Special Issue Article

Journal of CHEMOMETRICS

Received: 21 November 2011, Accepted: 19 December 2011, Published online in Wiley Online Library: 18 April 2012

(wileyonlinelibrary.com) DOI: 10.1002/cem.2416

Parameter estimation in differential equation models with constrained states

David A. Campbell[a]*, Giles Hooker[b] and Kim B. McAuley[c]

We introduce a method to estimate parameters and states from a differential equation model while enforcing interpretability constraints such as monotone or non-negative states. We motivate the methodology using a real data chemical engineering example and show that a variety of restrictive constraints from earlier analyses do not address the problem of interpretability. Our proposed method estimates parameters using a smoothing-based relaxation of the model to enforce interpretability of the observed and unobserved system states. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: nonlinear regression; functional data analysis; constrained smoothing; profiled likelihood estimation; nylon production

1. CHALLENGES OF PARAMETER ESTIMATION FROM DIFFERENTIAL EQUATION MODELS

Ordinary differential equations (ODEs) describe the rate of change of system states $x(t)$ with respect to time t (or another argument) as a function of $x(t)$ along with externally controlled inputs $u(t)$ and p model parameters θ. Using the differential operator $Dx(t) = \frac{d}{dt}x(t)$, the ODE model

$$Dx(t) = f(x[t], \theta, u[t]) \tag{1}$$

can be interpreted as a regression model with functional covariates $x(t)$. When $x(t)$ is a vector of chemical concentrations, model (1) describes the rate of change of chemical concentrations as a function of the chemicals in the system $x(t)$, rate parameters θ, and (time-varying) covariates such as temperature, pressure, and other catalytic inputs $u(t)$. Rate parameters θ are typically of main interest because they can be used to predict behavior of the system under different $u(t)$ and to optimize reactor settings. System states $x(t)$ are of interest to determine when the reactor holds the desired mixture of chemicals. This paper focuses on estimating both $x(t)$ and θ.

In general, for a nonlinear function $f(\cdot)$ in (1), there is no analytic form for $x(t)$, but given θ and initial system states $x(0)$, numerical methods can produce the solution $S(\theta, x[0], t) = x(t)$. However, the measurement error model, for example,

$$y(t) \sim N\left(S[\theta, x\{0\}, t], \sigma^2\right) \tag{2}$$

will not have a closed-form likelihood. Further, complicating the model statistically, observations $y(t)$ are often only available from a subset of the vector of states $x(t)$.

With $S(\theta, x[0], t)$ in (2), $\hat{\theta}$ can be estimated by nonlinear least squares (NLS) through gradient-based optimization, [1] or Markov chain Monte Carlo (MCMC) [2]. Knowing $x(0)$ improves the reliability of parameter estimates when the model trajectory is strongly impacted by changes in initial conditions [3], and therefore, the estimation problems are exacerbated when the unknown parameter vector from θ is augmented to $\psi = [\theta, x(0)]$. Complicating these estimation methods, a relatively compact ODE model can be used to describe complex dynamics such as limit cycles, exponential growth, or a strong tendency towards an equilibrium state depending on the values of ψ. The diversity in modeled dynamics gives rise to complex likelihood topology such as ripples, ridges, flat sections, and/or multiple local maxima [4].

The alternative to depending on $S(\theta, x[0], t)$ is to work with a relaxation of the ODE solution by smoothing the data to produce the approximation $\hat{x}(t) \approx S(\theta, x[0], t)$ and estimating θ via

$$\arg\min_{\theta} ||D\hat{x}(t) - f(\hat{x}[t], \theta, u[t])||^2$$

Some implementations use a nonparametric data smooth [5–8]; however, these methods are limited when the model of interest has discontinuous derivatives or mixtures of fast and slow changing dynamics. Other smoothing-based methods use

* Correspondence to: D. A. Campbell, Department of Statistics and Actuarial Science, Simon Fraser University, Surrey, BC, Canada.
E-mail: dac5@sfu.ca

a D. A. Campbell
Department of Statistics and Actuarial Science, Simon Fraser University, Surrey, BC, Canada

b G. Hooker
Department of Statistical Science and Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, NY, USA

c K. B. McAuley
Department of Chemical Engineering, Queen's University, Kingston, ON, Canada

iterative conditional updating of $\hat{\theta}$ and a parametric model-based smooth estimator of $x(t)$ via frequentist [9,10] or Bayesian methods [11]. This paper focuses on extending the generalized smoothing (GS) method of [12] to cases where the model-based data smooth is confined to a restricted function space meeting monotonicity or nonnegativity constraints. Using GS, $\hat{x}(t)$ and θ are estimated using a multilevel, multicriterion optimization and parametric ODE model-based smoothing. The related approximate maximum likelihood estimation (AMLE) method of [13] uses an unconstrained model-based smoothing strategy but estimates $x(t)$ and θ, simultaneously giving both types of parameters equal importance, whereas we consider $x(t)$ to be defined by a set of nuisance parameters. Furthermore, AMLE relies on the assumption that the measurement error variance σ^2 in (2) is known and the lack of model fit afforded by the smoothing process is therefore assumed to be caused by a stochastic forcing function. We do not assume that σ^2 is known, but we are also able to work with the assumption if it is reasonable.

Generalized smoothing is a collocation method; it uses a basis expansion with a vector of coefficients **b** and basis functions $\phi(t)$ to approximate the system states such that $\hat{x}(t) = b'\phi(t) \approx S(\theta, x[0], t)$. The data smooth $\hat{x}(t)$ is a model-based smooth because it is guided by the ODE by penalizing deviation from the residual of (1). The tradeoff between interpolating $y(t)$ and perfectly following $S(\theta, x[0], t)$ is controlled by the smoothing parameter λ, allowing $\hat{x}(t)$ to be interpreted as a relaxation of $S(\theta, x[0], t)$. GS requires estimation of $\hat{b}$, θ, and $\hat{\lambda}$, but through a multi-criterion three-level hierarchical optimization, the relationships $b(\theta, \lambda)$ and $\theta(\lambda)$ are defined such that the process reduces to estimating the degrees of freedom laden parameter λ. In this paper, we extend GS to accommodate interpretability constraints on the shape of $\hat{x}(t)$. In addition, we propose an intuitive estimation strategy for λ and highlight how to overcome challenges of a real data system.

Section 2 provides a description of the motivating data set from an industrial chemical engineering application with a nonlinear ODE model. The challenges of the motivating example include multiple experimental replications, step function changes in functional inputs, noisy system components observed at different time points, and precisions along with unobserved state variables within a nonlinear differential equation model where system states are only interpretable when taking on nonnegative values. Section 3 describes GS with the extension to ensure that state estimates are constrained to remain interpretable. Section 4 describes how GS is adapted to overcome the remaining challenges of the motivating data set. Section 5 provides an analysis of the real data set, highlighting how the assumptions and limitations of previous analyses are removed using the proposed methodology. Comparisons with alternative methods are also shown. Discussion about the method's performance and further adaptations and generalizations are given in Section 6.

2. NYLON SYSTEM

In a heated reactor, amine (A) and carboxyl (C) groups combine, producing polyamide links (L) and water (W), which escape as steam. Simultaneously, steam bubbled through molten nylon, decomposes L into A and C giving the symbolic competing reactions $A + C \rightleftharpoons L + W$. From any starting values, A, C, L, and W will adjust their concentrations until these competing reaction rates balance and equilibrium concentrations are achieved. In the experiment of [14], steam is bubbled through molten nylon to manipulate the concentration of W in the system. Consequently A, C, and L move towards equilibrium concentrations with the imposed level of W. Within each of the $i = 1, \ldots, 6$ experimental runs, the amount of input steam followed a step function. Initial high levels of steam were held constant until time τ_{i1}, then reduced until time τ_{i2}, and finally returned to the original level for the remainder of the experimental run. Each experimental run was performed at a constant temperature T_i, which, along with the input water pressure, determines the equilibrium concentration of water in the molten nylon mixture, W_{eq}. With reaction rates k_p and K_a, the reaction dynamics are described with differential equations

$$-DL = DA = DC = -k_p(CA - LW/K_a) \tag{3}$$

$$DW = k_p(CA - LW/K_a) - 24.3(W - W_{\text{eq}}) \tag{4}$$

Reaction rates k_p and K_a are allowed to change with T_i and W_{eq} according to parameter vector $\theta = [k_p, E, \gamma, \beta, K_{a0}, H]$ and the reference temperature $T_0 = 549.15$ K, through the following relationships:

$$k_p = \frac{k_{p0}}{1000}\exp\left(-E\frac{10^3}{8.314}\left\{\frac{1}{T_i} - \frac{1}{T_0}\right\}\right) \tag{5}$$

$$K_a = \left\{1 + W_{\text{eq}}\frac{\gamma}{1000}\exp\left(-\beta 10^3\left\{\frac{1}{T_i} - \frac{1}{T_0}\right\}\right)\right\}K_T[K_{a0}] \exp\left(-H\frac{10^3}{8.314}\left\{\frac{1}{T_i} - \frac{1}{T_0}\right\}\right) \tag{6}$$

$$K_T = 20.97\exp\left(-9.624 + \frac{3613}{T_i}\right) \tag{7}$$

These equations include scaling factors, making all initial arameter estimates used in [14] for θ of the same order of magnitude to ease estimation. A variety of alternative models for this system have been proposed [13,12,14,15], and the methodology and upcoming analysis could be applied to any of them. Equation (6) is altered from that of [14] to ensure consistency of parameter units. Alternatively, four or five parameters versions of (5)–(7) may also of interest and are described in Section 5.2. Given any three chemical components, the fourth can be determined algebraically using the mass balance of the system; however, it was only possible to measure *A* and *C*. Consequently, an unobserved state variable must also be estimated. Figure 1 shows the data for each of the experimental runs and includes vertical lines corresponding to τ_{i1} and τ_{i2}.

