Parameter Estimation for Differential Equations: A Generalized Smoothing Approach

J. O. Ramsay, G. Hooker, D. Campbell and J. Cao

J. O. Ramsay,
Department of Psychology,
1205 Dr. Penfield Ave.,
Montreal, Quebec,
Canada, H3A 1B1.
ramsay@psych.mcgill.ca

Summary. We propose a new method for estimating parameters in models defined by a system of non-linear differential equations. Such equations represent changes in system outputs by linking the behavior of derivatives of a process to the behavior of the process itself. Current methods for estimating parameters in differential equations from noisy data are computationally intensive and often poorly suited to the realization of statistical objectives such as inference and interval estimation. This paper describes a new method that uses noisy measurements on a subset of variables to estimate the parameters defining a system of nonlinear differential equations. The approach is based on a modification of data smoothing methods along with a generalization of profiled estimation. We derive estimates and confidence intervals, and show that these have low bias and good coverage properties, respectively, for data simulated from models in chemical engineering and neurobiology. The performance of the method is demonstrated using real-world data from chemistry and from the progress of the auto-immune disease lupus.

Keywords: Differential equation, dynamic system, functional data analysis, profiled estimation, parameter cascade, estimating equation, Gauss-Newton method

1. Challenges in dynamic systems estimation

1.1. Basic properties of dynamic systems

We have in mind a process that transforms a set of m input functions $\mathbf{u}(t)$ into a set of d output functions $\mathbf{x}(t)$. Dynamic systems model output change directly by linking the output derivatives $\dot{\mathbf{x}}(t)$ to $\mathbf{x}(t)$ itself, as well as to inputs $\mathbf{u}$:

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}, \mathbf{u}, t | \boldsymbol{\theta}), \quad t \in [0, T]. \tag{1}$$

Vector $\boldsymbol{\theta}$ contains any parameters defining the system whose values are not known from experimental data, theoretical considerations or other sources of information. Systems involving derivatives of x of order $n > 1$ are reducible to (1) by defining new variables, $x_1 = x, x_2 = \dot{x}_1, \ldots, x_n = \dot{x}_{n-1}$. Further generalizations of (1) are also candidates for the approach developed in this paper, but will not be considered. Dependencies of $\mathbf{f}$ on t other than through $\mathbf{x}$ and $\mathbf{u}$ arise when, for example, certain quantities defining the system are themselves time-varying.

Differential equations as a rule do not define their solutions uniquely, but rather as a manifold of solutions of typical dimension d. For example, $d^2x/dt^2 = -\omega^2 x(t)$, reduced to $\dot{x}_1 = x_2$ and $\dot{x}_2 = -\omega^2 x_1$, implies solutions of the form $x_1(t) = c_1 \sin(\omega t) + c_2 \cos(\omega t)$, where coefficients c_1 and c_2 are arbitrary; and at least $d = 2$ observations are required to identify the solution that best fits the data. *Initial value* problems supply $\mathbf{x}(0)$, while *boundary value* value problems require d values selected from $\mathbf{x}(0)$ and $\mathbf{x}(T)$.

However, we assume more generally that only a subset $\mathcal{I}$ of the d output variables $\mathbf{x}$ may be measured at time points $t_{ij}, i \in \mathcal{I} \subset \{1, \ldots, d\}; j = 1, \ldots, N_i$, and that y_{ij} is a corresponding measurement that is subject to measurement error $e_{ij} = y_{ij} - x_i(t_{ij})$. We may call such a situation a *distributed partial data* problem. If either there are no observations at 0 and T, or the observations supplied are subject to measurement error, then initial or boundary values may be considered as parameters that must be included in an augmented parameter vector $\boldsymbol{\theta}^* = (\mathbf{x}(0)', \boldsymbol{\theta}')'$.

Solutions of the ordinary differential equation (ODE) system (1) given initial values $\mathbf{x}(0)$ exist and are unique over a neighborhood of $(0, \mathbf{x}(0))$ if f is continuously differentiable or, more generally, Lipschitz continuous with respect to $\mathbf{x}$. However, most ODE systems are

not solvable analytically, which typically increases the computational burden of data-fitting methodology such as nonlinear regression. Exceptions are linear systems with constant coefficients, where the machinery of the Laplace transform and transform functions plays a role, and a statistical treatment of these is available in Bates and Watts (1988) and Seber and Wild (1989). Discrete versions of linear constant coefficient systems, that is, stationary systems of difference equations for equally spaced time points, are also well treated in the classical time series ARIMA and state-space literature, and will not be considered further in this paper.

