and the best-performing parameter set was returned. Although this method could effectively infer parameter sets with very low modeling error, the stochastic nature of this algorithm complicates error bound analysis. This report will attempt to remedy that by utilizing a deterministic, linear-algebraic method with more readily calculable error bounds. The genetic algorithm’s performance will be used as a guiding baseline for the deterministic algorithm’s error bound analysis.

The genetic algorithm was only applied to the two-population spheroid data (fibroblast and cancer cells). We will extend this model to three-population context by applying the generalized Lotka-Volterra equations to bioprinted spheroid data containing cell counts for fibroblast, cancer, and endothelial cells. The same genetic algorithm will again be used to produce baseline error rates.

2.7 Reference Formulas

This section serves as a reference for many of the formulas that will be commonly used throughout the report.

$$\frac{dx_i(t)}{dt} = x_i(t) \left(r_i + \sum_{j=1}^n \alpha_{ij} x_j(t) \right) \quad i = 1, \dots, n$$

Figure 2.4: Generalized Lotka-Volterra (gLV) equations for n populations.

This report will focus entirely on the two-population and three-populations versions of the gLV equations. These are listed here. Note that $x_i(t)$ is expressed as x_i because time dependence is implied:

$$\begin{aligned} \frac{dx_1}{dt} &= r_1 x_1 + \alpha_{11} x_1^2 + \alpha_{12} x_1 x_2 \\ \frac{dx_2}{dt} &= r_2 x_2 + \alpha_{12} x_1 x_2 + \alpha_{22} x_2^2 \end{aligned}$$

Figure 2.5: Generalized Lotka-Volterra (gLV) equations for 2 populations.

$$\begin{aligned} \frac{dx_1}{dt} &= r_1 x_1 + \alpha_{11} x_1^2 + \alpha_{12} x_1 x_2 + \alpha_{13} x_1 x_3 \\ \frac{dx_2}{dt} &= r_2 x_2 + \alpha_{21} x_1 x_2 + \alpha_{22} x_2^2 + \alpha_{23} x_2 x_3 \\ \frac{dx_3}{dt} &= r_3 x_3 + \alpha_{31} x_1 x_3 + \alpha_{32} x_2 x_3 + \alpha_{33} x_3^2 \end{aligned}$$

Figure 2.6: Generalized Lotka-Volterra (gLV) equations for 3 populations.

When performing convex optimization on the two-population model, we will use one of the two following matrix formulations. The first matrix equation corresponds exactly to the above equation expressed as a matrix product. This is known as the *integral formulation*. The second corresponds to a *logarithmic integral formulation* of the gLV equations:

$$\begin{bmatrix}
 x_1^{(1)} & 0 & x_1^{(1)}x_1^{(1)} & x_1^{(1)}x_2^{(1)} & 0 & 0 \\
 0 & x_2^{(1)} & 0 & 0 & x_1^{(1)}x_2^{(1)} & x_2^{(1)}x_2^{(1)} \\
 x_1^{(2)} & 0 & x_1^{(2)}x_1^{(2)} & x_1^{(2)}x_2^{(2)} & 0 & 0 \\
 0 & x_2^{(2)} & 0 & 0 & x_1^{(2)}x_2^{(2)} & x_2^{(2)}x_2^{(2)} \\
 & & & \vdots & & \\
 x_1^{(T)} & 0 & x_1^{(T)}x_1^{(T)} & x_1^{(T)}x_2^{(T)} & 0 & 0 \\
 0 & x_2^{(T)} & 0 & 0 & x_1^{(T)}x_2^{(T)} & x_2^{(T)}x_2^{(T)}
 \end{bmatrix}
 \begin{bmatrix}
 r_1 \\
 r_2 \\
 \alpha_{11} \\
 \alpha_{12} \\
 \alpha_{21} \\
 \alpha_{22}
 \end{bmatrix}
 =
 \begin{bmatrix}
 \frac{dx_1}{dt} \Big|_{t=t_1} \\
 \frac{dx_2}{dt} \Big|_{t=t_1} \\
 \frac{dx_1}{dt} \Big|_{t=t_2} \\
 \frac{dx_2}{dt} \Big|_{t=t_2} \\
 \vdots \\
 \frac{dx_1}{dt} \Big|_{t=t_T} \\
 \frac{dx_2}{dt} \Big|_{t=t_T}
 \end{bmatrix}
 \quad (2.3)$$

$$x_i^{(j)} := \text{value of population } i \text{ at time step } j \quad (2.4)$$

Figure 2.7: Integral formulation of 2 population generalized Lotka-Volterra equations.

$$\begin{bmatrix}
 1 & 0 & x_1^{(1)} & x_2^{(1)} & 0 & 0 \\
 0 & 1 & 0 & 0 & x_1^{(1)} & x_2^{(1)} \\
 1 & 0 & x_1^{(2)} & x_2^{(2)} & 0 & 0 \\
 0 & 1 & 0 & 0 & x_1^{(2)} & x_2^{(2)} \\
 & & & \vdots & & \\
 1 & 0 & x_1^{(T)} & x_2^{(T)} & 0 & 0 \\
 0 & 1 & 0 & 0 & x_1^{(T)} & x_2^{(T)}
 \end{bmatrix}
 \begin{bmatrix}
 r_1 \\
 r_2 \\
 \alpha_{11} \\
 \alpha_{12} \\
 \alpha_{21} \\
 \alpha_{22}
 \end{bmatrix}
 =
 \begin{bmatrix}
 \frac{d}{dt} \ln(x_1) \Big|_{t=t_1} \\
 \frac{d}{dt} \ln(x_2) \Big|_{t=t_1} \\
 \frac{d}{dt} \ln(x_1) \Big|_{t=t_2} \\
 \frac{d}{dt} \ln(x_2) \Big|_{t=t_2} \\
 \vdots \\
 \frac{d}{dt} \ln(x_1) \Big|_{t=t_T} \\
 \frac{d}{dt} \ln(x_2) \Big|_{t=t_T}
 \end{bmatrix}
 \quad (2.5)$$

$$x_i^{(j)} := \text{value of population } i \text{ at time step } j \quad (2.6)$$

Figure 2.8: Logarithmic integral formulation of 2 population generalized Lotka-Volterra equations.

Chapter 3

Parameter Inference

3.1 Problem Formulation

Overview

Given experimental time-series population count data, find a set of biologically feasible generalized Lotka-Volterra parameter values best describing it according to some predefined error metric (Figure 2.5). The methods used to infer these parameters from the experimental data should be deterministic and, ideally, robust to experimental noise. We will run an ensemble genetic algorithm on the same experimental data to obtain a baseline parameter set characterizing a baseline error to which to compare the errors produced by the deterministic parameters.

Population Data

We will be operating on experimental data from 4 cancer cell lines: SKBR3, MDAMB231, MCF7, and HCC1143. Each cancer cell line constitutes a single experiment. Parameter inference will be performed on each of these experiments separately as each cell line is expected to exhibit different ecological interactions.

As described in the background section, each experiment begins with a number of identical tumor spheroids with the specific line of cancer cells at its core. On days 4, 7, and 10 of the experiment, a subset of these tumor spheroids is fixed, embedded, and sectioned; these sections are then stained, imaged, and processed. These subsets must be non-overlapping because the spheroid cells die during this fixation process, so the same spheroid can't be opened at multiple different time points. As such, the data for each cell line is in the following form: on each of days 4, 7, 10, there are x_i counts of each of the cell types (fibroblast and cancer for the two-population analysis; fibroblast, cancer, and endothelial for the three-population) where x_i corresponds to the number of spheroids processed on day i .

Performance Metric

To measure how well a given parameter set describes the experimental data, we first numerically solve the system of generalized Lotka-Volterra equations defined by these parameters using the SciPy differential equation solver. The initial condition is defined to be the average of the cell type abundances across all cell slice images on day 4. Time points 4, 7, and 10 are then passed into the solution to the system of differential equations to obtain predicted fibroblast and cancer cell counts on each of the 3 days. For each day, a mean squared error is calculated by finding the average squared deviation between the predicted cell count for that specific cell type and the cell counts for each of the spheroids of the same cell type on the same day. The total error is the sum of all these mean squared errors across all cell types. The following algorithm summarizes the error calculation:

Figure 3.1 Error Function

```

1: procedure ERRORFUNCTION(MODELDATA, EXPDATA)
2:   totalError  $\leftarrow$  0
3:   cellTypes  $\leftarrow$  [Fibroblast, Cancer] ([Fibroblast, Cancer, Endothelial])
4:   days  $\leftarrow$  [4, 7, 10]
5:   for cellType c in cellTypes do
6:     cellTypeError  $\leftarrow$  0
7:     for day d in days do
8:       dayError  $\leftarrow$  0
9:       for image i in images of cellType c on day d do
10:        dayError  $\leftarrow$  dayError + (modelData[c][d] - expData[c][d][i])2
11:       cellTypeError  $\leftarrow$  cellTypeError + dayError / (# images of c on d)
12:   totalError  $\leftarrow$  totalError + cellTypeError
   return totalError

```

The goal of the inference method is to find a set of parameter values that yields solutions to the generalized Lotka-Volterra equations minimizing this error function.

3.2 Baseline: Genetic Algorithm

Overview

We will apply genetic optimization five times to the raw experimental data for each of the four cell lines (SKBR3, MCF7, HCC1143, MDAMB231). Due to the algorithm's stochastic nature, the parameter sets returned at the completion of each run should be different and will likely emit differing errors. The parameter set with the lowest error will be chosen as the representative baseline to which the parameters returned by the linear algebraic deterministic approach described in the next section will be compared.

Details

We will be using bluepyopt’s DeapOptimisation module to run these optimization. This package implements the genetic algorithm and allows the user to determine which values to use for each algorithm parameter.

The error function defined in Figure 3.1 will be used as a *fitness* measure for each of the generated gLV parameter sets (individuals) within the population for each iteration. Intuitively, individuals with a high *fitness* value relative to other members of the population will produce more similar offspring individuals in the next iteration. In other words, the lower the gLV parameter set’s error, the more likely it will be that the next iteration’s parameter sets will be similar to this one.

This gLV parameter values in this algorithm will respect the following biological feasibility constraints:

- $a_{ii} < 0$ - Cells of the same type compete with each other for the same resources, so the presence of other similar cells decreases the proliferation rate. All self-interactions are negative.
- $|a_{ij}| < 20$ - We assume that the interaction parameters aren’t massive. This constraint ensures that the resulting differential equations are not too stiff. Note that the upper bound of 20 is somewhat arbitrary and is a debated subject among experts.
- $r_i > 0$ - The innate growth rate is positive because cells of the same type should reproduce exponentially in the absence of influence from other populations, including self-interactions.

All individual parameter sets within the population will only contain parameters that satisfy these feasibility constraints.

3.3 Deterministic Approach

Overview

The deterministic approach to the problem will be two-pronged. The first method will revolve around optimizing the pure matrix formulation of the generalized Lotka-Volterra model (Figure 2.7). The feature matrix on the left hand side can be constructed directly from the experimental population counts. The derivative vector on the right hand side must be estimated. From this equation, we will take the $L2$ norm of the difference of the left and right-hand side of the equation to be the objective function. A convex optimizer is then used to minimize this objective over all values of the parameter vector within biological constraints.

The second approach uses the logarithmic integral formulation of the 2-population generalized Lotka-Volterra equations (Figure 2.8). All steps will be the same except for the

derivative vector estimation. Rather than directly estimating the derivatives from the data, the derivatives of the *logarithms* of the population counts will be calculated. This approach reduces the complexity of the feature matrix A and should therefore, in theory, be less sensitive to perturbations in the data [13]. We will verify this experimentally.

Approach 1: Integral Formulation

Populating the Feature Matrix and Derivative Vector

The feature matrix in the integral formulation is constructed entirely of values $x_i^{(j)}$, each representing the cell count for population i at time step j . The raw experimental data actually contains several values for $x_i^{(j)}$, one for each of the spheroid slice images taken on that day of the experiment. The average of the all these values is taken to be the representative for $x_i^{(j)}$, and it is this value that is used to populate the feature matrix:

$$x_i^{(j)} := \frac{1}{n} \sum_{k=0}^n x_i^{(j)}[k] \quad (3.1)$$

where n is the number of images of population i on day j , and $x_i^{(j)}[k]$ is the cell count of that specific image.

The derivative vector is much more difficult to construct because the paucity of data complicates accurate estimation for the slopes at each time point. The trapezoid method is the most common approximation for this problem, though interpolation techniques have also been used. Interpolation involves fitting a function through the data points (polynomial, spline, etc.), then taking the derivative of this interpolated function at the specified points. In separate trials, we will apply both trapezoid approximation and polynomial interpolation to construct this derivative vector, and their respective errors will be compared. For polynomial interpolation, we will experiment with polynomials of degrees 1 to 4 to see which interpolation yields the lowest modeling error.

Aside: Polynomial Interpolation for Derivative Estimation

Polynomial interpolation is the process of fitting a polynomial curve through a set of data points. Once the polynomial expression is determined, a closed-form expression for the derivative can be calculated by symbolically differentiating the expression using the power rule.