These data and differential equation model produce several challenges that are addressed in this paper:

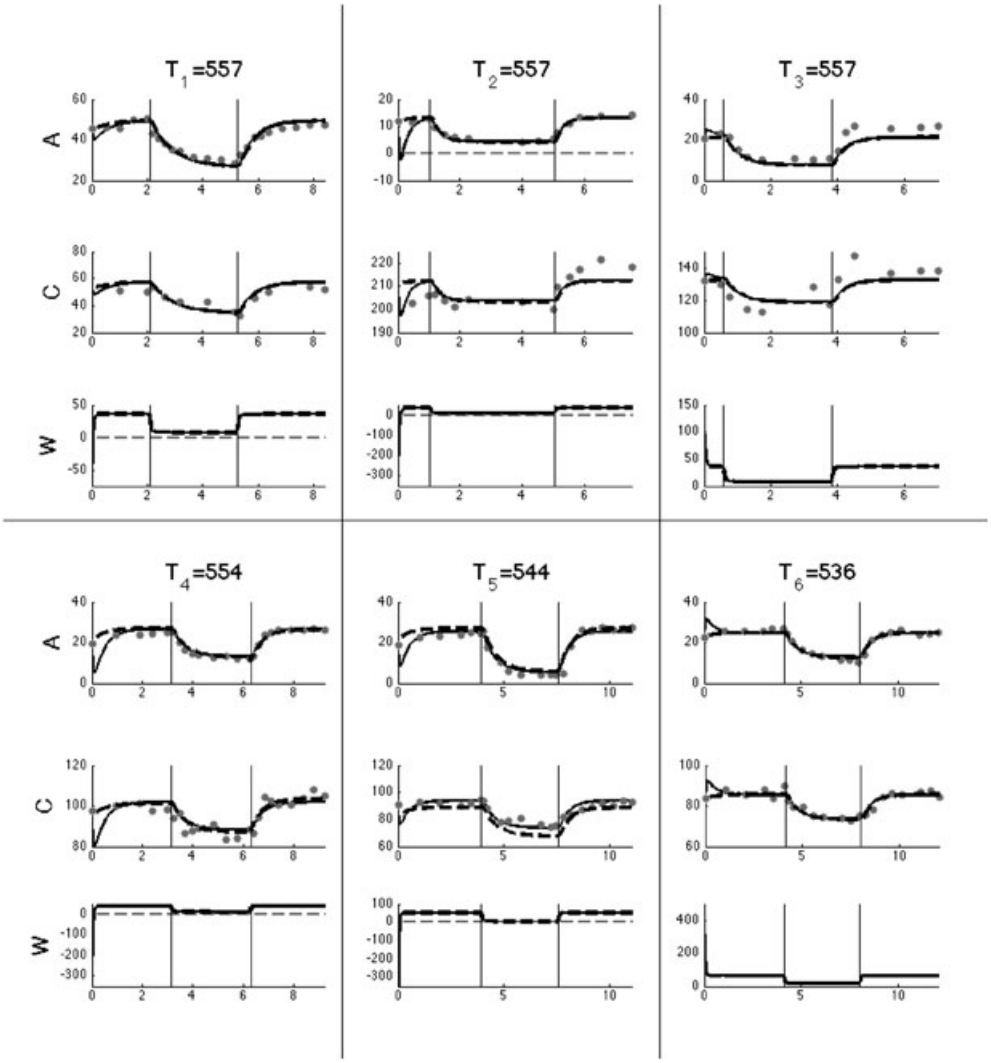

Figure 1. The nylon observations and data fit using unconstrained generalized smoothing (GS) (solid line) and the shape constrained GS (dashed line), both with dense bases. Temperatures of the experimental runs are given above component *A* in Kelvin. Vertical axes are in concentration units and horizontal axes are in hours. Horizontal dashed gray lines mark concentrations of zero. Vertical lines denote times τ_{i1} and τ_{i2}.

(1) This system describes chemical concentrations that are by definition constrained to take on nonnegative values. Using nonlinear regression, Zheng *et al*. [14] overcame this problem by assuming that the initial system states were known. However, along with restricting the set of possible ODE solutions, including strong assumptions about initial states can produce biased parameter estimates and artificially narrow confidence intervals [3]. Other analyses avoided negative states by using a basis incapable of matching the sharp modeled dynamics [12,13], producing state estimates that would not be considered optimal had they used NLS or a denser basis to obtain $\hat{\theta}$ and $\hat{x}(t)$. Using too few basis functions prevents $x(\hat{t})$ from following sharp dynamics and can introduce bias into the parameter estimates. Figure 1 shows the unconstrained optimal GS fit to the data using a much more flexible basis than that of [13] and [12]. By using a much more flexible basis, although $W(0)$ is negative in four cases, after a positive initial condition, $\hat{A}(t)$ becomes uninterpretably negative in the second experiment (top middle plot). Consequently, a constraint on initial conditions alone may not suffice to ensure that all states remain within the nonnegativity constraint. The parameter estimates that our paper produce are those obtained by constraining the states to be interpretable, without assuming x(0) is known and without limiting the dynamics by using too few basis functions.

(2) Information must be pooled across the six experimental runs. Using a method based on producing $S_i(\theta, x\{0\}, t)$ requires estimating the vector of $3 \times 6 = 18$ initial conditions and $p = 6$ ODE parameters. The increased parameter space and resulting likelihood difficulties were avoided in [14] by assuming that initial conditions were measured without observation error and that $W(0) = W_{eq}(0)$. We wish to estimate θ and $x(t)$ by removing these assumptions.

(3) Generally in ODE models, states $x(t)$ may be modeled in different units and have incomparable scales and precisions. Figure 1 shows that the variability in the measurement of *C* is larger than that of *A*, that is, $\sigma_A^2 \neq \sigma_C^2$. Although AMLE assumes that σ^2 is known from equipment manufacturers error suggestions or additional steady state experiments, σ^2 may vary depending on the practitioner who performs the chemical analysis, and as such, we wish to remove

assumptions about its value. In addition, the observations for components *A* and *C* are not evenly spaced, simultaneously measured or observed the same number of times.

(4) Although the ODE model is based on scientific theory, the model in (3)–(7) is believed to be useful but imperfect. The model may be subject to stochastic forcing modifying (1) to

$$dx(t) = f(x(t), \theta, u(t))dt + \xi d\omega(t) \tag{8}$$

where $\xi d\omega(t)$ could be, for example, a Weiner process with diffusion ξ [16]. Consequently, enforcing strict adherence to the model may not be reasonable.

3. GENERALIZED PROFILE ESTIMATION

Generalized smoothing estimates **b**, θ, and λ through a multi-criterion multilevel hierarchical optimization. A hierarchy is defined by rewriting some parameters as functions of others higher up the hierarchy, $b(\theta, \lambda)$ and $\theta(\lambda)$, such that the estimation process condenses to the estimation of the degrees of freedom heavy λ to define all parameters.

The incidental or local parameters in the sense of [17] are the basis coefficients **b** of the data smooth. There may be many more basis coefficients than data points, and the number of coefficients may grow as more observations are obtained. Representing the lowest level of the hierarchy, **b** is defined as a function of parameters θ and λ. For each λ and θ, the optimal b defines a data smooth, balancing the fit between interpolating the data and following the ODE model. The smoothing step described in Section 3.1 allows the data fit to deviate from the ODE model, potentially allowing process noise or other model imperfections.

The structural parameters θ define the behavior allowed by the ODE model. Changes in these parameters decide between limit cycles, exponential decay, or other behaviors from the ODE model. We are primarily interested in θ because of its interpretation and potential use for making decisions and predictions. As described in Section 3.2, for any λ, $\theta(\lambda)$ is estimated iteratively through a profile likelihood maintaining $b(\hat{\theta}[\lambda], \lambda)$ at its optimum at every iterated value of θ, defining the second level of the hierarchy.

The complexity or smoothing parameter λ defines the top level of parameters to estimate. It determines how closely the data follows the ODE model and allows for some model misspecification. One could interpret λ as ξ^2/σ^2, which is the ratio of process noise to measurement noise in the model formulation in (8) if $d\omega(t)$ is a Weiner process. For any λ, we can find $\theta(\lambda)$ and $b(\theta[\lambda], \lambda)$. Estimation of λ is detailed in Section 3.3.

