Applied Mathematics Letters 43 (2015) 10–18

Contents lists available at ScienceDirect

Applied Mathematics Letters

journal homepage: www.elsevier.com/locate/aml

Model comparison tests to determine data information content

H.T. Banks [a,*], J.E. Banks [b], Kathryn Link [a], J.A. Rosenheim [c,d], Chelsea Ross [a], K.A. Tillman [a]

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

[b] *Division of Sciences & Mathematics, School of Interdisciplinary Arts & Sciences, University of Washington, Tacoma, Tacoma, WA 98402, USA*

[c] *Department of Entomology and Nematology, University of California, Davis, Davis, CA 95616, USA*

[d] *Center for Population Biology, University of California, Davis, Davis, CA 95616, USA*

ARTICLE INFO

Article history:
Received 4 November 2014
Accepted 4 November 2014
Available online 28 November 2014

Keywords:
Ordinary least squares
Model comparison in inverse problems
Information content

ABSTRACT

In the context of inverse or parameter estimation problems we demonstrate the use of statistically based model comparison tests in several examples of practical interest. In these examples we are interested in questions related to information content of a particular given data set and whether the data will support a more complicated model to describe it. In the first example we compare fits for several different models to describe simple decay in a size histogram for aggregates in amyloid fibril formation. In a second example we investigate whether the information content in data sets for the pest *Lygus hesperus* in cotton fields as it is currently collected is sufficient to support a model in which one distinguishes between nymphs and adults. Finally in a third example with data for patients having undergone an organ transplant, we question whether the data content is sufficient to estimate more than 5 of the fundamental parameters in a particular dynamic model.

1. Introduction

Uncertainty quantification in the context of estimation of parameters has become a focus of increased attention in recent years. As mathematical models become more complex with multiple states and many parameters to be estimated using experimental data, there is a need for critical analytical tools in model validation related to the reliability of parameter estimates obtained in model fitting. Methodology is desirable to distinguish between lack of identifiability in a model (often formulated in a generalized algebraic context) vs. local insensitivity with respect to changes in particular parameters vs. lack of information content in a given data set. A recent concrete example involves previous HIV models [1,2] with 15 or more parameters to be estimated. In [3], using recently developed parameter selectivity tools [4] based on parameter sensitivity based scores, the authors showed that many of the parameters could not be estimated with any degree of reliability. Moreover, it was found that quantifiable uncertainty varies among patients depending upon the number of treatment interruptions (perturbations of therapy). This leads to a *fundamental question* of how much information with respect to model validation can be expected in a given data set or collection of data sets. In this note, we consider one tool that may be used in attempts to answer this question.

[*] Corresponding author.
E-mail address: htbanks@eos.ncsu.edu (H.T. Banks).

http://dx.doi.org/10.1016/j.aml.2014.11.002

Here we demonstrate the use of *statistically based model comparison tests* in several examples of practical interest. In these examples we are interested in questions related to information content of a particular given data set and whether the data will support a more detailed or sophisticated model to describe it. In the first example we compare fits for several different models to describe simple decay in a size histogram for aggregates in amyloid fibril formation. In a second example we investigate whether the information content in data sets for the pest *Lygus hesperus* in cotton fields as it is currently collected is sufficient to support a model in which one distinguishes between nymphs and adults. Finally in a third example with data for patients having undergone an organ transplant we question whether the data content is sufficient to estimate more than 5 of the fundamental parameters in a specific dynamic model. In the next section we recall the fundamental tests to be employed here.

2. Summary of ANOVA type statistical comparison tests

In general, assume that we have an inverse problem for the model observations $f(t, q)$ and are given n observations. We define

$$J_n(q) = J_n(\mathbf{Y}, q) = \frac{1}{n}\sum_{j=1}^{n}[Y_j - f(t_j, q)]^2 \tag{1}$$

where our statistical model has the form

$$Y_j = f(t_j, q_0) + \mathcal{E}_j, \quad j = 1, \dots, n.$$

Here, q_0 is the "true" value of q which we assume to exist. We use $\mathcal{Q}$ to represent the set of all the admissible parameters q.

We make the standard statistical assumptions [5–7]:

- (A1) The random variables $\{\mathcal{E}_j\}_{j=1}^{\infty}$ are independent and identically distributed with $\mathbb{E}(\mathcal{E}_j) = 0$ and $Var(\mathcal{E}_j) = \sigma^2$.
- (A2) $\mathcal{Q}$ is a compact subset of Euclidean space of R^p and $f(t, q)$ is continuous on $[0, T] \times \mathcal{Q}$.
- (A3) Observations are at $\{t_j\}_{j=1}^{n}$ in $[0, T]$. For some finite measure μ on $[0, T]$,

  $$\frac{1}{n}\sum_{j=1}^{n} h(t_j) \longrightarrow \int_0^T h(t)d\mu(t)$$

  as $n \to \infty$, for all continuous functions h.
- (A4) $J_0(q) = \int_0^T (f(t, q_0) - f(t, q))^2 d\mu(t) = \sigma^2$ has a unique minimizer in $\mathcal{Q}$ at q_0.

Let $q^n = q^n_{OLS}(\mathbf{Y})$ be the OLS estimator for J_n with corresponding estimate

$$\hat{q}^n = q^n_{OLS}(\{y_j\})$$

for a realization $\mathbf{y} = \{y_j\}$. That is,

$$q^n(\mathbf{Y}) = \arg\min_{q\in\mathcal{Q}} J_n(\mathbf{Y}, q)$$

and

$$\hat{q}^n = \arg\min_{q\in\mathcal{Q}} J_n(\mathbf{y}, q).$$

One can then establish a series of useful results (see [5,6] for detailed proofs; see also [8]).