To fit a polynomial curve of degree n through m points, we first formulate it as a least squares regressions problem with the exponentiated x values as the features and the y values as the \mathbf{b} vector:

$$\underbrace{\begin{bmatrix} 1 & x_0 & x_0^2 & x_0^3 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & x_1^3 & \dots & x_1^n \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_m & x_m^2 & x_m^3 & \dots & x_m^n \end{bmatrix}}_{\mathbf{A}} \underbrace{\begin{bmatrix} p_0 \\ p_1 \\ \vdots \\ p_n \end{bmatrix}}_{\mathbf{p}} = \underbrace{\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}}_{\mathbf{b}} \quad (3.2)$$

For $m \geq n$ i.e. the system is not underdetermined, we can solve this problem for the parameter vector by optimizing the following objective:

$$\min_p \|Ap - b\|_2 \quad (3.3)$$

which yields the solution:

$$p^* = (A^T A)^{-1} A^T b \quad (3.4)$$

If $m < n$, then the system is underdetermined, meaning that there are infinitely many p values that will satisfy the equality. We would like to choose the parameters with the smallest L2 norm, which is reduced to the following minimum L2 norm optimization problem:

$$\min_p \|p\|_2^2 \quad s.t. \quad Ap = b \quad (3.5)$$

The solution to this optimization is the following:

$$p^* = A^T (A A^T)^{-1} b \quad (3.6)$$

Once the optimal parameter set p^* has been obtained via equation 3.4 or 3.6, the derivative at any point on the curve can be calculated using the power rule:

$$\underbrace{\begin{bmatrix} \frac{dy}{dx} \Big|_{x=x_0} \\ \frac{dy}{dx} \Big|_{x=x_1} \\ \vdots \\ \frac{dy}{dx} \Big|_{x=x_m} \end{bmatrix}}_{\mathbf{v}} = \begin{bmatrix} 1 & x_0 & x_0^2 & \dots & x_0^{n-1} \\ 1 & x_1 & x_1^2 & \dots & x_1^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_m & x_m^2 & \dots & x_m^{n-1} \end{bmatrix} \begin{bmatrix} p_1 \\ 2p_2 \\ \vdots \\ np_n \end{bmatrix} \quad (3.7)$$

In the context of the generalized Lotka-Volterra model, the derivative vector v is what will be used on the right hand side of the matrix equation in figure 2.7.

Graphically, this process looks like the following. Consider this toy example with four points:

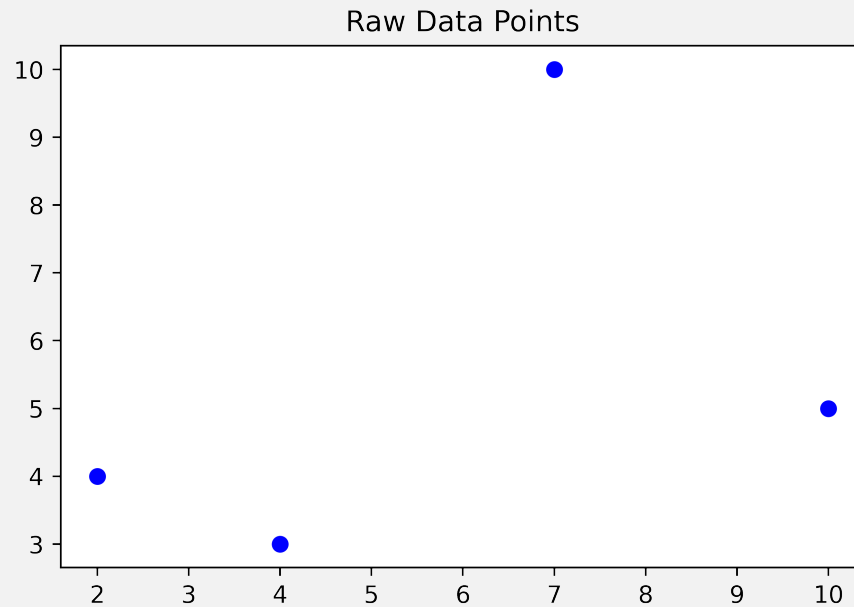
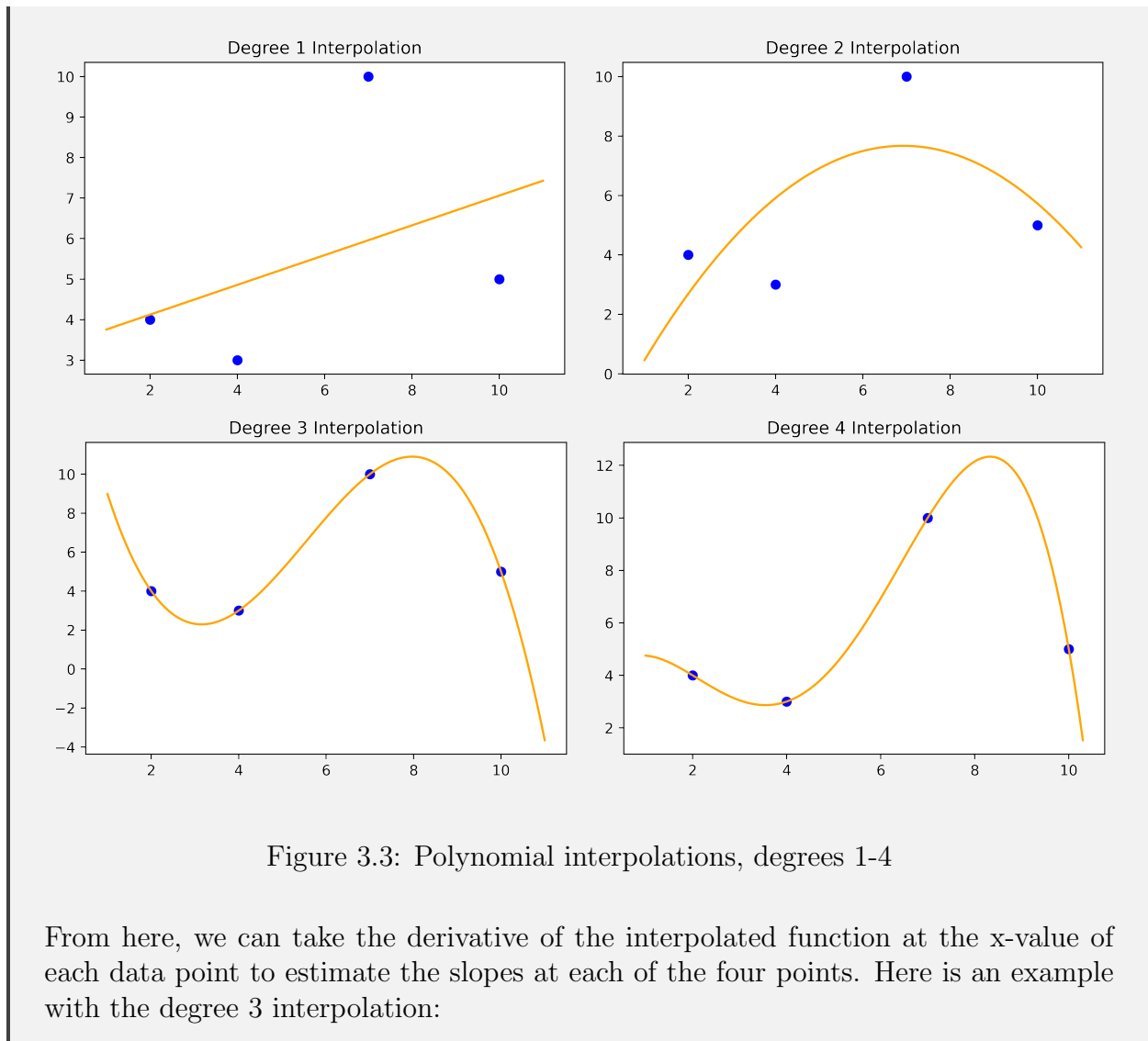


Figure 3.2: Raw data points: $(2, 4)$, $(4, 3)$, $(7, 10)$, $(10, 5)$

We first interpolate a polynomial function through these functions using equation 3.4 if $n < 4$ or equation 3.6 otherwise (underdetermined case). Below are polynomial interpolations of varying degrees:



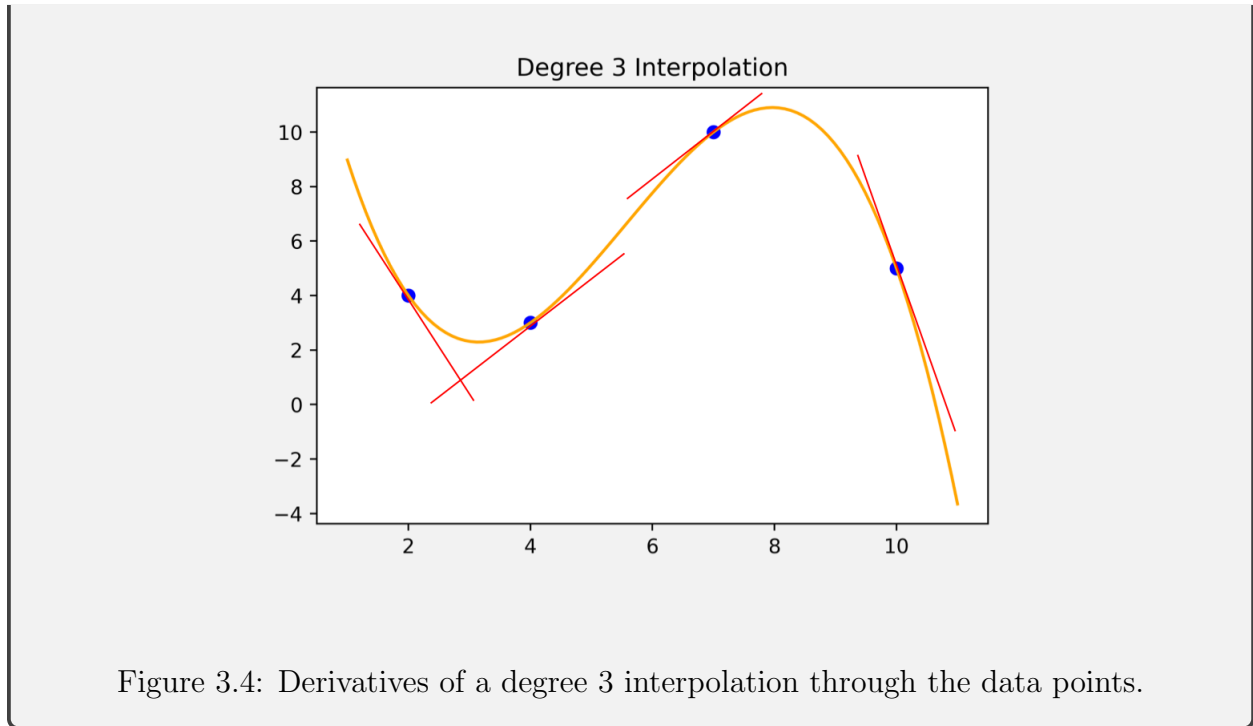


Figure 3.4: Derivatives of a degree 3 interpolation through the data points.

Optimizing for the Parameters

Once the feature matrix and derivative vector have been populated, recall that we obtain the following equation (2.7):

$$\underbrace{\begin{bmatrix} x_1^{(1)} & 0 & x_1^{(1)}x_1^{(1)} & x_1^{(1)}x_2^{(1)} & 0 & 0 \\ 0 & x_2^{(1)} & 0 & 0 & x_1^{(1)}x_2^{(1)} & x_2^{(1)}x_2^{(1)} \\ x_1^{(2)} & 0 & x_1^{(2)}x_1^{(2)} & x_1^{(2)}x_2^{(2)} & 0 & 0 \\ 0 & x_2^{(2)} & 0 & 0 & x_1^{(2)}x_2^{(2)} & x_2^{(2)}x_2^{(2)} \\ \vdots & & & & & \\ x_1^{(T)} & 0 & x_1^{(T)}x_1^{(T)} & x_1^{(T)}x_2^{(T)} & 0 & 0 \\ 0 & x_2^{(T)} & 0 & 0 & x_1^{(T)}x_2^{(T)} & x_2^{(T)}x_2^{(T)} \end{bmatrix}}_A \underbrace{\begin{bmatrix} r_1 \\ r_2 \\ \alpha_{11} \\ \alpha_{12} \\ \alpha_{21} \\ \alpha_{22} \end{bmatrix}}_p = \underbrace{\begin{bmatrix} \frac{dx_1}{dt} \Big|_{t=t_1} \\ \frac{dx_2}{dt} \Big|_{t=t_1} \\ \frac{dx_1}{dt} \Big|_{t=t_2} \\ \frac{dx_2}{dt} \Big|_{t=t_2} \\ \vdots \\ \frac{dx_1}{dt} \Big|_{t=t_T} \\ \frac{dx_2}{dt} \Big|_{t=t_T} \end{bmatrix}}_b \quad (3.8)$$

$$x_i^{(j)} := \text{value of population } i \text{ at time step } j \quad (3.9)$$

where A is defined to be the feature matrix; p , the gLV parameters; and b , the derivative vector. Since this equation is overdetermined, the parameter vector p that minimizes the L2 discrepancy between the two sides of the equation equates to the solution to the following optimization problem:

$$\min_p \|Ap - b\|_2 \quad (3.10)$$

This problem has a closed-form least-squares solution. Unfortunately the least-squares solution will not generally satisfy the biological feasibility constraints previously mentioned in the baseline algorithm section (innate growth rate must be positive, all self-interactions are negative, the magnitude of each interaction coefficient cannot be too high). Those same constraints will be applied here, resulting in a constrained least-squares formulation similar to the one above:

$$\begin{aligned} \min_p \quad & \|Ap - b\|_2 \\ \text{s.t.} \quad & r_i > 0 \text{ for } i = 1, 2 \\ & a_{ii} < 0 \text{ for } i = 1, 2 \\ & -20 < a_{ij} < 20 \text{ for } i, j = 1, 2 \end{aligned} \quad (3.11)$$

This formulation has a norm objective and affine constraints, meaning that it is a convex problem that can be solved using standard convex optimization methods. We use `cvxpy`, which applies the *interior point method* to solve for the parameter vector p .

Evaluation

We apply the Error Function (3.1) to the resulting parameter vector p , yielding an intuitive squared deviation between this parameter set's model predictions and the experimental population count values. This value will be used to compare this parameter set's performance to those obtained via the baseline genetic algorithm.