The estimation routine can be thought of as resulting from inner and outer loops to estimate $\hat{b}(\theta, \lambda)$ and $\hat{\theta}(\lambda)$ for fixed λ. Although the estimation process is described in [12], we extend these loops with interpretability constraints on the smooth and detail how to intuitively optimize λ in the remainder of this section. Software to perform the estimation is available on request.

3.1. Inner optimization: model-based interpretable smoothing

By using a piecewise differentiable one to one function $g(\cdot)$ to restrict the smooth to follow interpretable behavior, the basis expansion $x_{ik}(t) = g_{ik}\left(b'_{ik}\phi_{ik}(t)\right) \approx S(\theta, x[0], t)$ for the $i = 1, \ldots, I$ experimental runs and $k = 1, \ldots, K$ system components smooths the data and approximates the ODE trajectory. Some examples of $g(\cdot)$ from [10] include

$g(a) = exp(a)$	positive smooth
$g(a) = exp(a)/[1 + exp(a)]$	bounded smooth
$g(a) = \int_0^t exp(a)ds$	monotone smooth
$g(a) = a$	unconstrained smooth

L-spline smoothing in [10] estimates $\hat{b}$ from a (weighted) penalized likelihood, regularizing the data fit with the ODE model. Using weights $w_{ik} = \hat{\sigma}_{ik}-2$ and the measurement error modeling (2), the estimates $\hat{b}$ minimize the weighted penalized negative log likelihood

$$\begin{aligned}\hat{b}(\theta, \lambda) &= \arg\min_b J_i(b|y, \theta, \lambda) = \arg\min_b \sum_{k=1}^{K} w_{ik}(SSE_{ik} + \lambda \mathrm{PEN}_{ik}) \\ &= \arg\min_b \sum_{k=1}^{K} w_{ik}\left(\sum_{t \in t_{ik}} \left[y_{ik}(t) - g\left\{b'_{ik}\phi_{ik}(t)\right\}\right]^2 + \lambda \mathrm{PEN}_{ik}\right)\end{aligned} \tag{9}$$

In nonparametric smoothing, often $\mathrm{PEN} = \int (D^2 x[t])^2 dt$, implying a model with little curvature. In GS, PEN penalizes the residual of (1), which, hiding dependence on inputs $u[t]$, gives

$$\mathrm{PEN}_{ik} = \int (Dx_{ik}(s) - f_k(x_i[s], \theta))^2 ds \tag{10}$$

For fixed λ and θ and data y_{ik} observed at the vector of times t_{ik}, $\hat{b}(\theta, \lambda)$ is obtained through nonlinear regression. The integral in equation (10) is over the interval $[min_k(t_{ik}), max_k(t_{ik})]$, the maximum range of observation times over all K observed variables in the ith run. Section 4.2 describes how to approximate this integral.

3.2. The outer optimization: estimating ODE parameters

Using the error model $Y(t) \sim N(x[t], \sigma^2)$, the maximum profile likelihood estimate is used for $\hat{\theta}(\lambda)$, which is equivalent to minimizing the negative log profile likelihood

$$\begin{aligned}H(\theta, \mathrm{b}(\theta, \lambda)|\mathrm{y}, \lambda) = \sum_{i}^{I}\sum_{k=1}^{K} w_{ik} SSE_{ik} = \sum_{i}^{I}\sum_{k=1}^{K}\sum_{t \in t_{ik}} w_{ik} \\ \left(y_{ik}[t] - g_k\left[b_{ik}\left\{\theta, \lambda\right\}'\phi_{ik}\{t\}\right]\right)^2.\end{aligned} \tag{11}$$

A penalized likelihood is not used at this level of the parameter hierarchy because fidelity to the ODE model is already enforced in (9).

For fixed λ, finding $\hat{\theta}(\lambda)$ is a nonlinear regression problem applied to the likelihood while profiling over $b(\hat{\theta}, \lambda)$. Numerically, as $\hat{\theta}(\lambda)$ is updated through optimization iterations in the outer loop, $b(\hat{\theta}, \lambda)$ must be updated at each increment through the inner optimization step. Simplifying notation from $H(\theta, b[\theta, \lambda]|y, \lambda)$ to H, assuming that λ is fixed and notationally simplifying $\theta(\lambda)$ to θ, $b(\theta, \lambda)$ to b, and $g(b'\phi)$ to g, gives the total gradient for the profile likelihood

$$\frac{dH}{d\theta} = \frac{\partial H}{\partial \theta} + \frac{\partial H}{\partial g}\frac{dg}{db}\frac{db}{d\theta} \tag{12}$$

When $f(x[t], \theta)$ is a nonlinear function of $x(t)$, there is no explicit formula for $\hat{b}(\theta, \lambda)$, and consequently, $d\hat{b}/d\theta$ must be obtained using the implicit function theorem. Assume that H and J are twice continuously differentiable with respect to θ and b and that the Hessian matrices

$$\frac{\partial^2 H}{\partial \theta^2}, \frac{\partial^2 H}{\partial g^2}, \frac{\partial^2 J}{\partial \theta^2} \text{ and } \frac{\partial^2 J}{\partial g^2}$$

are positive definitive over a nonempty neighborhood of **y** in the data space. Using $dJ/db = 0$ at $b = \hat{b}$ from (9),

$$\begin{aligned}\frac{d^2 J}{db d\theta} &= \frac{d}{d\theta}\left(\frac{dJ}{db}\right) = \frac{d}{d\theta}\left(\frac{\partial J}{\partial g}\frac{dg}{b}\right) \\ &= \left[\frac{\partial^2 J}{\partial g \partial \theta}\frac{dg}{db} + \left\{\left(\frac{dg}{d\hat{b}}\right)'\frac{\partial^2 J}{\partial g^2}\frac{dg}{db} + \frac{\partial J}{\partial g}\frac{d^2 g}{db^2}\right\}\frac{db}{d\theta}\right]\end{aligned} \tag{13}$$

$$\left.\frac{db}{d\theta}\right|_{b=\hat{b}} = -\left\{\left(\frac{dg}{db}\right)\frac{\partial^2 J}{\partial g^2}\frac{dg}{db} + \frac{\partial J}{\partial g}\frac{d^2 g}{db^2}\right\}^{-1}\left\{\frac{\partial^2 J}{\partial g \partial \theta}\frac{dg}{db}\right\}\Big|_{b=\hat{b}} \tag{14}$$

Substituting (14) into (12) provides the total gradient for the maximum profile likelihood estimate

$$\frac{dH}{d\theta} = \frac{\partial H}{\partial \theta} - \frac{\partial H}{\partial g}\frac{dg}{db}\left\{\left(\frac{dg}{db}\right)\frac{\partial^2 J}{\partial g^2}\frac{dg}{db} + \frac{\partial J}{\partial g}\frac{d^2 g}{db^2}\right\}^{-1}\left\{\frac{\partial^2 J}{\partial g \partial \theta}\frac{dg}{db}\right\} \tag{15}$$

3.2.1. *Interval estimates for* $\hat{\theta}(\lambda)$

Interval estimates obtained using the delta method approximation

$$\text{var}(\theta) \approx \frac{d\theta}{dy}\text{var}(y)\frac{d\theta}{dy} \tag{16}$$

require the implicit function theorem once again to define $d\theta/dy$. Using the fact that at $\theta = \hat{\theta}$, $dH/d\theta = 0$ from (11), $d\theta/dy$ can be found using

$$\frac{d}{dy}\left(\frac{dH}{d\theta}\right)_{\theta=\hat{\theta}}\Big| = \left[\frac{d^2 H}{dy d\theta}\right] + \frac{d^2 H}{d\theta^2}\frac{d\theta}{dy}\Big|_{\theta=\hat{\theta}} = 0 \tag{17}$$

and therefore at $\theta = \hat{\theta}$,

$$\frac{d\theta}{dy} = -\left[\frac{d^2 H}{d\theta^2}\right]^{-1}\frac{d^2 H}{dy d\theta}$$

where

$$\begin{aligned}\frac{d^2 H}{d\theta^2} &= \frac{\partial^2 H}{\partial \theta^2} + 2\frac{\partial^2 H}{\partial \theta \partial g}\frac{dg}{db}\frac{db}{d\theta} + \left(\frac{dg}{db}\frac{db}{d\theta}\right)'\frac{\partial^2 H}{\partial g^2}\left(\frac{dg}{db}\frac{db}{d\theta}\right) \\ &+ \left(\frac{\partial b}{\partial \theta}\right)'\frac{\partial H}{\partial g}\frac{d^2 g}{db^2}\left(\frac{\partial b}{\partial \theta}\right) + \frac{\partial H}{\partial g}\frac{dg}{db}\frac{d^2 b}{d\theta^2}\end{aligned} \tag{18}$$

and

$$\begin{aligned}\frac{d^2 H}{d\theta dy} &= \frac{d^2 H}{d\theta dy} + \frac{d^2 H}{d\theta dg}\frac{dg}{db}\frac{db}{dy} + \frac{d^2 H}{dg dy}\frac{dg}{db}\frac{db}{d\theta} + \left(\frac{dg}{db}\frac{db}{dy}\right)'\frac{d^2 H}{dg^2}\frac{dg}{db}\frac{db}{d\theta} \\ &+ \left(\frac{db}{dy}\right)'\frac{dH}{dg}\frac{d^2 g}{db^2}\frac{db}{d\theta} + \frac{dH}{dg}\frac{dg}{db}\frac{d^2 b}{d\theta dy}\end{aligned} \tag{19}$$

These last two equations involve the terms $d^2 b/d\theta^2$, $d^2 b/d\theta dy$, and db/dy, all of which arise from further calls to the implicit function theorem. These terms are given in the Appendix.