Theorem 2.1. *Under* (A1)–(A4), $q^n = q^n_{OLS}(\mathbf{Y}) \longrightarrow q_0$ *as* $n \to \infty$ *with probability* 1.

We will need further assumptions to precede (these will be denoted by (A7)–(A11) to facilitate reference to [5,6]). These include:

- (A7) $\mathcal{Q}$ is finite dimensional in R^p and $q_0 \in \text{int}\,\mathcal{Q}$.
- (A8) $f : \mathcal{Q} \to C[0, T]$ is a C^2 function.
- (A10) $\mathcal{J} = \frac{\partial^2 J_0}{\partial q^2}(q_0)$ is positive definite.
- (A11) $\mathcal{Q}_H = \{q \in \mathcal{Q} | Hq = c\}$ where H is an $r \times p$ matrix of full rank, and c is a known constant.

In many instances, including the motivating examples discussed here, one is interested in using data to question whether the "true" parameter q_0 can be found in a subset $\mathcal{Q}_H \subset \mathcal{Q}$ which we assume for discussions here is defined by the constraints of assumption (A11). Thus, we want to test the *null hypothesis* H_0: $q_0 \in \mathcal{Q}_H$.

Define then

$$q^n_H(\mathbf{Y}) = \arg\min_{q\in\mathcal{Q}_H} J_n(\mathbf{Y}, q)$$

and

$$\hat{q}^n_H = \arg\min_{q\in\mathcal{Q}_H} J_n(\mathbf{y}, q)$$

and observe that $J_n(\mathbf{Y}, \hat{q}_H^n) \geq J_n(\mathbf{Y}, \hat{q}^n)$. We define the related non-negative test statistics and their realizations, respectively, by

$$T_n(\mathbf{Y}) = n(J_n(\mathbf{Y}, q_H^n) - J_n(\mathbf{Y}, q^n))$$

and

$$\hat{T}_n = T_n(\mathbf{y}) = n(J_n(\mathbf{y}, \hat{q}_H^n) - J_n(\mathbf{y}, \hat{q}^n)).$$

One can establish asymptotic convergence results for the test statistics $T_n(\mathbf{Y})$—see [5]. These results can, in turn, be used to establish a fundamental result about much more useful statistics for model comparison. We define these statistics by

$$U_n(\mathbf{Y}) = \frac{T_n(\mathbf{Y})}{J_n(\mathbf{Y}, q_n)}, \tag{2}$$

with corresponding realizations

$$\hat{u}_n = U_n(\mathbf{y}).$$

We then have the asymptotic result that is the basis of our ANOVA-type tests:

Theorem 2.2. *Under the assumptions* (A1)–(A4) *and* (A7)–(A11) *above and assuming the null hypothesis H_0 is true, then U_n converges in distribution (as $n \to \infty$) to a random variable $U(r)$*

$$U_n \xrightarrow{\mathcal{D}} U(r)$$

having a chi-square distribution $\chi^2(r)$ with r degrees of freedom.

We note that if one is dealing with vector observations with $n = n_1 + n_2$ total component observations as we do in two of the examples below, then asymptotic theory requires that $n_1 \to \infty$ and $n_2 \to \infty$. In any graph of a χ^2 density there are two parameters (τ, α) of interest. For a given value τ, the value α is simply the probability that the random variable U will take on a value greater than τ. That is, $Prob\{U > \tau\} = \alpha$ where in hypothesis testing, α is the *significance level* and τ is the *threshold*.

We then wish to use this distribution to test the null hypothesis, H_0, for $U_n \sim \chi^2(r)$. If the test statistic, $\hat{u}_n > \tau$, then we *reject H_0 as false* with confidence level $(1 - \alpha)100\%$. Otherwise, we *do not reject* H_0. For our examples below, we use a $\chi^2(1)$ table, which can be found in any elementary statistics text or online. Typical confidence levels of interest are 75%, 90%, 95%, 99% with corresponding (α, τ) values of $(.25, 1.32)$, $(.1, 2.71)$, $(.05, 3.84)$, $(.01, 6.63)$, respectively. To test the null hypothesis H_0, we choose a significance level α and use χ^2 tables to obtain the corresponding threshold $\tau = \tau(\alpha)$ so that $P(\chi^2(r) > \tau) = \alpha$. We next compute $\hat{u}_n = \overline{\tau}$ and compare it to τ. If $\hat{u}_n > \tau$, then we reject H_0 as false; otherwise, we do not reject the null hypothesis H_0.

2.1. Weighted least squares

The model comparison results outlined can be extended to deal with weighted least squares problems in which measurement errors are independent with $\mathbb{E}(\mathcal{E}_k) = 0$ and $Var(\mathcal{E}_k) = \sigma^2 w^2(t_k)$, $k = 1, 2, \ldots, n$, where w is some known real-valued function with $w(t) \neq 0$ for any t. This is achieved through rescaling the observations in accordance with their variance (as discussed in [6]) so that the resulting (transformed) observations are identically distributed as well as independent.

3. Size distribution of aggregates in amyloid fibril formation

3.1. Best fit to size distributions

In a recent paper [9], a question was addressed about size distribution of aggregates in amyloid fibril formation. While an exponential distribution was shown to provide a reasonable fit to the data depicted in Fig. 1, the question arose as to whether another distribution such as the Weibull or gamma distributions with more parameters might provide a better fit.

