International Journal of Pure and Applied Mathematics
Volume 109 No. 4 2016, 993-1014
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
doi: 10.12732/ijpam.v109i4.19

ANALYSIS OF A QUASI-CHEMICAL MODEL FOR BACTERIAL GROWTH

H.T. Banks[1 §], W.J. Browning[2], J. Catenacci[3], T.E. Wood[4]

[1,3]Center for Research in Scientific Computation
North Carolina State University
Raleigh, NC 27695-8212, USA

[3,4]Applied Mathematics, Inc.
Gales Ferry, CT 06335-0637, USA

Abstract: We analyze a quasi-chemical model for bacterial growth in the context of a parameter estimation problem using available experimental data sets. For many of the data sets the dynamics are simple and we show that the quasi-chemical model (QCM) is over parameterized in these cases. This results in larger uncertainty in the estimated parameters and is some cases instability in the inverse problem. We illustrate methods for reducing the uncertainty present in the estimated model parameters. We first consider a model reduction technique where subsets of the QCM parameters are fixed at nominal values and hypothesis testing is used to compare the nested models. An orthogonal decomposition of the sensitivity matrix is used to guide the choice of which parameters are fixed. Additionally, surrogate models are developed to compare to the QCM. In most cases one of the surrogate models is able to fit the data nearly as well as the QCM model while reducing the uncertainty in the model parameters.

Key Words: quasi-chemical model, parameter estimation, inverse problem, parameter reduction, parameter ranking, uncertainty quantification

1. Introduction

Predictive modeling of bacterial dynamics is an important step in the under-

Received: July 12, 2016
Revised: August 1, 2016
Published: October 8, 2016

§Correspondence author

standing and prediction of bacteria growth on food in various environmental settings. High-pressure processing (HPP) is one of the most widely used non-thermal food processing techniques. HPP processing results in different bacteria growth dynamics than would result from traditional thermal or chemical processing techniques. Often, this results in complex survival curves with at least one change in concavity [12]. This has led to the development of more sophisticated models, including a class of quasi-chemical models (QCM), that are capable of capturing the effects of microbial lag, inactivation and tailing [11, 8, 9]. These models have successfully been used to fit laboratory data through the means of an inverse problem. However, to the authors' knowledge, little effort has been made to account for the uncertainty present in the estimated model parameters. Using the data provided in [8] we will use various models to preform the inverse problem and quantify the uncertainty in the estimated parameters by constructing asymptotic confidence intervals.

In this work we will focus on the QCM originally derived in [11], but these results can be readily extended to other quasi-chemical models. It has been observed that the inverse problem using this model is ill-posed with the presence of many local minima. Additionally, there are subsets of parameter values which cause the differential equations to grow so rapidly that many standard ode solvers fail. Our focus in this work is not to improve the robustness of the optimization using the current bacterial growth model. Rather, we aim to illustrate that for some data sets the resulting estimates have very large uncertainty bounds. We will consider methods aimed at improving the confidence in the estimated parameters. One such method is a model reduction in which the number of estimated parameters is reduced. We will detail this approach in Section 2.1. For some of the data sets where the uncertainty is large, we will propose simple models capable of achieving similar model fits as the bacterial growth model, but have better uncertainty properties. Since our goal is not to develop a robust optimization of the bacterial growth model, we will typically choose a starting point for the optimization which is near a local minima that obtains a model fit similar to what is shown in [8]. Additionally, for a particular data set, if the optimization routine searches a region of parameter space which causes the model to grow unboundedly, we will enforce *ad hoc* bounds on the parameters.

1.1. Quasi-Chemical Model

The original quasi-chemical model proposed in [11, 8] is given by

$$\begin{aligned}
\frac{dM}{dt} &= -k_1 M \\
\frac{dM^*}{dt} &= k_1 M + (k_2 - k_4)M^* - hk_3 M^* A \\
\frac{dA}{dt} &= k_2 M^* - hk_3 M^* A \\
\frac{dD}{dt} &= k_4 M^* + hk_3 M^* A.
\end{aligned} \tag{1.1}$$

In the above equation M is the concentration of lag phase cells, M^* the growth phase cells, A the antagonist cells, and D the dead cells, the values k_j, j = 1, 2, 3, 4, are the rate constants, and h is a scaling factor set to $h = 10^{-9}$. The initial conditions are given by $[M, M^*, A, D]^T(0) = [I, 0, 0, 0]^T$. We can calculate the solution for the lag phase cells as $M(t) = Ie^{-k_1 t}$. Then, the total microbial population is given by

$$U(t) = M(t) + M^*(t) = Ie^{-k_1 t} + M^*(t). \tag{1.2}$$

Notice that the concentration of dead cells D is uncoupled from the other equations, and since we have no measurement data on the number of dead cells, we may ignore this equation. Furthermore, the term $(k_2 - k_4)M^*$ present in the first equation will lead to a correlation between the parameters k_2 and k_4. In consideration of this, we will define $\alpha = k_2 - k_4$. Hence, we now can consider (1.1) as a system of two differential equations given by

$$\begin{aligned}
\frac{dM^*}{dt} &= k_1 Ie^{-k_1 t} + \alpha M^* - hk_3 M^* A \\
\frac{dA}{dt} &= k_2 M^* - hk_3 M^* A \\
M^*(0) &= A(0) = 0.
\end{aligned} \tag{1.3}$$

For notational convenience, let $\mathbf{x} = [M^*, A]^T$ and let $\boldsymbol{\theta} = [k_1, k_2, k_3, \alpha]^T$. Then (1.1) can be written compactly as

$$\frac{d\mathbf{x}}{dt} = \mathbf{f}(\mathbf{x}(t; \boldsymbol{\theta}); \boldsymbol{\theta}), \quad \mathbf{x}(0) = [0, 0]^T, \tag{1.4}$$

where $\mathbf{f}(\mathbf{x}; \boldsymbol{\theta})$ represents the right hand side of (1.3).