Approach 2: Logarithmic Integral Formulation

Derivation

The logarithmic integral formulation of the generalized Lotka-Volterra equations can be derived from the regular integral formulation. We start with the general equations for n populations (Figure 2.4). Note that $x_i(t)$ is represented as x_i because time dependence is implied:

$$\frac{dx_i}{dt} = x_i \left(r_i + \sum_{j=1}^n \alpha_{ij} x_j \right) \quad i = 1, \dots, n \quad (3.12)$$

$$\frac{1}{x_i} \frac{dx_i}{dt} = \left(r_i + \sum_{j=1}^n \alpha_{ij} x_j \right) \quad i = 1, \dots, n \quad (3.13)$$

Using the following derivative of a natural logarithm:

$$\frac{d}{dt} \ln(x) = \frac{1}{x} \frac{dx}{dt} \quad (3.14)$$

We can represent the left-hand side of equation 3.12 as the derivative of a logarithm:

$$\frac{d}{dt} \ln(x_i) = \left(r_i + \sum_{j=1}^n \alpha_{ij} x_j \right) \quad i = 1, \dots, n \quad (3.15)$$

Equation 3.15 is the basis of the generalized Lotka-Volterra model's logarithmic integral formulation. Notice that the derivative now represents the change in the *logarithm* of a population count x_i . More importantly, the value of this derivative (the right-hand side of equation 3.15) now no longer contains quadratic terms with respect to the populations values x_i . This latter property may have stability implications that will be experimentally explored later in this report.

Populating the Feature Matrix and Derivative Vector

The same raw experimental data used in the integral formulation approach will also be used for the logarithmic integral formulation optimization. Recall that the matrix equation for the logarithmic integral formulation is the following (Figure 2.8):

$$\underbrace{\begin{bmatrix} 1 & 0 & x_1^{(1)} & x_2^{(1)} & 0 & 0 \\ 0 & 1 & 0 & 0 & x_1^{(1)} & x_2^{(1)} \\ 1 & 0 & x_1^{(2)} & x_2^{(2)} & 0 & 0 \\ 0 & 1 & 0 & 0 & x_1^{(2)} & x_2^{(2)} \\ & & & \vdots & & \\ 1 & 0 & x_1^{(T)} & x_2^{(T)} & 0 & 0 \\ 0 & 1 & 0 & 0 & x_1^{(T)} & x_2^{(T)} \end{bmatrix}}_{A_l} \underbrace{\begin{bmatrix} r_1 \\ r_2 \\ \alpha_{11} \\ \alpha_{12} \\ \alpha_{21} \\ \alpha_{22} \end{bmatrix}}_{p_i} = \underbrace{\begin{bmatrix} \frac{d}{dt} \ln(x_1)|_{t=t_1} \\ \frac{d}{dt} \ln(x_2)|_{t=t_1} \\ \frac{d}{dt} \ln(x_1)|_{t=t_2} \\ \frac{d}{dt} \ln(x_2)|_{t=t_2} \\ \vdots \\ \frac{d}{dt} \ln(x_1)|_{t=t_T} \\ \frac{d}{dt} \ln(x_2)|_{t=t_T} \end{bmatrix}}_{b_l} \quad (3.16)$$

$$x_i^{(j)} := \text{value of population } i \text{ at time step } j \quad (3.17)$$

The feature matrix from A_l will be populated using the same average values as the integral formulation. To calculate the derivative vector, the natural logarithm will be applied to each of the population values x_i before averaging and interpolation will be used to estimate the slopes at each point as before.

Optimizing for Parameters

The optimization objective and constraints are identical to that of the integral formulation except for the fact that the feature matrix and derivative vector now come from the

logarithmic integral formulation:

$$\begin{aligned} \min_{p_l} \quad & \|A_l p_l - b_l\|_2 \\ \text{s.t.} \quad & r_i > 0 \text{ for } i = 1, 2 \\ & a_{ii} < 0 \text{ for } i = 1, 2 \\ & -20 < a_{ij} < 20 \text{ for } i, j = 1, 2 \end{aligned} \tag{3.18}$$

Cvxpy will be applied to solve for the optimal p_l^* .

Chapter 4

Experimental Evaluation

4.1 Two-Population Analysis

Baseline: Genetic Algorithm

Five runs of the genetic algorithm were applied to the experimental data for each cell line to establish a baseline error to which to compare the results of the linear algebra deterministic method. The specific parameter values used in the genetic algorithm for the runs are specified in Table 4.1. Identical parameters were used for all four cell lines.

Parameter	Value
Generations	100
# Offspring	500
Mutation Probability	100 %
Crossover Probability	100 %

Table 4.1: Genetic algorithm parameters for two-population analysis. The generation number is the number of iterations for which to run the algorithm. The offspring number is the population size at each step. The mutation and crossover probabilities define how often an individual changes its individual parameters according to some stochastic process between iterations and the frequency with which it gets its parameters stochastically affected by other members of the population, respectively [3].

The error for each iteration of the algorithm is plotted with respect to the generation number for each cell type in Figures A.1 and A.2. Within each cell line, the errors for each run converged to the same value, indicating that, for this specific set of data, the genetic algorithm will consistently output sets of gLV parameters with nearly identical errors. Note that this does not mean that the gLV parameters themselves will be identical; the values of these parameters do vary, but not nearly enough to change the sign of the parameter

value. Thus, the types of interactions this algorithm infers is identical across all runs since the interaction type between two populations depends only on the sign of the gLV coefficient describing that interaction (Table 4.2). This observation also corroborates the generalized Lotka-Volterra model’s practical identifiability even with the presence of noise in the data.

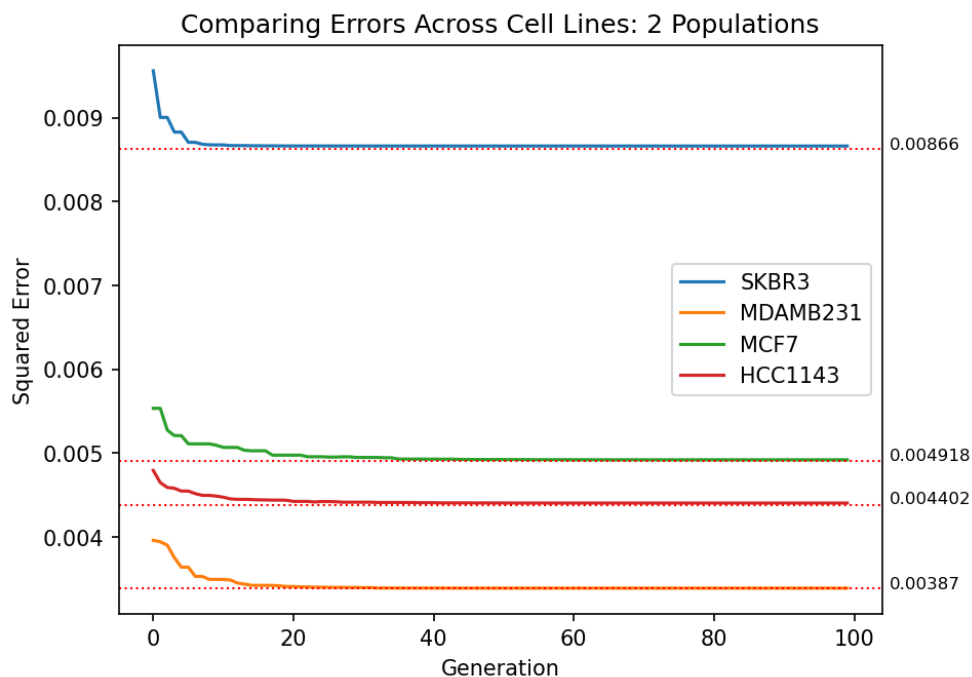


Figure 4.1: An error comparison between runs from each of the cell lines for the two-population case. Note the different convergence values.

Figure 4.1 compares the error profiles across different cell lines. Note that the convergence values are distinct by cell type. The SKBR3 cell line has the largest model error followed by MCF7, HCC1143, and MDAMB231, in that order. Since all errors were normalized by the number of experimental data points, this disparity likely represents the upper limits of how well this method can describe the particular set of experimental data. When we perform gLV parameter inference using the deterministic method in the next section, we will be comparing the *relative* errors between the cell lines to identify whether this same trend is reflected.

Deterministic Algorithm

Integral Formulation

The linear algebra optimization algorithm was applied to the same data to infer the same generalized Lotka-Volterra parameters. While the algorithm as previously introduced framed this as a pure constrained least-squares problem, in practice, an L_2 normalization is placed on

Cell Line	Cancer's effect on Fibroblast	Fibroblast's effect on Cancer
SKBR3	-	-
MDAMB231	+	+
MCF7	-	+
HCC1143	+	+

Table 4.2: Inter-population interactions for two populations as predicted from the genetic algorithm. An interaction between population A and B indicated by a negative sign indicates that population A inhibits the growth of population B. A positive sign indicates the opposite. Cancer has a negative effect on the fibroblasts in two of these lines while the presence of fibroblasts promotes the growth of cancer cells in three of the cell types.

the generalized Lotka-Volterra parameters and added to the objective function because the feature matrix A is underdetermined, yielding multiple minima. This leads to the following formulation:

$$\begin{aligned}
\min_p \quad & \|Ap - b\|_2 + \rho\|p\|_2 \\
\text{s.t.} \quad & r_i > 0 \text{ for } i = 1, 2 \\
& a_{ii} < 0 \text{ for } i = 1, 2 \\
& -20 < a_{ij} < 20 \text{ for } i, j = 1, 2
\end{aligned} \tag{4.1}$$

where ρ is a non-negative constant. For this specific run, ρ was set to 10^{-7} because it led to the least overall error when compared to experimental data. The appendix contains a record of the errors attained for each cell type using this optimization method compared to the respective error from the genetic algorithm. Importantly, the relative ordering of the errors for each cell type is the same as in the genetic algorithm case: that is, SKBR3, MDAMB231, MCF7, and HCC1143, in that order from greatest to lowest error. This strongly suggests that the relative ordering of the error is caused by noise inherent to the data from each cell line and that both the above inference methods uncover it.

Table 4.3 summarizes the interactions predicted by this optimization. With the exception of the predicted effect of cancer on fibroblast cells in the SKBR3 cell line, which changes from a negative to a positive interaction, it is encouraging to see that the interactions predicted by the deterministic optimization is nearly identical to the types predicted by the stochastic genetic algorithm (Table 4.2), once again implying the generalized Lotka-Volterra model's practical identifiability and the deterministic algorithm's viability as an inference method even for noisy experimental data.

Cell Line	Cancer's effect on Fibroblast	Fibroblast's effect on Cancer
SKBR3	+	-
MDAMB231	+	+
MCF7	-	+
HCC1143	+	+

Table 4.3: Inter-population interactions for two populations as predicted by the integral formulation of the deterministic algorithm. The logarithmic formulation yielded the same results.

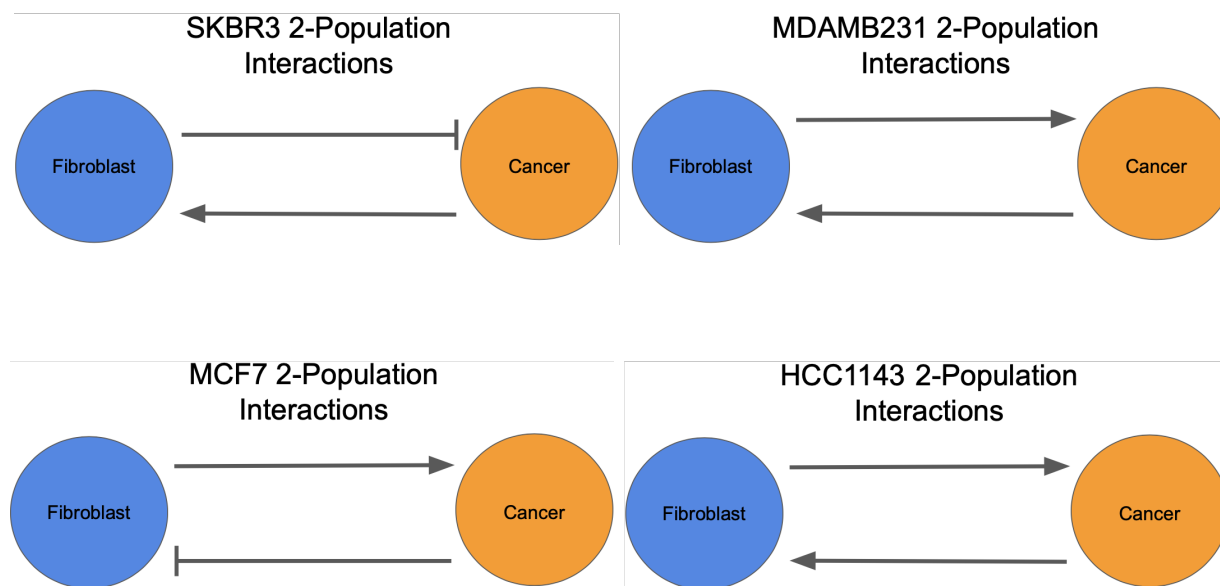


Figure 4.2: 2-population Network Interaction Graphs. An arrow from node A to node B indicates a positive interaction. A closed line from A to B indicates an inhibitory interaction.