3.2. The exponential, Weibull and gamma distributions

On initial observation, the data appears to be well suited to an exponential distribution. The exponential distribution probability density function is defined as $E(x; \lambda) = \lambda e^{-\lambda x}$. Note that when fitting the data, an additional parameter A was added to the exponential function resulting in a total of two parameters and the function to be defined for these purposes as

$$E(x; A, \lambda) = A\lambda e^{-\lambda x}.$$

The Weibull distribution probability density function is defined as (for the purposes of modeling the data we again add the additional parameter A)

$$W(x; A, \lambda, k) = Ak\lambda(\lambda x)^{k-1} e^{-(\lambda x)^k}, \quad x \geq 0.$$

Note that if we take $k = 1$ we have that $W(x; A, \lambda, 1) = E(x; A, \lambda)$. This function is shown plotted below with several values of k. We can see that when $k = 2$ or $k = 1$ the function also bears a resemblance to the shape of our data.

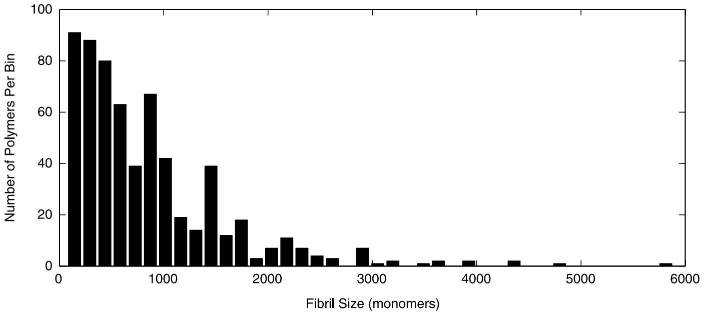
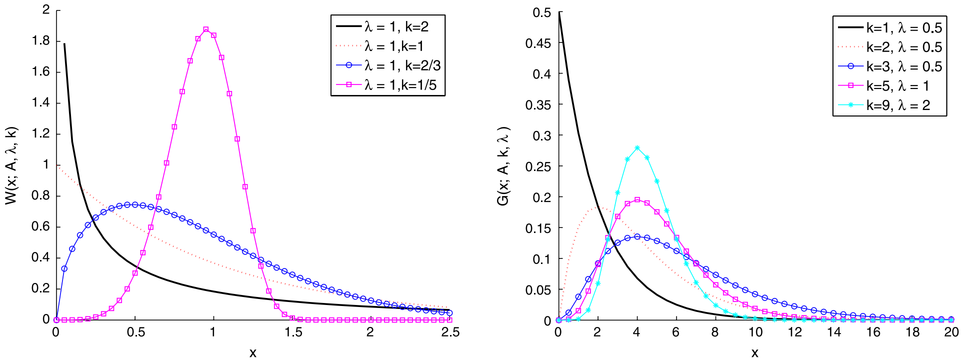


Fig. 1. Experimental distribution of the sample x_i, $1 \leq i \leq n$, representing the measured sizes of polymers (the number of polymers below a certain size (145 monomers) is unknown). The total size of the sample is $n = 626$.



(a) Weibull probability density function.

(b) Gamma probability density function.

Fig. 2. Graphical comparisons of the Weibull and gamma with different values of λ and k.

The probability density function of the gamma distribution is defined as (we again include the additional parameter A for modeling purposes)

$$G(x; A, k, \lambda) = A\frac{\lambda^k}{\Gamma(k)}x^{k-1}e^{-\lambda x} \quad \text{for } x > 0 \text{ and } k, \lambda > 0,$$

where $\Gamma(k)$ is the gamma function evaluated at k. We can see in Fig. 2 that when $k = 1$ and $\lambda = 0.5$, the gamma probability density function again has a similar shape to the data. Since we know that $\Gamma(1) = 1$, we can see that when we take $k = 1$ we have that $G(x; A, 1, \lambda) = E(x; A, \lambda)$. Thus an interesting question is whether we can obtain an statistically better fit to the data in Fig. 1 by allowing an additional free parameter k in either the Weibull or gamma distribution in comparison to the two parameter (A, λ) exponential model.

3.3. Results using the comparison tests

We tested the following hypothesis and alternative for two different alternative models: a Weibull and a gamma distribution:

- $\mathbf{H_0}$: The fit provided by an alternative model is not significantly different from the fit with an exponential distribution.
- $\mathbf{H_A}$: The alternative model with an unrestricted additional parameter k provides a significantly better fit than the exponential model (corresponding to the restriction $k = 1$).

When comparing the best fits of the exponential vs. the Weibull distributions we obtained the following results: $J_n^w = 1.4359 \times 10^{-4}$, $J_n^e = 1.6081 \times 10^{-4}$, $\hat{T}_n^{ew} = 4.6495 \times 10^{-4}$, and $\hat{u}_n^{ew} = \bar{\tau} = 3.2381$. In this case we cannot reject the null hypothesis at the 95% or higher level. We can reject at the 90% confidence level.

When comparing the best fits of the exponential vs. the gamma distribution we obtained the following results: $J_n^g = 1.4277 \times 10^{-4}$, $J_n^e = 1.6081 \times 10^{-4}$, $\hat{T}_n^{eg} = 4.8693 \times 10^{-4}$, and $\hat{u}_n^{eg} = \bar{\tau} = 3.4105$. Again in this case we cannot reject the null hypothesis at the 95% or higher level but we can reject at the 90% confidence level.