The total microbial population can be measured by plate counts that do not distinguish between cells in the growth and lag phases. Furthermore, the data collected is typically presented on a log scale. Thus, we will fit the above model to the log-scaled data. In light of this we assume a statistical model of the form

$$Y_j = g(t_j; \boldsymbol{\theta}_0) + \mathcal{E}_j, \quad j = 1, ..., n, \tag{1.5}$$

where Y_j is a random variable which is composed of the log-scaled total microbial population count, $g(t, \boldsymbol{\theta}) = \log(U(t; \boldsymbol{\theta}))$, at the sampling time t_j with $\boldsymbol{\theta}_0$, the "true" or nominal parameters, and the measurement error $\mathcal{E}_j$. With this assumption, the estimators $\boldsymbol{\theta}$ can be found using an ordinary least squares formulation

$$\boldsymbol{\theta} = \underset{\boldsymbol{\theta} \in \Omega}{\operatorname{argmin}} \sum_{j=1}^{n} \left(Y_j - g(t_j; \boldsymbol{\theta})\right)^2, \tag{1.6}$$

with realizations

$$\widehat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Omega}{\operatorname{argmin}} \sum_{j=1}^{n} \left(y_j - g(t_j; \boldsymbol{\theta})\right)^2, \tag{1.7}$$

where y_j is a realization of Y_j, $j = 1, ..., n$ in (1.5). That is,

$$y_j = g(t_j; \boldsymbol{\theta}_0) + \varepsilon_j, \quad j = 1, ..., n, \tag{1.8}$$

with ε_j a realization of $\mathcal{E}_j$.

2. Sensitivity Analysis and Uncertainty Quantification

In this section we investigate the sensitivity of the QCM with respect to the parameters θ_j, $j = 1, 2, 3$, and 4. Since the model is relatively simple, we can compute the traditional sensitivities using the well-known sensitivity equations, which are given by

$$\frac{d}{dt}\left(\frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}}\right) = \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \frac{\partial \mathbf{x}}{\partial \boldsymbol{\theta}} + \frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}}. \tag{2.1}$$

The individual traditional sensitivity function is then given by

$$\mathbf{s}_{\theta_j}(t) = \frac{d\mathbf{x}}{d\theta_j}. \tag{2.2}$$

Let $\mathbf{s} = [s_{\theta_1}, s_{\theta_2}, s_{\theta_3}, s_{\theta_4}]$ be a 2×4 matrix, then $\mathbf{s}$ is the solution to the matrix system

$$\frac{d\mathbf{s}}{dt} = \frac{\partial \mathbf{f}}{\partial \mathbf{x}} \mathbf{s}(t) + \frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}}, \quad \mathbf{s}(0) = \mathbf{0}_{2\times 4}, \tag{2.3}$$

where

$$\frac{\partial \mathbf{f}}{\partial \mathbf{x}} = \begin{bmatrix} \alpha - hk_3x_2 & -hk_3x_1 \\ k_2 - hk_3x_2 & -hk_3x_1 \end{bmatrix} \tag{2.4}$$

and

$$\frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}} = \begin{bmatrix} Ie^{-k_1t} - k_1Ite^{-k_1t} & 0 & -hx_1x_2 & x_1 \\ 0 & x_1 & -hx_1x_2 & 0 \end{bmatrix}. \tag{2.5}$$

We are also concerned with the sensitivity of the total microbial population $U(t)$. We find that

$$\frac{\partial U}{\partial k_j} = \mathbf{s}_{1j} - Ite^{-k_1t}\delta_{1j}, \tag{2.6}$$

where δ_{1j} is the Kronecker delta function. Since we are fitting the log-scaled total microbial population, we must consider the sensitivity of $g(t;\boldsymbol{\theta})$

$$\frac{\partial g}{\partial \theta_j} = \left(U(t;\boldsymbol{\theta})\ln 10\right)^{-1} \frac{\partial U}{\partial \theta_j}. \tag{2.7}$$

In order to give a level of confidence in the parameter estimates, we will compute the confidence intervals. The standard least squares estimator $\boldsymbol{\theta}$ has the asymptotic properties [4, 6, 7]:

$$\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}_0, \Sigma_0), \tag{2.8}$$

where $\boldsymbol{\theta}_0$ is the nominal or "true" parameter vector, and the 4×4 covariance matrix is given by the approximation

$$\Sigma_0 \approx \sigma_0^2 \left(F^T(\boldsymbol{\theta}_0)F(\boldsymbol{\theta}_0)\right)^{-1}.$$

Here the $n \times 4$ sensitivity matrix is given by

$$F(\boldsymbol{\theta}) = \left(\frac{\partial g(t_i;\boldsymbol{\theta})}{\partial \theta_j}\right). \tag{2.9}$$

Since $\boldsymbol{\theta}_0$ is unknown, we will use the estimates

$$\Sigma_0 \approx \widehat{\Sigma} = \widehat{\sigma}_0^2 \left(F^T(\widehat{\boldsymbol{\theta}})F(\widehat{\boldsymbol{\theta}})\right)^{-1},$$

with the approximation for the variance

$$\widehat{\sigma}_0^2 = \frac{1}{n-\kappa} \sum_{j=1}^{n} \left(y_j - g(t_j;\widehat{\boldsymbol{\theta}})\right)^2, \tag{2.10}$$

where κ is the number of estimated parameters.

We can then construct the $100(1-\rho)\%$ level confidence intervals by

$$\left[\widehat{\theta}_j - t_{1-\rho/2} SE_j,\ \widehat{\theta}_j + t_{1-\rho/2} SE_j\right], \tag{2.11}$$

where $SE_j = \sqrt{\widehat{\Sigma}_{jj}}$, and the critical value $t_{1-\rho/2}$ is determined by $\text{Prob}\{S > t_{1-\rho/2}\} = \rho/2$, where S has a students's t distribution with $n-\kappa$ degrees of freedom.

2.1. Model Reduction via Parameter Ranking and Model Comparison Testing

As we will see in the next section, the data sets typically do not support a reliable estimation of all of the model parameters. For this reason we will perform a parameter ranking based on the orthogonalization of the sensitivity matrix $F(\boldsymbol{\theta})$. This is implemented by using an "economy size" QR decomposition of the matrix F, so that $FP = QR$, where P is a permutation matrix. The order of the permutations give the ranking of the parameters where the rankings are chosen according the 2-norm of the sensitivity with respect to the parameter (see [10] for a detailed description). In [1, 5] global techniques similar to those presented here were considered. At this time, we do not have reliable ranges for the parameter values, and so we do not consider global methods.

Our goal is to fit the data with the minimal set of parameters that can be reliably estimated. To accomplish this we will take the following approach. We will first estimate only the most important parameter (determined from the sensitivity-based parameter ranking scheme) and fix the three remaining parameters at a nominal value. Then we will estimate the two most important parameters and use a statistically-based model comparison test to determine if the data can be adequately described by the single parameter model compared to the two parameter model. If it is determined that the second parameter significantly improves the model fit to the data, then we will estimate the three most important parameters and compare with the two parameter model, again using the a model comparison test. In this way we have an automated method for ranking the parameters and determining the minimal set of parameters needed to describe the data.