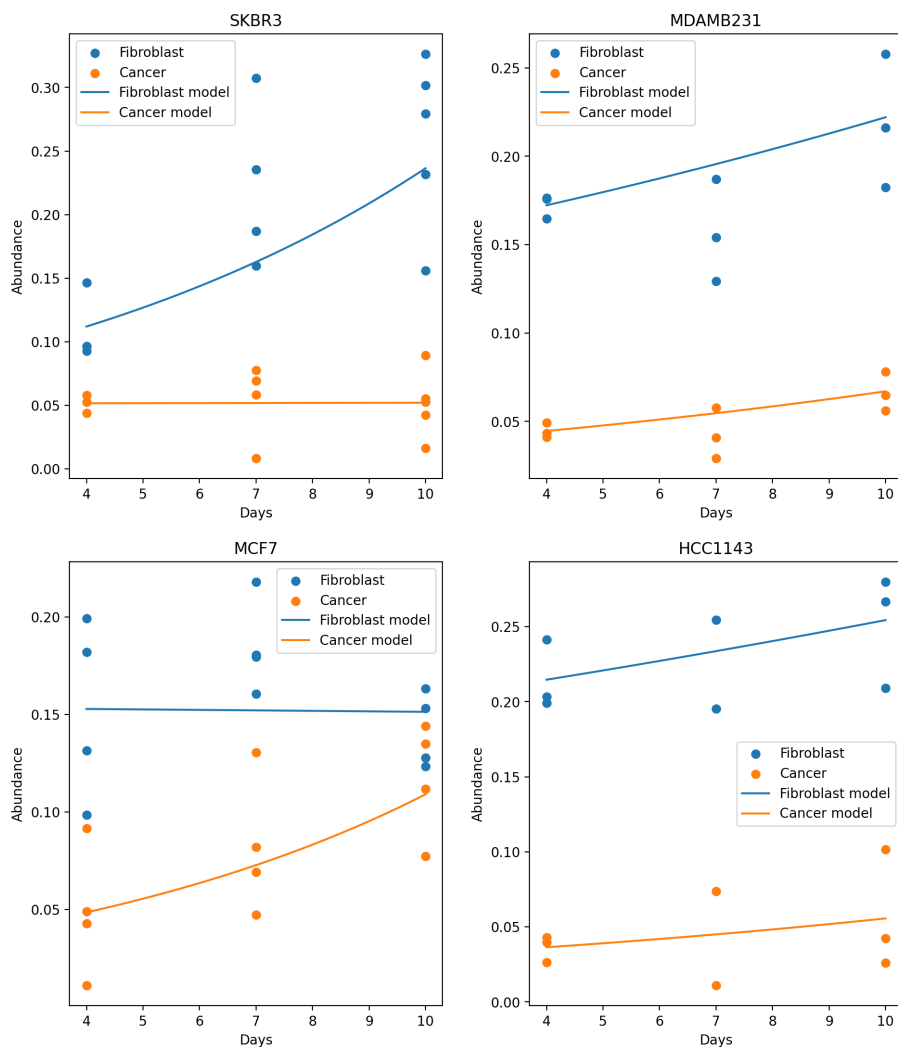


Figure 4.3: A comparison between the experimental data and the trend predicted by optimizing the generalized Lotka-Volterra model on two-population data using the **integral formulation**.

Logarithmic Integral Formulation

The results from the logarithmic integral formulation of the generalized Lotka-Volterra optimization problem were nearly identical to those inferred from the simple integral formulation. The interactions this variant of the algorithm infers is identical to the ones predicted by the integral formulation in Table 4.3. The trend graphs are also quite similar to those from the integral formulation with negligible differences (Figure 4.4).

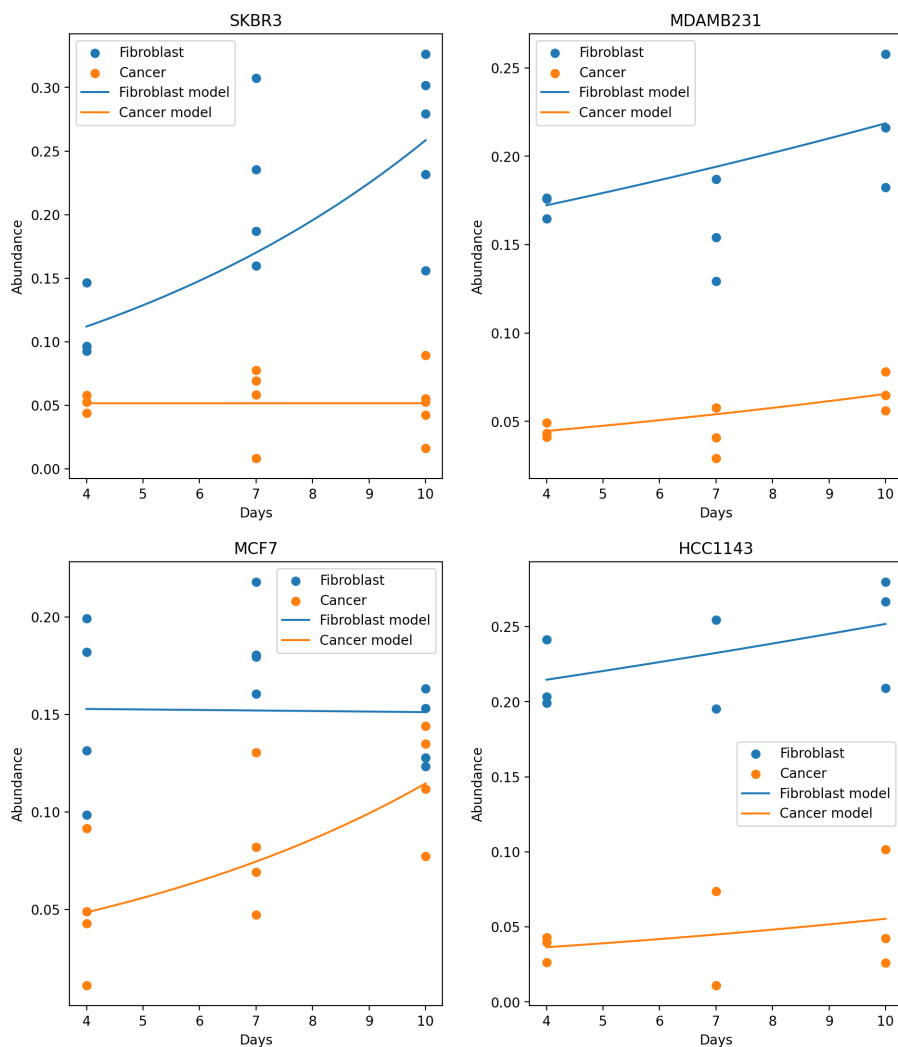


Figure 4.4: A comparison between the experimental data and the trend predicted by optimizing the generalized Lotka-Volterra model on two-population data using the **logarithmic integral formulation**.

4.2 Three-Population Analysis

Baseline: Genetic Algorithm

A nearly identical analysis was performed on the three-population experimental (fibroblast, cancer, and endothelial). Not all time points had identical numbers of experimental abundance data points. The MCF7 line, for example, only had one data point for day 10 while day 4 had four points. The lack of data at these time points for some of the cell lines reduces

Parameter	Value
Generations	100
# Offspring	3000
Mutation Probability	100 %
Crossover Probability	100 %

Table 4.4: Genetic algorithm parameters for three populations. The only notable difference between this configuration and that of the two-population model is the six-fold increase in offspring size to stabilize error convergence.

the confidence in which we place on the models optimized on these data points and should be the subject of further study. Again, five runs of the genetic algorithm was applied to the experimental data for each cell line. Identical parameter sets were used for each line, the specifics of which are recorded in Table 4.4. The only configuration difference between that specified for the two-population model is the increased offspring size. Intuitively, the three-population model is more complex and requires a more fine-grained offspring set to capture any minimum directions. Increasing the offspring size raises the probability that some individual in the population pool will emit an error better than that of the best-performing individual from the previous generation, leading to more consistent convergence.

The error plots are very similar to those in the two-population analysis, though convergence is slightly less smooth, as expected. Again, the convergence value for each cell line is different. Interestingly, with the exception of the HCC1143 cell line, the relative ordering of the other cell lines by convergence error is analogous to the two-population case: MDAMB231 has the lowest error, and SKBR3 has the highest error with MCF7 in between. The HCC1143 exception may likely have been caused by the small number of tissue samples collected on day 10 for this cell line, leading to higher error rates.

Table 4.4 summarizes the interactions between the fibroblast, cancer, and endothelial cells for each cell line. Results were very mixed. The only strong trend was the fibroblast cell population’s inhibitory effect on endothelial cell growth. Slightly weaker patterns include the cancer population’s positive effect on the fibroblast population and the endothelial population’s inhibitory effect on cancer growth.

Deterministic Algorithm

Integral Formulation

The same constrained least-squares optimization with regularization was performed on the three-population experimental data. The interactions predicted by the deterministic algorithm recorded in Table 4.6 agree very little with those inferred by the baseline genetic algorithm, likely due to the increased data and model complexity imposed with the addition

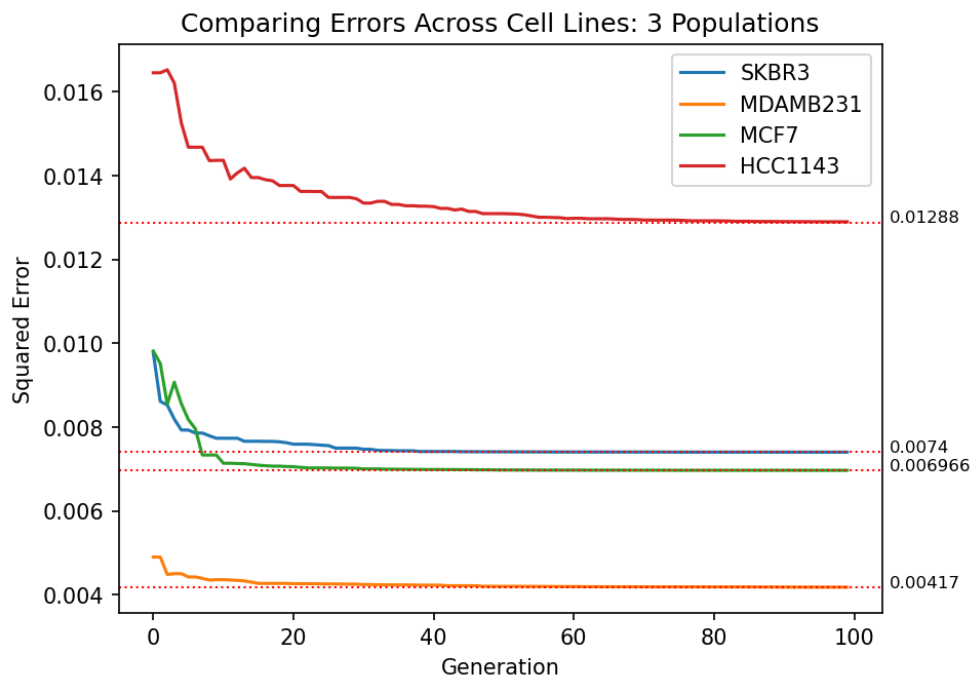


Figure 4.5: An error comparison between runs from each of the cell lines for the three-population case. Note the different convergence values.

of the third endothelial population. The deterministic algorithm actually reveals several much stronger trends in the data. Across all four cell lines, cancer actually benefits fibroblast growth while suppressing endothelial cell proliferation. Endothelial cells consistently promote the fibroblast population's growth, and fibroblast proliferation decreases endothelial cell growth. It is worth noting that the same interaction type that the baseline genetic algorithm between cancer and fibroblast cells and the interaction between fibroblast and endothelial cells was also reflected here in the deterministic inference, but as a much stronger trend.

Cell Line	Cancer's effect on Fibroblast	Cancer's effect on Endothelial
SKBR3	+	-
MDAMB231	+	+
MCF7	-	-
HCC1143	+	+
Cell Line	Endothelial's effect on Fibroblast	Endothelial's effect on Cancer
SKBR3	+	-
MDAMB231	-	+
MCF7	+	-
HCC1143	-	-
Cell Line	Fibroblast's effect on Cancer	Fibroblast's effect on Endothelial
SKBR3	+	-
MDAMB231	-	-
MCF7	+	-
HCC1143	-	-

Table 4.5: Inter-population interactions for three populations as predicted by the genetic algorithm.

Cell Line	Cancer's effect on Fibroblast	Cancer's effect on Endothelial
SKBR3	+	-
MDAMB231	+	-
MCF7	+	-
HCC1143	+	-
Cell Line	Endothelial's effect on Fibroblast	Endothelial's effect on Cancer
SKBR3	+	-
MDAMB231	+	+
MCF7	+	+
HCC1143	+	-
Cell Line	Fibroblast's effect on Cancer	Fibroblast's effect on Endothelial
SKBR3	-	-
MDAMB231	+	-
MCF7	+	-
HCC1143	-	-

Table 4.6: Inter-population interactions for three populations as predicted by the deterministic algorithm.