The insolvability of most ODEs has meant that statistical science has had comparatively little impact on the fitting of dynamic systems to data. Current methods for estimating ODEs from noisy data, reviewed below, are often slow, uncertain to provide satisfactory results, and do not lend themselves well to follow-up analyses such as interval estimation and inference. Moreover, when only a subset of variables in a system are actually measured, the remainder are effectively functional latent variables, a feature that adds further challenges to data analysis. For example, in systems describing chemical reactions, the concentrations of only some reactants are easily measurable and inference may be based on measurements of external quantities such as the temperature of the system.

This paper describes an extension of data smoothing methods along with a generalization of profiled estimation to estimate the parameters $\boldsymbol{\theta}$ defining a system of nonlinear differential equations. High dimensional basis function expansions are used to represent the outputs $\mathbf{x}$, and our approach depends critically on considering the coefficients of these expansions as nuisance parameters. This leads to the notion of a *parameter cascade*, and the impact of nuisance parameters on the estimation of structural parameters is controlled through a multi-criterion optimization process rather than the more usual marginalization procedure.

1.2. Two test-bed problems

1.2.1. FitzHugh-Nagumo equations

These equations were developed by FitzHugh (1961) and Nagumo et al. (1962) as simplifications of the Hodgkin and Huxley (1952) model of the behavior of spike potentials in the giant axon of squid neurons:

$$\begin{aligned}
\dot{V} &= c\left(V - \frac{V^3}{3} + R\right) \\
\dot{R} &= -\frac{1}{c}\left(V - a + bR\right)
\end{aligned} \tag{2}$$

The system describes the reciprocal dependencies of the voltage V across an axon membrane and a recovery variable R summarizing outward currents. Although not intended to provide a close fit to neural spike potential data, solutions to the FitzHugh-Nagumo ODEs do exhibit features common to elements of biological neural networks (Wilson (1999)).

The parameters are $\boldsymbol{\theta} = \{a, b, c\}$, to which we will assign values $(0.2, 0.2, 3)$, respectively. The R equation is the simple constant coefficient linear system $\dot{R} = -(b/c)R$ with linear inputs V and a. However, the V equation is nonlinear; when $V > 0$ is small, $\dot{V} \approx cV$ and consequently exhibits nearly exponential increase, but as V passes $\pm\sqrt{3}$, the influence of $-V^3/3$ takes over and turns V back toward 0. Consequently, solutions corresponding to a range of initial values quickly settle down to alternate between the smooth evolution and the sharp changes in direction shown in Figure 1.
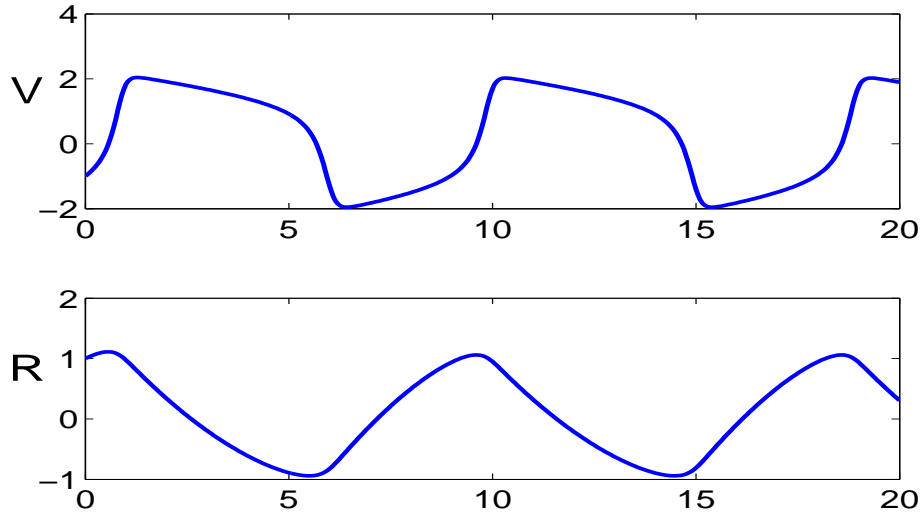
Fig. 1. The limiting behavior of voltage V and recovery R variables defined by the FitzHugh-Nagumo equations (2) with parameter values $a = 0.2, b = 0.2$ and $c = 3.0$ and initial conditions $(V_0, R_0) = (-1, 1)$. The horizontal axis is time in milliseconds.
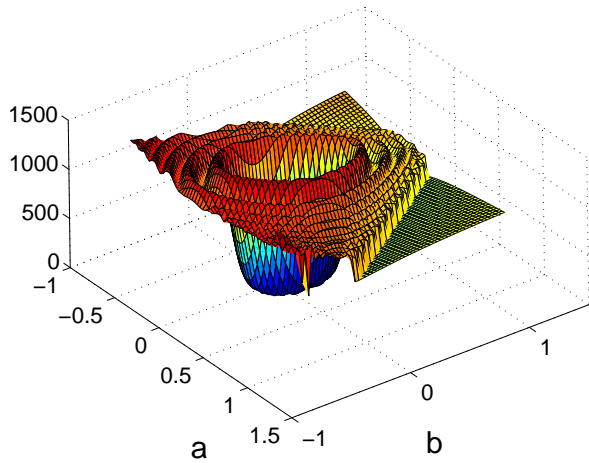