Since we are comparing nested models, we are testing the null hypothesis that the constrained model provides an adequate fit to the data. Let $J(\boldsymbol{\theta}; \mathbf{y})$ and $J(\boldsymbol{\theta}_H; \mathbf{y})$ denote the value of the objective function, where $\boldsymbol{\theta}_H \subset \boldsymbol{\theta}$ and $\dim(\boldsymbol{\theta} - \boldsymbol{\theta}_H) = 1$ (i.e. $\boldsymbol{\theta}_H$ contains 1 less parameter than $\boldsymbol{\theta}$). Then the test

statistic can by computed by

$$T(\mathbf{y}) = J(\boldsymbol{\theta}_H; \mathbf{y}) - J(\boldsymbol{\theta}; \mathbf{y}) \geq 0,$$

and we define

$$\mathcal{U}(\mathbf{y}) = \frac{nT(\mathbf{y})}{J(\boldsymbol{\theta}; \mathbf{y})}.$$

Then, it is known [4] that $\mathcal{U}(\mathbf{y})$ converges to $\mathcal{U}$ in distribution, where $\mathcal{U} \sim \chi^2(1)$, $\chi^2(1)$ a chi-square distribution with 1 degree of freedom. Thus, if the statistic $\mathcal{U}(\mathbf{y}) > \tau$, then we reject the null hypothesis as false with confidence level $(1-\beta)100\%$, where $\text{Prob}\{\chi(1) > \tau\} = \beta$.

3. Aggregate vs. Individual Data

Here we discuss potential issues that can occur with mathematical modeling and estimation in connection with bacteria kinetics experiments. Consider an experiment in which independent samples of inoculated food are prepared and incubated. At each sample time, a sample is extracted and the bacteria are recovered and plate counted. If the sample is discarded after enumeration, then we are not repeatedly measuring the same sample over time. This type of data is called *aggregate data* which is similar to the type of data one would obtain in ecological catch and release experiments see [4, Chapter 5]. Repeated measurements of the same sample over a period of time would result in *individual data*. Aggregate data is often treated as individual data, and modeling and estimation techniques derived for the purpose of data assimilation with individual data are applied. When this is done, the effects of variability across samples is ignored, which can lead to over confidence in estimation results and overly conservative predictions. Techniques designed to handle aggregate data have been developed [4], but require sufficient replicates at each sampling time to be collected which can greatly increase the cost of an experiment. Examples of the pitfalls which can occur when aggregate data is treated as individual data, and a more detailed discussion on aggregate data approaches applied in a pharmacokinetic setting can be found in [2]. We briefly discuss potential problems that may arise when utilizing an individual data-based mathematical modeling framework to describe an aggregate data experiment by considering the following hypothetical scenario.

Suppose we have an experiment involving two different strains of a microorganism with different growth rates. In Figure 1 we present two simulations of the model, in one strain $\alpha = 2.7$ (green circles) and in the other strain $\alpha = 3.7$

(blue dots). In this depiction we have created data which matches our hypothetical scenario.

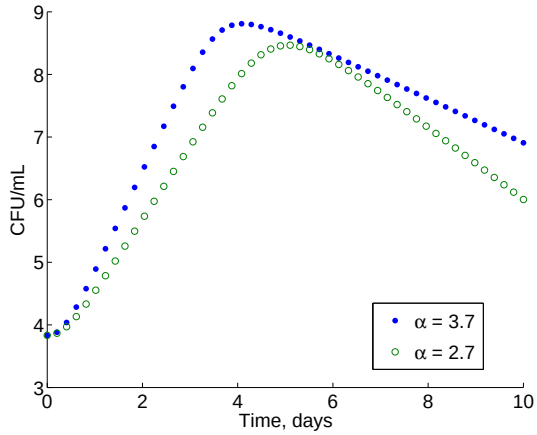

Figure 1: Example of how the data differs using a growth rate of $\alpha = 3.7$ (blue dots) and $\alpha = 2.7$ (green circles).

Assume that there are microorganisms from several different strains, and that any sample is randomly selected from this data. In Figure 2 we depict how a data set collected in this manner might appear. To construct this data set at each time point we randomly select a data point from one of the two trajectories shown in Figure 1. If we look at the data set given in Figure 2, it appears to resemble a noisy data set representing a population with constant growth rate. Therefore, if we use this data set to estimate α, we will estimate a value for α somewhere roughly midway between $\alpha = 2.7$ and 3.7 (the two values used to produce the data set).

Ideally, what we would like to do is be able to use the data shown in Figure 2 to infer that the samples contain microorganisms that grow at two distinctly different rates. In this way, we can consider alpha as an individual based parameter which has an associated distribution. If we impose a distribution on the model parameter α to obtain a model of the form

$$v(t;G) = \int_{\Omega} g(t;\alpha)dG(\alpha), \tag{3.1}$$

we will *not* be able to determine that two growth rates produced the data regardless of the assumptions we place on the distribution $G(\alpha)$. Even if we assume G is a discrete distribution such that the Prob$\{G(\alpha) = 2.7\} =$

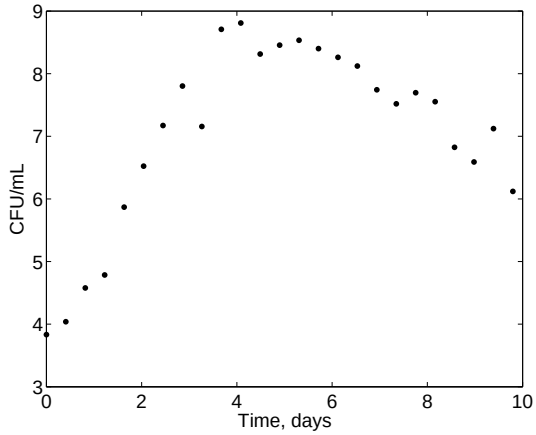

Figure 2: Data set produced by randomly sampling the two trajectories in Figure 1.

Prob$\{G(\alpha) = 3.7\} = 0.5$, the resulting model $v(t; G)$ will only exhibit the *average* behavior. In order to gain more insight into the underlying distribution $G(\alpha)$ we would need to acquire replicate data at each sampling time where the number of replicates counted at each sampling time is sufficient to identify the distribution G (see [2] for more details).

4. Data Fitting Results Using the Quasi-chemical Model

In spite of the above warning regarding potential difficulties with using the data of [8] in estimating parameters with confidence, we also treat the aggregate data as individual data. We investigate the inverse problem using the four parameter QCM given by (1.3) on various data sets taken from [8].

