Estimation of Probability Measures Using Aggregate Population Data: Analysis of a CAR T-cell Cancer Model and Optimal Design of Experiments

March 28, 2019

Celia Schacht[1], **Annabel Meade**[1], **H.T. Banks**[1], **Heiko Enderling**[2], **Daniel Abate-Daga**[2]

[1] *North Carolina State University, Raleigh, NC*
cmschach@ncsu.edu
aemeade@ncsu.edu
htbanks@ncsu.edu

[2] *Moffitt Cancer Center, Tampa, FL*
heiko.enderling@moffitt.org
daniel.abatedaga@moffitt.org

Abstract In this effort we explain fundamental formulations for aggregate data inverse problems requiring estimation of probability distribution parameters. We use as a motivating example a class of CAR T-cell cancer models in mice. After ascertaining results on model stability and sensitivity with respect to parameters, we carry out first elementary computations on the question how much data is needed for successful estimation of probability distributions. This is followed by presentation of an advanced framework for optimal design of experiments to be employed in future efforts.

Key words: aggregate data, CAR T-cell cancer model, inverse problems, design of experiments

Mathematics Subject Classification: 31B20, 65L09, 46N60, 62K86

1 Aggregate Data

In the mathematical modeling of physical and biological systems, situations often arise where some facet of the underlying dynamics (in the form of a parameter) is not constant but rather may be distributed probabilistically within the system or across the population under study. While traditional inverse problems involve the estimation, given a set of data/observations, of a fixed set of parameters contained within some finite dimensional admissible set, models with parameters distributed across the population require the estimation of a probability measure or distribution over the set of admissible parameters. The techniques for measure estimation (along with their theoretical justifications) are widely scattered throughout the literature in applied mathematics and statistics, often with few cross references to related ideas; a review is given in [BKTReview]. Of course, it is highly likely that individual parameters might vary from one individual to the next within the sampled population. Thus, our goal in this case is to use the sample of individuals to estimate the probability measure describing the distribution of certain parameters in the full population.

In some situations (e.g., in certain pharmacokinetics and/or pharmacodynamics examples) one is able to follow each individual separately and collect longitudinal time course data for each individual. In other investigations, situations arise in which one is only able to collect *aggregate data*. Such might be the case, for example, in marine or insect catch and release experiments in which one cannot be certain of measuring the same individual multiple times. In this case one does not have individual data, but rather histograms showing the aggregate number of individuals sampled from the population at a given time having a given size/weight or other characteristic of interest ([BBKW] and Chapter 9 of [BT2009]). The goal then is to estimate the probability distributions describing the variability of the parameters across the population.

These two situations lead to fundamentally different estimation problems. In the first case, one attempts to use a mathematical model describing a single individual along with data collected from multiple individuals (but data in which subsets of longitudinal data points could be identified with a *specific individual*) in order to determine population level characteristics. In the second case, while one again has an individual model, the data collected *cannot* be identified with individuals and is considered to be *sampled longitudinally* from the *aggregate population*. It is worth noting that special care must be taken in this latter case to identify the model such as that introduced below as an individual model in the sense that it describes an individual subpopulation. That is, we have all "individuals" (i.e., shrimp [BDEHAD] or mosquitofish [BBKW]) described by the model share a common growth/birth/death rate function g. Mathematically, here we define the "individual" in terms of the underlying parameters, using 'individual' to describe the unit characterized by a single parameter set (excretion rate, growth/birth/death rate, damping rate, relaxation times, etc.).

One can also distinguish two generic estimation problems. In the first example, we consider the case, as in a structured density model, that one has a mathematical model for individual dynamics but only aggregate data (we refer to this as the *individual dynamics/aggregate data* problem). The second possibility–that one has only an aggregate model (i.e., the dynamics depend explicitly on a distribution of parameters across the population) with aggregate data–can also be examined (we refer to this as the *aggregate dynamics/aggregate data* problem). Such examples arise in electromagnetic models with a distribution of polarization relaxation times for molecules (e.g., [BG1, BG2]); in biology with HIV cellular models [BanksBortz, BBH, BBPP]; and in wave propagation in viscoelastic materials such as biotissue [Stenosis1, Stenosis2, Stenosis3, BanksPinter]. The measure estimation problem for such examples is sufficiently similar to the individual dynamics/aggregate data situation and accordingly we do not consider aggregate dynamics models as a separate case.

In both generic estimation problems mentioned above, the underlying goal is the determination of the probability measure which describes the distribution of parameters across all members of the

population. Thus, two main issues are of interest. First, a sensible framework (the *Prohorov Metric Framework* as we have developed it–see Chapter 14 of [B2012] and Chapter 5 of [BHT2014]) must be established for each situation so that the estimation problem is meaningful. Thus we must decide what type of information and/or estimates are desired (e.g., mean, variance, complete distribution function, etc.) and determine how these decisions will depend on the type of data available. Second, we must examine what techniques are available for the computation of such estimates. Because the space of probability measures is an infinite dimensional space, we must make some type of finite dimensional approximations so that the estimation problem is amenable to computation. Moreover, here we are interested in frameworks for *approximation and convergence* which are not restricted to classic parametric estimates. Thus as discussed in [BHT2014], while the individual dynamics/individual data and individual dynamics/aggregate data problems are fundamentally different, there are notable similarities in the approximation techniques (for example, use of discrete measure approximates) common to both problems.

In light of the dispersion in growth discussed in the mosquitofish problems of [BBKW], we replace the growth rate g by a family $\mathcal{G}$ of growth rates and reconsider the model with a probability distribution P on this family. The population density is then given by summing "cohorts" of subpopulations where individuals belong to the same subpopulation if they have the same growth rate [BBKW, BF1991, BFPZ1998]. Thus, in what has been termed as *Growth Rate Distribution (GRD)* models, the population density $u(t, x; P)$, first discussed in [BBKW] for mosquitofish, developed more fully in [BF1991] and subsequently used for shrimp models, is actually given by

$$u(t, x; P) = \int_{\mathcal{G}} v(t, x; g) dP(g) = \mathbb{E}(v(t, x; \cdot)|P), \tag{1}$$

where $\mathcal{G}$ is a collection of admissible growth rates, P is a probability measure on $\mathcal{G}$, and $v(t, x; g)$ is the solution of the model equation

$$\frac{dx}{dt} = g(t, x)$$

for individual growth rate g. This model assumes the population is made up of *collections of subpopulations* with individuals in the same subpopulation if they possess the same size/weight dependent growth rate.

Many inverse problems, such as those discussed in [BHT2014], involve individual dynamics and individual data. For example, if one had individual longitudinal data y_{ij} corresponding to the structured population density $v(t_i, x_j; g)$, where v is the solution corresponding to an (individual) cohort all with the same individual growth rate, one could then formulate a standard ordinary least squares (OLS) problem for estimation of g. This would entail finding

$$\hat{g} = \arg\min_{g \in \mathcal{G}} \sum_{i,j} |v_{ij} - v(t_i, x_j; g)|^2, \tag{2}$$

where $\mathcal{G}$ is a family of admissible growth rates. However for many problems tracking of individuals is often impossible. Thus we turn to methods where one can use aggregate data.

As outlined above, we find that the expected weight density $u(t, x; P) = \mathbb{E}(v(t, x; \cdot)|P)$ is described by a general probability distribution P where the density $v(t, x; g)$ satisfies, for a given g, the dynamics (an ODE, a PDE or a Delay Differential Equation). In these problems, even though one has "individual dynamics" (the $v(t, x; g)$ for a given cohort with growth rate g), one has only *aggregate* or population level longitudinal data available. Again, this is common in marine, insect, etc., *catch and release* experiments [BK1989] where one samples at different times from the same population but cannot be guaranteed of observing the same set of individuals at each sample time. This type of data is also

quite typical in experiments where the organism or population member being studied is sacrificed in the process of making a single observation as in the mouse cancer data motivating this paper and described more fully below. In these cases one may still have dynamic (i.e., time course) models for individuals, but no individual data is available.

Since we must use aggregate population data $\mathbf{u} = \{u_{ij}\}$ to estimate P itself in the inverse problems of interest to us here, we are therefore required to understand the qualitative properties (continuity, sensitivity, etc.) of $u(t, x; P)$ as a function of P. (These issues are discussed in some detail in [BHT2014].) The data for the resulting parameter estimation problems are u_{ij}, which are observations for $u(t_i, x_j; P)$. The corresponding OLS inverse problem consists of minimizing

$$J(P; \mathbf{u}) = \sum_{ij} |u_{ij} - u(t_i, x_j; P)|^2 \tag{3}$$

over $P \in \mathbb{P}(\mathcal{G})$, the set of probability distributions on $\mathcal{G}$, or over some suitably chosen subset of $\mathbb{P}(\mathcal{G})$.

2 Model Description

We now turn to two important questions: the first is how to deal with inverse problems for dynamical systems when longitudinal data does not exist? A second question that we shall consider below entails how to efficiently design experiments to collect data (how much data and when to collect it?, etc,) that is necessary to validate models in such aggregate data situations. Our efforts here are motivated by a set of data collected with NSG (NOD/SCID/GAMMA), or NOD.Cg-Prkdcscidll2rgtm1Wjl/SzJ mice injected with with cancer and then sequentially sacrificed to collect the needed data about tumor growth. The resulting aggregate data is absolutely common in biological experiments where data collection requires sacrifice of the subjects. Colleagues at Moffitt Cancer Center have carried out preliminary trial experiments (data collection already completed) as follows: At t=-14 (14 days before the trial begins) NSG mice are injected with cancer, which should take 12-14 days to begin growing. At t=0 (as determined by tumor volume) the mice are further injected with chimeric antigen receptor [CAR] or CAR T-cells (engineered T-cells to specifically target cancer cells). Mice are divided into four groups with different treatments. Autopsies are performed on each of 5 mice sacrificed at t=5, 10, and 15 days respectively, to determine the concentration of the engineered T-cells within the blood, spleen, and within the tumor. Because of nature of data collection, longitudinal data is not possible as mice must be killed for data to be collected. A major goal for these scientists: Between the four different treatments, which treatment is the best and why is it the best?

Here we demonstrate use of the Prohorov metric in formulating and carrying out least squares parameter estimation problems when only aggregate data is available. In our discussion here we present results using inverse problem techniques for estimation of growth/death distributions in ordinary differential population models using aggregate population data. The models employed here are based on ideas initially discussed in [BBKW] which entail models wherein growth rates may vary across individuals of the population as well as with size and time to maturity. We first investigate with simple heuristic simulations how much data is needed determine the underlying distributions. We then turn to significant optimal design questions and outline some of the substantial literature which is available.

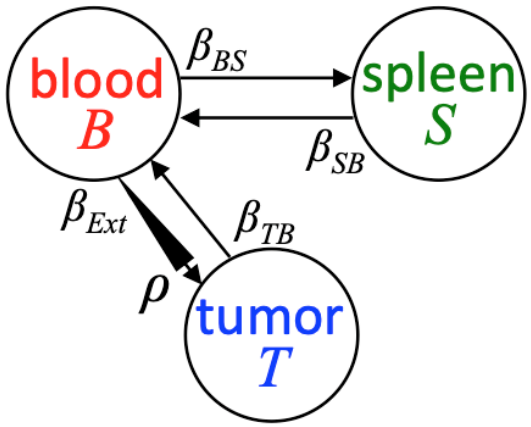


Figure 1: Schematic of a simple compartmental model.