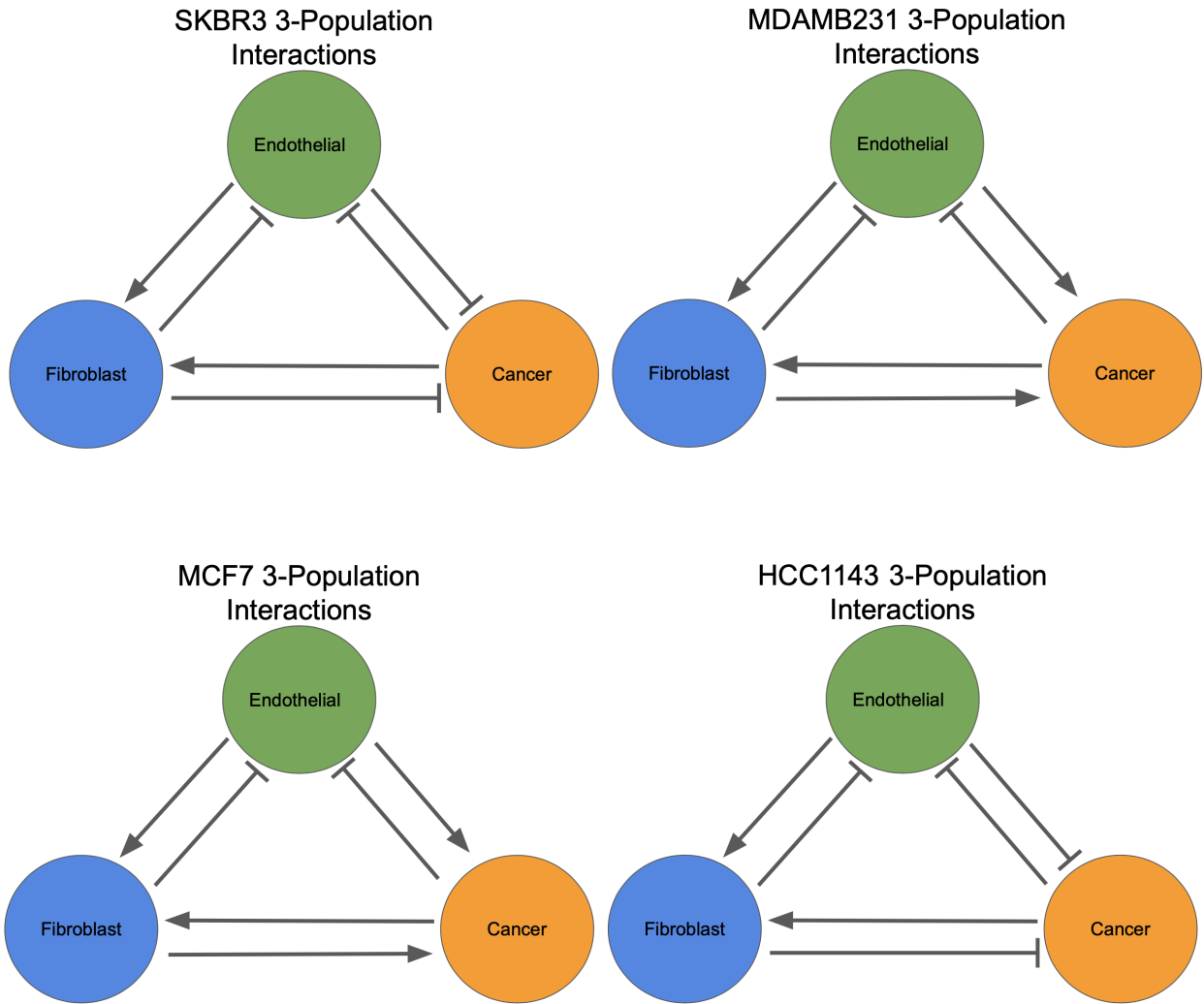


Figure 4.6: 3-population Network Interaction Graphs. An arrow from node *A* to node *B* indicates a positive interaction. A closed line from *A* to *B* indicates an inhibitory interaction.

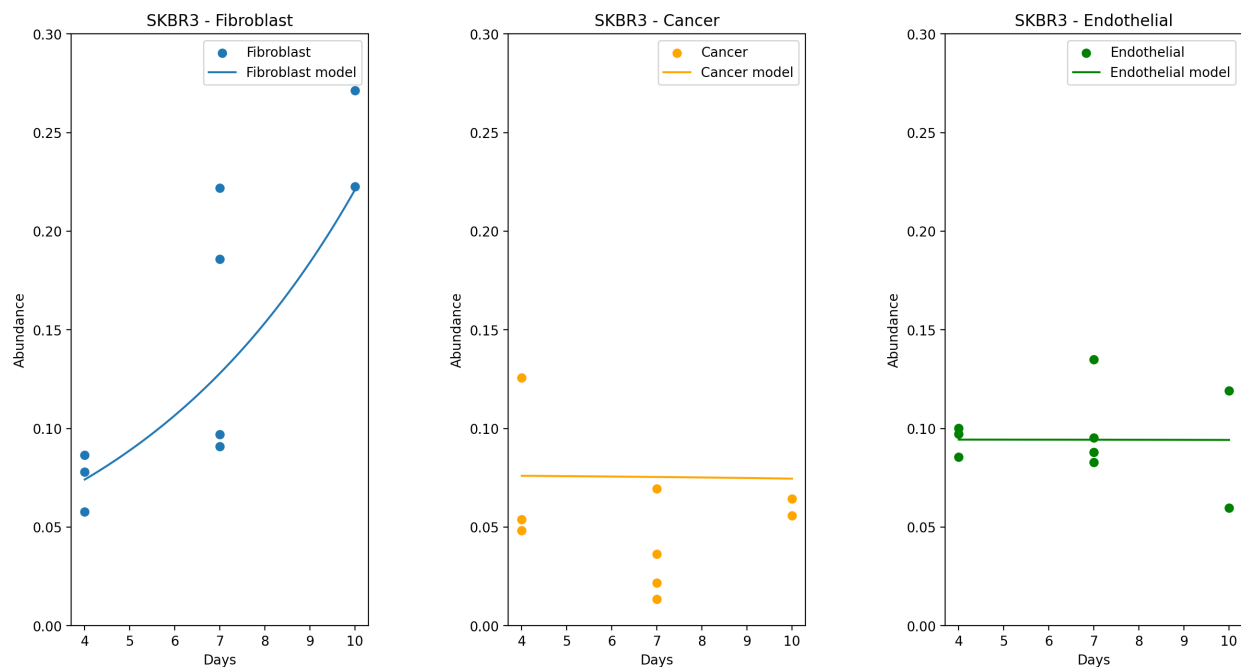


Figure 4.7: SKBR3 Data vs. Model, three-population, integral formulation

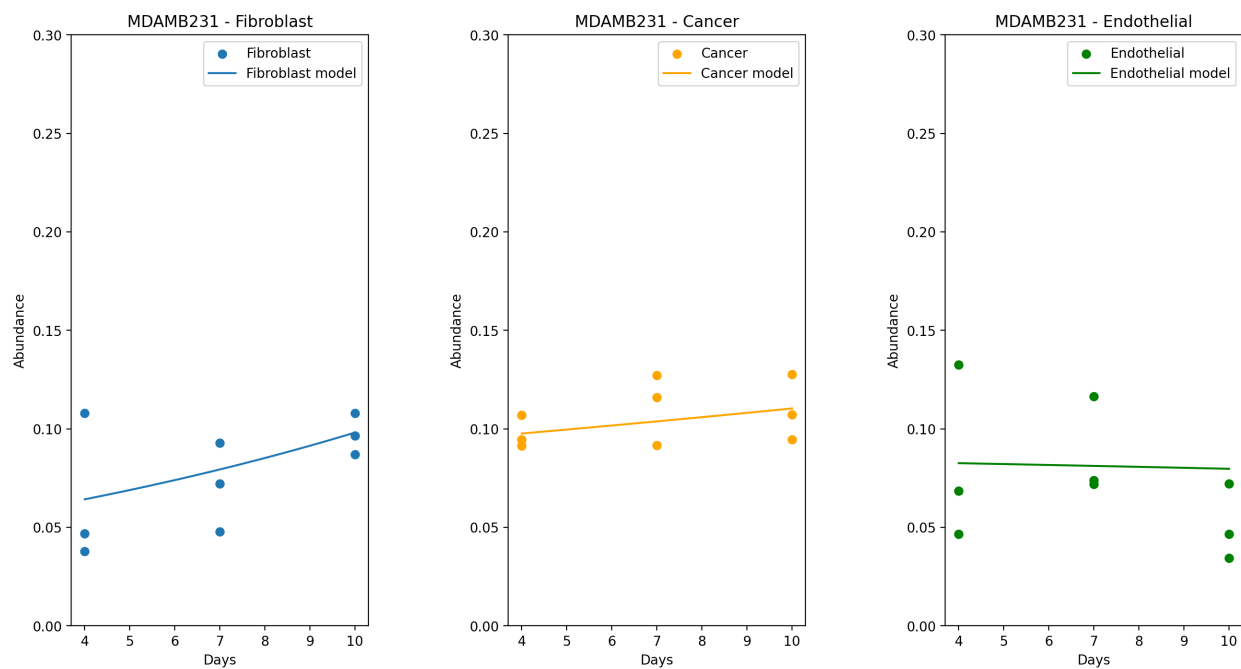


Figure 4.8: MDAMB231 Data vs. Model, three-population, integral formulation

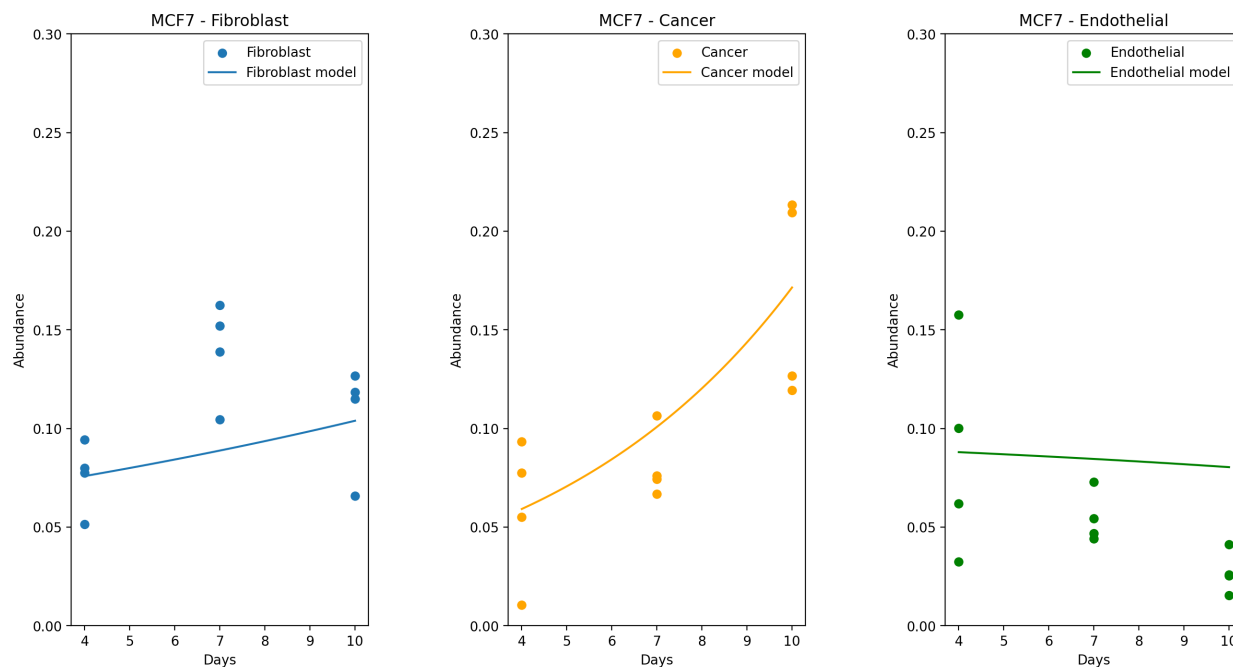


Figure 4.9: MCF7 Data vs. Model, three-population, integral formulation

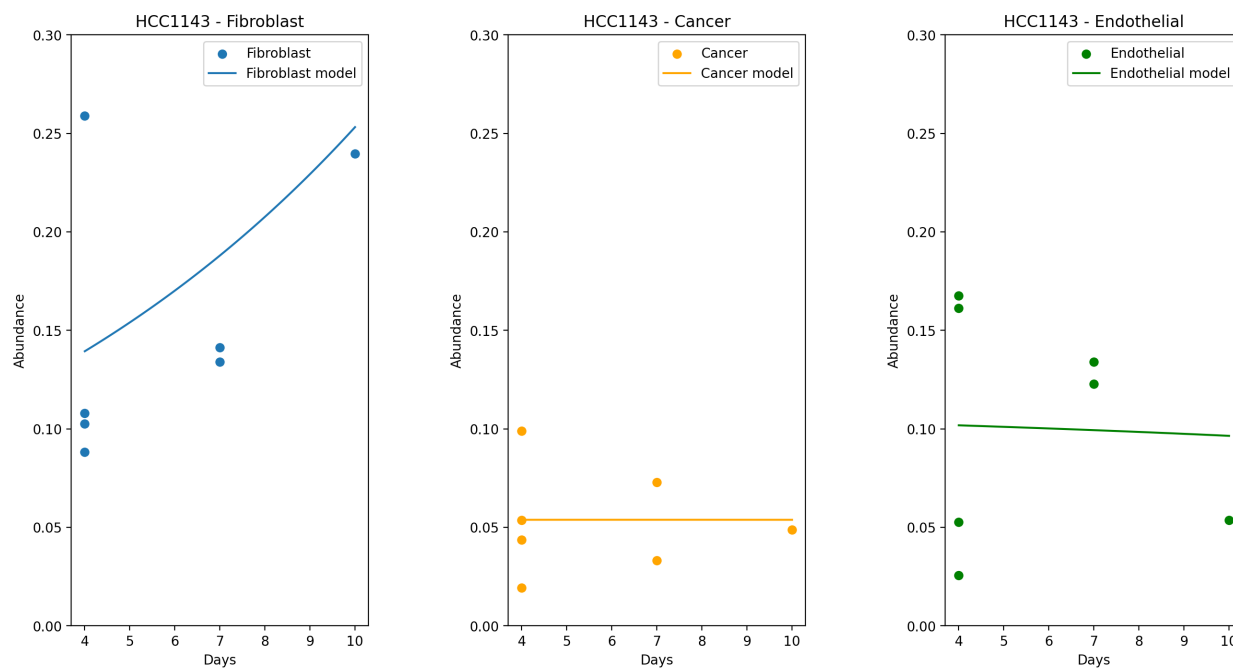


Figure 4.10: HCC1143 Data vs. Model, three-population, integral formulation

Logarithmic Integral Formulation

Optimizing the logarithmic integral formulation of this problem yields the same as interactions as predicted by the vanilla integral formulation (Table 4.6). As in the two-population case, there are no notable differences between the trend graphs of the two different problem formulations, as expected.