4.1. Inverse Problem

Here we present the results of the inverse problem using the counts of *S. aureus* in bread at a pH of 5.4, temperature of $T = 35°$C, with a wide range of environmental conditions of water activity, a_W, see Figure 1 in [8]. For each data set we give the model fits to the data, the sensitivity of $g(t; \theta)$ with respect to the parameters, and the residual plots (see Figures 3–6). The estimated values for the parameters along with the standard errors (SE) are given in Table 1. While the uncertainty (SE) in the estimated parameters varies widely in the examples,

for each data set we achieve a reasonable model fit to the data, similar to those presented in [8].

To implement the inverse problem, we used the MATLAB routine `fmincon` with the interior-point algorithm, with the function tolerance set to 10^{-13} and the step size tolerance set to 10^{-8}. The system of differential equations (1.3) couple with equations (2.3) was solved using `ode15s` with the relative tolerance set to 10^{-4}. The sensitivity equations (2.3) are used to compute the gradient of the objective function using

$$\nabla J(\theta) = [J_{\theta_1}, ..., J_{\theta_\kappa}]^T,$$
$$J_{\theta_i} = \frac{\partial J(\theta)}{\partial \theta_i} = 2\sum_{j=1}^{n}(g(t_j;\theta) - y_j)\frac{\partial g(t_j;\theta)}{\partial \theta_i},$$

where

$$J(\theta) = \sum_{j=1}^{n}\left(y_j - g(t_j;\theta)\right)^2.$$

The gradient is supplied to the optimization routine `fmincon`.

Below we present the results of inverse problem on a selection of data sets. As we mentioned previously, the goal in the subsequent section is not to improve upon the estimation results given in [8], but to illustrate how for many of the data sets considered the uncertainty present in the estimated values is so large that it is difficult to make confident conclusions based on the estimates. Particularly if one wishes to use the calibrated parameters to provide predictions on either the interior of the sampling domain (interpolation) or outside the sampling domain (extrapolation).

4.1.1. *S. aureus*: $\mathsf{a}_W = 0.79$

For the case of $a_W = 0.79$, from Table 1 we see that the standard errors are unreasonably large for all of the parameters, particularly for k_2 and k_3. This indicates that we have very little confidence in our estimated values for θ. From Figure 3 we see that we have a good fit to the data, however the sensitivity with respect to k_2 and k_3 is near zero for the entire sampling interval. Observe also that the residual plot illustrates an approximately random pattern, indicating that we have correctly specified the statistical model. That is, there is no need at this stage to consider more sophisticated error models, such as a relative error model [4].

For different starting points chosen for the optimization algorithm, the fit of the model to the data was consistent. However, we saw considerable variation

in the parameter estimates of k_2 and k_3 (as much as an order of magnitude). This is consistent with the fact that we have such large standard errors (a direct consequence of the lack of sensitivity). The estimated values for k_1 and α were consistent. The maximum difference between estimated values was observed to be 0.015 for k_1 and 0.002 for α.

For this data set we only have 7 data points (excluding the data point at $t = 0$ which is not considered since we have assumed that the data point at $t = 0$ is the initial condition for M, that is $M(0) = I = y_0$). Therefore, in this case we have less than double the number of data points as parameters, so we cannot reasonably expect to obtain a large degree of confidence in our estimates.

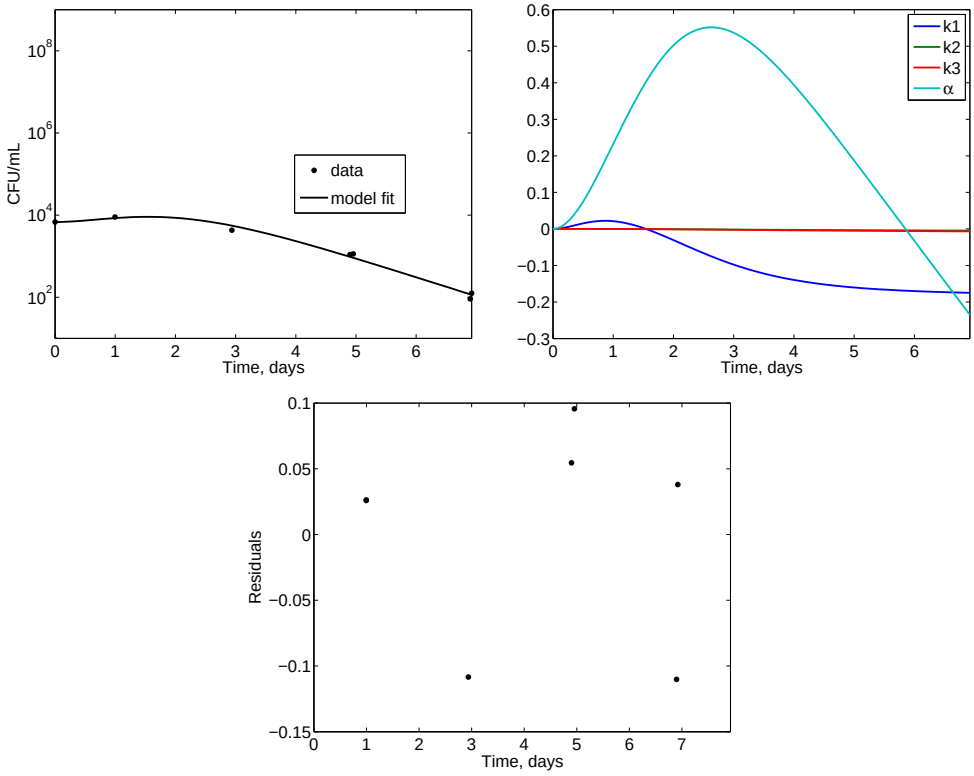

Figure 3: (left, top) The model fit (solid line) to the colony counts (circles) with $a_W = 0.79$, the sensitivity of $g(t; \theta)$ with respect to the parameters (right, top), and the residuals (bottom).

4.1.2. *S. aureus*: $\mathsf{a}_W = 0.84$

For the case of $a_W = 0.84$, we again see from Table 1 that the standard errors are unreasonably large for all of the parameters, particularly for k_2 and k_3. Again indicating that we have very little confidence in our estimated values for θ. In Figure 4 we see that the sensitivity of k_2 and k_3 are again near zero over the entire sampling region, leading to the large standard errors. Here, we obtained consistent estimates with regards to the initial starting values for the optimization.