We use the system of ordinary differential equations,

$$\frac{dT}{dt} = \rho\beta_{Ext}B - \beta_{TB}T \tag{4}$$

$$\frac{dB}{dt} = (\beta_{TB}T - \rho\beta_{Ext}B) + (\beta_{SB}S - \beta_{Ext}B) \tag{5}$$

$$\frac{dS}{dt} = \beta_{Ext}B - \beta_{SB}S, \tag{6}$$

based on simple mass balance as described in ([BT2009], [de Vries etal], [Rubinow]) and depicted in Figure 1 to model the flow of T-cells in the tumor T, blood B, and spleen S in a cancerous body. The number of T-cells in the tumor, T, travel to the blood, B, at rate β_{TB}, and flow from the blood to the spleen, S, at a rate β_{Ext}. The number of T-cells in the spleen flow back to the blood at rate β_{SB}, and travel from the blood to the tumor at a rate $\rho_{Tx} = \rho\beta_{Ext}$. For each T-cell that leaves the blood, there may be a transient expansion, ρ, in the tumor due to antigen recognition. If there is no antigen recognition, $\rho_{Tx} = \beta_{Ext}$ and $\rho = 1$, and if there is any antigen recognition, $\rho_{Tx} > \beta_{Ext}$ and $\rho > 1$. It is important to note that β_{Ext}=β_{BS}, that is, the rate of travel from blood to spleen and from blood to tumor is equivalent without the added ρ. Our four parameters of interest are thus ρ, β_{Ext}, β_{TB}, and β_{SB}.

In the experiment of interest, cancerous mice are separated into four treatment categories: untransduced T-cells (UT), chimeric antigen receptor therapy (CAR), CAR treatment with added CXCR1 chemokine receptors (CAR+CXCR1), and CAR treatment with added CXCR2 chemokine receptors (CAR+CXCR2). Each treatment has a different effect on antigen recognition in the tumor, or parameter ρ. The parameters and their values are listed in Table 1 and are partially motivated by reference to other tumor-related experiments [Moon]. The ρ values for each treatment are estimations around which we can study the behavior for each system. The T-cell movement rates, β_{EXT}, β_{TB}, and β_{SB}, are estimated based on knowledge that the exit spleen rate is considerably lower than the exit tumor rate, such that $\beta_{EXT} > \beta_{TB} > \beta_{SB}$. It is also known that the initial T-cell counts in the tumor and spleen, T_0 and S_0, respectively, are zero, and the initial T-cell count in the blood, B_0, ranges from 0 to 10 million. Thus, $B_0 = 10^6$ is chosen.

Table 1: Parameters and descriptions of parameters in equations (4)-(6).

$\boldsymbol{\theta}$	**Definition**	**Chosen Values**	**Units**
β_{Ext}	Rate at which T-cells exit blood	0.01	1/day
β_{TB}	Rate at which T-cells exit tumor and enter blood	0.001	1/day
β_{SB}	Rate at which T-cells exit spleen and enter blood	0.0001	1/day
$\boldsymbol{\rho}$	**Transient expansion factor of T-cells in the tumor**	–	**1**
ρ_{UT}	With no treatment	14	1
ρ_{CAR}	With CAR treatment	15	1
ρ_{CXCR1}	With CAR+CXCR1 treatment	20	1
ρ_{CXCR2}	With CAR+CXCR2 treatment	30	1
T_0	Initial condition of T at the first time point ($T_0 = T(t_1)$)	0	T-cells
B_0	Initial condition of B at the first time point ($B_0 = B(t_1)$)	10^6	T-cells
S_0	Initial condition of S at the first time point ($S_0 = S(t_1)$)	0	T-cells

3 Stability Analysis

Before comparing data to a mathematical model, it is important to understand the behavior of the mathematical model, especially the limiting behavior. Although in reality we will not observe the number of T-cells for longer than a few hundred days, it is important to understand what happens to these state values in the limit as time becomes large. To understand the stability of this system, we first examine the Jacobian matrix,

$$J = \begin{bmatrix} -\beta_{TB} & \rho\beta_{Ext} & 0 \\ \beta_{TB} & -\rho\beta_{Ext} - \beta_{Ext} & \beta_{SB} \\ 0 & \beta_{Ext} & -\beta_{SB} \end{bmatrix}. \tag{7}$$

Next, we find the eigenvalues of the Jacobian using the characteristic polynomial,

$$-\lambda^3 + m\lambda^2 + p\lambda + q = 0 \tag{8}$$

where

$$\begin{aligned} m &= -(\beta_{SB} + \beta_{TB} + \rho\beta_{TB} + \beta_{Ext}) \\ p &= \beta_{SB}\beta_{TB} - \beta_{SB}\beta_{Ext} - \beta_{TB}\beta_{Ext} \\ q &= \beta_{Ext}\beta_{SB}\beta_{TB}(\rho - 1). \end{aligned}$$

By Decartes' Rule of Signs, the signs of the eigenvalues, λ, change depending on the value of ρ. We observe that if $m > 0$, then $(\beta_{SB} + \beta_{TB} + \rho\beta_{TB} + \beta_{Ext}) < 0$, which is not plausible since all of our parameters non-negative, so $m \leq 0$. Now, when $q > 0$, then we see that $\rho > 1$. So we will look at long term behavior of our model, holding all other parameters constant, but varying ρ.

In the following figures, we see the different behavior of the system as time goes to infinity. In reality, the data we collect should not span more than 100 days at most, but it is important to know the long-term behavior of our model. In Figures 2 and 3, when $\rho < 1$ and when $\rho = 1$, the number of T-cells in the spleen reach a positive steady state, while the number of T-cells in the blood and tumor spike quickly and then approach zero. T-cells are not travelling to the tumor quickly enough and T-cells are leaving the spleen at a lower rate. In Figure 4, we see that when $\rho > 1$, T-cells in the spleen become unbounded, and T-cells in the blood and tumor approach zero but at a much slower rate.

In Figure 2, we see the behavior of the system for $\rho < 1$ and $B_0 = 10^6$. It should be noted that for all B_0 values, the behavior is the same. Figure 2A shows behavior within 100 days. We see that T-cells in the blood steadily decrease, while T-cells in the tumor rise initially. T-cells in the spleen grow even faster. Figure 2B depicts long term behavior. We see that T-cells in the blood decrease sharply and then go toward zero. T-cells in the spleen rise and appear to go to a steady state. T-cells in the tumor rise initially, but then decline.

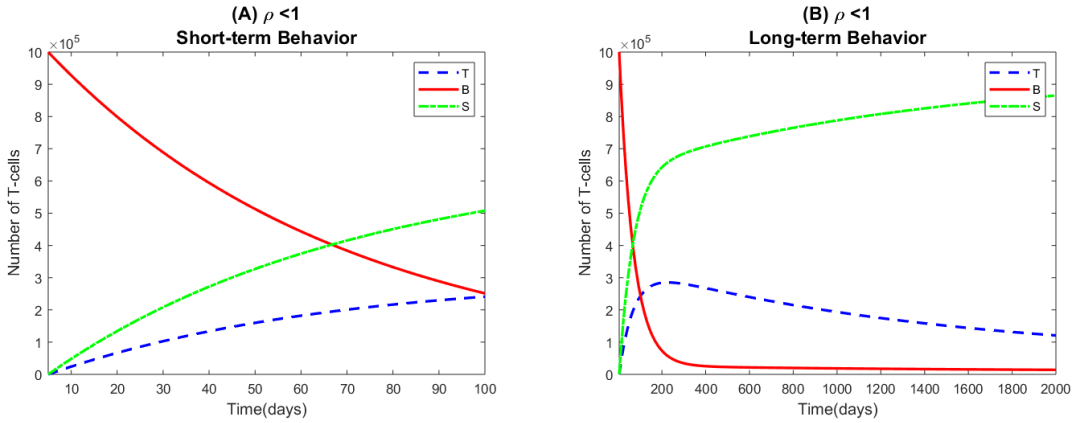


Figure 2: Numerical solutions to our model while setting $\rho = 0.5$. The state variables, T, B, and S are the number of T-cells in the tumor, blood, and spleen, respectively. The other parameter values are fixed at values from Table 1.

In Figure 3, we see the behavior of the system for $\rho = 1$. Figure 3A shows behavior within 100 days. Again, T-cells in the blood decrease. T-cells in the spleen and blood grow almost at an identical rate, although the T-cell rate in the spleen is slightly higher. Figure 3B shows long term behavior. Again, T-cells in the spleen grow considerably and approach a steady state, but we notice that it takes much longer than when $\rho < 1$. Again, T-cells in the tumor decrease and approach zero, and T-cells in the tumor spike initially, before slowly decreasing.

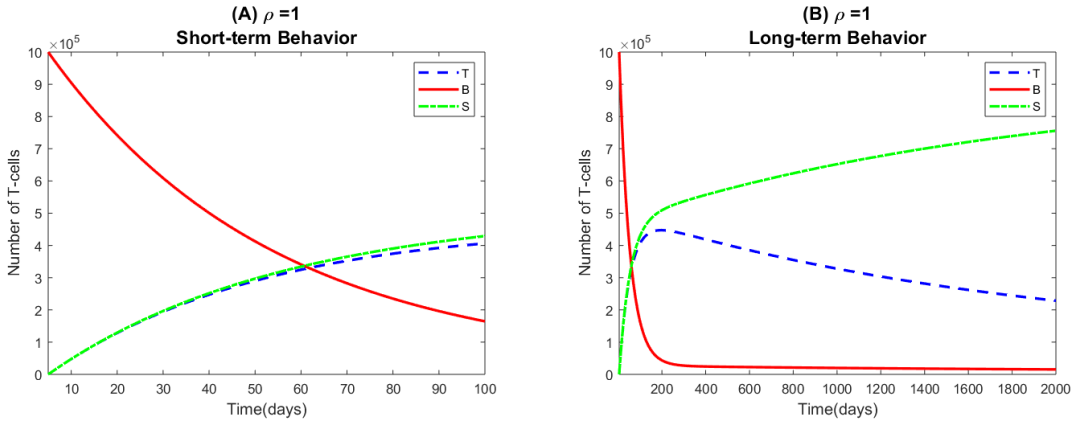


Figure 3: Numerical solutions to our model while setting $\rho = 1$. The state variables, T, B, and S are the number of T-cells in the tumor, blood, and spleen, respectively. The other parameter values are fixed at values from Table 1.

Next we consider the behavior of the model when $\rho > 1$, plotted in Figure 4. Figure 4A displays

short-term behavior. With a value of $\rho = 5$, T-cells leave the blood rapidly and are shuttled to the tumor and spleen. The rise and decline of the T-cells in the tumor and spleen is so slow that it takes much longer to reach a steady state, as can be seen from Figures 4B and 4C.