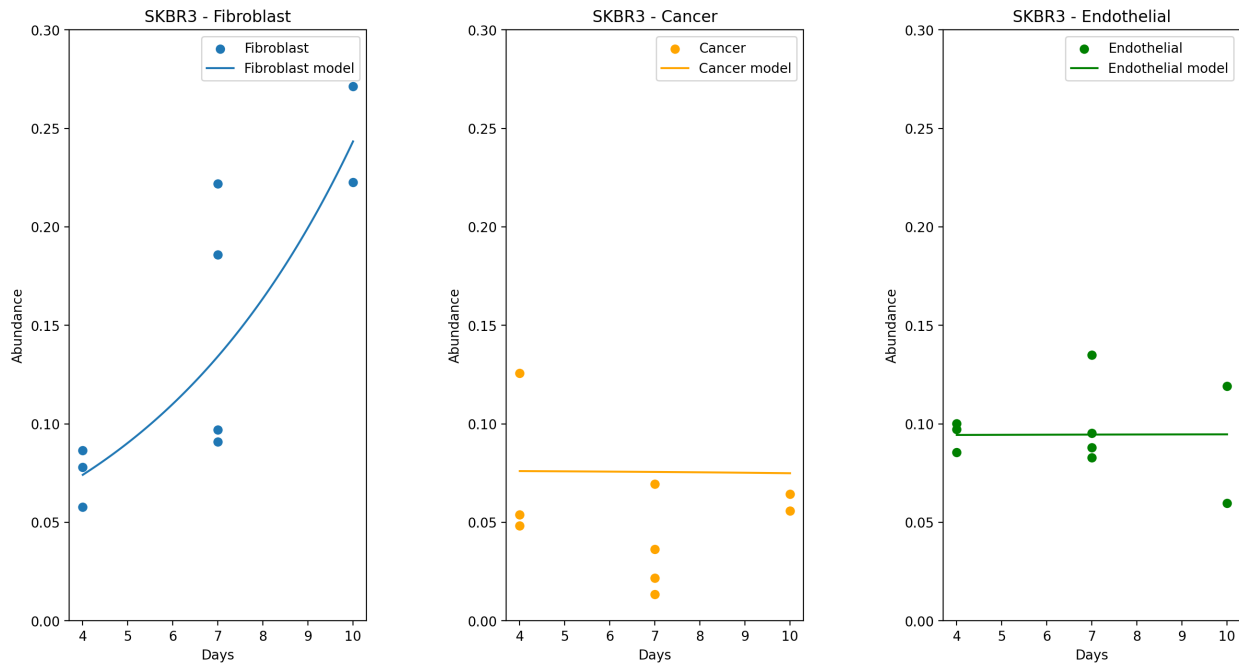


Figure 4.11: SKBR3 Data vs. Model, three-population, logarithmic integral formulation

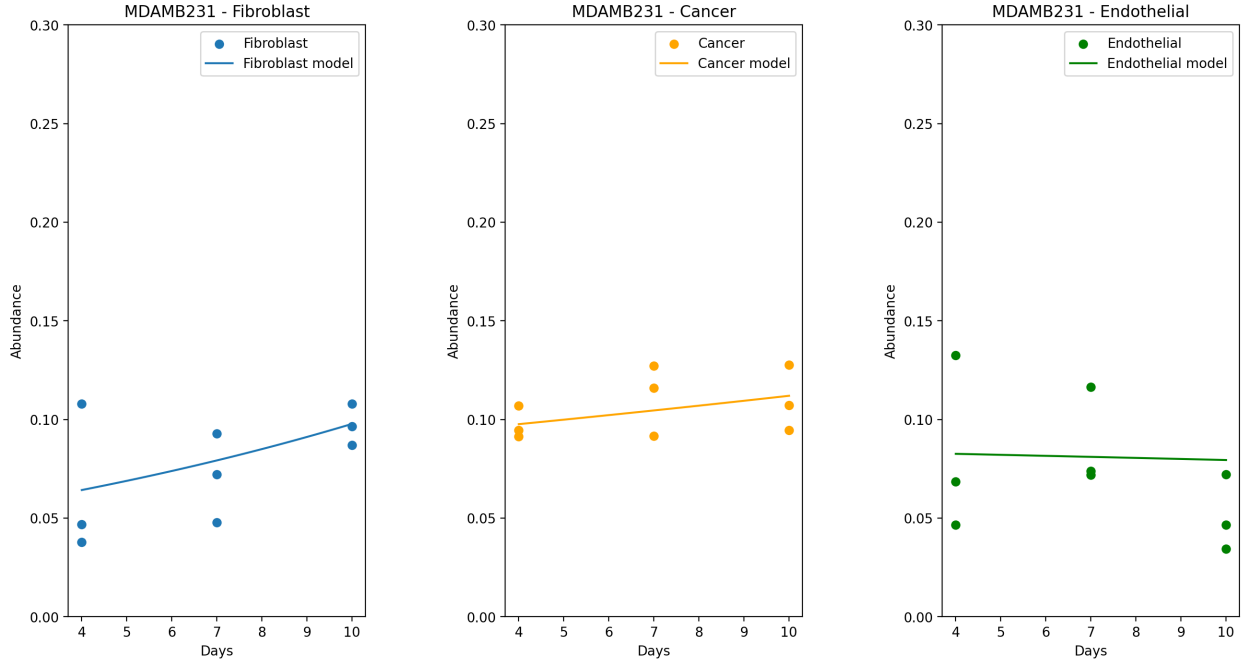


Figure 4.12: MDAMB231 Data vs. Model, three-population, logarithmic integral formula

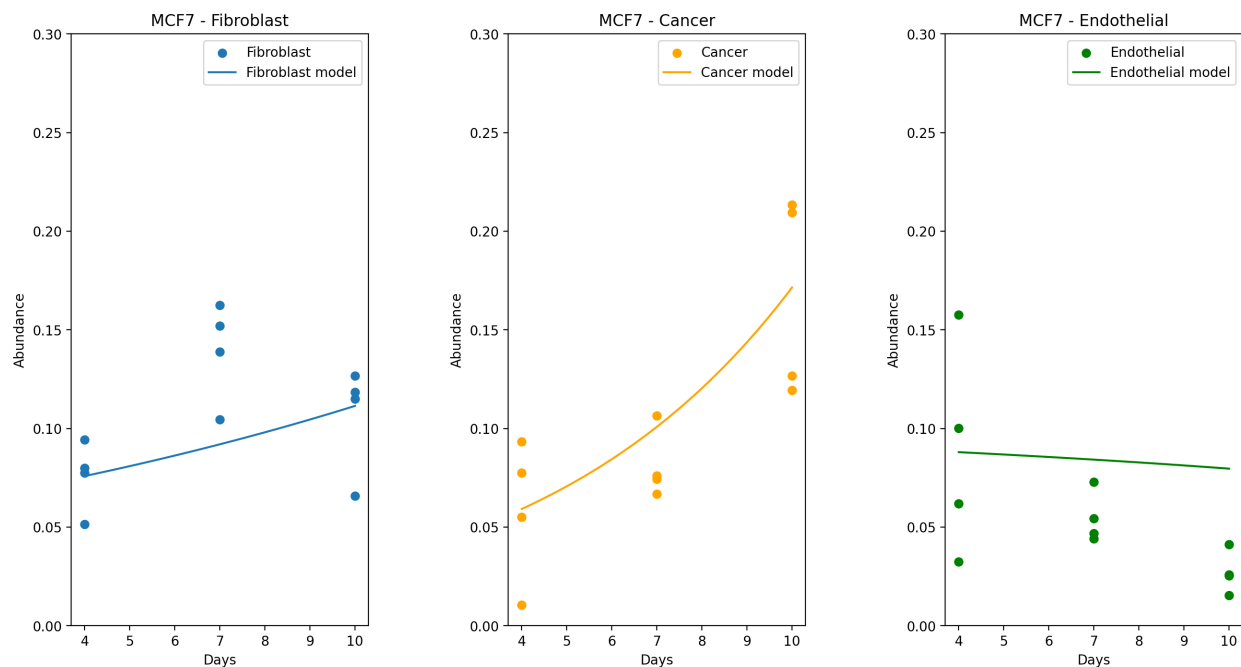


Figure 4.13: MCF7 Data vs. Model, three-population, logarithmic integral formulation

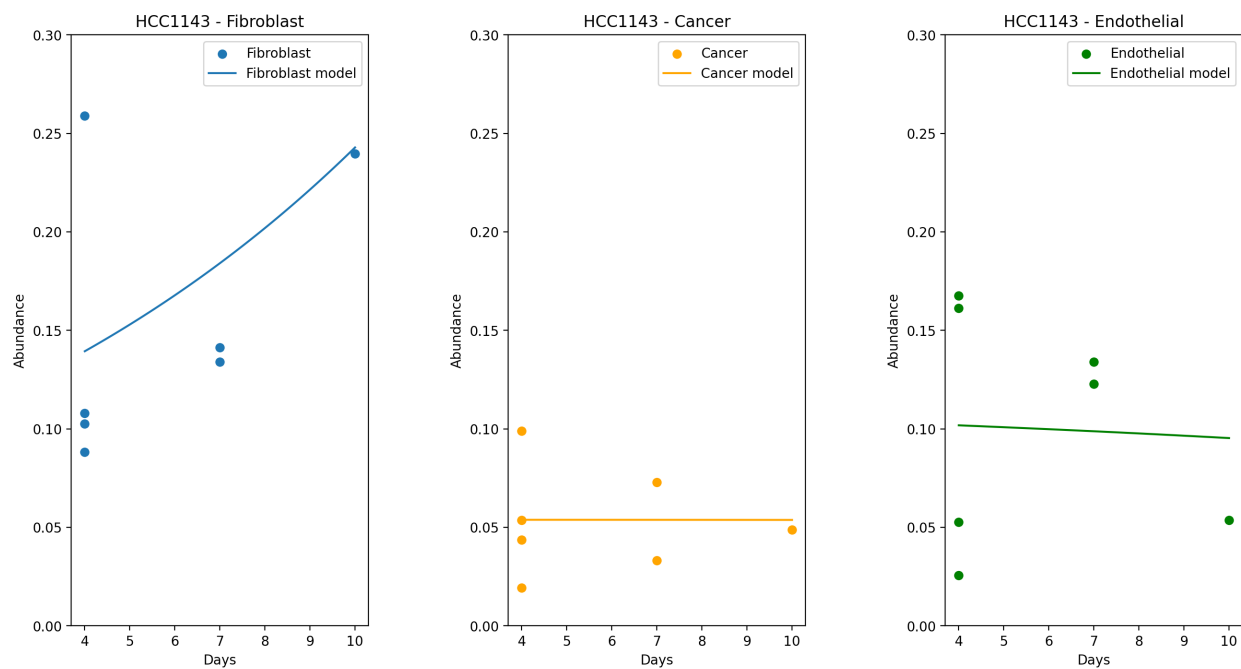


Figure 4.14: HCC1143 Data vs. Model, three-population, logarithmic integral formulation

Chapter 5

Model Robustness

5.1 Integral Formulation

Ideally, the linear algebra deterministic approach should be robust to experimental noise. In the presence of noise within some predefined absolute magnitude, the optimization should still emit interactions types between the populations similar to what is inferred in its absence. To analyze this, we start with this formulation of the generalized Lotka-Volterra optimization:

$$\begin{aligned}
 \min_p \quad & \|Ap - b\|_2 \\
 \text{s.t.} \quad & r_i > 0 \text{ for } i = 1, 2 \\
 & a_{ii} < 0 \text{ for } i = 1, 2 \\
 & -20 < a_{ij} < 20 \text{ for } i, j = 1, 2
 \end{aligned} \tag{5.1}$$

Recall that A is our feature matrix constructed entirely from the experimental data and the structure of the generalized Lotka-Volterra equations. Suppose that the experimental values have been perturbed, leading to an overall perturbation of ξ of the A matrix. This yields the following objective expression:

$$\begin{aligned}
 \text{objective} &= \|(A + \xi)p - b\|_2 \\
 &= \|Ap + \xi p - b\|_2 \\
 &\leq \|Ap - b\|_2 + \|\xi p\|_2 \\
 &\leq \|Ap - b\|_2 + \|\xi\|_F \|p\|_2 \\
 &\leq \|Ap - b\|_2 + \rho \|p\|_2
 \end{aligned} \tag{5.2}$$

where we define ρ to be a constant upper-bounding the norm of the perturbation matrix ξ [7]. Note that since ξ represents the total perturbation in the feature matrix A , it captures both the experimental error and the generalized Lotka-Volterra *modeling error*, the deviation caused by using a model that doesn't necessarily best describe the experimental data.

Optimizing the objective with the same constraints and increasing values of ρ will test the robustness of the deterministic algorithm's ability to infer the generalized Lotka-Volterra

parameters. Here, robustness is defined by how resistant the parameters p are resistant to sign changes as the parameter’s *sign* is what determines the interaction, not the *value* (the norm of p will converge to 0 as ρ grows large and the model loss $\|(A + \xi)p - b\|_2$ becomes comparatively trivial). Note that this formulation is trivially *robustly feasible* because the constraints do not depend on the experimental data, which matches the intuition that these constraints represent hard biological boundaries [4].

Tables 5.1 and 5.2 summarize how robustly the deterministic algorithm predicts HCC1143 spheroid cell interactions for the two-population and three-population cases, respectively. Tables for all cell lines can be found in Appendix B. Across all cell lines, the method can robustly predict interactions across up to approximately seven orders of magnitude greater than the baseline, up to a Frobenius norm magnitude of 0.1.

ρ	Cancer’s effect on Fibroblast	Fibroblast’s effect on Cancer
0.0000001	+	+
0.000001	+	+
0.00001	+	+
0.0001	+	+
0.001	+	+
0.01	+	+
0.1	+	+
1	-	+

Table 5.1: HCC1143 robustness results for two populations. ρ is the magnitude of the Frobenius norm of the noise matrix and captures the overall noise of the experimental data and model. The optimization was run on increasing values of ρ to simulate increasing amounts of added noise, and the predicted cell interactions are recorded here. Interaction types that deviate from the baseline prediction, which typically only occurs at high ρ (noise) values, is shaded red. ρ is capped at 1 because a value of 1 means that the gLV model loss is weighted equally with the magnitude of the parameter norm; any higher ρ would imply that there is so much noise that minimizing the magnitude of parameters to mitigate the effects of noise is more “important” than minimizing the gLV model loss, which doesn’t make sense in this context.