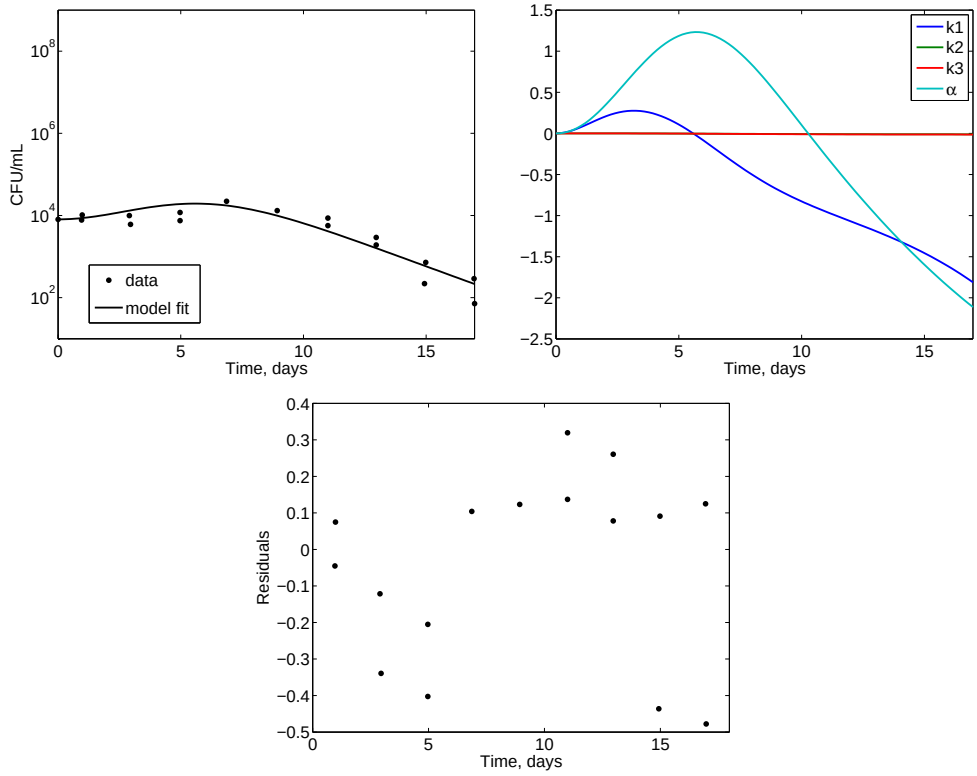

Figure 4: (left, top) The model fit (solid line) to the colony counts (circles) with $a_W = 0.84$, the sensitivity of $g(t;\theta)$ with respect to the parameters (right, top), and the residuals (bottom).

4.1.3. *S. aureus*: $a_W = 0.87$

For the data set with $a_W = 0.87$, the standard errors are much smaller (see Table 1), however they remain close to (larger for k_2 and k_3) the actual values of the parameter estimates. We see from Figure 5 that k_3 now has a nonzero sensitivity at the end of the sampling region. This undoubtedly aids in the reduction of the standard error values. The residual plots also exhibit a independent and identically distributed (i.i.d.) pattern, leading us to again conclude that the statistical model is specified correctly. The estimates were observed to be robust with respect to the initial starting points selected for the optimization.

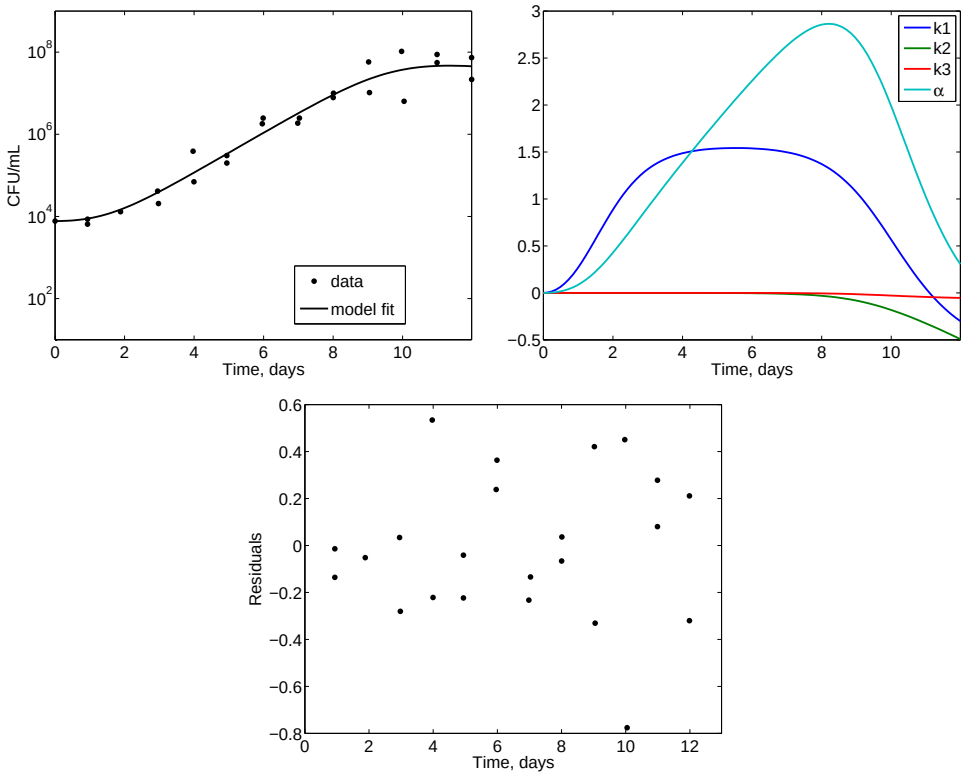

Figure 5: (left, top) The model fit (solid line) to the colony counts (circles) with $a_W = 0.87$, the sensitivity of $g(t;\theta)$ with respect to the parameters (right, top), and the residuals (bottom).

4.1.4. *S. aureus*: $\mathsf{a}_W = 0.91$

For this data set, the inverse problem was found to be severely ill-posed. This was resolved by enforcing a stringent upper bound for α, namely $\alpha \leq 6$. If α was taken to be a value larger than 6, than the system of differential equations, became very stiff, and the integration could not be completed. Previously, all of the estimated values for α were less than 1.2, so we may be justified in our upper bound. Yet, more attention is required on both the biological justification for such a bound, and on understanding the reason for the ill effects observed with this data set. Alternatively, if an initial guess was chosen near the estimated value reported in Table 1, then the inverse problem was carried out with no issue.

In Table 1 we give the results of the inverse problem. For $a_W = 0.91$ the standard error for k_1 is approximately the same value as the k_1. The standard errors for the remaining parameters are reasonable in this case. Overall, the results presented on these data sets indicate that the model *may* be over-parameterized. That is, given the data available, the data sets may not contain enough information in order to estimate all of the parameters with any significant degree of confidence.