3.3. Estimating the complexity parameter

3.3.1. *When* σ *is known*

When experiments are repeated frequently, for example, to assess quality control, the measurement noise variance σ^2 may be reasonably well known. Consequently, the discrepancy between the data and the original model should be partitioned into a measurement error component of known magnitude and a stochastic forcing term, such as ξ in (8), the latter being absorbed into the model relaxation afforded by the data smoothing process. Estimating the complexity parameter is then equivalent to estimating ξ with fixed σ^2. Following [18], the optimal $\hat{\lambda}$ is

$$\hat{\lambda} = \min_{\lambda}\left(\frac{\hat{\sigma}_{\lambda}^2}{\sigma^2} - 1\right)^2 \tag{20}$$

where σ^2 is the known value and $\hat{\sigma}_{\lambda}2$ is the estimated residual variance based on the value of λ and corresponding estimates for $\theta(\lambda)$ and $b(\theta, \lambda)$.

3.3.2. *When* σ *is unknown*

Smoothing parameter λ controls the flow of information between $y(t)$ and $\theta(\lambda)$ such that small λ allows limited influence from $\theta(\lambda)$ on the shape of $x(t)$. When $\lambda = 0$, the differential equation model is ignored, and (9) is minimized nonuniquely along a data-interpolating manifold in the function space of $\psi(t)$. Small values of λ limit the impact of $\theta(\lambda)$ on $x(t)$, producing a wide basin of attraction for $\hat{\theta}(\lambda)$, allowing the optimization increments to traverse difficult likelihood surfaces and avoid local likelihood maxima. As λ increases, the SSE term in (9) increases, as shown in the left plot of Figure 2 using 100 simulated data sets based on the four-parameter variant of the nylon model described in Section 5.2.1. Simulated data based on the error structure in (2) were obtained by using additive noise $\left[\sigma_A^2, \sigma_C^2\right] = \left[0.6^2, 2.4^2\right]$. The right column of Figure 2 shows how PEN decreases with increasing λ. Increasing λ increases the impact of changes in θ on $x(t)$ and reintroduces the sharper profile likelihood features. Consequently, λ has a similar role to a temperature parameter in annealing or parallel tempering by easing movement around the parameter space and refining point and interval estimates.

When $\lambda = \infty$, PEN = 0 by forcing the residual from (1) to 0. Because $x(t)$ is then a solution to the ODE, GS is equivalent to NLS while profiling over $x(0)$. However, setting $\lambda = \infty$ requires a basis capable of perfectly matching $S(\theta, x[0], t)$ in a neighborhood of the θ_{true}.

For most practical choices of $\phi(t)$, as $\lambda \to \infty$, the range of values of θ for which PEN $\to 0$ is limited. For instance, when λ becomes "too large" for $\phi(t)$, $x(t)$ is forced to find reductions in PEN at the resolution demanded by λ by moving $x(t)$ towards a trajectory that is closer to the function space of ODE solutions. This results in $x(t)$ moving away from the dynamics in the data

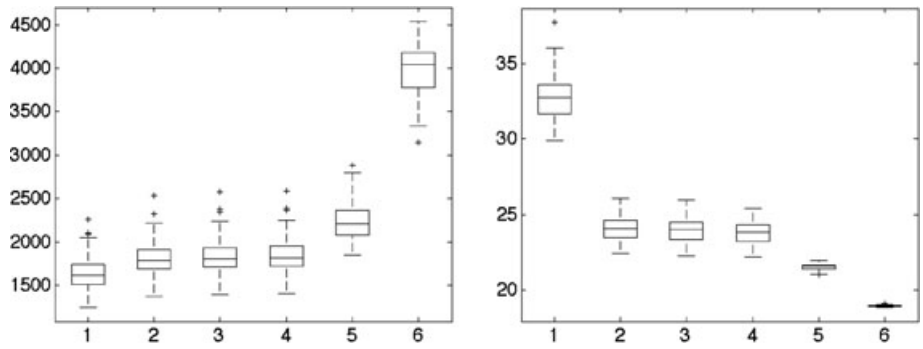

Figure 2. SSE (left) and PEN (right) as a function of $log_{10}(\lambda)$ for 100 simulated data sets.

but reduces PEN by moving $\theta(\lambda)$ towards a value where the function spaces of the basis and ODE are closer to intersecting. This happened in [12] in order to ensure nonnegative $\hat{x}(t)$ with the potential side effect of introducing bias in θ and $x(t)$.

As an extreme example, consider the situation where the chosen basis is the set of Haar wavelets and the ODE model describes a sinusoid with frequency θ. As λ increases, to reduce PEN, θ must move towards a frequency of 0 to eliminate the periodicity so that the model function space can match that of the basis. This will occur as $\lambda \to \infty$ regardless of the dynamics expressed by the data. Increasing the number of basis functions will postpone the problem towards a larger λ if the basis and the model do not span intersecting function spaces. More generally, as $\lambda \to \infty$, the problem of basis-induced bias manifests by forcing $\theta(\lambda)$ and $b(\theta, \lambda)$ to take values that describe dynamics inconsistent with the data. The parameter estimates from 100 simulated data sets show the onset of the basis-induced bias in Figure 3 using the B-spline basis of Section 5.1. Figure 2 also shows a corresponding sharp increase in SSE and an abrupt decrease in PEN at $\lambda > 10^4$, which is too large for the basis.

Although [12] admit that choosing an optimal λ remains an open problem, Figures 2 and 3 show that in this example, $\lambda \in (10^2, 10^4)$ produces values of $\theta(\hat{\lambda})$ close to θ_{true} and gives stable values of SSE and PEN. Within this range of λ, $x(t)$ is not visually different from $S(\theta, x[0], t)$. Consequently, we propose the following intuitive estimation strategy:

(1) Initialize: start with a small λ_j to obtain a low bias but potentially high variance initial estimate $\hat{\theta}(\lambda_j)$.
(2) Anneal: increase $\lambda_{j+1} = \lambda_j * 10$ to remove smoothness in the likelihood and obtain the refined estimate $\hat{\theta}(\lambda_{j+1})$.
(3) Repeat step 2 until parameter estimates converge or destabilize at λ_J. Keep the estimates from $\theta(\lambda_{J-1})$ because larger λ enforces the ODE model as much as possible for the given basis.

This estimation method can be further refined to account for states observed with different levels of unknown precision through incorporating an iterative weighting as demonstrated in Section 4.3. When the model is not accurate, this approach permits model relaxations, balancing the data fitting and model fitting criteria without interpolating or extending beyond the capabilities of the basis as seen in the analysis of the real data and the simulation study of Section 4.4.

4. OVERCOMING CHALLENGES OF A REAL DATA SYSTEM

4.1. Unobserved outputs

The model (1) depends on observed and unobserved system states. The unobserved states in GS are estimated by $\hat{x}(t)$ while following the ODE model through PEN. Estimated unobserved states are data regularized by their impact on the fit to $y(t)$ through the observed states in (9). Therefore, the shape of $\hat{x}(t)$ for the unobserved states is the function that helps to produce the best fit to the observed system states while penalized to follow the model dynamics.

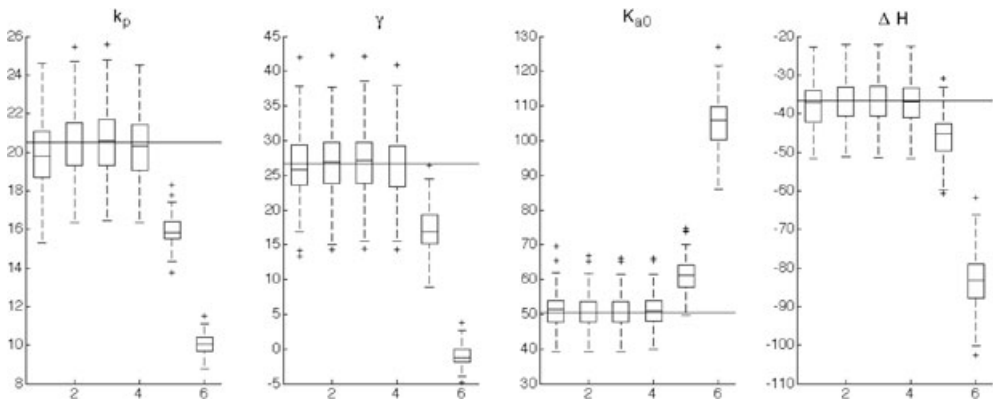

Figure 3. The four parameter estimates from the four-parameter nylon model from 100 simulated data sets at different values of $log_{10}(\lambda)$. The horizontal line is the true parameter value.