5.2 Logarithmic Integral Formulation

The two-population optimization exhibited much more robustness when using the logarithmic integral formulation. As shown in tables B.13, B.14, B.15, and B.16, there was no change in the predicted interaction types all the way until the threshold of $\rho = 1$, implying that the interaction prediction is robust to noise up until the point where the magnitude of the noise

ρ	Cancer's effect on Fibroblast	Cancer's effect on Endothelial
0.0000001	+	-
0.000001	+	-
0.00001	+	-
0.0001	+	-
0.001	+	-
0.01	+	-
0.1	+	-
1	+	+
ρ	Endothelial's effect on Fibroblast	Endothelial's effect on Cancer
0.0000001	+	-
0.000001	+	-
0.00001	+	-
0.0001	+	-
0.001	+	-
0.01	+	-
0.1	+	-
1	+	+
ρ	Fibroblast's effect on Cancer	Fibroblast's effect on Endothelial
0.0000001	-	-
0.000001	-	-
0.00001	-	-
0.0001	-	-
0.001	-	-
0.01	-	-
0.1	-	-
1	+	+

Table 5.2: HCC1143 robustness results for 3 populations.

makes it such that objective prioritizes minimizing the parameter's magnitude rather than the gLV model error. The robustness of this formulation for the three-population case over the integral formulation is less pronounced. Both formulations became much less consistent as the value of ρ approached 1, but it's worth noting that, in the logarithmic integral formulation, deviations from the noiseless predictions only occurred when the noise was at its maximum ($\rho = 1$) while the integral formulation showed some signs of deviation when the ρ was around 0.1.

The stronger robustness exhibited by this formulation over the simple integral formulation is likely related to the logarithmic integral formulation's simpler feature matrix, which only contains constant and linear terms with respect to the experimental population data. The integral formulation contains quadratic terms, which intuitively increases the total uncertainty characterizing the feature matrix with respect to the noise of each individual data point. More studies are required to analyze the specific effects that noise has on each specific inferred gLV parameter.

Chapter 6

Conclusions and Future Work

6.1 Summary of Results

Both formulations of the linear algebra deterministic algorithm perform similarly to the baseline genetic algorithm on the two-population case, but with the added benefit that this former's robustness can be evaluated by upper-bounding the magnitude of the deviation from the algorithm's feature matrix. The deterministic algorithm infers a mutualistic interaction between cancer and fibroblasts where both promote the growth of the other.

The deterministic algorithm agreed much less with the genetic algorithm in the three-population case, likely due to the increase in model complexity. The only strong patterns inferred were the inhibitory effects cancer and fibroblast cells had on endothelial cells as well as the positive effects cancer and endothelial cells had on fibroblast proliferation. There was much more noise in the three-population experimental data than the two-population set, which likely impacted the parameter inference.

Both formulations of the deterministic algorithm are robustly feasible, and the inferred interactions, indicated by the interaction parameters' signs, are robust to experimental noise as well as modeling error. Overall, the logarithmic integral formulation was slightly more robust than the integral formulation for both two-population and three-population inference, and the optimization algorithms for both formulations were more robust on the two-population data than for the three-population set. Due to the lack of an adequate amount of data, it is difficult to determine whether the deterministic algorithm or stochastic genetic algorithm yielded more trustworthy inferences, though the deterministic method's demonstrated robustness does provide a greater degree of confidence. Further analogous studies with longer-term time series data are required to further verify the reliability of these methods.

6.2 Future Directions

Increasing the Amount of Data

The paucity of data was the main limiting factor for most of the analysis in this report. With data for just three time points across both two-population and three-population inference, it was extremely difficult to estimate derivatives, predict trends, etc. With spheroid data for just two more time points, techniques like spline interpolation could be used for better derivative estimation. The polynomial interpolation used in this report would also be more accurate with more data points.

Model Discovery

In this report, we assumed the populations within the tumor spheroid would exhibit predator-prey behavior, which was the main motivation for applying the generalized Lotka-Volterra model for inference. With enough data, it would be interesting to do the reverse: analyzing the “strength” of this model compared to a multitude of other models. Previous work by Hayden, Chang, Goncalves, and Tomlin has suggested that the latter problem can be reduced to the optimization of an $L1$ norm via *compressive sensing*, a procedure that optimizes the fitting of data over the coefficients of a number of different functions [11]. The functions whose coefficients have the largest magnitudes typically correspond to functions that can best describe the data [6]. Theoretically, placing the functions involved in the generalized Lotka-Volterra model alongside other potential descriptive functions in a sort of “bag of functions” would allow for the comparison of the relative strength of the generalized Lotka-Volterra functions in describing the experimental data with the descriptive strength of the other functions within the “bag”.

Bibliography

- [1] Ellen M. Langer et al. “Modeling Tumor Phenotypes In Vitro with Three- Dimensional Bioprinting”. In: *Cell Reports* 26.288 (Feb. 2019), pp. 608–623. DOI: <https://doi.org/10.1016/j.celrep.2018.12.090>.
- [2] Margaret P. Chapman et al. “Modeling differentiation-state transitions linked to therapeutic escape in triple-negative breast cancer”. In: (2016). DOI: 10.1101/071134. arXiv: 2210.03198v1 [q-bio.QM]. URL: <https://doi.org/10.1101/071134>.
- [3] Roy Ben-Shalom and Michael Lam. <https://github.com/roybens/SimulatingVariants.git>. Accessed: 2023-02-26.
- [4] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust Optimization*. Princeton: Princeton University Press, 2009.
- [5] James Brunner and Nicholas Chia. “Metabolic Model-based Ecological Modeling for Probiotic Design”. In: (2022). arXiv: 2210.03198v1 [q-bio.QM].
- [6] Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. “Discovering governing equations from data by sparse identification of nonlinear dynamical systems”. In: *The Proceedings of the National Academy of Sciences* 113.15 (Mar. 2016). DOI: <https://doi.org/10.1073/pnas.151738411>.
- [7] Giuseppe Calafiore and Laurent El Ghaoui. *Robust Optimization*. Cambridge: Cambridge University Press, 2014.
- [8] Margaret Chapman. “Risk-Sensitive Safety Analysis and Control for Trustworthy Autonomy”. PhD thesis. EECS Department, University of California, Berkeley, May 2020. URL: <http://www2.eecs.berkeley.edu/Pubs/TechRpts/2020/EECS-2020-66.html>.
- [9] Margaret Chapman and Claire Tomlin. “Ordinary Differential Equations in Cancer Biology”. In: (2016). DOI: 10.1101/071134. arXiv: 2210.03198v1 [q-bio.QM]. URL: <https://doi.org/10.1101/071134>.
- [10] Heyrim Cho et al. “Designing experimental conditions to use the Lotka-Volterra model to infer tumor cell line interaction types”. In: (2022). arXiv: 2209.08402v1 [q-bio.PE].
- [11] David Hayden et al. “Compressed Sensing for Network Reconstruction”. In: (2014). arXiv: 1411.4095v1 [math.DS].

- [12] Sharon Kingsland. “Alfred J. Lotka and the origins of theoretical population ecology”. In: (2015). DOI: <https://doi.org/10.1073/pnas.1512317112>.
- [13] P.H. Kloppers and J.C. Greeff. “Lotka–Volterra model parameter estimation using experiential data”. In: *Applied Mathematics and Computation* 224 (2013), pp. 817–825. ISSN: 0096-3003. DOI: <https://doi.org/10.1016/j.amc.2013.08.093>. URL: <https://www.sciencedirect.com/science/article/pii/S0096300313009806>.
- [14] Marcin et al. Paczkowski. “Reciprocal interactions between tumour cell populations enhance growth and reduce radiation sensitivity in prostate cancer”. In: (2021). DOI: <https://doi.org/10.1038/s42003-020-01529-5>.
- [15] Christopher H. Remien, Mariah J. Eckwright, and Benjamin J. Ridenhour. “Parameter Identifiability of the Generalized Lotka-Volterra Model for Microbiome Studies”. In: (Nov. 2018). DOI: <https://doi.org/10.1101/463372>.
- [16] William Sharpless and Michael Lam. <https://github.com/mikelam-us/Cancer-Biology.git>. Accessed: 2023-04-2.
- [17] Guillaume Thibault. <https://www.thibault.biz/Research/cycIFAAP/cycIFAAP.html>. Accessed: 2023-03-1.