	$a_W = 0.79$		$a_W = 0.84$		$a_W = 0.87$		$a_W = 0.91$	
$\boldsymbol{\theta}$	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
k_1	1.609	13.13	0.3712	0.8025	0.2299	0.1901	1.265	1.250
k_2	307.3	5.7×10^5	97.66	1.6×10^4	1.7704	3.2420	4.523	0.3752
k_3	209.5	3.9×10^5	77.06	1.3×10^4	9.7030	28.5216	3.550	0.7740
α	0.5417	1.998	0.4136	0.2976	1.1188	0.1372	3.800	0.3835

Table 1: The parameter estimates and standard error (SE) for each data set.

4.2. Parameter Ranking

In this section we illustrate by example *one* possible technique for obtaining improved uncertainty bounds on the estimated parameters. This will be accomplished by the method outlined in Section 2.1. For all of the following results, the parameters which are set to fixed values are set to be the value of the corresponding starting point used in the optimization.

For the case of $a_W = 0.79$, the parameters were ranked in order of decreasing sensitivity, given as: α, k_1, k_3, k_2. We begin by comparing the situation when only α is estimated compared to estimating both α and k_1. Thus, $\alpha = \theta_H =$

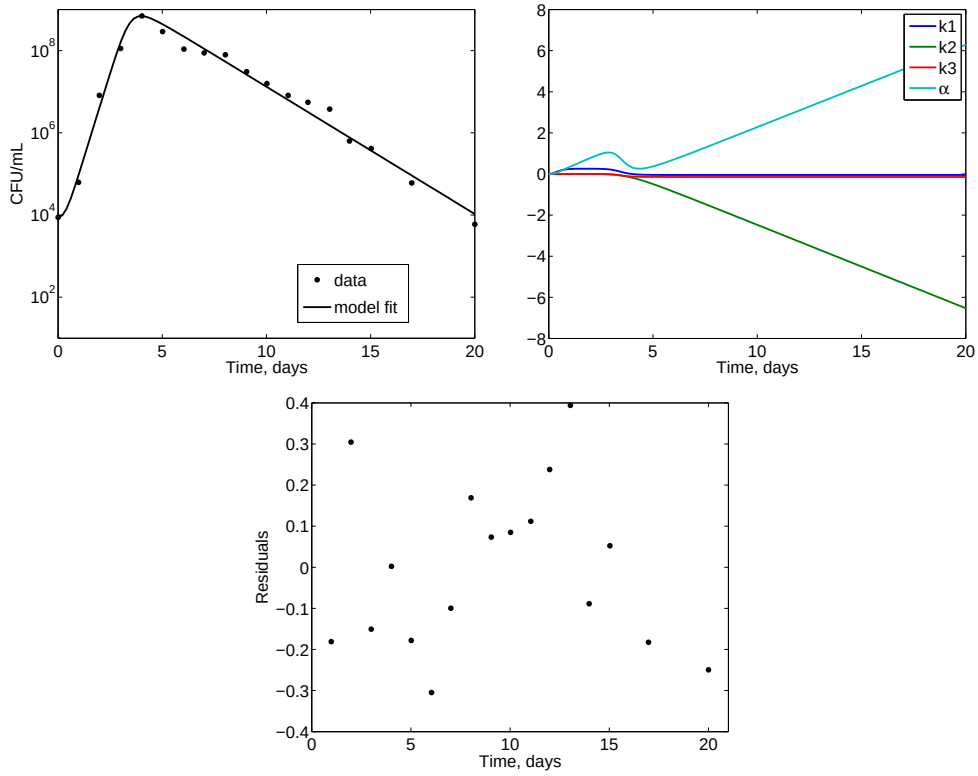

Figure 6: (left, top) The model fit (solid line) to the colony counts (circles) with $a_W = 0.91$, the sensitivity of $g(t;\theta)$ with respect to the parameters (right, top), and the residuals (bottom).

-0.2759 and $(\alpha, k_1) = \theta = (0.5990, 2.4122)$. From Table 2, we see that we reject the null hypothesis at a 99.9% confidence level. That is, the data set *does* support at a significantly statistical level the model with the estimation of both α and k_1.

θ_H	θ	$J(\theta_H;\mathbf{y})$	$J(\theta;\mathbf{y})$	$T(\mathbf{y})$	$\mathcal{U}(\mathbf{y})$
α	α, k_1	0.6036	0.0393	0.5644	100.6
α, k_1	α, k_1, k_3	0.0393	0.0388	4.399×10^{-4}	0.0793

Table 2: Model comparison results for the data set with $a_W = 0.79$.

Next we test to see if the data will support the estimation of the third parameter k_3. Hence, we now compare estimating $(\alpha, k_1) = \theta_H$ and $(\alpha, k_1, k_3) =$

$\theta = (0.5411, 1.591, 215.1)$. From the values presented in Table 2, we see that we fail to reject the null hypothesis. That is, we conclude that the data *does not* support the estimation of the third parameter k_3.

Therefore, we conclude that the model can sufficiently describe this data set where only the parameters α and k_1 are estimated. The standard errors for α and k_1 are 0.0982 and 0.5615 which are now at a much more acceptable level compared to the standard errors obtained using the full four-parameter model.

We now consider the data set where $a_W = 0.84$. In this case the parameters were ranked in decreasing importance given by: α, k_1, k_3, k_2. We again attempt to determine the minimal number of parameters which are needed to describe the data. We first consider only estimating α, compared with estimating both α and k_1. From Table 3, we conclude that we can reject the null hypothesis with a 95% confidence level. Next, we test if the third parameter k_3 provides a significantly better fit to the data. We see from Table 3 that we fail to reject the null hypothesis. Again, we conclude that the data *does not* support the estimation of the third parameter k_3. The estimated values were $\alpha = 0.4082$ and $k_1 = 0.4333$ with standard errors of 0.1037 and 8.9582×10^{-2}, respectively. Again we achieve an increased confidence in the parameter estimates, while sacrificing very little with regards to the model fit to the data.

θ_H	θ	$J(\theta_H; \mathbf{y})$	$J(\theta; \mathbf{y})$	$T(\mathbf{y})$	$\mathcal{U}(\mathbf{y})$
α	α, k_1	1.193	1.015	0.1776	2.799
α, k_1	α, k_1, k_3	1.015	1.007	8.78×10^{-3}	0.1399

Table 3: Model comparison results for the data set with $a_W = 0.84$.

Next we consider the case of $a_W = 0.87$. The parameters were ranked in descending order as: α, k_1, k_2, k_3. For this data set, we reject the null hypothesis until we compare the three and four parameter models, as can be seen in Table 4. Hence, the three-parameter model can adequately describe the data. The estimated values were found to be $\alpha = 1.1199$, $k_1 = 0.2286$, and $k_2 = 1.7384$ with standard errors of 0.1036, 0.1584 and 0.4397, respectively. We again observe that the standard errors decrease using the three-parameter model. The decrease in the standard errors is less significant here, most likely due in part to the fact that we only reduced the model by a single parameter.