Fig. 2. A response surface for solutions of the FitzHugh-Nagumo equations (2) as parameters a and b are varied. Surface values give the integrated squared difference between solutions at parameters $a = 0.2$, $b = 0.2$ with solutions at the values of a and b given on the x and y axes, respectively; $c = 3$ and initial conditions $V(0) = -1$, $R(0) = 1$ are held constant.

A concern in dynamic systems modeling is the possibly complex nature of the fit surface. The existence of many local minima has been commented on in Esposito and Floudas (2000); and a number of computationally demanding algorithms, such as simulated annealing, have been proposed to overcome this problem. For example, Jaeger et al. (2004) reported using weeks of computation to compute a point estimate. Figure 2 displays the integrated squared difference between the paths in Figure 1 and those resulting from varying only the parameters a and b. The features of this surface include "ripples" due to changes in the shape and period of the limit cycle and breaks due to bifurcations, or sharp changes in behavior.

1.2.2. Tank reactor equations

The chemical engineering concept of a continuously stirred tank reactor (CSTR) consists of a tank surrounded by a cooling jacket containing an impeller which stirs its contents. A fluid containing a reagent with concentration C_{in} enters the tank at a flow rate F_{in} and temperature T_{in}. A reaction produces a product that leaves the tank with concentration C and temperature T. A coolant in the cooling jacket has temperature T_{co} and flow rate F_{co}.

The differential equations used to model a CSTR, simplified by setting the volume of the tank to one, are

$$
\begin{aligned}
\dot{C} &= -\beta_{CC}(T, F_{in})C + F_{in}C_{in} \\
\dot{T} &= -\beta_{TT}(F_{co}, F_{in})T + \beta_{TC}(T, F_{in})C + F_{in}T_{in} + \alpha(F_{co})T_{co}.
\end{aligned}
\tag{3}
$$

The input variables play two roles in the right sides of these equations: through added terms such as $F_{in}C_{in}$ and $F_{in}T_{in}$, and via the weight functions $\beta_{CC}, \beta_{TC}, \beta_{TT}$ and α that multiply the output variables and T_{co}, respectively. These time-varying multipliers depend on four system parameters as follows:

$$
\begin{aligned}
\beta_{CC}(T, F_{in}) &= \kappa \exp[-10^4 \tau(1/T - 1/T_{ref})] + F_{in} \\
\beta_{TT}(F_{co}, F_{in}) &= \alpha(F_{co}) + F_{in} \\
\beta_{TC}(T, F_{in}) &= 130\beta_{CC}(T, F_{in}) \\
\alpha(F_{co}) &= aF_{co}^{b+1}/(F_{co} + aF_{co}^b/2),
\end{aligned}
\tag{4}
$$

where T_{ref} is a fixed reference temperature within the range of the observed temperatures, and in this case was 350 deg K. These functions are defined by two pairs of parameters: (τ, κ) defining coefficient β_{CC} and (a, b) defining coefficient α. The factor 10^4 in β_{CC} rescales τ so that all four parameters are within $[0.4, 1.8]$. These parameters are gathered in the vector $\boldsymbol{\theta}$ in (1), and determine the rate of the chemical reactions involved, or the reaction kinetics.