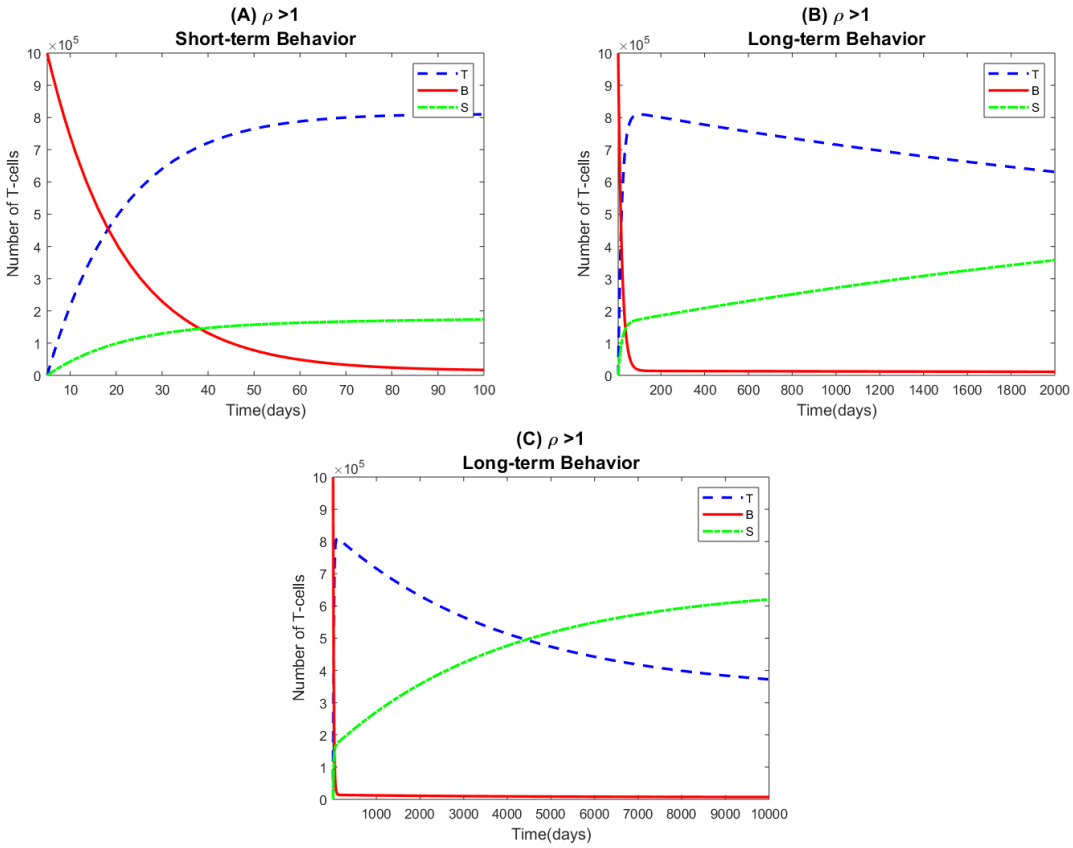


Figure 4: Numerical solutions to our model while setting $\rho = 5$. The state variables, T, B, and S are the number of T-cells in the tumor, blood, and spleen, respectively. The other parameter values are fixed at values from Table 1.

4 Parameter Selection via Sensitivity Analysis

Statistical significance of an inverse problem (fitting data to a mathematical model) depends largely on the significance of the parameter chosen to be estimated. Utilizing sensitivity analysis, we estimate which parameters are most significant in affecting the behavior of the model. That is, parameters with high sensitivities dramatically affect the solution, as the observations (T-cell concentrations in the blood, B, tumor, T, and spleen, S) are most sensitive to those parameters, while parameters with low sensitivities have little influence on the model. The sensitivity of observation f to parameter estimates θ is

$$\chi(\theta, t) = \frac{\partial f(t;\theta)}{\partial \theta} \tag{9}$$

where t is time, and $f = T$, $f = B$, or $f = S$. Since many of the parameters have different orders of magnitude, (for example, $\rho \approx 10^1$ while $\beta_{SB} \approx 10^{-4}$) it is useful to observe the normalized sensitivities,

$$\chi_n(\theta, t) = \frac{\partial f(t;\theta)}{\partial \theta} \frac{\theta}{f(t;\theta)}. \tag{10}$$

While general sensitivities, χ look at how the data reacts to the parameters as a whole by taking the partial derivative of the output factor with respect to the input factor, normalized sensitivities, χ_n, are scaled down by dividing the derivative by the observation in order to compare each parameter against the other.

For all four categories (UT, CAR, CAR+CXCR1, and CAR+CXCR2) the concentration of T-cells in the tumor, blood, and spleen, are most sensitive to parameters β_{Ext} and ρ. Since graphs of the sensitivities look very similar for the different treatment categories, we only show graphs of the sensitivities for the last treatment, CAR+CXCR2, where $\rho = 30$.

4.1 Sensitivity Analysis with $\rho = 30$ (CAR+CXCR2 treatment)

In order to better compare the sensitivities to each parameter, for each observation (T, B, and S) and parameter (see Table 1) we find the maximum sensitivity and maximum normalized sensitivity over time and plot these results in Figure 5. The full time-dependent sensitivities are plotted in Figures 6 and 7. Since the observations T, B, and S are consistently not sensitive to the initial conditions, sensitivities to T_0, B_0, and S_0 are not plotted in Figures 6 and 7.

The observations of T-cells in the tumor, blood, and spleen in Figure 5A are most sensitive to the T-cell movement rates, β_{Ext}, β_{TB}, and β_{SB}, followed by antigen recognition in the tumor, ρ. The observations of T-cells in the tumor and spleen are most sensitive to parameter β_{Ext}, which controls the rate at which T-cell leave the tumor and spleen. However, the observation of T-cells in the blood, B, is most sensitive to parameter β_{TB}, which controls the rate at which T-cells leave the blood to go to a detected tumor. T-cell counts, are not very sensitive to the initial conditions, B_0, T_0, and S_0. These results are consistent with our model design.

When sensitivities are normalized, the T-cell counts in Figure 5B are most sensitive to the initial T-cell count in the blood, B_0, followed by the rate at which T-cell exit the blood, β_{Ext}, and antigen recognition in the tumor, ρ. Since $B_0 = 10^6$ is so large compared to the other parameters, its normalized sensitivity is over inflated and can be ignored. Thus, according to the normalized sensitivities, T-cell counts are most effected by parameters β_{Ext} and ρ. Since ρ is the transient expansion factor of antigen recognition in the tumor and changes depending on the treatment, we choose to estimate this parameter.

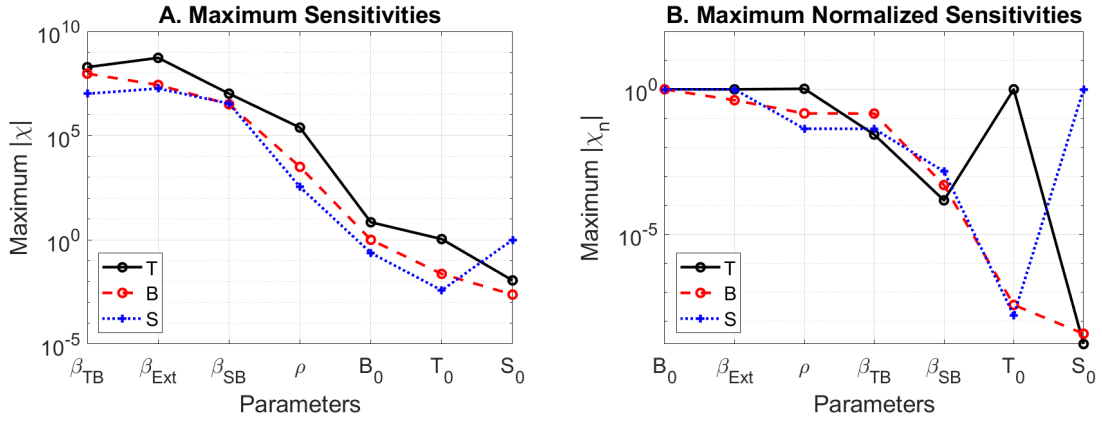


Figure 5: Maximum sensitivities (A) and maximum normalized sensitivities (B) of each observation, T, B, and S, to each of the parameters over a time period of 30 days. The expansion factor of T-cells in the tumor $\rho = 30$, and the initial number T-cells in the tumor, blood, and spleen $[T_0, B_0, S_0] = [1, 10^6, 1]$, since this is data from the CAR+CXCR2 treatment group.

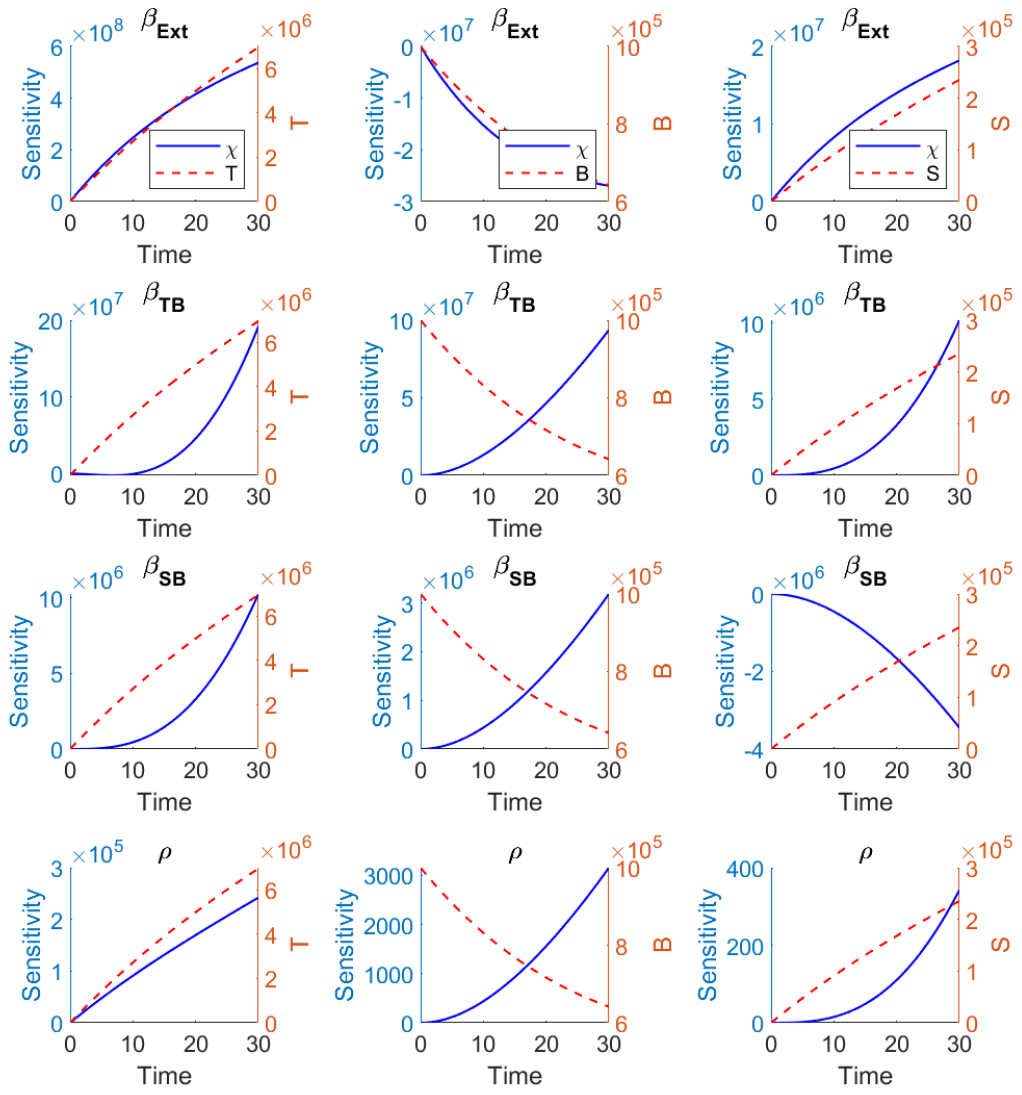


Figure 6: Sensitivities, $\chi(t)$, of observation T, B, and S to parameters β_{Ext}, β_{TB}, β_{SB}, and ρ over a time period of 30 days. These sensitivities are calculated at parameter values from Table 1 with $\rho = 30$ to represent the CAR+CXCR2 treatment.