4. *Lygus hesperus* population dynamics: model comparison and parameter estimation

Lygus hesperus is a prevalent insect in California which feeds on cotton and other plants [10]. Given a robust data set of *L. hesperus* counts from over 500 Californian fields over several years, we aim to gain more information about *L. hesperus* and direct future research relating to its effects on crops. We propose 2 ordinary differential equation models, estimate parameters for each model, and perform model comparison techniques to determine which model is more appropriate, given the population dynamics and the nature of the data.

4.1. Data

Our main database consists of over 1500 data sets (comprising over 500 distinct fields) of *L. hesperus* counts. One data set is characterized by the following: a designated pest control advisor (PCA) counts the number of *L. hesperus* found in a sample of field sweeps (50 large net sweeps = 1 sample) at intermittent times from early June to early August. We assume that field counts are independent between years (i.e., if one field is sampled in 2004 and 2005, we consider these data sets to be independent).

To narrow down this vast collection of data, and to start with the simplest case, we choose a sub-collection of the data consisting only of data sets corresponding to fields that were untreated by pesticides for a minimum of 2 uninterrupted months, in which PCAs counted both nymphs AND adults. There were at least 40 data sets of this nature. By starting with this sub-collection, we are able to study the insect population dynamics which are not directly affected by pesticides. We note that pesticide usage on nearby crops can have both direct and indirect effects on pests in neighboring fields, but choose to ignore this potential effect for now, as it is largely unknown and variable. In addition, this allows us to propose a 2-dimensional population model. In this model, we choose 6 of these data sets as a preliminary study. An example of one data set can be seen in Table 4 of [11]. Note that there are several data points where adult and nymph counts are non-integer values. This is due to the fact that several fields were so large that PCAs chose to do a number of samples within one field on one particular observation day and averaged the results.

4.2. Model

We assume that there are 2 distinct population classes: nymphs and adults. We will denote their populations as $x_1(t)$ and $x_2(t)$ respectively, where t is time measured in months ($t \geq 0$). Given this particular insect and data collection scheme, we consider $t = 0$ to mean June 1 (as no observations in our data sets are made before this date). For now, we will ignore the effect of pesticides on the population, and consider the population dynamics of *L. hesperus* in an untreated environment. We do not assume a closed population (i.e. $\frac{dX}{dt} \equiv \frac{dx_1}{dt} + \frac{dx_2}{dt} \neq 0$.) In addition, it is assumed that there are at least 3 generations per year. We first consider a simple 2-dimensional ordinary differential equation model. Model A is as follows:

$$\begin{aligned} \frac{dx_1(t)}{dt} &= \beta x_2(t) - \gamma x_1(t) \\ \frac{dx_2(t)}{dt} &= \gamma x_1(t) - \mu_2 x_2(t), \end{aligned} \tag{3}$$

where β is the birth rate of nymphs, γ is the transition rate of nymphs into adulthood, and μ_2 is the adult death rate, all with unit $[1/t]$. Clearly, Model A assumes that there is no (or trivial) nymph mortality. However, Model B assumes a non-trivial nymph mortality:

$$\begin{aligned} \frac{dx_1(t)}{dt} &= \beta x_2(t) - (\gamma + \mu_1) x_1(t) \\ \frac{dx_2(t)}{dt} &= \gamma x_1(t) - \mu_2 x_2(t), \end{aligned} \tag{4}$$

where μ_i is the death rate for x_i, $i = 1, 2$. For both models A and B, initial conditions

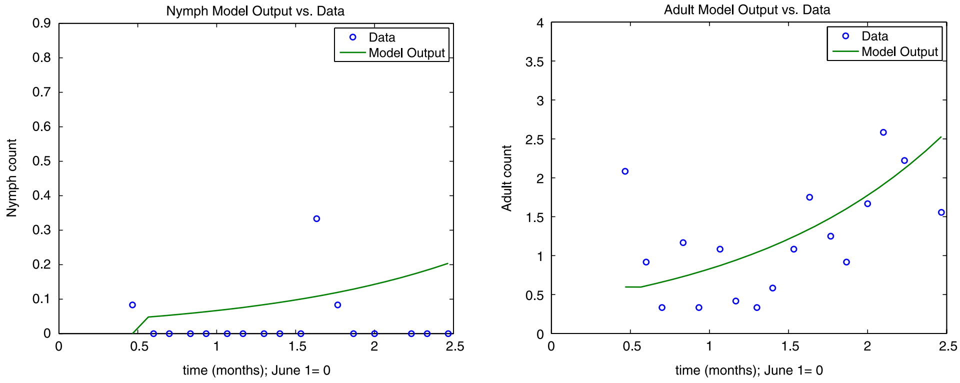
$$\mathbf{X}_1 = (x_1(t_1), x_2(t_1)) := (x_{1,1}, x_{2,1})$$

are first estimated for a given data set and then fixed (see [11] for discussions). Note that t_1, the time of the first observation, varies between data sets. Our goal is to estimate parameters $q = \{\beta, \gamma, \mu_1, \mu_2\}$ in Model B using our chosen data sets (note that the parameters in Model A are equivalent to those in Model B, with the constraint that $\mu_1 = 0$). We will use MATLAB's constrained optimization tool, `fmincon` and both ordinary least squares (OLS) and weighted least squares (WLS) techniques [6].