4.2. Computing PEN and discontinuous derivatives

The integral in (10) is evaluated using a numerical quadrature approximation, which allows each component to have a unique differentiable basis and does not require variables to have been measured at the same times. It does however require that the quadrature points be the same for all components within the kth experimental run so that information from each component is available to compute PEN_{ik} over the kth quadrature grid. Basic Simpson's rule quadrature for approximating PEN is explained in [12]. Here, we highlight how to overcome numerical difficulties associated with discontinuous derivatives. A quadrature approximation to PEN is set up by dividing T_i into a set of small intervals whose boundaries are the unique knot locations compiled over the bases of all k components. By denoting the location of the ℓth such knot location by ξ_ℓ, intervals are then split into four equal-sized subintervals, and Simpson's rule weights $(1,4,2,4,1)(\xi_{\ell+1}-\xi_\ell)/5$ are used to approximate the integral over each interval.

At points of discontinuity in the first derivative τ_i, the integral in (10) is undefined. At τ_i, the left or right hand derivatives could be defined at τ_i. Alternatively, the integration can be set up to avoid τ_i by integrating over the intersection of T_i and the compliment of a small δ sized neighborhood, τ_i^δ:

$$\begin{aligned}\text{PEN}_{ik} &= \int_{T_i \cap \bar{\tau}_i^\delta} (Dx_{ik}(s) - f_k(x_i, u_i, s|\theta))^2 ds \\ &= \int_{T_i^{min}}^{\tau_i - \delta/2} (Dx_{ik}(s) - f_k(x_i, u_i, s|\theta))^2 ds \\ &\quad + \int_{\tau_i + \delta/2}^{T_i^{max}} (Dx_{ik}(s) - f_k(x_i, u_i, s|\theta))^2 ds\end{aligned} \tag{21}$$

This integral is approximated by shifting the quadrature interval boundaries at times τ_i to the points defining the boundaries of τ_i^δ and omitting quadrature weights across τ_i^δ. To avoid enabling the basis to push an extremely poor fit into τ_i^δ, effectively allowing a discontinuous smooth across τ_i^δ, we recommend using a small neighborhood such as $10^{-6} \times min_\ell(\xi_{\ell+1} - \xi_\ell)$.

4.3. Outputs measured with different precision

In ODE systems, often components are measured in different units, scales, and precisions; consequently, it is important for parameter estimation that weights w_{ij} bring SSE_k to approximately the same scale. In some cases, this may include using vectors for w_{ij} to accommodate autocorrelations in the data. For a Gaussian likelihood, the optimal choice is $w_{ki} = 1/\sigma_{ki}^2$ [19]. Iterative reweighting can be applied to estimate w_{ki} as follows:

(1) At iteration $m=0$, initialize $w_{ki}^{(m)} = 1$ or use another value consistent with prior information.
(2) Perform the profile estimation to obtain $\hat{b}, \hat{\theta}, \hat{\lambda}$ using weights $w_{ki}^{(m)}$.
(3) Obtain the vectors of residuals $r_{ki} = \hat{y}_{ki} - y_{ki}$, and estimate new weights $\hat{w}_{ki}(m+1) = \left(\frac{n_{ki}}{r'_{ki} r_{ki}}\right)$. Other robust weight estimators may also be appropriate.
(4) If $\left|\hat{w}_{ki}^{(m+1)} - \hat{w}_{ki}^{(m)}\right| > \epsilon$ for some convergence tolerance $\epsilon > 0$, set $m = m+1$ and return to step 2.

4.4. Model imperfection

When the model is not correct, $\lambda \to \infty$ is not reasonable because a smaller λ balances the fit to the model dynamics with the fit to the data. We used a simulation study to explore model imperfections. We altered models (3) and (4) such that when W_{eq} decreases after τ_{i1}, it moves to a level $\tilde{W}_{\text{eq}} = W_{\text{eq}} \times \xi_i$, where ξ_i $N(1,.2^2)$, but we continue to model it assuming that it is equal to W_{eq}. At time $\tau_{i,2}$, the level is returned to its original known value. This models the situation that could arise if the gauge used to calibrate W_{eq} was malfunctioning. Allowing $\tilde{W}_{\text{eq}}$ to differ from W_{eq} by 20% or 40% results in the model misspecifying $A(t)$ and $C(t)$ over the intervals (τ_{i1}, τ_{i2}).

Generalized smoothing was performed using the four-parameter nylon model of Section 5.2.1. Simulated data have Gaussian measurement error noise $\left[\sigma_A^2, \sigma_C^2\right] = \left[2.26^2, 4^2\right]$. The resulting parameter estimates are in Figure 4. Parameter estimates remain close to the true values despite the misspecification in the value of W_{eq}. The parameters in this model are involved in determining the rate at which the system approaches equilibrium after τ and in determining the equilibrium concentration levels. Although $\tilde{W}_{\text{eq}}$ alters some of this information, the equilibrium levels before τ_{i1} and after τ_{i2} provide accurate information but with wider spread than in Figure 3. Naturally, the quality of the estimation is dependent on the type of model misspecification, and more extreme cases will require major model refinements such as those described in [20].

5. ESTIMATION DETAILS AND RESULTS FOR THE NYLON DATA

5.1. Details of the basis and weights

Data smoothing was performed using a fifth-order B-spline basis with knots linearly spaced at a rate of 10 per experiment hour

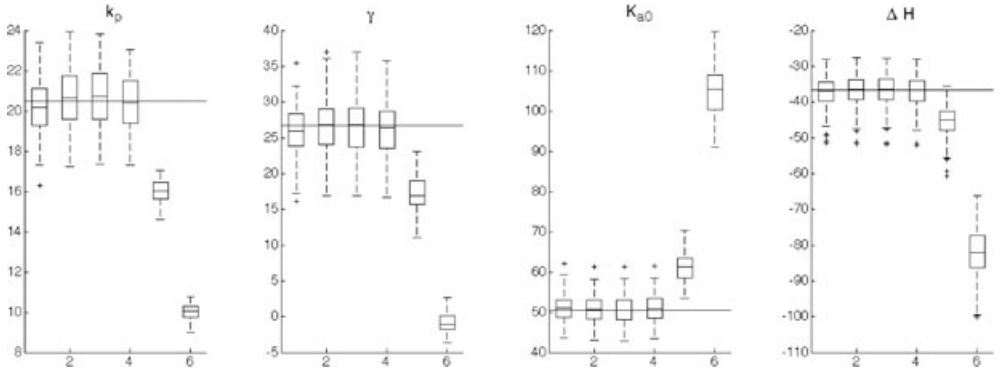

Figure 4. The four parameter estimates from the misspecified four-parameter nylon model from 100 simulated data sets at different values of $log_{10}(\lambda)$. The horizontal line is the true parameter value.

and one at each observation of component *A*. Additional knots at the times of changes in input W_{eq} were used to allow a discontinuity in the first derivative at τ. This same basis was also used for *C* and unobserved *W*. This strategy produced between 84 and 142 unique interior knots per component. Although there are at most 23 observations for any component, the density of this basis is necessary to capture the fast-changing dynamics immediately after τ_{i1} and τ_{i2}. A denser basis should be used if the ODE solution is expected to exhibit finer resolution features or faster dynamics. Because all of the components are concentrations, negative estimates of concentration have no interpretation and we use the positively constrained data smooth $x(t) = g(b'\phi(t)) = exp(b'\phi(t))$.

Alternative basis set up is worth considering on the basis of the structure of this particular system. The ODE in (3) and (4) can be reformulated by noting that *A* and *C* have identical derivatives producing a constant the difference in concentration $C_i(t) - A_i(t) = \delta_i$ within each experimental run. Consequently, the ODE system could be rewritten as a differential algebraic equation (DAE):

$$\begin{aligned} -DL = DA &= -k_p(CA - LW/K_a) \\ C &= \delta + A \\ DW &= k_p(CA - LW/K_a) - 24.3(W - W_{eq}) \end{aligned} \tag{22}$$

The DAE model requires the augmenting θ by including $\delta = [\delta_1, \ldots, \delta_I]$. Using GS, in the ODE version of the nylon system, δ is estimated implicitly by the smooth in the same way that the smooth estimates $x(0)$. However, these values are profiled out in the multilevel optimization. As δ and $x(0)$ are not of primary interest, we prefer to keep this system as an ODE model.

An alternative basis system exploiting the DAE structure uses the smooth for *A* as one of two basis functions for *C*. The second basis function is equivalent to estimating the constant δ_i. At large values of λ, this accelerates the computation by reducing the number of basis coefficients. However, at small values of λ this basis hinders the flexibility of the smooth to simultaneously interpolate the data for *A* and *C* and therefore the estimation process may lose robustness to poor initial parameter estimates. While there is little lost by using too many basis functions, too few will not permit the model dynamics to be adequately represented by the basis expansion.