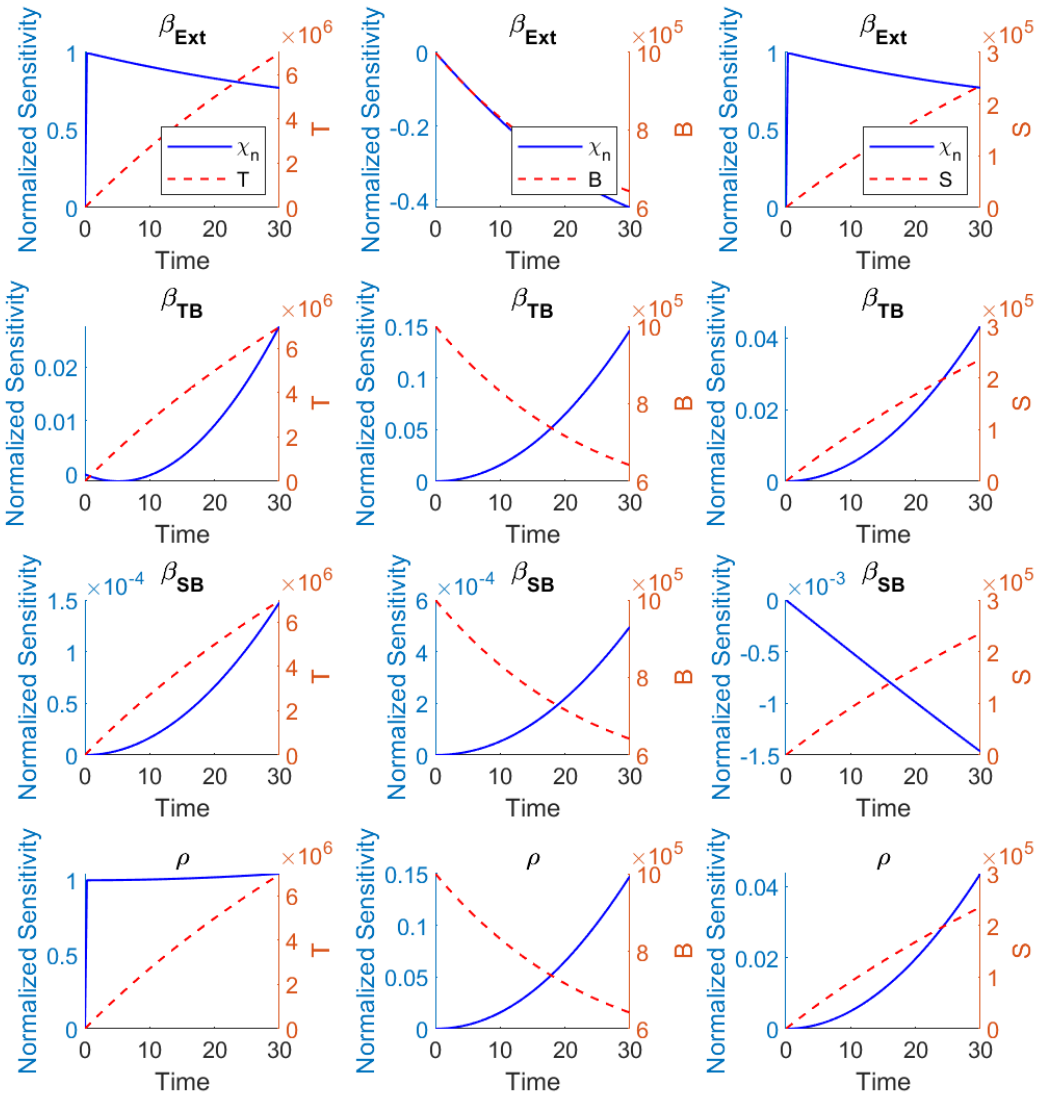


Figure 7: Normalized sensitivities, $\chi_n(t)$, of observation T, B, and S to parameters β_{Ext}, β_{TB}, β_{SB}, and ρ over a time period of 30 days. These sensitivities are calculated at parameter values from Table 1 with $\rho = 30$ to represent the CAR+CXCR2 treatment.

5 Steps to Create an Accurate Inverse Problem

Now that we have ascertained some specific features of the behavior of the mathematical model and its parameters, we carry out a series of inverse problems to estimate the desired parameter ρ for a given number of time observations. Based on the sensitivity analysis and our interest in the different types of treatment, we attempt to estimate the probability distributions for the parameter ρ using observations of the aggregate engineered T-cell concentrations in the tumor T, blood B, and spleen S compartments. However, since currently available data sets contain data at only three distinct time points, our inverse problem might not be feasible. Nonetheless we proceed in our efforts using a rather straightforward if unsophisticated approach to the question of how many mice must be sacrificed to reliably determine the desired parameter distribution for ρ.

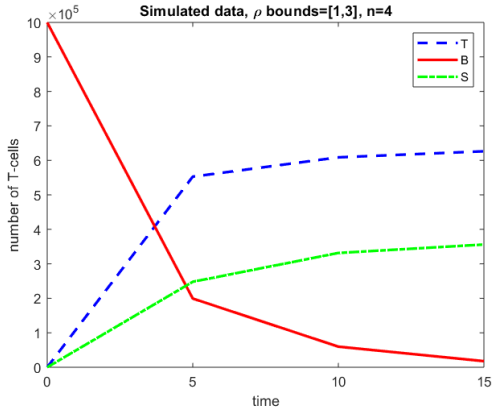
Because the dynamics of the system are difficult to see when ρ is large (since ρ drives the solutions to change very quickly), we looked at a distribution with values in $1 \leq \rho \leq 3$. We then run the inverse problems to solve for a distribution of possible parameter values for ρ.

We first simulate data based on the model, using random values of ρ from a normal distribution with mean $\mu = 2$ and standard deviation $\sigma = 0.2$. Since our actual mice data comes from a time of 5 to 15 days, we look at a time span of 0 to 15 days, assuming the initial conditions from Table 1. We then investigate the results of the inverse problems assuming availability of different numbers of equally spaced time point observations of the aggregate T-cell concentrations in the tumor, blood, and spleen.

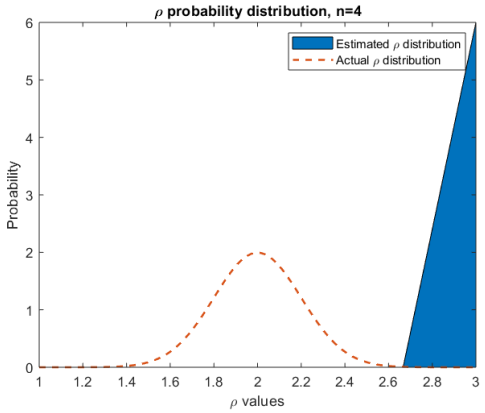
Case 1: $n = 4$ time points

In Case 1, we simulate as few as 4 time points and attempt to estimate the probability distribution. Figure 8(a) shows $n = 4$ time points of simulated aggregate data, which is then used to run the inverse problem in which we assume our parameter ρ is randomly and normally distributed with mean $\mu = 2$ and standard deviation $\sigma = 0.2$. This normal distribution, the "actual" distribution of ρ is graphed in Figure 8(b) along with the estimated distribution from the inverse problem. As we see from Figure 8(b), the estimated ρ distribution, which is shaded in, does not overlap with the "actual" distribution of ρ, which was previously assumed.

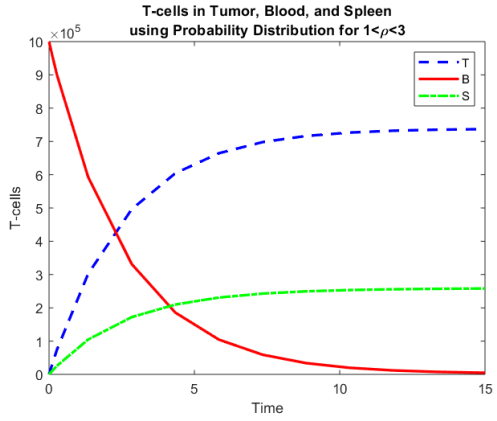
This is most likely due to the fact that 4 time points is too few to glean enough information. As such, we get a skewed distribution of the parameter ρ, which does not provide an accurate result as to the true dynamics of the system. Figure 8(c) shows the aggregate solution curves of T-cell concentrations with the new, estimated distribution. The condition number of the Fisher Information Matrix (FIM) for this inverse problem is 3.54×10^{16}, which is used to determine the uncertainty in the estimated probability distribution and compare the uncertainty in the different cases.



(a) $n = 4$ timepoints of simulated aggregate data, assuming that ρ is normally distributed with mean $\mu = 2$ and standard devation $\sigma = 0.2$.



(b) Estimated probability distribution of ρ, compared to the "actual" normal probability distribution of $\rho \sim \mathcal{N}(\mu = 2, \sigma = 0.2)$.

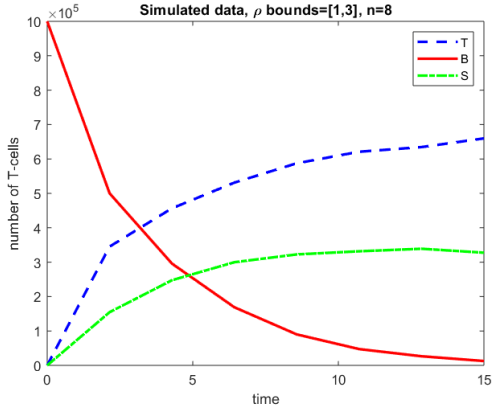


(c) Aggregate solution curves, with ρ set at the probability distribution estimated in (b).

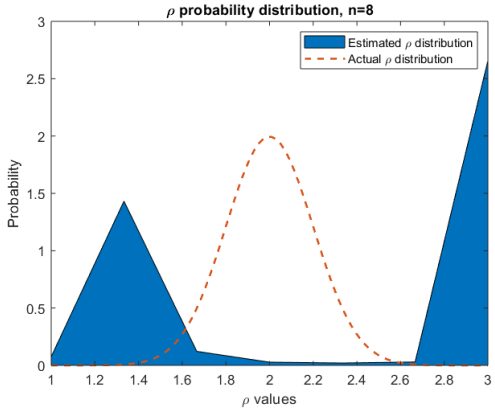
Figure 8: Case 1: $n = 4$ time points of simulated aggregate observations (a) of the number of T-cells in the tumor T, blood B, and spleen S used to estimate the probability distribution of ρ (b), and compared to estimated aggregate observations (c) of T, B, and S given the estimated distribution. All parameter values except for ρ are set at values from Table 1.

Case 2: $n = 8$ time points

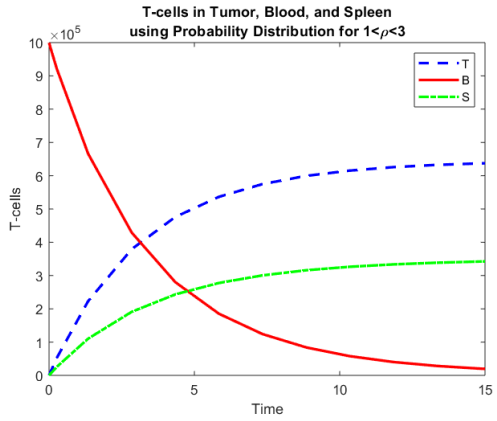
We now consider $n = 8$ time points of simulated aggregate data. In this case, the estimated distribution of ρ is bimodal and still fails to capture the "actual" probability distribution that was assumed, as can be seen in Figure 9. The condition number of the FIM is 6.3×10^{14}, and is much smaller than in Case 1, indicating that the inverse problem is improving.