For a subset of the data, our team used a more thorough method to collect data on the same fields at roughly the same time as the PCAs to be used for comparison purposes. In comparing these data (see [11]), we found higher variability in the nymph counts than in the adult counts. This leads us to believe that using weighted least squares in our parameter estimation is important. To estimate parameters, one must search within an admissible parameter space, $\mathcal{Q}$, for the model parameters that produce a model output most similar to the data. In other words, one must minimize the cost functional, J_n defined to be

$$J_n = J_n(\mathbf{y}, q) = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{k} \omega_j (y_{ij} - m_{ij})^2, \tag{5}$$

where $y_{ij} =$ is the data point from the jth class at the ith time point, and $m_{ij} =$ is the model output for the jth class at the ith time point, given a parameter estimate. Between fields, n (the number of *vector observations* in a sample) is variable. Note



(a) Model fit vs. DS 1—nymphs.

(b) Model fit vs. DS 1—adults.

Fig. 3. Model fit vs. DS1.

that $k = 2$ (the total number of classes within the data), and $j = 1$ corresponds to the nymph class and $j = 2$ corresponds to the adult class so that the total number of data points is $2n$. Let $\Omega = \{\omega_1, \omega_2\}$. There are formal ways of choosing Ω, but we will start with some basic choices. If we choose $\Omega = \{0, 1\}$, we are ignoring the nymph counts in the search for the best parameter estimates for the model. If we choose $\Omega = \{0.5, 1\}$, we are giving less weight to the nymph class than to the adult class. Note that if we choose $\Omega = \{1, 1\}$, we return to an OLS method.

4.3. Parameter estimates and model comparison test

There are differing opinions among PCAs and researchers about whether both nymphs and adults need to be counted. The reasons for these differences are varying beliefs regarding the effect of pesticides and other factors on the *L. hesperus* populations. To the extent that accurately counting both nymphs and adults is more time-intensive than simply regarding adult *L. hesperus*, we seek a quantitative measure to determine whether counting both nymphs and adults (in the manner in which it is presently done) is necessary, or if it is sufficient to simply count the total number of insects. We see that the sole difference between Models A and B ((3) and (4), respectively) is the assumption of no nymph mortality in Model A. Note that model A can be more simply written as

$$\frac{dX}{dt} = \alpha x_2(t), \tag{6}$$

where $X(t) =$ the total number of *L. hesperus* at time t ($X = x_1 + x_2$), and $\alpha = \beta - \mu_2$. This simpler model is exponential in nature. One may wonder how this model could possibly be exponential in nature, when there are 2 state variables, X and x_2 in one differential equation. We found consistently among PCA-collected data that the nymph counts were almost always zero. Therefore, given the current collection strategies, $X \approx x_2$, and (6) truly becomes an exponential growth model. A natural question is the following: by allowing nymph mortality to be non-zero, does our model better fit the data? To address this question we use the model comparison results outlined above to test the null hypothesis: is the true set of parameter values, q_0, in a constrained subset $\mathcal{Q}_H$ of $\mathcal{Q}$, which requires that $\mu_1 = 0$, or do we obtain a statistically significant better fit allowing $\mu_1 \neq 0$? Here $q = \{\beta, \gamma, \mu_1, \mu_2\} \in \mathcal{Q} \equiv [-\delta, 100] \times [-\delta, 100] \times [-\delta, 100] \times [-\delta, 100]$ where $\delta > 0$ is very small and $\mathcal{Q}_H \equiv \{q \in \mathcal{Q} \mid \mu_1 = 0\}$.

Therefore, by testing the null hypothesis $H_0 : q_0 \in \mathcal{Q}_H$, we can determine with a definitive amount of confidence whether we can assume no nymph mortality and thus use a simple model such as Model A to describe the data.

4.4. Results

We chose to perform this analysis on 6 data sets, with 4 choices of Ω: $\Omega_1 = \{1, 1\}$, $\Omega_2 = \{0.5, 1\}$, $\Omega_3 = \{0.2, 1\}$, and $\Omega_4 = \{0, 1\}$. As seen in Table 5 of [11] (an abbreviated version of this table for data set 1 is given in Table 1), for all cases (except for data set 4 with Ω_3), the confidence to reject H_0 is less than 19%. In [11], we see that many estimates for μ_1 returned values relatively small and/or close to zero. This is further evidence that it may be acceptable to assume no nymph mortality.

We have also included plots of model fits vs. data for data set 1, as these were illustrative of the results we found across the 6 data sets used in the previous analysis. As one can see in Fig. 3, the model fits the adult data well, while the model fits the nymph data poorly.

4.5. Conclusions

Overall, we find compelling evidence for the untreated fields, by the model comparison test, that we should NOT reject the null hypothesis. In other words, it may be reasonable to ignore nymph mortality (i.e., just count total number of *L. hesperus*

Table 1
Model comparison test results for data set 1, with several weights, where “Confid” denotes the confidence to reject the null hypothesis, H_0.