For the final case of $a_W = 0.91$, the parameters were ranked in descending order given by: k_2, α, k_3, k_1. From Table 5 it is seen that we fail to reject the null hypothesis when comparing the three and four parameter models. The estimated values were determined to be $k_2 = 4.6079, \alpha = 3,8890$, and $k_3 = 0.7537$ and the standard errors were 0.1544, 0.1568, and 0.7537, respectively.

θ_H	θ	$J(\theta_H;\mathbf{y})$	$J(\theta;\mathbf{y})$	$T(\mathbf{y})$	$\mathcal{U}(\mathbf{y})$
α	$\alpha,\ k_1$	3.905	2.825	1.08	8.791
$\alpha,\ k_1$	$\alpha,\ k_1, k_2$	2.825	2.078	0.7473	8.272
$\alpha,\ k_1, k_2$	$\alpha,\ k_1, k_2, k_3$	2.078	2.076	0.0015	0.0169

Table 4: Model comparison results for the data set with $a_W = 0.87$.

This situation is similar to the $a_W = 0.87$ case in that we do not observe a large decrease in the standard errors.

θ_H	θ	$J(\theta_H;\mathbf{y})$	$J(\theta;\mathbf{y})$	$T(\mathbf{y})$	$\mathcal{U}(\mathbf{y})$
$k_2,\ \alpha$	$k_2,\ \alpha, k_3$	41.22	1.685	39.53	398.8
$k_2,\ \alpha$	$k_2,\ \alpha, k_3$	1.685	0.6571	1.028	26.59
$k_2,\ \alpha, k_3$	$k_2,\ \alpha, k_3, k_1$	0.6571	0.6544	2.754×10^{-3}	0.0715

Table 5: Model comparison results for the data set with $a_W = 0.91$.

One major drawback with this approach is the need for reasonable values for the fixed parameters. Even if a parameter has a low sensitivity, the possible range of parameter values is generally so large that a poor choice for a nominal value will not allow the reduced model to provide a reasonable approximation. Thus, this approach does not alleviate one of the major shortcomings of the QCM which is the sensitivity to the starting point for the optimization in the data fitting procedure. Because of this, we will next focus on alternative (surrogate) models which can be used to describe the same, or similar, dynamics captured by the quasi-chemical model.

4.3. Alternative Candidate Models

Several of the available data sets exhibit primarily only a growth or a death phase. The QCM provides adequate fits to the data in these cases, yet the uncertainty in the estimated parameters becomes extremely large. This is due to the fact that the model is over parameterized with respect to the rather simple dynamics observed in the data. For such cases it may be sufficient to describe the data using a simple exponential model,

$$x(t;\mathbf{q}) = Ie^{at}, \tag{4.1}$$

or an exponential model with a lag phase,

$$x(t;\mathbf{q}) = Ie^{-k_1 t} + \frac{k_1 I}{a + k_1}(e^{at} - e^{-k_1 t}). \tag{4.2}$$

Let $x(t)$ denote the total concentration of cells alive at time t. Note that the exponential lag model can be considered as a subset of the QCM (i.e., set $k_3 = k_2 = 0$).

The basic dynamics that the QCM was derived to model is a lag phase which transitions into a growth phase, and is then followed by a death phase. In this section we propose a simple model derived from first principles which can also be used to describe this two stage growth-death cycle.

Assume that over the interval $0 \leq t \leq \tau$, the cells reproduce at a rate proportional to the total concentration at time t. This gives

$$\begin{aligned} \frac{dx}{dt} &= ax, \quad 0 \leq t \leq \tau \\ x(0) &= x_0, \end{aligned} \tag{4.3}$$

where a is the growth rate, and x_0 is the initial cell concentration. Solving the above equation, we obtain $x(t) = x_0 e^{at}$ for $t \in [0, \tau]$. Now suppose that at time $t = \tau$ the cells enter the death phase. During the death phase we assume that the cells die at a rate proportional to the total concentration at time t. This gives

$$\begin{aligned} \frac{dx}{dt} &= -bx, \quad t > \tau \\ x(\tau) &= x_0 e^{a\tau}, \end{aligned} \tag{4.4}$$

where b is the death rate, and the initial condition at $t = \tau$ is chosen so the the cell concentration is continuous. Putting the two stages together, we arrive at the simple piecewise defined model

$$x(t; \mathbf{q}) = \begin{cases} x_0 e^{at} & 0 \leq t \leq \tau \\ x_0 e^{-b(t-\tau)+a\tau} & t > \tau, \end{cases} \tag{4.5}$$

where $\mathbf{q} = (a, b, \tau)^T$. This model is of course a crude approximation of the bacteria dynamics. However, it is appealing because of its simplicity. Notice that we do not attempt to separate the bacteria concentration into two sub-populations as is done in the QCM. Thus, we take as our initial condition $x_0 = I$.

If we wish to incorporate the dynamic that the cells initially start off in a lag phase, transitioning at a constant rate to the growth phase, we can modify

the two-state exponential model as follows,

$$\begin{aligned} \frac{dx_1}{dt} &= -k_1 x_1 \\ \frac{dx_2}{dt} &= \begin{cases} k_1 x_1 + a x_2 & 0 < t < \tau \\ k_1 x_1 - b x_2 & t \geq \tau \end{cases} \end{aligned} \tag{4.6}$$

where $x_1(0) = I$ and $x_2 = 0$, and we impose that the solution is continuous at $t = \tau$. In this case the total bacteria population is given by $x(t) = x_1(t) + x_2(t)$, which can explicitly be given by

$$x(t;\mathbf{q}) = \begin{cases} Ie^{-k_1 t} + \frac{k_1 I}{a+k_1} e^{at} - \frac{k_1 I}{a+k_1} e^{-k_1 t} & 0 < t < \tau \\ Ie^{-k_1 t} + \left(x_\tau - \frac{k_1 I}{b-k_1} e^{-k_1 \tau} \right) e^{b(\tau - t)} + \frac{k_1 I}{b-k_1} e^{-k_1 t} & t \geq \tau \end{cases} \tag{4.7}$$

where

$$x_\tau = \frac{k_1 I}{a+k_1} e^{a\tau} - \frac{k_1 I}{a+k_1} e^{-k_1 \tau},$$

and $\mathbf{q} = (k_1, a, b, \tau)$. We will refer to this model as the two-stage exponential lag model. This model has the ability to represent many of the same general dynamics as the QCM. Both the two-stage exponential lag model and the QCM have 4 unknown parameters. However, the two-stage exponential lag model has an explicit solution, and has no indication of being ill-posed with regards to the inverse problem.