(a) $n = 8$ timepoints of simulated aggregate data, assuming that ρ is normally distributed with mean $\mu = 2$ and standard devation $\sigma = 0.2$.



(b) Estimated probability distribution of ρ, compared to the "actual" normal probability distribution of $\rho \sim \mathcal{N}(\mu = 2, \sigma = 0.2)$.

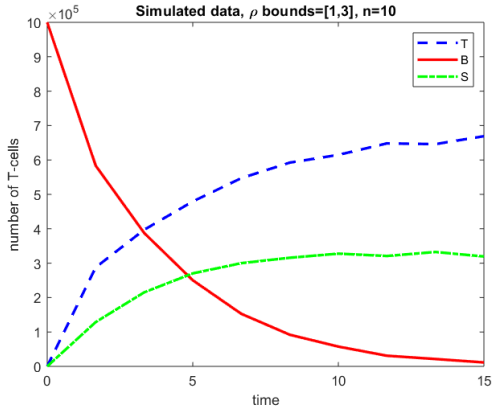


(c) Aggregate solution curves, with ρ set at the probability distribution estimated in (b).

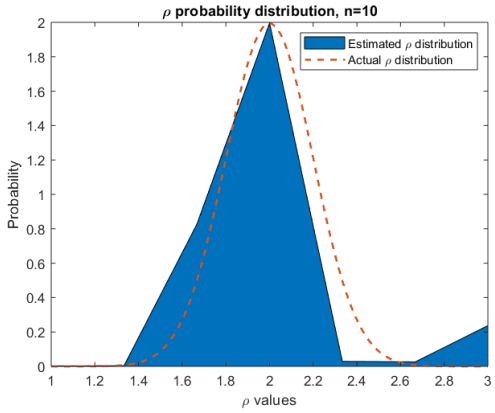
Figure 9: Case 2: $n = 8$ time points of simulated aggregate observations (a) of the number of T-cells in the tumor T, blood B, and spleen S used to estimate the probability distribution of ρ (b), and compared to estimated aggregate observations (c) of T, B, and S given the estimated distribution. All parameter values except for ρ are set at values from Table 1.

Case 3: n=10 time points

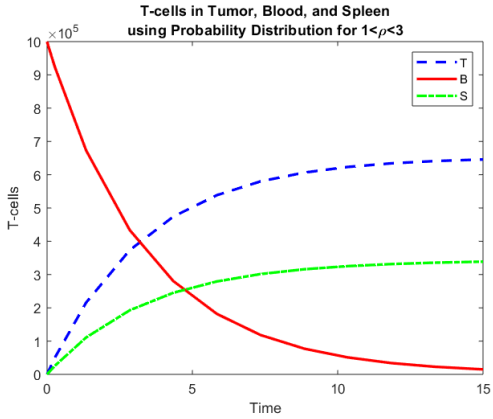
In Case 3, we increase the number of simulated time points to 10. From Figure 10, we see that the estimated probability distribution for the parameter ρ is starting to shift more towards the center of the assumed normal distribution. Even though the estimated distribution is still slightly skewed and bimodal, the estimate is very close to the "actual" distribution. The condition number of the FIM is 5.84×10^{14}, which is slightly smaller compared to the FIM in Case 2, indicating that the inverse problem is still improving.



(a) $n = 10$ timepoints of simulated aggregate data, assuming that ρ is normally distributed with mean $\mu = 2$ and standard devation $\sigma = 0.2$.



(b) Estimated probability distribution of ρ, compared to the "actual" normal probability distribution of $\rho \sim \mathcal{N}(\mu = 2, \sigma = 0.2)$.

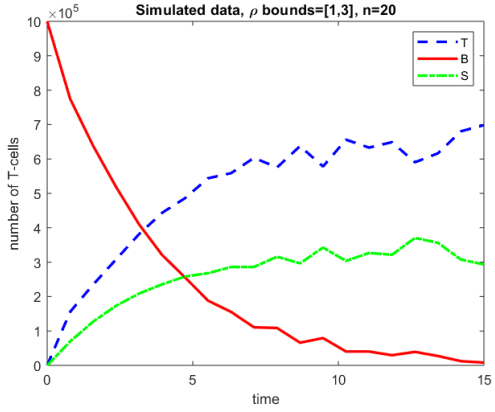


(c) Aggregate solution curves, with ρ set at the probability distribution estimated in (b).

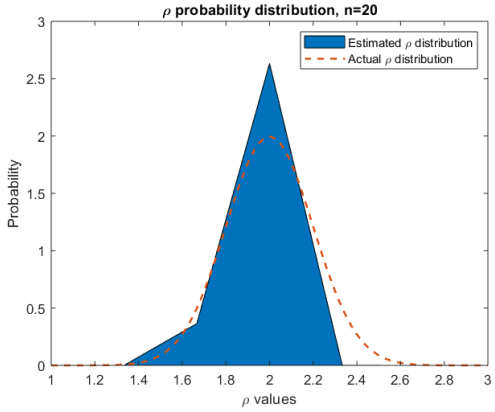
Figure 10: Case 3: $n = 10$ time points of simulated aggregate observations (a) of the number of T-cells in the tumor T, blood B, and spleen S used to estimate the probability distribution of ρ (b), and compared to estimated aggregate observations (c) of T, B, and S given the estimated distribution. All parameter values except for ρ are set at values from Table 1.

Case 4: n=20 time points

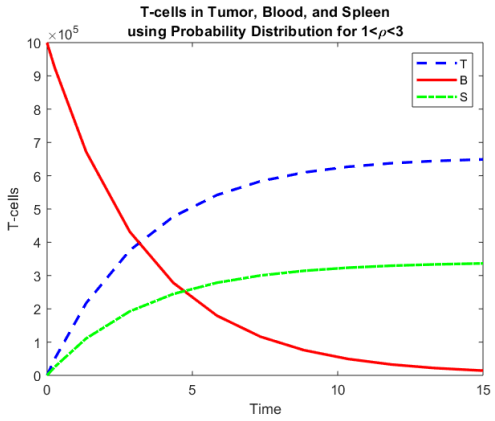
When we consider 20 time points of simulated data, the estimated probability distribution is almost as accurate as a spline estimation can be to the assumed normal distribution. See Figure 11. The condition number of the FIM is 2.43×10^{14}, which is even smaller than the previous case, indicating that the inverse problem is still improving.



(a) $n = 20$ timepoints of simulated aggregate data, assuming that ρ is normally distributed with mean $\mu = 2$ and standard devation $\sigma = 0.2$.



(b) Estimated probability distribution of ρ, compared to the "actual" normal probability distribution of $\rho \sim \mathcal{N}(\mu = 2, \sigma = 0.2)$.

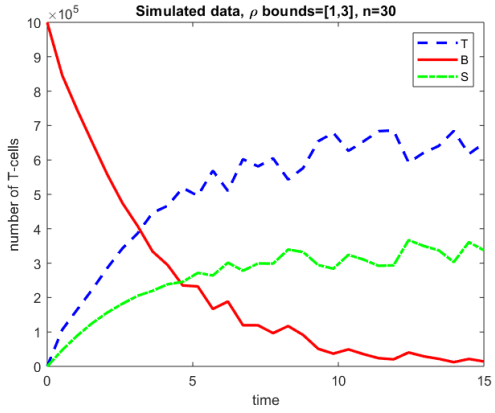


(c) Aggregate solution curves, with ρ set at the probability distribution estimated in (b).

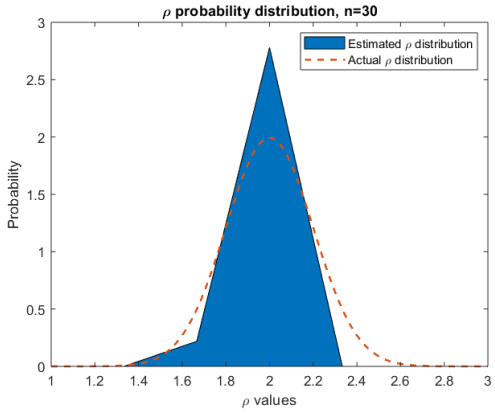
Figure 11: Case 4: $n = 20$ time points of simulated aggregate observations (a) of the number of T-cells in the tumor T, blood B, and spleen S used to estimate the probability distribution of ρ (b), and compared to estimated aggregate observations (c) of T, B, and S given the estimated distribution. All parameter values except for ρ are set at values from Table 1.

Case 5: n=30 time points

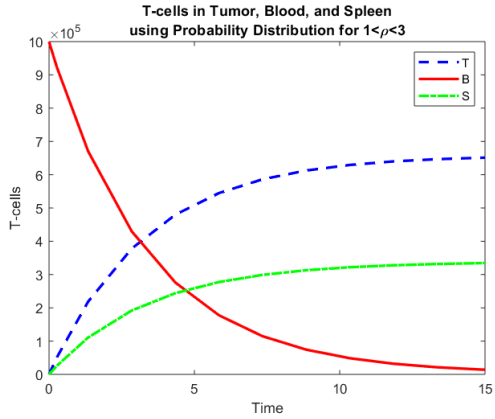
Case 5 utilizes 30 simulated time points of simulated aggregate data, and the estimated distribution in Figure 12(b) is becoming longer and narrower compared to the previous case. That is, the estimated ρ distribution is starting to cluster around $\rho = 2$, with a higher probability of ρ values occurring closest to $\mu = 2$, the mean of the assumed probability distribution. The condition number for the FIM is 2.43×10^{14}, which is the same as in the previous case, indicating that additional time points are no longer significantly improving the results of the inverse problem.



(a) $n = 30$ timepoints of simulated aggregate data, assuming that ρ is normally distributed with mean $\mu = 2$ and standard devation $\sigma = 0.2$.



(b) Estimated probability distribution of ρ, compared to the "actual" normal probability distribution of $\rho \sim \mathcal{N}(\mu = 2, \sigma = 0.2)$.

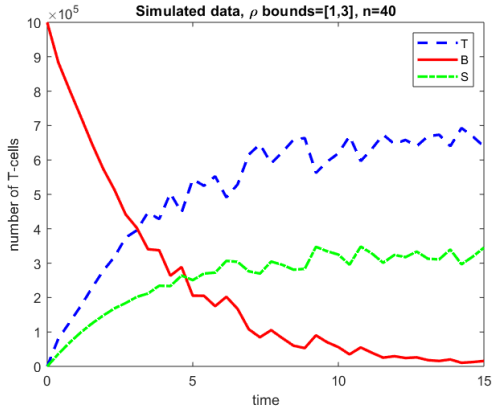


(c) Aggregate solution curves, with ρ set at the probability distribution estimated in (b).

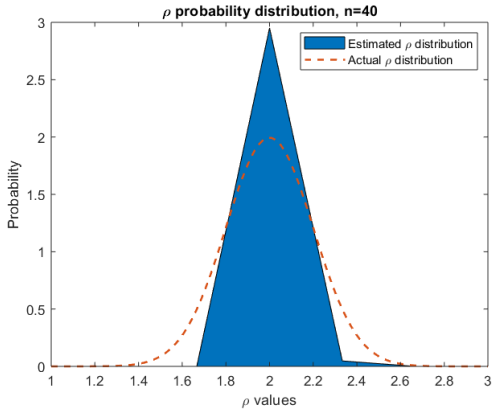
Figure 12: Case 5: $n = 30$ time points of simulated aggregate observations (a) of the number of T-cells in the tumor T, blood B, and spleen S used to estimate the probability distribution of ρ (b), and compared to estimated aggregate observations (c) of T, B, and S given the estimated distribution. All parameter values except for ρ are set at values from Table 1.