Data set 1 with estimated initial conditions $\{x_{1,1}, x_{2,1}\} = \{0, 0.06\}$						
Ω	Confid	$\hat{q}^n$	$\hat{q}^n_H$	$\hat{u}_n$	$J_n(\mathbf{y}, \hat{q}^n)$	$J_n(\mathbf{y}, \hat{q}^n_H)$
{1,1}	3.57%	{8.14, 99.74, 0.00, 7.32}	{7.59, 92.97, 0, 6.77}	.0020	.2410	.2410
{0.5,1}	0%	{9.47, 94.14, 0.00, 8.62}	{8.97, 89.35, 0, 8.13}	.0000	.2309	.2309
{0.2,1}	2.98%	{12.79, 88.69, 0.00, 11.88}	{12.02, 83.84, 0, 11.12}	.0014	.2241	.2241

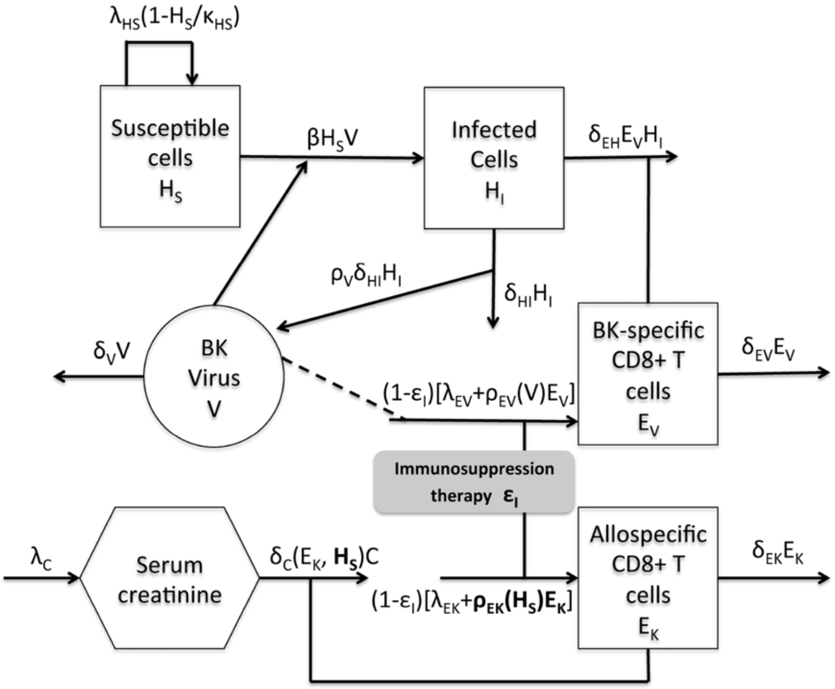


Fig. 4. BKV model.

and not distinguish between nymphs and adults), which would greatly simplify the model, as given in (6), *as well as the data collection process*. It is important to note that this conclusion may not be reasonable for data sets in which pesticide treatment was used, as we have not yet performed analysis on data sets of that nature. While our earlier findings suggest that it may be sufficient to only count the total number of *L. hesperus*, rather than distinguish between adults and nymphs, we must in future efforts proceed to use similar analyses with data from treated fields.

5. Model comparison in organ transplant modeling

5.1. Mathematical model description and data

We focus on modeling of the BK virus, a common pathogen (and major threat) found in kidney transplant patients—see [12] and references therein. We describe the dynamics of the viral load V, susceptible H_S and infected H_I host cells, BKV-specific E_V and allospecific E_K effector CD8+ T cells and serum creatinine C with a brief description of the underlying biological model for which we base our mathematical model. In Fig. 4 we illustrate the intracellular dynamics embodied in the model.

Active BKV infection targets both renal tubular epithelial cells and urothelial cells. For this model, however, we focus on one BKV target, the renal tubular epithelial cells. Susceptible host cells, the uninfected kidney tubular epithelial cells, H_S, in the absence of infection, are assumed to proliferate through the term $\lambda_{HS}\big(1 - \frac{H_S}{\kappa_{HS}}\big)H_S$, indicating that new epithelial cells are derived from proliferation of existing H_S. Proliferation is modeled by logistic dynamics with λ_{HS} being the maximum proliferation rate and κ_{HS} is the cell density at which proliferation shuts off. A loss of susceptible cells, H_S, due to viral infection which occurs by cell-to-cell transmission, is represented by the term $\beta H_S V$. Here we assume that one copy of virion infects one cell. Infected host cells or BKV replicating cells, H_I, lyse due to the cytopathic effect of BK virus, represented by the term $\delta_{HI} H_I$ and produce ρ_V virions before death. In addition, infected host cells are eliminated by the BK-specific effector

CD8+ T cells with rate term $\delta_{EH}E_VH_I$. Free virus is naturally cleared at the rate δ_V by the body and a loss of viral concentration is experienced through the infection of susceptible host cells.

The cellular immune response is the key defense against the BK-viral infection. The terms λ_{EV} and δ_{EV} represent the source and death rates of BK-specific effector CD8+ T cells. The concentration of BK-specific CD8+ T cells increases in response to the presence of free virus through the term $\rho_{EV}E_V$, where ρ_{EV} is a bounded positive increasing function of free virus concentration. Specifically, $\rho_{EV}(V) = (\bar{\rho}_{EV}V)/(V+\kappa_V)$ is a saturating nonlinearity with positive constants $\bar{\rho}_{EV}$ and κ_V. The alloreactive immune response to the transplanted kidney is incorporated into the model via a state variable, E_K, which denotes the effector CD8+ T cells that specifically target the transplant. The source rate for E_K, λ_{EK}, is dependent upon the HLA donor/recipient matching conducted prior to transplantation. Similar to the BK-specific CD8+ T cells, the concentration of allospecific CD8+ T cells increases in response to the presence of susceptible host cells H_S, since BK-virus targets kidney cells, represented by the term $\rho_{EK}E_K$, where $\rho_{EK}(H_S) = (\bar{\rho}_{EK}H_S)/(H_S+\kappa_{KH})$ with positive constants $\bar{\rho}_{EV}$ and κ_{KH}. The death rate of E_K is represented by δ_{EK}.