We can further compare the QCM and all of the surrogate models by using the Akaike Information Criteria (AIC), a model selection criteria which allows for the comparison between models without the necessary condition that the models are nested. The AIC is based on the Kullback-Leibler information and maximum likelihood estimation as a way to balance model bias and variance (see [4] for details). The AIC is given by

$$\mathrm{AIC} = n \log \left(\frac{J(\theta)}{n} \right) + 2(\kappa + 1),$$

where n is the number of observations, J is the residual sum of squares between the model and the data, and κ is the number of estimated parameters in the model. Since we have relatively few data points, we will use the small sample AIC (AIC_c) which is given by

$$\mathrm{AIC}_c = \mathrm{AIC} + \frac{2(\kappa + 1)(\kappa + 2)}{n - \kappa - 2}.$$

For a given data set, among the competing models, the best approximating model is the one with the minimum AIC_c.

In Table 6 we report the frequency that each of the models achieved the minimum AIC_c score for a total of 39 data sets. The exponential model, the 2-stage exponential model, and the QCM model were most commonly found to be the best model according to the AIC_c. When the QCM achieved the minimum AIC_c score using data sets which exhibited primarily only growth or only death dynamics the uncertainty was unreasonably large. However, if the data exhibited both growth and death phases and the QCM was selected as the best model, then it was consistently observed that the uncertainty in the estimated parameters remained reasonable. A detailed discussion regarding the pros and cons of each model applied to the individual data sets is given in [3].

Model	Frequency of minimum AIC_c
Exponential	11
Exponential with lag phase	1
2-stage exponential	14
2-stage exponential with lag	2
QCM	11

Table 6: A summary of the frequency that each of the models achieved the minimum AIC_c value for a total of 39 data sets.

To be clear, we are not proposing that the surrogate models are superior to the QCM. But it is important to acknowledge the shortcomings of the QCM in the context of the given parameter estimation problem. The inverse problem using the QCM is quite ill-posed, and we have seen that for some of the data sets the estimated parameters contain uncertainty bounds many orders of magnitude larger than the estimate. Given the current data sets, it is useful to consider alternative, less complex models. The appeal of the class of quasi-chemical kinetic models is that they are sufficiently robust to model many different experimental results. Yet, the cost of the robustness is often directly seen in large uncertainties in the parameter estimates for cases where the data does not contain dynamics which require fitting by a sophisticated model.

5. Concluding Remarks

In summary, we have analyzed a quasi-chemical kinetic model which provides very good fits to experimental data collected from various bacteria species under

different environmental conditions. However, we have shown that although the QCM model has great flexibility, it often results in an ill-posed problem, or estimated parameter values with little statistical significance. We illustrated that use of statistical model reduction techniques provides a way in which one might reduce the uncertainty in model parameters for cases where the data does not exhibit the complex behavior the QCM was designed to capture. Additionally, we developed surrogate models, which at the very least provide alternatives for the QCM model, and in the case of the so-called hybrid quasi-chemical model developed in [3], might be used to estimate a subset of the QCM model parameters. Indeed focus on the hybrid quasi-chemical model in future efforts may lead to reduced uncertainty and better initial guesses for the associated inverse problems and subsequently less ill-posedness and more readily identifiable parameters.

Acknowledgements

This work has been supported in part by the US Department of Education Graduate Assistance in Areas of National Need (GAANN) under grant number P200A120047, in part by the Air Force Office of Scientific Research under grant numbers AFOSR FA9550-12-1-0188 and AFOSR FA9550-15-1-0298, and in part by the Army Research Office under contract number W911NF-13-P-0017.

References

[1] H.S. Abdel-Khalik, Y. Bang, and C. Wang. Overview of hybrid subspace methods for uncertainty quantification, sensitivity analysis. *Annals of Nuclear Engineering*, 52:28–46, 2013.

[2] H.T. Banks, R. Baraldi, J. Catenacci, and N. Myers. Parameter estimation using unidentified individual data in individual based models. *Center for Research in Scientific Computation Technical Report CRSC-TR16-04, NC State University, Raleigh, NC*, 2016.

[3] H.T. Banks, W.J. Browning, J. Catenacci, and T.E. Wood. Analysis of a quasi-chemical kinetic food chemistry model. *Center for Research in Scientific Computation Technical Report CRSC-TR16-05, NC State University, Raleigh, NC*, 2016.

[4] H.T. Banks, S. Hu, and W.C. Thompson. *Modeling and Inverse Problems in the Presence of Uncertainty*. Taylor/Francis-Chapman/Hall-CRC Press, Boca Raton, FL, 2014.

[5] P.G. Constantine, E. Dow, and Q. Wang. Active subspace methods in theory and practice: applications to kriging surfaces. *SIAM Journal on Scientific Computing*, 36(4):A1500–A1524, 2014.

[6] M. Davidian. Nonlinear models for univariate and multivariate response. `http://www4.stat.ncsu.edu/davidian/courses.html`, 2007.

[7] M. Davidian and D. Giltinan. Nonlinear models for repeated measurement data: An overview and update. *Journal of Agricultural, Biological, and Environmental Statistics*, 8:387–419, 2003.

[8] C.J. Doona, F.E. Feeherry, and E.W. Ross. A quasi-chemical model for the growth and death of microorganisms in foods by non-thermal and high-pressure processing. *International Journal of Food Microbiology*, 100(1):21–32, 2005.

[9] C.J. Doona, F.E. Feeherry, E.W. Ross, and K. Kustin. Inactivation kinetics of listeria monocytogenes by high-pressure processing: Pressure and temperature variation. *Journal of Food Science*, 77(8):M458–M465, 2012.

[10] B.L. Lund, H.E. Berntsen, and B.A. Foss. Methods for parameter ranking in nonlinear, mechanistic models. In *IFAC World Congress, Prague*. Citeseer, 2005.

[11] E.W. Ross, I.A. Taub, C.J. Doona, F.E. Feeherry, and K. Kustin. The mathematical properties of the quasi-chemical model for microorganism growth–death kinetics in foods. *International Journal of Food Microbiology*, 99(2):157–171, 2005.

[12] V. Serment-Moreno, G. Barbosa-Cánovas, J.A. Torres, and J. Welti-Chanes. High-pressure processing: kinetic models for microbial and enzyme inactivation. *Food Engineering Reviews*, 6(3):56–88, 2014.