In the nylon example, all components have the same scale and units; however, *A* was measured with more precision than *C*. In what we will refer to as method 1 below, we use the weights determined by additional experiments and used by [14] in their weighted NLS. In accordance with their weights, for all $i = 1, \ldots, 6$ experimental runs, we use $w_{iA} = 1/\sigma_A^2 = 1/0.6^2$ and $w_{iC} = 1/\sigma_C^2 = 1/2.4^2$ held constant across all runs. In method 2 (Section 5.3), we challenge this assumption and estimate σ_A^2 and σ_C^2 through iterative reweighting.

5.2. Parameter estimation details

The parameter estimation process for the nylon model was initialized with the parameters all set to the value 10. This choice is discussed in Section 6. GS was performed using the iterative refinement of $\hat{\theta}(\lambda)$ following the strategy of Section 3.3 initialized with $\lambda^{(1)} = 10$.

5.2.1. Results

The nylon model in (3)–(7) differs from that of [14] to correct an inconsistency in the balance of units. However, point estimates in our model were plagued by extremely high correlation ($\hat{r}^2 > 0.99999$) between β and γ, removing the ability to statistically uniquely identify these parameters. Consequently, β was eliminated from the model so that instead of (6), we use

$$K_a = \left(1 + \frac{\gamma}{1000} W_{eq}\right) K_T K_{a0} \exp\left[-\frac{H}{R}\left(\frac{1}{T} - \frac{1}{T_0}\right)\right] \tag{23}$$

A similar simplification was also recently developed by [13] to address the high correlations of β and γ in a stochastic differential equation version of this system. In this reduced model using the same incrementally increasing λ procedure, the final 95% confidence interval estimates for *E* overlap zero, suggesting that a simpler model would be just as effective at fitting the data. Although *E* is interpreted as an activation energy parameter, it does not make interpretative sense to have a value of 0; we reduce the model to focus the estimation abilities on parameters for which the model is informative, and note that further experiments are necessary to determine the impact of temperature on the rate parameter k_{p0}. Setting *E* to zero in (5) is equivalent to removing the effect of temperature on k_p in (3) and (4), replacing (5) with

$$k_p = \frac{k_{p0}}{1000} \tag{24}$$

For comparison, NLS was also performed with $x(0)$ held fixed; however, [14] assumed that the first observation was measured without error. Here, $x(0)$ was set equal to values estimated by the smoothing step of GS. Final point and interval estimates for the four-parameter model are given under method 1 in Table I using GS and NLS. Note that the point estimates for NLS and GS are very similar, but the assumed knowledge about the initial states reduces the overall uncertainty in $\hat{\theta}(\lambda)$. Estimating the initial system states in NLS adds three new parameters to the model for each experimental run and produces some negative estimates for $W(0)$. Furthermore, with $x(0)$ unknown, NLS was more sensitive to the choice of parameters used to initialize the algorithm.

Figure 1 shows the fit to the data. Because of the weighting, the fit to *A* is slightly better than the fit to *C* because the latter is presumed to have a larger measurement variance. The difference between the smooth at $\lambda = 10^3$ and the solution to the differential equation based on the smooth estimates $\hat{x}(0)$ is less than 1% of the variability of the data fit at its worst point, suggesting reasonable agreement between the observed and modeled dynamics.

5.3. Estimating unknown weights

In this section, we use method 2 to challenge the assumption of method 1 that σ_A^2 and σ_C^2, as determined from additional experiments, are indicative of the relative precisions of the additive error component. Here, we use iterative reweighting as described in Section 4.3, assuming that σ_A^2 and σ_C^2 are unknown but constant across experimental runs.

For comparison with model 2, the table includes Bayesian point and interval estimates using a likelihood centered on the solution to the ODE model $S_i(\theta, x_i(0), t)$ [2,21]:

Table I. 95% confidence intervals for the nylon data using method 1 with values from nonlinear least squares with $X(0)$ fixed at their generalized smoothing estimates in brackets and method 2 values with Markov chain Monte Carlo results in brackets

Parameter	Lower	Point	Upper
Method 1 (fixed weights), $\sigma_A = 0.6$ and $\sigma_C = 2.4$			
k_p	14.63 (20.54)	20.50 (20.67)	26.37 (20.80)
γ	14.44 (21.86)	26.71 (26.91)	38.99 (31.97)
K_{a0}	39.04 (45.58)	50.35 (50.23)	61.66 (54.87)
H	−49.91 (−43.48)	−36.56 (−37.06)	−23.20 (−30.64)
Method 2 (two weights)$\sigma_A = 2.13(2.26)$ and $\sigma_C = 4.37(4.00)$			
k_p	15.95 (15.64)	18.72 (17.72)	21.48 (20.16)
γ	14.92 (17.09)	22.10 (22.66)	29.29 (28.67)
K_{a0}	47.46 (49.51)	55.68 (55.65)	63.90 (62.49)
H	−47.08 (−44.28)	−36.74 (−35.18)	−26.40 (−26.57)

$$P\left(y_{i,k}[t]\middle|\theta, x_{i,k}[0], \sigma_A^2, \sigma_C^2\right) \sim N\left(S_{i,k}(\theta, x_i(0)), \sigma_{i,k}^2\right)$$

The Bayesian model uses the variance structure of method 2 and the following priors, which include the indicator function I to assign zero probability to negative system states:

$$\begin{aligned} P(k_{p0}) &= P(K_{a0}) = \Gamma(4,8) I_{S(\theta,x[0])>0}, \\ P(\gamma) &= N(0, 15^2) I_{S(\theta,x[0])>0}, \\ P(H) &= N(0, 50^2) I_{S(\theta,x[0])>0}, \\ P(\sigma_A^2) &= P(\sigma_C^2) = \Gamma(3,3), \\ P(X_i(0)) &= N(X_{\text{observed}}(0), \sigma_A^2), \\ P(W_i(0)) &= N(W_{\text{eq},i}(0), 25^2) \end{aligned} \tag{25}$$

These priors represent the expected order of magnitude along with considerable uncertainty. While Bayesian estimates depend on the prior specification, these priors and the Bayesian results are included for illustrative purposes and comparison with the results of GS. Large discrepancies between MCMC and GS results would suggest cause for concern. MCMC was performed using the previous model with 85,000 iterations, where the first 25,000 were discarded as burn in.

Method 2 estimated standard deviations $\left(\hat{\sigma}_k = 1/\sqrt{w_k}\right)$ using iteratively reweighted GS, and the corresponding marginal posterior means of the Bayesian model are included in Table I. Point and interval estimates for GS and MCMC are close with differences in interval estimates because, in part, of the GS intervals reflecting uncertainty in parameters due to the model relaxation, whereas the Bayesian model assumes that the model is perfect.

6. DISCUSSION

In this paper, we introduce an extension of GS to enforce interpretability constraints on state estimates such as monotonicity or nonnegativity. Furthermore, we develop an intuitive method for estimating the model discrepancy smoothing parameter, which performs well with simulated data both when the model is correct and when it is subject to misspecification. Finally, we showcase practical considerations for implementing the methodology with a real data example where we are able to remove assumptions or model perfection and known initial conditions of [14] and reduce restrictions on potentially fast-moving dynamics of [12].

Although (1) uses a measurement error model, using a finite λ in estimating $\hat{x}(t)$ allows some deviation from the deterministic system that may prove useful when applied to SDE models. Furthermore, although this paper assumes a Gaussian error structure, alternative error distributions can be accommodated by altering the log likelihood in (9) and (11).

The rough topologies associated with dynamic system models necessitate examination of a methods performance with poorly chosen parameter values used to initialize the iterative algorithm. Parameter estimation was attempted with the poor choice of initial parameter estimates $k_{p0} < 0$ and/or $K_{a0} < 0$. Negative values of these parameters alter (3) and (4) from describing exponential decay to unbounded exponential growth, ignoring the mass balance of the system. When θ prevents the dynamics of (1) from matching the dynamics of the data, a small value of λ reduces the influence of $f(\theta, x)$ and improves the basin of attraction for θ by shifting $x(t)$ towards the data. Decreasing the starting value of λ provides considerable robustness to poor initial parameter estimates by allowing $x(t)$ to approach a data interpolant.

Another type of poor initial parameter estimates also arises when γ is large and consequently in (23) the term

$$\left(1 + \frac{\gamma}{1000} W_{\text{eq}}\right) K_{a0} \approx \frac{\gamma}{1000} W_{\text{eq}} K_{a0}$$

eliminating the ability to uniquely determineγ and K_{a0}. This feature of the parametrization may not be overcome by altering λ and highlights the potential problem of the parameterization of ODE models. A poor choice of initial parameter estimates did not affect the converged values if GS was initialized in a region of the parameter space where the parameters were identifiable.

APPENDIX
A. Additional implicitly defined derivatives

This section provides implicitly defined derivatives for confidence intervals of Section 3.2.1.