Finally, we discuss the role of creatinine in the model. Creatinine is a waste product in the blood resulting from muscle activity and is removed by the healthy kidney. Therefore, serum creatinine concentration C is used as a surrogate for glomerular filtration rate (GFR), a commonly used index of kidney function [12]. The production rate of C is represented by λ_C and when the kidney is impaired, creatinine is not effectively filtered and its concentration increases. (Recall that the renal allograft is a site of replication. Hence, the concentration of susceptible cells reflects the health of the kidney.) To account for the negative effect of the alloreactive immune response E_K on the kidney and the positive effect of susceptible cells H_S, the clearance rate δ_C is defined as follows

$$\delta_C(E_K, H_S) = \frac{\delta_{C0}\kappa_{EK}}{E_K+\kappa_{EK}} \cdot \frac{H_S}{H_S+\kappa_{CH}}.$$

Based on the above discussions and those in [12], the model is given as follows:

$$\dot{H_S} = \lambda_{HS}\Big(1-\frac{H_S}{\kappa_{HS}}\Big)H_S - \beta H_S V, \tag{7}$$

$$\dot{H_I} = \beta H_S V - \delta_{HI}H_I - \delta_{EH}E_V H_I, \tag{8}$$

$$\dot{V} = \rho_V\delta_{HI}H_I - \delta_V V - \beta H_S V, \tag{9}$$

$$\dot{E_V} = (1-\epsilon_I)[\lambda_{EV} + \rho_{EV}(V)E_V] - \delta_{EV}E_V, \tag{10}$$

$$\dot{E_K} = (1-\epsilon_I)[\lambda_{EK} + \rho_{EK}(H_S)E_K] - \delta_{EK}E_K, \tag{11}$$

$$\dot{C} = \lambda_C - \delta_C(E_K, H_S)C, \tag{12}$$

with initial conditions $(H_S(0),\ H_I(0),\ V(0),\ E_V(0),\ E_K(0),\ C(0)) = (H_{S0},\ H_{I0},\ V_0,\ E_{V0},\ E_{K0},\ C_0)$.

We note that (7)–(10) describe the immune response to the viral infection coupled with (11) and (12) describing the immune response to the transplanted kidney. Here ϵ_I represents the efficacy of immunosuppressive drugs and is assumed to be scaled to less than or equal to 1. This variable serves as the controller of the system to achieve balance between under-suppression and over-suppression of the patient's immune system.

In order to compare the effectiveness of various model components, we again used the statistical model comparison test described earlier to test the null hypothesis, H_0, that an additional 6th parameter is not needed to describe the system. Among the parameters we focus on here are β, δ_{EK}, λ_C, ρ_V, δ_V, $\bar{\rho}_{EK}$, δ_{EV}, $\bar{\rho}_{EV}$, and ϵ_I. If the null hypothesis is rejected, we determine that the parameter in question is needed to better describe the data. The parameter vector q belongs to the parameter set $\mathcal{Q}$, and the restricted parameter set $\mathcal{Q}_H \subset \mathcal{Q}$ is defined for each model comparison test by fixing the parameter in question. The observed amount of free virus (DNA) in the blood is represented by $\bar{y_i}^1$, with corresponding measured time point t_i^1, $i = 1, 2, \dots, n_1$, and $\bar{y_i}^2$ is the observed amount of serum creatinine at time point t_i^2, $i = 1, 2, \dots, n_2$. We define $y_i^1 = \log_{10}(\bar{y_i}^1)$, $i = 1, 2, \dots, n_1$, and $y_i^2 = \bar{y_i}^2$, $i = 1, 2, \dots, n_2$, with $n = n_1 + n_2 = 8 + 16$ data points in the data considered here and in [12]. Let $\mathbf{y} = [y_1^1, \dots, y_{n_1}^1, y_1^2, \dots, y_{n_2}^2]^T$. We define the OLS cost to be

$$J_n(\mathbf{y}, q) = \frac{1}{n_1+n_2}\Big(\sum_{i=1}^{n_1}|f_1(t_i^1; q) - y_i^1|^2 + \sum_{i=1}^{n_2}|f_2(t_i^2; q) - y_i^2|^2\Big).$$

5.2. Comparison of 5 vs. 6 parameters

We tested whether the immune response to BK virus infection and donor kidney in renal transplant recipients could be more accurately described estimating six vs. five parameters, using the model found in previous work. Based on our sensitivity analysis in [12], we felt we could reliably estimate 5 parameters including $\mathcal{Q}_H = \{\beta,\ \bar{\rho}_{EV},\ \delta_{EV},\ \delta_{EK},\ \bar{\rho}_{EK}\}$. We chose an additional sixth parameter to estimate to form $\mathcal{Q}$ and ran the corresponding inverse problems. Here we refer to the case of estimating 5 parameters as "Model $\mathcal{Q}_H$" and the case associated with 6 estimated parameters as "Model $\mathcal{Q}$". To

Table 2
Ordinary least squares costs and model comparison test statistics for the BK virus models. Model $\mathcal{Q}_H$ is the case in which we estimate 5 parameters and Model $\mathcal{Q}$ is the case in which we estimate 6 parameters. We used the statistical model comparison techniques given above to test whether the OLS cost was significantly lower for Model $\mathcal{Q}$. The resulting model comparison test statistics were not significant at the 95% level, indicating that estimating 6 parameters does not yield a statistically better data fit.