Case 6: n=40 time points

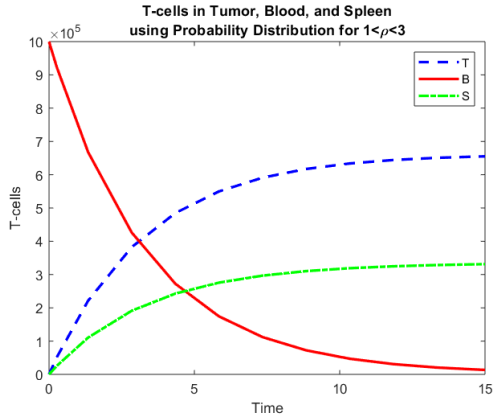
With 40 time points of simulated aggregate data, the highest number of time points we consider, the estimated distribution of ρ is almost symmetrically centered around $\rho = 2$. The estimated probability distribution does not appear to have significantly improved compared to the previous two cases. Furthermore, the condition number for the FIM is 2.06×10^{14}, which is still very close to the value in the previous case, indicating that the inverse problem may not be significantly improved further by adding more time points.



(a) $n = 40$ timepoints of simulated aggregate data, assuming that ρ is normally distributed with mean $\mu = 2$ and standard devation $\sigma = 0.2$.



(b) Estimated probability distribution of ρ, compared to the "actual" normal probability distribution of $\rho \sim \mathcal{N}(\mu = 2, \sigma = 0.2)$.



(c) Aggregate solution curves, with ρ set at the probability distribution estimated in (b).

Figure 13: Case 6: $n = 40$ time points of simulated aggregate observations (a) of the number of T-cells in the tumor T, blood B, and spleen S used to estimate the probability distribution of ρ (b), and compared to estimated aggregate observations (c) of T, B, and S given the estimated distribution. All parameter values except for ρ are set at values from Table 1.

6 Conclusions and Next Steps

Sensitivity Analysis Takeaway: For all four categories (UT, CAR, CAR+CXCR1, and CAR+CXCR2) the model observations T, B, and S (the number of T-cells in the tumor, blood, and spleen, respectively) are most sensitive to parameters ρ and β_{Ext}. Because of this, we can consider these parameters the most important when comparing the data to the model.

Stability Analysis Takeaway: At different values of ρ, the transient expansion factor of tumor antigen recognition by the immune system, the mathematical model described in (4)-(6) has different long-term behaviors. When $\rho = 1$ (assuming parameter values β_{TB}=0.001, β_{Ext}=0.01, and β_{SB}=0.0001), the number of T-cells in the spleen eventually increase to a steady state, the number of T-cells in the tumor initially grow and then decrease to some constant value, and the number of T-cells in the blood decrease to zero. When $\rho < 1$ (which is biologically irrelavent, but interesting mathematically), the number of T-cells in the spleen grows to some constant, the number of T-cells in the tumor grow initially and then decrease to zero, and the number of T-cells in the blood immediately decrease to zero. When $\rho > 1$ (biologically relavent and most common case), the number of T-cells in the spleen increase without bound, while the T-cells in the tumor and blood decrease slowly. The behavior of T-cell concentrations in biologically relavent case, in which $\rho > 1$, is close to reality. In actuality, our specific model will only consider a time-line of, at most, a few years (several hundred days). However, it is important to understand long term behavior of the system.

Parameter Estimation Takeaway: In order to save on costly experiments, we should use the minimum number of data points necessary to feasibly estimate the probability distribution of our parameter of interest, ρ. Utilizing our estimation methods and the aggregate version of our model with fixed parameters set at biologically relavent values (see Table 1), we find that $n = 20$ time points results in the most accurate estimated distribution of ρ. However, with $n = 10$ time points of aggregate data, we still get relatively accurate results, and collecting 10 time points of this type of data over a 15 day period is much more feasible than collecting 20 time points. With $n < 10$ time points, we have inaccurate results, and with $n > 20$, our results do not significantly improve. If it is feasible to collect more than $n = 10$ time points, however, we may improve our results by increasing the number of splines used to estimate the probability distribution.

Although these methods lead to satisfactory results that inform future data collection, our method of experimental design is still very restricted. (For example, the aggregate time points are assumed to be equally spaced, which limits the possibilities for experimental design.) None of these efforts involve use of the sophisticated optimal experimental design formulations outlined in the Sections below. An obvious next step would include use of these design ideas to attempt to further refine the number of and specific times of the needed observations to successfully carry out the needed distributional estimations with aggregate data.

7 Design of Experiments

We turn to the second question of how to best design experiments to collect data (how much data? and when to collect it?) necessary to validate models in models with only aggregate data available. To this point we have discussed various aspects of uncertainty arising in inverse problem techniques. All discussions have been in the context of *a given set* or *sets* of data carried out under various assumptions on how (e.g., independent sampling, absolute measurement error, relative measurement error) the data were collected. For many years [AB, AD, BW, Fed, Fed2, MS, PB, UA] now scientists (and especially engineers) have been actively involved in designing experimental protocols to best study engineering systems including parameters describing mechanisms. Recently with increased involvement of scientists

working in collaborative efforts with ecologists, biologists, and quantitative life scientists, renewed interest in design of "best" experiments to elucidate mechanisms has been seen [AB]. Thus, a major question that experimentalists and inverse problem investigators alike often face is how to best collect the data to enable one to efficiently and accurately estimate model parameters. This is the well-known and widely studied *optimal design* problem. A rather through review is given in [BHT2014]. Briefly, traditional optimal design methods (D-optimal, E-optimal, c-optimal) [AD, BW, Fed, Fed2] use information from the model to find the sampling distribution or mesh for the observation times (and/or locations in spatially distributed problems) that minimizes a design criterion, quite often a function of the Fisher Information Matrix (FIM). Experimental data taken on this optimal mesh are then expected to result in accurate parameter estimates. We outline a framework based on the FIM for a system of ordinary differential equations (ODEs) to determine *when an experimenter should take samples* and *what variables to measure* when collecting information on a physical or biological process modeled by a dynamical system.

Inverse problem methodologies are often discussed in the context of a dynamical system or mathematical model where a sufficient number of observations of one or more states (variables) are available. The choice of method depends on assumptions the modeler makes on the form of the error between the model and the observations (the statistical model). The most prevalent source of error is observation error, which is made when collecting data. (One can also consider model error, which originates from the differences between the model and the underlying process that the model describes. But this is often quite difficult to quantify.) Measurement error is most readily discussed in the context of statistical models. The three techniques commonly addressed are *maximum likelihood estimation (MLE)*, used when the probability distribution form of the error is known; *ordinary least squares (OLS)*, for error with constant variance across observations; and *generalized least squares (GLS)*, used when the variance of the data can be expressed as a nonconstant function. Uncertainty quantification is also described for optimization problems of this type, namely in the form of observation error covariances, standard errors, residual plots, and sensitivity matrices. Techniques to approximate the variance of the error are also included in these discussions.

In [BHK], the authors develop an experimental design theory using the FIM to identify optimal sampling times for experiments on physical processes (modeled by an ODE system) in which scalar or vector data is taken.

In addition to when to take samples, the question of what variables to measure is also very important in designing effective experiments, especially when the number of state variables is large. Use of such a methodology to optimize what to measure would further reduce testing costs by eliminating extra experiments to measure variables neglected in previous trials [BCK]. In [ABCEKPR], the best set of variables for an ODE system modeling the Calvin cycle is identified using two methods. The first, an ad-hoc statistical method, determines which variables directly influence an output of interest at any one particular time. Such a method does not utilize the information on the underlying time-varying processes given by the dynamical system model. The second method is based on optimal design ideas. Extension of this method is developed in [BR1, BR2]. Specifically, in [BR1] the authors compare the SE-optimal design introduced in [BDEK] and [BHK] with the well-known methods of D-optimal and E-optimal design on a six-compartment HIV model [ABDR] and a thirty-one dimensional model of the Calvin Cycle. Such models where there may be a wide range of variables to possibly observe are not only ideal on which to test the proposed methodology, but also are widely encountered in applications. We turn to an outline of this methodology to make observation for best times and best variables.

8 Mathematical and Statistical Models: Formulation of an Optimal Design Problem

We consider inverse or parameter estimation problems in the context of a parameterized (with vector parameter $\boldsymbol{q} \in \mathbb{R}^{\kappa_q}$) n-dimensional vector dynamical system or **mathematical model**

$$\begin{aligned} \frac{d\boldsymbol{x}}{dt}(t) &= \boldsymbol{g}(t, \boldsymbol{x}(t), \boldsymbol{q}), \quad (11) \\ \boldsymbol{x}(t_0) &= \boldsymbol{x}_0, \quad (12) \end{aligned}$$

with **observation process**

$$\boldsymbol{f}(t; \boldsymbol{\theta}) = \mathcal{C}\boldsymbol{x}(t; \boldsymbol{\theta}), \quad (13)$$

where $\boldsymbol{\theta} = \text{column}\,(\mathbf{q}, \tilde{\mathbf{x}}_0) \in \mathbb{R}^{\kappa_q + \tilde{n}} = \mathbb{R}^{\kappa_\theta}, \tilde{n} \leq n$, and the observation operator $\mathcal{C}$ maps $\mathbb{R}^n$ to $\mathbb{R}^m$. In most of the discussions below we assume without loss of generality that some $\tilde{\boldsymbol{x}}_0$ of the initial values $\boldsymbol{x}_0$ are also unknown.

If we were able to observe all states, each measured by a possibly different sampling technique, then $m = n$ and $\mathcal{C} = \mathbf{I}_n$ would be a possible choice; however, this is most often not the case because of the impossibility of or the expense in measuring all state variables. In other cases we may be able to directly observe only combinations of the states. In the formulation below we will be interested in collections or sets of up to K (where $K \leq n$) one-dimensional observation operators or maps so that sets of $m = 1$ observation operators will be considered. In order to discuss the amount of uncertainty in parameter estimates, one can formulate a **statistical model** of the form

$$\boldsymbol{Y}(t) = \boldsymbol{f}(t; \boldsymbol{\theta}_0) + \boldsymbol{\mathcal{E}}(t), \quad t \in [t_0, t_f], \quad (14)$$

where $\boldsymbol{\theta}_0$ is the hypothesized true values of the unknown parameters and $\boldsymbol{\mathcal{E}}(t)$ is a random vector that represents observation error for the measured variables at time t. Realizations of the statistical model (7) are written

$$\boldsymbol{y}(t) = \boldsymbol{f}(t; \boldsymbol{\theta}_0) + \boldsymbol{\epsilon}(t), \quad t \in [t_0, t_f].$$

It is standard to make certain assumptions:

$$\mathbb{E}(\boldsymbol{\mathcal{E}}(t)) = \mathbf{0}, \quad \text{Var}(\boldsymbol{\mathcal{E}}(t)) = V_0(t) = \text{diag}(\sigma_{0,1}^2(t), \sigma_{0,2}^2(t), \ldots, \sigma_{0,m}^2(t)), \quad t \in [t_0, t_f].$$

It is usual to assume further that

$$\text{Cov}\{\mathcal{E}_i(t), \mathcal{E}_i(s)\} = 0, \quad s \neq t, \text{ and } \text{Cov}\{\mathcal{E}_i(t), \mathcal{E}_j(s)\} = 0, \quad i \neq j, \quad s, t \in [t_0, t_f].$$

When collecting experimental data, it is often difficult to take continuous measurements of the observed variables. Instead, we assume that we have N observations at times t_j, $j = 1, \ldots, N$, $t_0 \leq t_1 < t_2 < \cdots < t_N \leq t_f$. We then write the observation process as

$$\boldsymbol{f}(t_j; \boldsymbol{\theta}) = \mathcal{C}\boldsymbol{x}(t_j; \boldsymbol{\theta}), \quad j = 1, 2, \ldots, N, \quad (15)$$

the discrete statistical model as

$$\boldsymbol{Y}_j = \boldsymbol{f}(t_j; \boldsymbol{\theta}_0) + \boldsymbol{\mathcal{E}}(t_j), \quad j = 1, 2, \ldots, N, \quad (16)$$

with realizations $\boldsymbol{y}_j = \boldsymbol{f}(t_j; \boldsymbol{\theta}_0) + \boldsymbol{\epsilon}(t_j), \quad j = 1, 2, \ldots, N.$

We can use this mathematical and statistical framework to develop a methodology to identify *sampling variables* (for a fixed number K of variables or combinations of variables) and the most

informative *times* (for a fixed number N) at which the samples should be taken so as to provide the most information pertinent to estimating a given set of parameters.