A.1. $\partial^2 \mathbf{b}/\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}_K$

The implicit function theorem is required to define $\partial^2 b/\partial\theta\,\partial\theta_k$ in (18). The term $\partial^2 b/\partial\theta\,\partial\theta_k$ comes from the fact that $\partial J/\partial b = 0$

at the optimal choice of $b = \hat{b}$. Then, differentiating twice with respect to θ and θ_k, equivalent to differentiating (13) with respect to θ_k, produces (26), which is then rearranged to give the derivative in (27).

$$\begin{aligned}\frac{\partial}{\partial\theta_k}\left(\frac{\partial^2 J}{\partial\hat{b}\partial\theta}\right) &= \frac{\partial}{\partial\theta_k}\left(\frac{\partial^2 J}{\partial g\partial\theta}\frac{dg}{d\hat{b}} + \left\{\left(\frac{dg}{d\hat{b}}\right)'\frac{\partial^2 J}{\partial g^2}\frac{dg}{d\hat{b}} + \frac{\partial J}{\partial g}\frac{d^2 g}{d\hat{b}^2}\right\}\frac{\partial\hat{b}}{\partial\theta}\right)\\ &= \frac{\partial^3 J}{\partial g\partial\theta\partial\theta_k}\frac{dg}{d\hat{b}} + \left(\frac{dg_\ell}{d\hat{b}_\ell}\frac{d\hat{b}_\ell}{d\theta_k}\right)'\frac{\partial^3 J}{\partial g\partial\theta\partial g_\ell}\frac{dg}{d\hat{b}}\\ &\quad + \frac{\partial^2 J}{\partial g\partial\theta}\frac{d^2 g}{d\hat{b}d\hat{b}_\ell}\frac{d\hat{b}_\ell}{d\theta_k} + \left\{\left(\frac{d^2 g}{d\hat{b}d\hat{b}_\ell}\frac{d\hat{b}_\ell}{d\theta_k}\right)\frac{\partial^2 J}{\partial g^2}\frac{dg}{d\hat{b}}\right.\\ &\quad + \left(\frac{dg}{d\hat{b}}\right)'\frac{\partial^3 J}{\partial g\partial\theta_k\partial g}\frac{dg}{d\hat{b}} + \left(\frac{dg}{d\hat{b}}\right)'\frac{\partial^3 J}{\partial g^2 dg_\ell}\frac{dg}{d\hat{b}}\frac{dg_\ell}{d\hat{b}_\ell}\frac{d\hat{b}_\ell}{d\theta_k}\\ &\quad + \left(\frac{dg}{d\hat{b}}\right)'\frac{\partial^2 J}{\partial g^2}\frac{d^2 g}{d\hat{b}d\hat{b}_\ell}\frac{d\hat{b}_\ell}{d\theta_k} + \frac{\partial^2 J}{\partial g\partial\theta_k}\frac{d^2 g}{d\hat{b}^2}\frac{\partial^2 J}{\partial g\partial g_\ell}\frac{d^2 g}{d\hat{b}^2}\frac{dg_\ell}{db_\ell}\frac{db_\ell}{d\theta_k}\\ &\quad \left.+\frac{\partial J}{\partial g}\frac{d^3 g}{d\hat{b}^2 d\hat{b}_\ell}\frac{d\hat{b}_\ell}{d\theta_k}\right\}\frac{\partial\hat{b}}{\partial\theta} + \left\{\left(\frac{dg}{d\hat{b}}\right)'\frac{\partial^2 J}{\partial g^2}\frac{dg}{d\hat{b}} + \frac{\partial J}{\partial g}\frac{d^2 g}{d\hat{b}^2}\right\}\frac{\partial^2\hat{b}}{\partial\theta\partial\theta_k}\end{aligned} \tag{26}$$

$$\begin{aligned}\frac{\partial^2\hat{b}}{\partial\theta\partial\theta_k} = -+ &\left[\left(\frac{dg}{d\hat{b}}\right)'\frac{\partial^2 J}{\partial g^2}\frac{dg}{d\hat{b}} + \frac{\partial J}{\partial g}\frac{d^2 g}{d\hat{b}^2}\right]^{-1}\\ &\left[\frac{\partial^3 J}{\partial g\partial\theta\partial\theta_k}\frac{dg}{d\hat{b}} + \left(\frac{dg_\ell}{d\hat{b}_\ell}\frac{d\hat{b}_\ell}{d\theta_k}\right)'\frac{\partial^3 J}{\partial g\partial\theta\partial g_\ell}\frac{dg}{d\hat{b}} + \frac{\partial^2 J}{\partial g\partial\theta}\frac{d^2 g}{d\hat{b}d\hat{b}_\ell}\frac{d\hat{b}_\ell}{d\theta_k}\right.\\ &3 + \left\{\left(\frac{d^2 g}{d\hat{b}d\hat{b}_\ell}\frac{d\hat{b}_\ell}{d\theta_k}\right)'\frac{\partial^2 J}{\partial g^2}\frac{dg}{d\hat{b}}\left(\frac{dg}{d\hat{b}}\right)'\frac{\partial^3 J}{\partial g\partial\theta_k\partial g}\frac{dg}{d\hat{b}}\right.\\ &+\left(\frac{dg}{d\hat{b}}\right)'\frac{\partial^3 J}{\partial g^2 dg_\ell}\frac{dg}{d\hat{b}}\frac{dg_\ell}{d\hat{b}_\ell}\frac{d\hat{b}_\ell}{d\theta_k} + \left(\frac{dg}{d\hat{b}}\right)'\frac{\partial^2 J}{\partial g^2}\frac{dg}{d\hat{b}d\hat{b}_\ell}\frac{d\hat{b}_\ell}{d\theta_k}\\ &\left.\left.+\frac{\partial^2 J}{\partial g\partial\theta_k}\frac{d^2 g}{d\hat{b}^2} + \frac{\partial^2 J}{\partial g\partial g_\ell}\frac{d^2 g}{d\hat{b}^2}\frac{dg_\ell}{d\hat{b}_\ell}\frac{d\hat{b}_\ell}{d\theta_k} + \frac{\partial J}{\partial g}\frac{d^3 g}{d\hat{b}^2 d\hat{b}_\ell}\frac{d\hat{b}_\ell}{d\theta_k}\right\}\frac{\partial\hat{b}}{\partial\theta}\right]\end{aligned} \tag{27}$$

A.2. $\partial\hat{\boldsymbol{b}}/\partial y$

The implicit function theorem is required to define $\partial\hat{b}/\partial y$ in (19). This derivative again uses the fact that $\partial J/\partial b = 0$ at the optimal choice of $b = \hat{b}$. Then, differentiating twice with respect to **y** produces (28), which is then rearranged to give the derivative in (29).

$$\frac{\partial}{\partial y}\left(\frac{\partial J}{\partial\hat{b}}\right) = \frac{\partial^2 J}{\partial g\partial y}\frac{dg}{d\hat{b}} + \left(\frac{dg}{d\hat{b}}\right)'\frac{\partial^2 J}{\partial g^2}\frac{dg}{d\hat{b}}\frac{\partial\hat{b}}{\partial y} + \frac{\partial J}{\partial g}\frac{d^2 g}{d\hat{b}^2}\frac{\partial\hat{b}}{\partial y} \tag{28}$$

$$\frac{\partial\hat{b}}{\partial y} = \left\{\left(\frac{dg}{d\hat{b}}\right)'\frac{\partial^2 J}{\partial g^2}\frac{dg}{d\hat{b}} + \frac{\partial J}{\partial g}\frac{d^2 g}{d\hat{b}^2}\right\}^{-1}\left\{\frac{\partial^2 J}{\partial g dy}\frac{dg}{d\hat{b}}\right\} \tag{29}$$

A.3. $\partial^2\hat{\boldsymbol{b}}/\partial\boldsymbol{y}\partial\theta$

We obtain this derivative by differentiating (28) with respect to θ_k to produce Equation (30). Solving for $\partial^2\hat{b}/\partial y\partial\theta$ gives us the results in (31).