Appendix A

Training Curves

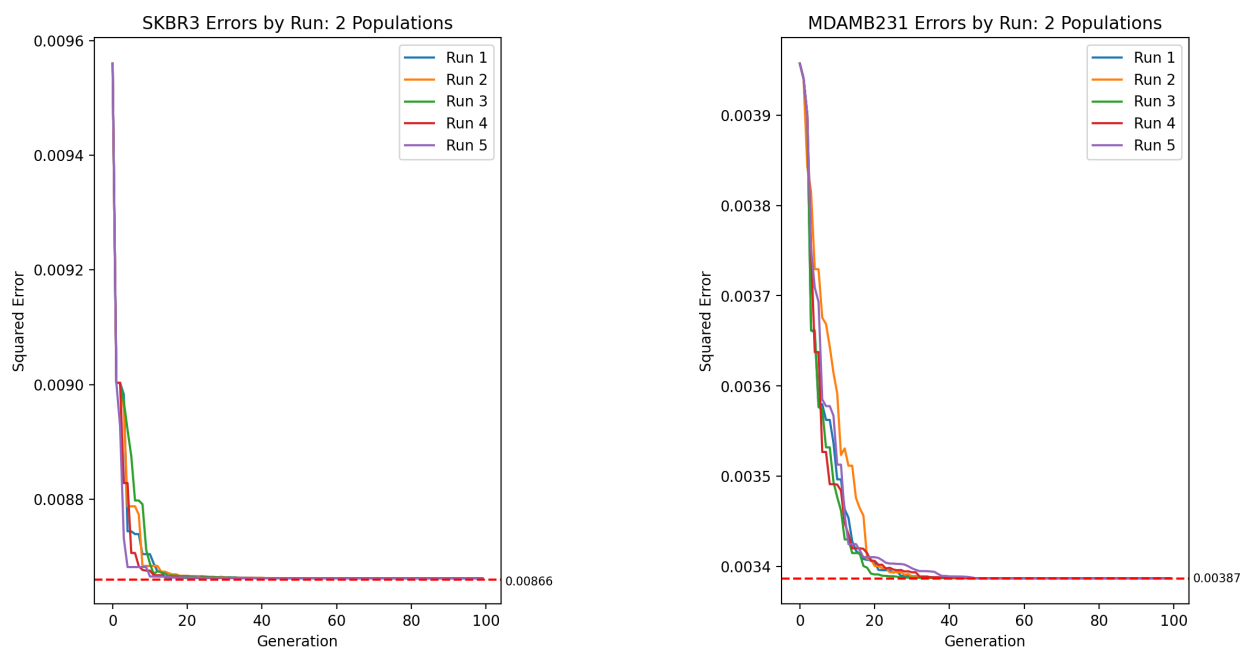


Figure A.1

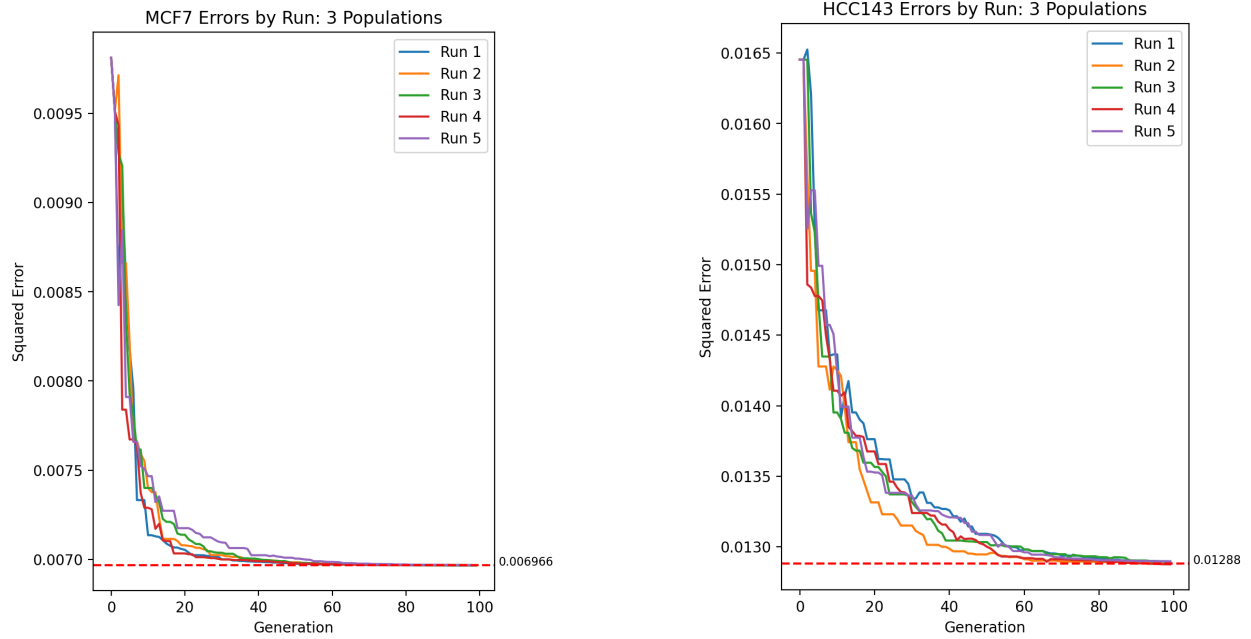


Figure A.2

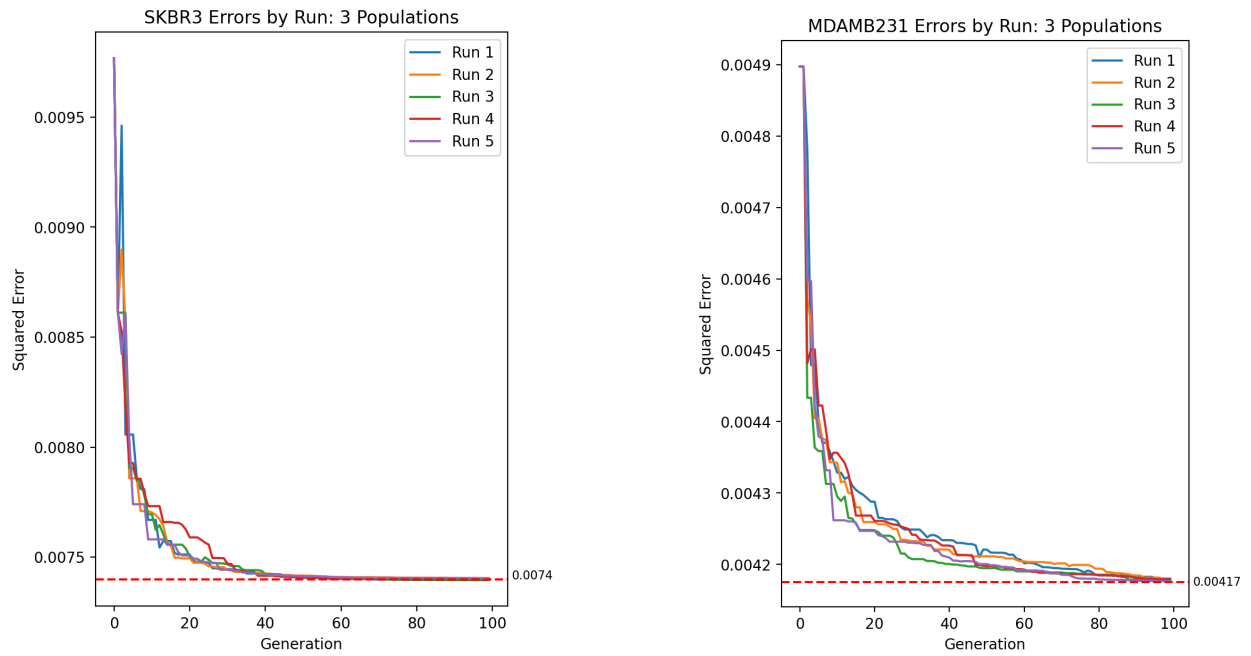


Figure A.3

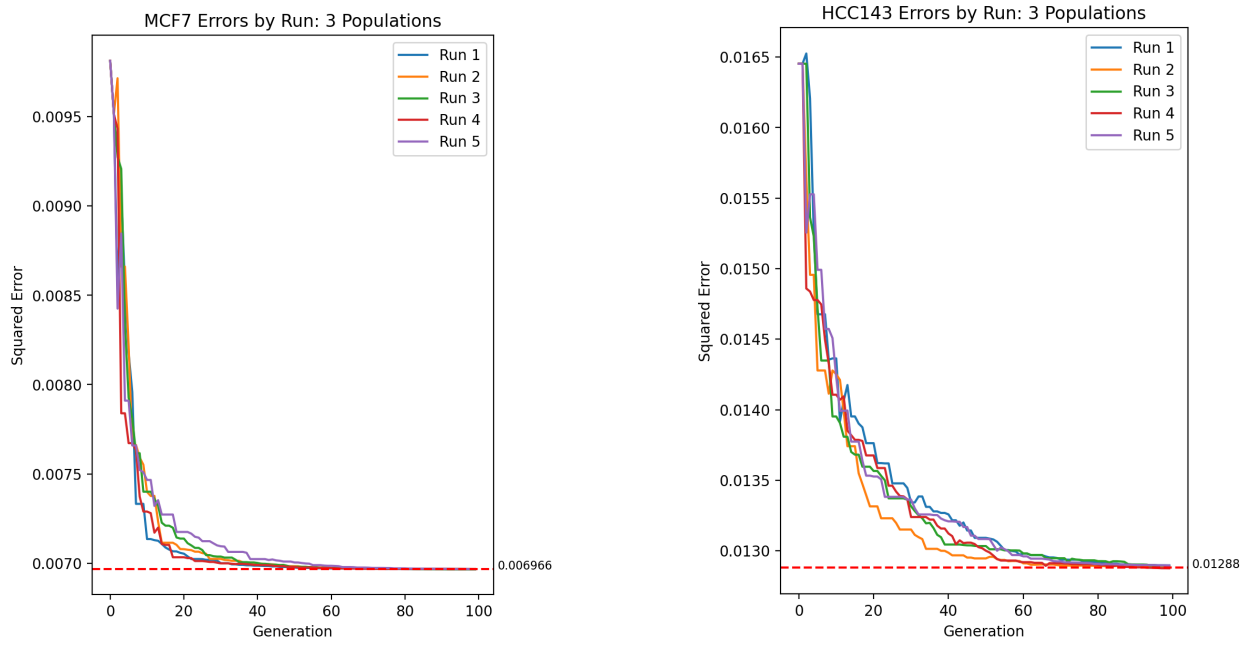


Figure A.4

Appendix B

Parameter Robustness Charts

B.1 Two Populations

Integral Formulation

ρ	Cancer's effect on Fibroblast	Fibroblast's effect on Cancer
0.0000001	+	-
0.000001	+	-
0.00001	+	-
0.0001	+	-
0.001	+	-
0.01	+	-
0.1	+	-
1	-	+

Table B.1: SKBR3 Robustness, Integral Formulation, 2-population

ρ	Cancer's effect on Fibroblast	Fibroblast's effect on Cancer
0.0000001	+	+
0.000001	+	+
0.00001	+	+
0.0001	+	+
0.001	+	+
0.01	+	+
0.1	+	+
1	+	-

Table B.2: MDAMB231 Robustness, Integral Formulation, 2-population

ρ	Cancer's effect on Fibroblast	Fibroblast's effect on Cancer
0.0000001	-	+
0.000001	-	+
0.00001	-	+
0.0001	-	+
0.001	-	+
0.01	-	+
0.1	-	+
1	+	+

Table B.3: MCF7 Robustness, Integral Formulation, 2-population

ρ	Cancer's effect on Fibroblast	Fibroblast's effect on Cancer
0.0000001	+	+
0.000001	+	+
0.00001	+	+
0.0001	+	+
0.001	+	+
0.01	+	+
0.1	+	+
1	-	+

Table B.4: HCC1143 Robustness, Integral Formulation, 2-population

Logarithmic Integral Formulation

ρ	Cancer's effect on Fibroblast	Fibroblast's effect on Cancer
0.000001	+	-
0.00001	+	-
0.0001	+	-
0.001	+	-
0.01	+	-
0.1	+	-
1	+	-

Table B.5: SKBR3 Robustness, Logarithmic Integral Formulation, 2-population

ρ	Cancer's effect on Fibroblast	Fibroblast's effect on Cancer
0.000001	+	+
0.00001	+	+
0.0001	+	+
0.001	+	+
0.01	+	+
0.1	+	+
1	+	+

Table B.6: MDAMB231 Robustness, Logarithmic Integral Formulation, 2-population

ρ	Cancer's effect on Fibroblast	Fibroblast's effect on Cancer
0.000001	-	+
0.00001	-	+
0.0001	-	+
0.001	-	+
0.01	-	+
0.1	-	+
1	-	+

Table B.7: MCF7 Robustness, Logarithmic Integral Formulation, 2-population

ρ	Cancer's effect on Fibroblast	Fibroblast's effect on Cancer
0.000001	+	+
0.00001	+	+
0.0001	+	+
0.001	+	+
0.01	+	+
0.1	+	+
1	+	+

Table B.8: HCC1143 Robustness, Logarithmic Integral Formulation, 2-population

B.2 Three Populations

Integral Formulation

ρ	Cancer's effect on Fibroblast	Cancer's effect on Endothelial
0.0000001	+	-
0.000001	+	-
0.00001	+	-
0.0001	+	-
0.001	+	-
0.01	+	-
0.1	+	-
1	+	-
ρ	Endothelial's effect on Fibroblast	Endothelial's effect on Cancer
0.0000001	+	-
0.000001	+	-
0.00001	+	-
0.0001	+	-
0.001	+	-
0.01	+	-
0.1	+	-
1	+	-
ρ	Fibroblast's effect on Cancer	Fibroblast's effect on Endothelial
0.0000001	-	-
0.000001	-	-
0.00001	-	-
0.0001	-	-
0.001	-	-
0.01	-	-
0.1	-	-

Table B.9: SKBR3 Robustness, Integral Formulation, 3-population

ρ	Cancer's effect on Fibroblast	Cancer's effect on Endothelial
0.0000001	+	-
0.000001	+	-
0.00001	+	-
0.0001	+	-
0.001	+	-
0.01	+	-
0.1	+	+
1	-	-
ρ	Endothelial's effect on Fibroblast	Endothelial's effect on Cancer
0.0000001	+	+
0.000001	+	+
0.00001	+	+
0.0001	+	+
0.001	+	+
0.01	+	+
0.1	+	+
1	-	-
ρ	Fibroblast's effect on Cancer	Fibroblast's effect on Endothelial
0.0000001	+	-
0.000001	+	-
0.00001	+	-
0.0001	+	-
0.001	+	-
0.01	+	-
0.1	+	-
1	-	-

Table B.10: MDAMB231 Robustness, Integral Formulation, 3-population

ρ	Cancer's effect on Fibroblast	Cancer's effect on Endothelial
0.0000001	+	-
0.000001	+	-
0.00001	+	-
0.0001	+	-
0.001	+	-
0.01	+	-
0.1	+	-
1	-	-
ρ	Endothelial's effect on Fibroblast	Endothelial's effect on Cancer
0.0000001	+	+
0.000001	+	+
0.00001	+	+
0.0001	+	+
0.001	+	+
0.01	+	+
0.1	+	+
1	-	-
ρ	Fibroblast's effect on Cancer	Fibroblast's effect on Endothelial
0.0000001	+	-
0.000001	+	-
0.00001	+	-
0.0001	+	-
0.001	+	-
0.01	+	-
0.1	+	-
1	-	-

Table B.11: MCF7 Robustness, Integral Formulation, 3-population

ρ	Cancer's effect on Fibroblast	Cancer's effect on Endothelial
0.0000001	+	-
0.000001	+	-
0.00001	+	-
0.0001	+	-
0.001	+	-
0.01	+	-
0.1	+	-
1	-	-
ρ	Endothelial's effect on Fibroblast	Endothelial's effect on Cancer
0.0000001	+	-
0.000001	+	-
0.00001	+	-
0.0001	+	-
0.001	+	-
0.01	+	-
0.1	+	-
1	-	-
ρ	Fibroblast's effect on Cancer	Fibroblast's effect on Endothelial
0.0000001	-	-
0.000001	-	-
0.00001	-	-
0.0001	-	-
0.001	-	-
0.01	-	-
0.1	-	-
1	-	-

Table B.12: HCC1143 Robustness, Integral Formulation, 3-population

Logarithmic Integral Formulation

ρ	Cancer's effect on Fibroblast	Cancer's effect on Endothelial
0.0000001	+	-
0.000001	+	-
0.00001	+	-
0.0001	+	-
0.001	+	-
0.01	+	-
0.1	+	-
1	+	-
ρ	Endothelial's effect on Fibroblast	Endothelial's effect on Cancer
0.0000001	+	-
0.000001	+	-
0.00001	+	-
0.0001	+	-
0.001	+	-
0.01	+	-
0.1	+	-
1	+	-
ρ	Fibroblast's effect on Cancer	Fibroblast's effect on Endothelial
0.0000001	-	-
0.000001	-	-
0.00001	-	-
0.0001	-	-
0.001	-	-
0.01	-	-
0.1	-	-

Table B.13: SKBR3 Robustness, Logarithmic Integral Formulation, 3-population

ρ	Cancer's effect on Fibroblast	Cancer's effect on Endothelial
0.0000001	+	-
0.000001	+	-
0.00001	+	-
0.0001	+	-
0.001	+	-
0.01	+	-
0.1	+	-
1	-	-
ρ	Endothelial's effect on Fibroblast	Endothelial's effect on Cancer
0.0000001	+	+
0.000001	+	+
0.00001	+	+
0.0001	+	+
0.001	+	+
0.01	+	+
0.1	+	+
1	-	-
ρ	Fibroblast's effect on Cancer	Fibroblast's effect on Endothelial
0.0000001	+	-
0.000001	+	-
0.00001	+	-
0.0001	+	-
0.001	+	-
0.01	+	-
0.1	+	+
1	-	-

Table B.14: MDAMB231 Robustness, Logarithmic Integral Formulation, 3-population

ρ	Cancer's effect on Fibroblast	Cancer's effect on Endothelial
0.0000001	+	-
0.000001	+	-
0.00001	+	-
0.0001	+	-
0.001	+	-
0.01	+	-
0.1	+	-
1	-	-
ρ	Endothelial's effect on Fibroblast	Endothelial's effect on Cancer
0.0000001	+	+
0.000001	+	+
0.00001	+	+
0.0001	+	+
0.001	+	+
0.01	+	+
0.1	+	+
1	-	+
ρ	Fibroblast's effect on Cancer	Fibroblast's effect on Endothelial
0.0000001	+	-
0.000001	+	-
0.00001	+	-
0.0001	+	-
0.001	+	-
0.01	+	-
0.1	+	-
1	+	-

Table B.15: MCF7 Robustness, Logarithmic Integral Formulation, 3-population

ρ	Cancer's effect on Fibroblast	Cancer's effect on Endothelial
0.0000001	+	-
0.000001	+	-
0.00001	+	-
0.0001	+	-
0.001	+	-
0.01	+	-
0.1	+	-
1	+	+
ρ	Endothelial's effect on Fibroblast	Endothelial's effect on Cancer
0.0000001	+	-
0.000001	+	-
0.00001	+	-
0.0001	+	-
0.001	+	-
0.01	+	-
0.1	+	-
1	+	+
ρ	Fibroblast's effect on Cancer	Fibroblast's effect on Endothelial
0.0000001	-	-
0.000001	-	-
0.00001	-	-
0.0001	-	-
0.001	-	-
0.01	-	-
0.1	-	-
1	+	+

Table B.16: HCC1143 Robustness, Logarithmic Integral Formulation, 3-population