Several methods exist to solve the inverse problem. A major factor in determining which method to use is additional assumptions made about $\boldsymbol{\mathcal{E}}(t)$. It is common practice to make the assumption that realizations of $\boldsymbol{\mathcal{E}}(t)$ at particular time points are independent and identically distributed (i.i.d.). If, additionally, the distributions describing the behavior of the components of $\boldsymbol{\mathcal{E}}(t)$ are known, then a maximum likelihood estimation method may be used to find an estimate of $\boldsymbol{\theta}_0$. On the other hand, if the distributions for $\boldsymbol{\mathcal{E}}(t)$ are not known but the covariance matrix $V_0(t)$ (also unknown) is assumed to vary over time, weighted least squares methods are often used. We propose an optimal design problem formulation using a general weighted least squares criterion.

Let $\mathbb{P}_1([t_0, t_f])$ denote the set of all bounded distributions on the interval $[t_0, t_f]$. We consider the *generalized weighted least squares cost functional* for systems with vector output

$$J_{\mathrm{WLS}}(\boldsymbol{\theta}; \boldsymbol{y}) = \int_{t_0}^{t_f} [\boldsymbol{y}(t) - \boldsymbol{f}(t; \boldsymbol{\theta})]^T V_0^{-1}(t) [\boldsymbol{y}(t) - \boldsymbol{f}(t; \boldsymbol{\theta})] dP_1(t), \tag{17}$$

where $P_1 \in \mathbb{P}_1([t_0, t_f])$ is a general measure on the interval $[t_0, t_f]$. For a given continuous data set $\boldsymbol{y}(t)$, we search for a parameter $\hat{\boldsymbol{\theta}}$ that minimizes $J_{\mathrm{WLS}}(\boldsymbol{\theta}; \boldsymbol{y})$.

We next consider the case of observations collected at discrete times. If we choose a set of N time points $\tau = \{t_j\}_{j=1}^N$, where $t_0 \leq t_1 < t_2 < \cdots < t_N \leq t_f$ and take

$$P_1 = P_\tau = \sum_{j=1}^{N} \Delta_{t_j}, \tag{18}$$

where Δ_a represents the Dirac measure with atom at a, then the weighted least squares criterion (10) for a finite number of observations becomes

$$J_{\mathrm{WLS}}^N(\boldsymbol{\theta}; \boldsymbol{y}) = \sum_{j=1}^{N} [\boldsymbol{y}(t_j) - \boldsymbol{f}(t_j; \boldsymbol{\theta})]^T V_0^{-1}(t_j) [\boldsymbol{y}(t_j) - \boldsymbol{f}(t_j; \boldsymbol{\theta})].$$

Note here we do not normalize the time "distributions" by a factor of $\frac{1}{N}$ so that they are not the usual cumulative distribution functions but would be if we normalized each distribution by the integral of its corresponding density to obtain a true probability measure. A similar remark holds for the "variables" observation operator distributions introduced below where without loss of generality we could normalize by a factor of $\frac{1}{K}$ when using K 1-dimensional sampling maps.

To select a useful distribution of time points and set of observation variables, we introduce the m by κ_θ sensitivity matrices $\frac{\partial \boldsymbol{f}(t;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ and the n by κ_θ sensitivity matrices $\frac{\partial \boldsymbol{x}(t;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ that are determined using the differential operator in row vector form $(\partial_{\theta_1}, \partial_{\theta_2}, \ldots, \partial_{\theta_{\kappa_\theta}})$ represented by $\nabla_{\boldsymbol{\theta}}$ and the observation operator defined in (8)

$$\nabla_{\boldsymbol{\theta}} \boldsymbol{f}(t; \boldsymbol{\theta}) \quad = \mathcal{C} \nabla_{\boldsymbol{\theta}} \boldsymbol{x}(t; \boldsymbol{\theta}). \tag{19}$$

Using the sensitivity matrix $\nabla_{\boldsymbol{\theta}} \boldsymbol{f}(t; \boldsymbol{\theta}_0)$, we may formulate the Generalized Fisher Information Matrix (GFIM). Consider the set (assumed compact) $\Omega_C \subset \mathbb{R}^{1 \times n}$ of admissible observation maps and let $\mathbb{P}_2(\Omega_C)$ represent the set of all bounded distributions P_2 on Ω_C. Then the GFIM may be written

$$F(P_1, P_2, \boldsymbol{\theta}_0) \quad = \quad \int_{t_0}^{t_f} \int_{\Omega_C} \frac{1}{\sigma^2(t, c)} \nabla_{\boldsymbol{\theta}}^T \left(\rfloor \boldsymbol{x}(t; \boldsymbol{\theta}_0) \right) \nabla_{\boldsymbol{\theta}} \left(\rfloor \boldsymbol{x}(t; \boldsymbol{\theta}_0) \right) dP_2(c) dP_1(t).$$

In fact we shall be interested in collections of K 1-dimensional "variable" observation operators and a choice of which K variables provide best information to estimate the desired unknown parameters in a given model. Thus taking K different sampling maps in Ω_C represented by the $1 \times n$-dimensional matrices $\mathcal{C}_k$, $k = 1, 2, \ldots, K$, we construct the discrete distribution on $\Omega_C^K = \bigotimes_{i=1}^K \Omega_C$ (the k-fold cross products of Ω_C)

$$P_{\mathcal{S}} = \sum_{k=1}^{K} \Delta_{\mathcal{C}_k}, \tag{20}$$

where Δ_a represents the Dirac measure with atom at a. Using $P_{\mathcal{S}}$ in (13), one can argue that the GFIM for multiple discrete observation methods taken continuously over $[t_0, t_f]$ is given by

$$F(P_1, P_{\mathcal{S}}, \boldsymbol{\theta}_0) \;\; = \;\; \int_{t_0}^{t_f} \nabla_{\boldsymbol{\theta}}^T \boldsymbol{x}(t; \boldsymbol{\theta}_0) \left(\mathcal{S}^T V_K^{-1}(t) \mathcal{S} \right) \nabla_{\boldsymbol{\theta}} \boldsymbol{x}(t; \boldsymbol{\theta}_0) dP_1(t),$$

where $\mathcal{S} = \text{column}(\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_K) \in \mathbb{R}^{K \times n}$ is the set of observation operators defined above and $V_K(t) = \text{diag}(\sigma^2(t, \mathcal{C}_1), \ldots, \sigma^2(t, \mathcal{C}_K))$ is the corresponding covariance matrix for K 1-dimensional observation operators. Applying the distribution P_τ as described in (11) to the GFIM (13) for discrete observation operators measured continuously yields the discrete $\kappa_\theta \times \kappa_\theta$ Fisher Information Matrix (FIM) for discrete observation operators measured at discrete times

$$F(\tau, \mathcal{S}, \boldsymbol{\theta}_0) \equiv F(P_\tau, P_{\mathcal{S}}, \boldsymbol{\theta}_0) = \sum_{j=1}^{N} \nabla_{\boldsymbol{\theta}}^T \boldsymbol{x}(t_j; \boldsymbol{\theta}_0) \mathcal{S}^T V_K^{-1}(t_j) \mathcal{S} \nabla_{\boldsymbol{\theta}} \boldsymbol{x}(t_j; \boldsymbol{\theta}_0). \tag{21}$$

This describes the amount of information about the κ_θ parameters of interest that is captured by the observed quantities described by the sampling maps $\mathcal{C}_k$, $k = 1, 2, \ldots, K$, defining $\mathcal{S}$, when they are measured at the time points in τ.

The questions of determining the best (in some sense) $\mathcal{S}$ and τ are the important questions in the optimal design of an experiment. Note that the set of time points τ has an associated distribution $P_\tau \in \widetilde{\mathbb{P}}_1([t_0, t_f])$, where $\widetilde{\mathbb{P}}_1([t_0, t_f])$ is the set of all bounded discrete distributions on $[t_0, t_f]$. Similarly, the set of sampling maps $\mathcal{S}$ has an associated bounded discrete distribution $P_{\mathcal{S}} \in \widetilde{\mathbb{P}}_2(\Omega_C^K)$. Define the space of bounded discrete distributions $\widetilde{\mathbb{P}}([t_0, t_f] \times \Omega_C^K) = \widetilde{\mathbb{P}}_1([t_0, t_f]) \times \widetilde{\mathbb{P}}_2(\Omega_C^K)$ with elements $P = (P_\tau, P_{\mathcal{S}}) \in \widetilde{\mathbb{P}}$. We assume, without loss of generality, that $\Omega_C^K \subset \mathbb{R}^{K \times n}$ is closed and bounded, and assume that there exists a functional $\mathcal{J} : \mathbb{R}^{\kappa_\theta \times \kappa_\theta} \to \mathbb{R}^+$ of the GFIM (15). Then the **optimal design problem** associated with $\mathcal{J}$ is selecting a discrete distribution $\hat{P} \in \widetilde{\mathbb{P}}([t_0, t_f] \times \Omega_C^K)$ such that

$$\mathcal{J}\left(F(\hat{P}, \boldsymbol{\theta}_0) \right) = \min_{P \in \widetilde{\mathbb{P}}([t_0, t_f] \times \Omega_C^K)} \mathcal{J}\left(F(P, \boldsymbol{\theta}_0) \right), \tag{22}$$

where $\mathcal{J}$ depends continuously on the elements of $F(P, \boldsymbol{\theta}_0)$.

The Prohorov metric discussed in [BHT2014, B2012] provides a general theoretical framework for the existence of $\hat{P}$ and approximation in $\mathbb{P}([t_0, t_f] \times \Omega_C)$ (a general theoretical framework with proofs is developed in [BBi, BDEK]). The application of the Prohorov metric to optimal design problems formulated as (16) is explained more fully in [BDEK, BHK]. For example, one can argue

an optimal distribution P^* exists in $\widetilde{\mathbb{P}}([t_0, t_f] \times \Omega_C^K)$ and may be approximated by an optimal discrete distribution $\hat{P}$ in $\widetilde{\mathbb{P}}([t_0, t_f] \times \Omega_C^K)$.