$$\begin{aligned}\frac{\partial}{\partial\theta_k}\left(\frac{\partial^2 J}{\partial\hat{b}\partial y}\right) &= \frac{\partial}{\partial\theta_k}\left(\frac{\partial^2 J}{\partial g dy}\frac{dg}{d\hat{b}} + \left(\frac{dg}{d\hat{b}}\right)'\frac{\partial^2 J}{\partial g^2}\frac{dg}{d\hat{b}}\frac{\partial\hat{b}}{\partial y} + \frac{\partial J}{\partial g}\frac{d^2 g}{d\hat{b}^2}\frac{\partial\hat{b}}{\partial y}\right)\\ &= \frac{\partial^3 J}{\partial g\partial y\partial\theta_k}\frac{dg}{d\hat{b}} + \frac{\partial^3 J}{\partial g\partial y\partial g}\frac{dg}{d\hat{b}}\frac{dg}{d\hat{b}_\ell}\frac{d\hat{b}_\ell}{d\theta_k} + \frac{\partial^2 J}{\partial g\partial y}\frac{d^2 g}{d\hat{b}d\hat{b}_\ell}\frac{d\hat{b}_\ell}{d\theta_k}\\ &\quad + \left(\frac{d^2 g}{d\hat{b}d\hat{b}_\ell}\frac{d\hat{b}_\ell}{d\theta_k}\right)'\frac{\partial^2 J}{\partial g^2}\frac{dg}{d\hat{b}}\frac{\partial\hat{b}}{\partial y} + \left(\frac{dg}{d\hat{b}}\right)'\frac{\partial^3 J}{\partial g^2\partial\theta_k}\frac{dg}{d\hat{b}}\frac{\partial\hat{b}}{\partial y}\\ &\quad + \left(\frac{dg}{d\hat{b}}\right)'\frac{\partial^3 J}{\partial g^3}\frac{dg}{d\hat{b}}\frac{\partial\hat{b}}{\partial y}\frac{dg}{d\hat{b}_\ell}\frac{\partial\hat{b}_\ell}{\partial\theta_k} + \left(\frac{dg}{d\hat{b}}\right)'\frac{\partial^2 J}{\partial g^2}\frac{d^2 g}{d\hat{b}d\hat{b}_\ell}\frac{\partial\hat{b}}{\partial y}\frac{\partial\hat{b}_\ell}{\partial\theta_k}\\ &\quad + \left(\frac{dg}{d\hat{b}}\right)'\frac{\partial^2 J}{\partial g^2}\frac{dg}{d\hat{b}}\frac{\partial^2\hat{b}}{\partial y\partial\theta_k} + \frac{\partial^2 J}{\partial g\partial\theta_k}\frac{d^2 g}{d\hat{b}^2}\frac{\partial\hat{b}}{\partial y}\\ &\quad + \left(\frac{dg}{d\hat{b}_\ell}\frac{\partial\hat{b}_\ell}{\partial\theta_k}\right)'\frac{\partial^2 J}{\partial g^2}\frac{d^2 g}{d\hat{b}^2}\frac{\partial\hat{b}}{\partial y} + \frac{\partial J}{\partial g}\frac{d^3 g}{d\hat{b}^2 d\hat{b}_\ell}\frac{\partial\hat{b}}{\partial y}\frac{\partial\hat{b}_\ell}{\partial\theta_k}\\ &\quad + \frac{\partial J}{\partial g}\frac{d^2 g}{d\hat{b}^2}\frac{\partial^2\hat{b}}{\partial y\partial\theta_k}\end{aligned} \tag{30}$$

$$\begin{aligned}\frac{\partial^2\hat{b}}{\partial y\partial\theta} = -&\left\{\left(\frac{dg}{d\hat{b}}\right)'\frac{\partial^2 J}{\partial g^2}\frac{dg}{d\hat{b}} + \frac{\partial J}{\partial g}\frac{d^2 g}{d\hat{b}^2}\right\}^{-1}\\ &\left\{\frac{\partial^3 J}{\partial g\partial y\partial\theta_k}\frac{dg}{d\hat{b}} + \frac{\partial^3 J}{\partial g\partial y\partial g}\frac{dg}{d\hat{b}}\frac{dg}{d\hat{b}_\ell}\frac{d\hat{b}_\ell}{d\theta_k} + \frac{\partial^2 J}{\partial g\partial y}\frac{d^2 g}{d\hat{b}d\hat{b}_\ell}\frac{d\hat{b}_\ell}{d\theta_k}\right.\\ &+\left(\frac{d^2 g}{d\hat{b}d\hat{b}_\ell}\frac{d\hat{b}_\ell}{d\theta_k}\right)'\frac{\partial^2 J}{\partial g^2}\frac{dg}{d\hat{b}}\frac{\partial\hat{b}}{\partial y} + \left(\frac{dg}{d\hat{b}}\right)'\frac{\partial^3 J}{\partial g^2\partial\theta_k}\frac{dg}{d\hat{b}}\frac{\partial\hat{b}}{\partial y}\\ &+\left(\frac{dg}{d\hat{b}}\right)'\frac{\partial^3 J}{\partial g^3}\frac{dg}{d\hat{b}}\frac{\partial\hat{b}}{\partial y}\frac{dg}{d\hat{b}_\ell}\frac{d\hat{b}_\ell}{d\theta_k} + \left(\frac{dg}{d\hat{b}}\right)'\frac{\partial^2 J}{\partial g^2}\frac{d^2 g}{d\hat{b}d\hat{b}_\ell}\frac{\partial\hat{b}}{\partial y}\frac{\partial\hat{b}_\ell}{\partial\theta_k}\\ &+\frac{\partial^2 J}{\partial g\partial\theta_k}\frac{d^2 g}{d\hat{b}^2}\frac{\partial\hat{b}}{\partial y} + \left(\frac{dg}{d\hat{b}_\ell}\frac{\partial\hat{b}_\ell}{\partial\theta_k}\right)'\frac{\partial^2 J}{\partial g^2}\frac{d^2 g}{d\hat{b}^2}\frac{\partial\hat{b}}{\partial y}\\ &\left.+\frac{\partial J}{\partial g}\frac{d^3 g}{d\hat{b}^2 d\hat{b}_\ell}\frac{\partial\hat{b}}{\partial y}\frac{\partial\hat{b}_\ell}{\partial\theta_k}\right\}\end{aligned} \tag{31}$$

REFERENCES

1. Bates DM, Watts DB. *Nonlinear Regression Analysis and Its Applications*. Wiley: New York, 1988.
2. Gelman A, Bois FY, Jiang J. Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *J. Am. Stat. Assoc.* 1996; **91**(436): 1400–1412.
3. Wu H, Zhu H, Miao H, Perelson AS. Identifiability and statistical estimation of dynamic parameters in HIV/AIDS dynamic models. *Bull. Math. Biol.* 2008; **70**: 785–799.
4. Esposito WR, Floudas C. Deterministic global optimization in nonlinear optimal control problems. *J. Glob. Optim.* 2000; **17**: 97–126.
5. Varah JM. A spline least squares method for numerical parameter estimation in differential equations. *SIAM J. Sci. Stat. Comput.* 1982; **3**(1): 28–46.
6. Voit EO, Savageau MA. Power-law approach to modeling biological-systems. 3. Methods of analysis. *J. Ferment. Technol.* 1982; **60**: 233–241.
7. Chen J, Wu H. Efficient local estimation for time-varying coefficients in deterministic dynamic models with applications to HIV-1 dynamics. *J. Am. Stat. Assoc.* 2008; **103**(481): 369–384.
8. Calderhead B, Girolami M, Lawrence N. Accelerating Bayesian inference over nonlinear differential equations with Gaussian processes. *Adv. Neural Inf. Process. Syst.* 2009; **21**: 219–224.

9. Poyton AA, Varziri MS, McAuley KB, McLellan PJ, Ramsay JO. Parameter estimation in continuous-time dynamic models using principal differential analysis. *Comput. Chem. Eng.* 2006; **30**: 698–708.
10. Ramsay JO, Silverman BW. *Functional Data Analysis*, (2nd edn). Springer: New York, 2005.
11. Campbell D, Steele RJ. Smooth functional tempering for nonlinear differential equation models. *Stat. Comput.* 2011, DOI: 10.1007/s11222-011-9234-3.
12. Ramsay JO, Hooker G, Campbell D, Cao J. Parameter estimation for differential equations: a generalized smoothing approach (with discussion). *J. Roy. Stat. Soc., Series B* 2007; **69**(part 5): 1–30.
13. Varziri MS, McAuley KB, McLellan PJ. Approximate maximum likelihood parameter estimation for nonlinear dynamic models: application to a laboratory-scale nylon reactor model. *Ind. Eng. Chem. Res.* 2008; **47**: 7274–7283.
14. Zheng W, McAuley KB, Marchildon EK, Zhen Yao K. Effects of end-group balance on melt-phase nylon 612 polycondensation: experimental study and mathematical model. *Ind. Eng. Chem. Res.* 2005; **44**: 2675–2686.
15. Schaffer MA, McAuley KB, Cunningham MF, Marchildon EK. Experimental study and modeling of nylon polycondensation in the melt phase. *Ind. Eng. Chem. Res.* 2003; **42**: 2946.
16. Iacus SM. *Simulation and Inference for Stochastic Differential Equations*. Springer-Verlag New York: Springer e-books: New York, NY, 2008.
17. Neyman J, Scott EL. Consistent estimates based on partially consistent observations. *Econometrika* 1948; **16**: 1–32.
18. Varziri MS, McAuley KB, McLellan PJ. Approximate maximum likelihood parameter estimation for nonlinear dynamic models: application to a laboratory-scale nylon reactor model. *Ind. Eng. Chem. Res.* 2008; **47**: 7274–7283.
19. Sen A, Srivastava M. *Regression Analysis: Theory, Methods and Applications*. Springer-Verlag: New York, 1990.
20. Hooker G. Forcing function diagnostics for nonlinear dynamics. *Biometrics* 2009; **65**(3): 928–936. DOI: 10.1111/j.1541-0420.2008.01172.x.
21. Huang Y, Wu H. A Bayesian approach for estimating antiviral efficacy in HIV dynamic models. *J. Appl. Stat.* 2006; **33**(2): 155–174.