$\mathcal{Q}$	Model $\mathcal{Q}_H$ cost	Model $\mathcal{Q}$ cost	$\hat{u}_n$
$\{q = (\beta,\ \delta_V,\ \bar{\rho}_{EV},\ \delta_{EV},\ \delta_{EK},\ \bar{\rho}_{EK})\}$	0.0031	0.0030	0.8000
$\{q = (\beta,\ \bar{\rho}_{EV},\ \delta_{EV},\ \delta_{EK},\ \rho_V,\ \bar{\rho}_{EK})\}$	–	0.0031	0
$\{q = (\beta,\ \bar{\rho}_{EV},\ \delta_{EV},\ \delta_{EK},\ \lambda_C,\ \bar{\rho}_{EK})\}$	–	0.0030	0.8000
$\{q = (\beta,\ \bar{\rho}_{EV},\ \delta_{EV},\ \delta_{EK},\ \bar{\rho}_{EK},\ \epsilon_I)\}$	–	0.0030	0.8000

estimate the parameters in the BKV model, we first fixed the remaining parameters, using the parameter estimates found in [12] for the 10 estimated parameters case. We note that the forward simulations were run using *ode15s* and the inverse problems were solved using *lsqnonlin* with various parameter bounds found in [12]. We obtained the results given in Table 2.

To validate our use of 5 parameters, we then tested the model with these five parameters against five reduced models, each with one parameter removed (see [11] for detailed results). Despite previous sensitivity analysis leading us to believe that we could reliably estimate five parameters, the results of these tests indicated with 95% confidence that four parameters were sufficient.

6. Concluding remarks

The diversity of the examples described above are ample evidence of the wide applicability of the methodology we have proposed here. These known [5] statistically-based model comparison tests add to a growing list of tools including the parameter subset/parameter selectivity tools based on parameter sensitivity based scores [3], and other Fisher Information Matrix, Akaike Information Criteria based techniques [6,13] that may be used to better understand information content in data sets.

Acknowledgments

This research was supported in part by grant number NIAID R01AI071915-10 from the National Institute of Allergy and Infectious Diseases, in part by the Air Force Office of Scientific Research under grant number AFOSR FA9550-12-1-0188, and in part by the National Science Foundation under Research Training Grant (RTG) DMS-1246991. The authors are grateful to Dr. S. Prigent for providing the data used in Section 3.

References

[1] B.M. Adams, H.T. Banks, M. Davidian, E.S. Rosenberg, Model fitting and prediction with HIV treatment interruption data, in: Center for Research in Scientific Computation Technical Report CRSC-TR05-40, NC State Univ., October, 2005, Bull. Math. Biol. 69 (2007) 563–584.

[2] H.T. Banks, M. Davidian, S. Hu, G.M. Kepler, E.S. Rosenberg, Modeling HIV immune response and validation with clinical data, J. Biol. Dyn. 2 (2008) 357–385.

[3] H.T. Banks, R. Baraldi, K. Cross, K. Flores, C. McChesney, L. Poag, E. Thorpe, Uncertainty quantification in modeling HIV viral mechanics, in: CRSC-TR13-16, N. C. State University, Raleigh, NC, December, 2013, Math. Biosci. Engr. (2014) submitted for publication.

[4] H.T. Banks, A. Cintron-Arias, F. Kappel, Parameter selection methods in inverse problem formulation, CRSC-TR10-03, N.C. State University, February, 2010, Revised, November, 2010, in: J.J. Batzel, M. Bachar, F. Kappel (Eds.), Mathematical Modeling and Validation in Physiology: Application to the Cardiovascular and Respiratory Systems, in: Lecture Notes in Mathematics, vol. 2064, Springer-Verlag, Berlin, 2013, pp. 43–73.

[5] H.T. Banks, B.G. Fitzpatrick, Statistical methods for model comparison in parameter estimation problems for distributed systems, J. Math. Biol. 28 (1990) 501–527.

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

[7] H.T. Banks, H.T. Tran, Mathematical and Experimental Modeling of Physical and Biological Processes, CRC Press, New York, 2009.

[8] H.T. Banks, K. Kunisch, Estimation Techniques for Distributed Parameter Systems, Birkhauser, Boston, 1989.

[9] S. Prigent, H.W. Haffaf, H.T. Banks, M. Hoffmann, H. Rezaei, M. Doumic, Size distribution of amyloid fibrils: Mathematical models and experimental data, in: CRSC TR14-04, N. C. State University, Raleigh, NC, April, 2014, Int. J. Pure Appl. Math. 93 (2014) 845–878.

[10] W.H. Day, C.R. Baird, S.R. Shaw, New native species of parasitizing *Lygus hesperus* in Idaho: Biology, importance and description, Ann. Entomol. Soc. Am. 92 (3) (1999) 370–375.

[11] H.T. Banks, J.E. Banks, Kathryn Link, J.A. Rosenheim, Chelsea Ross, K.A. Tillman, Model comparison tests to determine data information content, Center for Research in Scientific Computation Technical Report CRSC-TR14-13, N. C. State University, Raleigh, NC, October, 2014.

[12] H.T. Banks, S. Hu, K. Link, E.S. Rosenberg, S. Mitsuma, L. Rosario, Modeling immune response to BK virus infection and donor kidney in renal transplant recipients, in: CRSC Technical Report CRSC-TR14-09, NCSU, June 2014, J. Inverse Probl. Sci. Eng. (2014) submitted for publication.

[13] Kenneth P. Burnham, David R. Anderson, Model Selection and Multimodal Inference, second ed., Springer, New York, 2002.