As explained in [BHT2014], in SE-optimal design, the cost functional $\mathcal{J}_{SE}$ is a sum of the elements on the diagonal of $(F(\tau, \mathcal{S}, \boldsymbol{\theta}_0))^{-1}$ weighted by the respective parameter values [BDEK, BHK], written

$$\mathcal{J}_{SE}(F) = \sum_{i=1}^{\kappa_\theta} \frac{(F(\tau, \mathcal{S}, \boldsymbol{\theta}_0)))^{-1}_{i,i}}{\theta_{0,i}^2}.$$

Thus in SE-optimal design, the goal is to minimize the standard deviation of the parameters, normalized by the true parameter values. As the diagonal elements of F^{-1} are all positive and all parameters are assumed non-zero in $\boldsymbol{\theta} \in \mathbb{R}^{\kappa_\theta}$, $\mathcal{J}_{SE} : \mathbb{R}^{\kappa_\theta \times \kappa_\theta} \to (0, \infty)$.

in [BHK], it is shown that the D-, E-, and SE-optimal design criteria select different time grids and in general yield different standard errors. As we might expect these design cost functionals will also generally choose different observation variables (maps) [BR1] in order to minimize different aspects of the confidence interval ellipsoid.

Acknowledgements

This research was supported in part by the Air Force Office of Scientific Research under grant number AFOSR FA9550-18-1-0457.

References

[CAR] D. Abate-Daga and Marco L Davila, CAR models: next-generation CAR modifications for enhanced T-cell function, Mol Ther Oncolytics May 18;3(2016) :16014. doi: 10.1038/mto.2016.14. eCollection 2016.

[Stenosis1] H.T. Banks, J.H. Barnes, A. Eberhardt, H. Tran and S. Wynne, Modeling and computation of propagating waves from coronary stenosis, Comput. Appl. Math., 21 (2002), 767–788.

[BanksBortz] H. T. Banks and D.M. Bortz, Inverse problems for a class of measure dependent dynamical systems, J. Inverse and Ill-posed Problems, 13 (2005), 103–121.

[BBH] H.T. Banks, D.M. Bortz and S.E. Holte, Incorporation of variability into the mathematical modeling of viral delays in HIV infection dynamics, Mathematical Biosciences, 183 (2003), 63–91.

[BBPP] H.T. Banks, D.M. Bortz, G.A. Pinter and L.K. Potter, Modeling and imaging techniques with potential for application in bioterrorism, CRSC-TR03-02, January 2003; Chapter 6 in Bioterrorism: Mathematical Modeling Applications in Homeland Security (H.T. Banks and C. Castillo-Chavez, eds.), Frontiers in Applied Math, FR28, SIAM, Philadelphia, 2003, 129–154.

[BBKW] H.T. Banks, L.W.Botsford, F. Kappel and C. Wang, Modeling and estimation in size structured population models, LCDS/CCS Rep. 87-13, March, 1987, Brown Univ.; *Proc. 2nd Course on Math. Ecology* (Trieste, December, 1986), World Scientific Press, Singapore (1988), 521-541.

[BBKW1] H. T. Banks, L.W. Botsford, F. Kappel and C. Wang, Estimation of growth and survival in size-structured cohort data: An application to larval striped bass (Morone saxatilis), CAMS Tech. Rep. 89-10, University of Southern California, 1989; J. Math. Biol., 30 (1991), 125–150.

[BF1991] H. T. Banks and B. G. Fitzpatrick, Estimation of growth rate distributions in size-structured population models, CAMS Tech. Rep. 90-2, University of Southern California, January, 1990, Quarterly of Applied Mathematics, 49 (1991), 215–235.

[BFPZ1998] H. T. Banks, B. G. Fitzpatrick, L. K. Potter, and Y. Zhang, Estimation of probability distributions for individual parameters using aggregate population data; CRSC-TR98-6, January, 1998; In Stochastic Analysis, Control, Optimization and Applications, (W. McEneaney, G. Yin, and Q. Zhang, eds.), Birkhauser, Boston, 1999, pp. 353-371.

[BDEHAD] H.T. Banks, J.L. Davis, S.L. Ernstberger, S. Hu, E. Artimovich and A.K. Dhar, Experimental design and estimation of growth rate distributions in size-structured shrimp populations, CRSC-TR08-20, November, 2008; Inverse Problems, 25 (2009), 095003 (28 pp), Sept.

[BKTReview] H.T. Banks, Z.R. Kenz and W.C. Thompson, A review of selected techniques in inverse problem nonparametric probability distribution estimation, CRSC-TR12-13, May 2012; J. Inverse and Ill-Posed Problems, 20 (2012), 429–460

[BG1] H.T. Banks and N.L. Gibson, Well-posedness in Maxwell systems with distributions of polarization relaxation parameters, CRSC-TR04-01, January, 2004; Applied Math. Letters, 18 (2005), 423–430.

[BG2] H.T. Banks and N.L. Gibson, Electromagnetic inverse problems involving distributions of dielectric mechanisms and parameters, CRSCTR05-29, August, 2005; Quarterly of Applied Mathematics, 64 (2006), 749–795.

[BT2009] H.T. Banks and H.T. Tran, Chapter 9.7 of *Mathematical and Experimental Modeling of Physical and Biological Processes,* CRC Press, Boca Raton, FL, July, 2008, 308pp. published, January 2, 2009.

[B2012] H.T. Banks, Chapter 14.4 of *A Functional Analysis Framework for Modeling, Estimation and Control in Science and Engineering,* Taylor and Frances Publishing, 2012. (258 pages).

[BHT2014] H.T. Banks, Shuhua Hu, and W. Clayton Thompson, Chapter 5 of *Modeling and Inverse Problems in the Presence of Uncertainty,* Taylor and Frances Publishing, (411 pages), April, 2014.

[Stenosis2] H. T. Banks, S. Hu, Z.R. Kenz, C. Kruse, S. Shaw, J.R. Whiteman, M.P. Brewin, S.E. Greenwald and M.J. Birch, Material parameter estimation and hypothesis testing on a 1D viscoelastic stenosis model: methodology, CRSC-TR12-09, April, 2012; J. Inverse and Ill-posed Problems, 21 (2013), 25–57.

[Stenosis3] H.T. Banks, S. Hu, Z.R. Kenz, C. Kruse, S. Shaw, J.R. Whiteman, M.P. Brewin, S.E. Greenwald and M.J. Birch, Model validation for a noninvasive arterial stenosis detection problem, CRSC-TR12-22, December, 2012; Mathematical Biosciences and Engr., 11 (2013), 427–448, doi:10.3934/mbe.2014.11.427.

[BKTReview] H.T. Banks, Z.R. Kenz and W.C. Thompson, A review of selected techniques in inverse problem nonparametric probability distribution estimation, CRSC-TR12-13, May 2012; J. Inverse and Ill-Posed Problems, 20 (2012), 429–460.

[BK1989] H. T. Banks and K Kunisch, *Estimation Techniques for Distributed Parameter Systems*, Birkhasser, Boston, 1989.

[BanksPinter] H.T. Banks and G.A. Pinter, A probabilistic multiscale approach to hysteresis in shear wave propagation in biotissue, CRSC-TR04-03, January, 2004; SIAM J. Multiscale Modeling and Simulation, 3 (2005), 395–412.

[de Vries etal] Gerda de Vries, Thomas Hillen, Mark Lewis, Johannes Muller, and Birgitt Schonfisch, *A Course in Mathematical Biology: Quantitative Modelling with Mathematical and Computational Methods*, SIAM, Philadephia, 2006.

[Rubinow] S. I. Rubinow, *Introduction to Mathematical Biology*, John Wiley & Sons, New York, 1975.

[AB] A.C. Atkinson and R.A. Bailey, One hundred years of the design of experiments on and o the pages of Biometrika, Biometrika, 88 (2001), 53–97.

[ABCEKPR] M. Avery, H.T. Banks, K. Basu, Y. Cheng, E. Eager, S. Khasawinah, L. Potter and K.L. Rehm, Experimental design and inverse problems in plant biological modeling, CRSC-TR11-12, October, 2011; J. Inverse and Ill-posed Problems, DOI 10.1515/jiip-2012-0208.

[ABDR] B.M. Adams, H.T. Banks, M. Davidian and E.S. Rosenberg, Model fitting and prediction with HIV treatment interruption data, CRSC TR05-40, October, 2005; Bulletin of Math. Biology, 69 (2007), 563–584.

[AD] A.C. Atkinson and A.N. Donev, *Optimum Experimental Designs*, Oxford University Press, New York, 1992.

[BBi] H.T. Banks and K.L. Bihari, Modeling and estimating uncertainty in parameter estimation, CRSC-TR99-40, December, 1999; Inverse Problems, 17 (2001), 95–111.

[BCK] H.T. Banks, A. Cintron-Arias and F. Kappel, Parameter selection methods in inverse problem formulation, CRSC-TR10-03, revised November 2010; in Mathematical Model Development and Validation in Physiology: Application to the Cardiovascular and Respiratory Systems, Lecture Notes in Mathematics, Vol. 2064, Mathematical Biosciences Subseries, Springer-Verlag, Berlin, 2013.

[BDEK] H.T. Banks, S. Dediu, S.L. Ernstberger and F. Kappel, Generalized sensitivities and optimal experimental design, CRSC-TR08-12, September, 2008, Revised November, 2009; J. Inverse and Ill-posed Problems, 18 (2010), 25–83.

[BHK] H.T. Banks, K. Holm and F. Kappel, Comparison of optimal design methods in inverse problems, CRSC-TR10-11, July, 2010; Inverse Problems, 27 (2011), 075002.

[BR1] H.T. Banks and K.L. Rehm, Experimental design for vector output systems, CRSC-TR12-11, April, 2012; Inverse Problems in Sci. and Engr., (2013), 1–34. DOI: 10.1080/17415977.2013.797973.

[BR2] H.T. Banks and K.L. Rehm, Experimental design for distributed parameter vector systems, CRSC-TR12-17, August, 2012; Applied Mathematics Letters, 26 (2013), 10–14; http://dx.doi.org/10.1016/ j.aml.2012.08.003.

[BW] M.P.F. Berger and W.K. Wong (Editors), *Applied Optimal Designs,* John Wiley & Sons, Chichester, UK, 2005.

[Fed] V.V. Fedorov, *Theory of Optimal Experiments,* Academic Press, New York and London, 1972.

[Fed2] V.V. Fedorov and P. Hackel, *Model-Oriented Design of Experiments,* Springer-Verlag, New York, 1997.

[Moon] Edmund K. Moon, Carmine Carpenito, Jing Sun, L.C. Wang, V. Kapoor, J. Predina, D.J. Powell Jr, J.L. Riley, C.H. June, S. M. Albelda, Expression of a functional CCR2 receptor enhances tumor localization and tumor eradication by retargeted human T cells expressing a mesothelin-specific chimeric antibody receptor, Clinical Cancer Research, 17(14) (2011), 4719-4730.

[MS] W. Muller and M. Stehlik, Issues in the optimal design of computer simulation experiments, Appl. Stochastic Models in Business and Industry, 25 (2009), 163–177.

[PB] M. Patan and B. Bogacka, Optimum experimental designs for dynamic systems in the presence of correlated errors, Computational Statistics and Data Analysis, 51 (2007), 5644–5661.

[UA] D. Ucinski and A.C. Atkinson, Experimental design for time-dependent models with correlated observations, Studies in Nonlinear Dynamics and Econometrics, 8(2) (2004), Article 13: The Berkeley Electronic Press.