The plant engineer needs to understand the dynamics of the two output variables C and T as determined by the five inputs $C_{in}, F_{in}, T_{in}, T_{co}$ and F_{co}. A typical experiment designed to reveal these dynamics is illustrated in Figure 3, where we see each input variable stepped up from a baseline level, stepped down, and then returned to baseline. Two baseline levels are presented for the most critical input, the coolant temperature T_{co}.

The behaviors of output variables C and T under the two experimental regimes, given values 0.833, 0.461, 1.678 and 0.5 for parameters τ, κ, a and b, respectively, are shown in Figure 4. When the reactor runs in the cool mode, where the baseline coolant temperature is 335 degrees Kelvin, the two outputs respond smoothly to the step changes in all inputs.
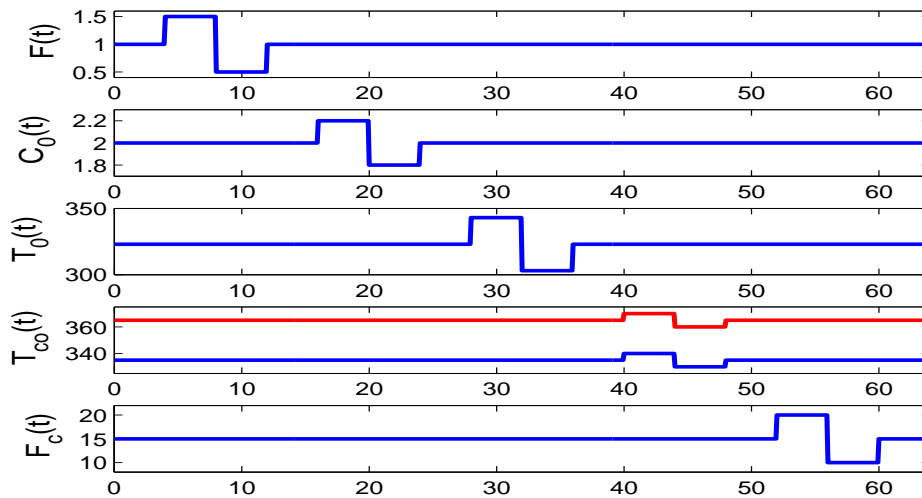
Fig. 3. The five inputs to the chemical reactor modeled by the equations (3) and (4): flow rate $F(t)$, input concentration $C_0(t)$, input temperature $T_0(t)$, coolant temperature $T_{co}(t)$ and coolant flow $F_0(t)$. Coolant temperature $T_{co}(t)$ was set at two baseline levels, cool and hot.
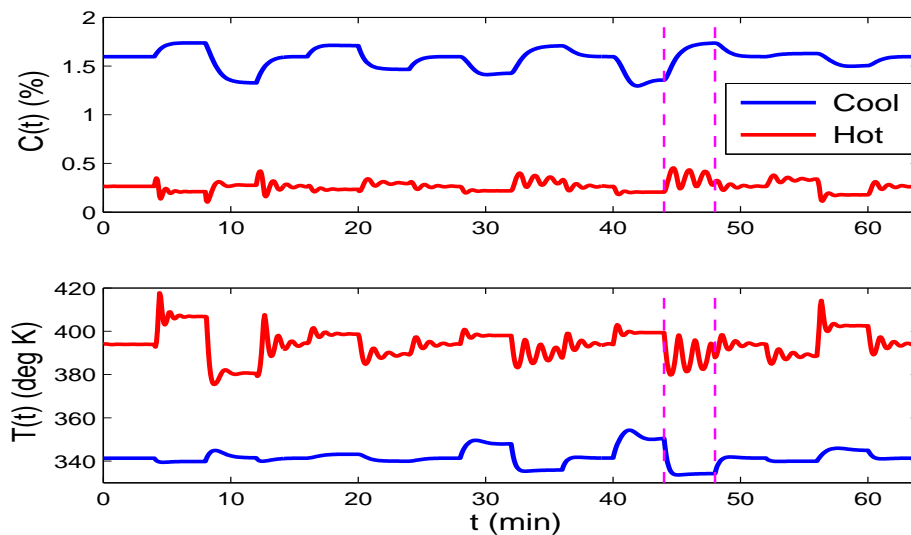

Fig. 4. The two outputs, for each of baseline coolant temperatures T_{co} of 335 and 365 deg. K, from the chemical reactor modeled by the two equations (3): concentration $C(t)$ and temperature $T(t)$. The input functions are shown in Figure 3. Times at which an input variable $T_{co}(t)$ was stepped down and then up are shown as vertical dotted lines.

However, an increase in baseline coolant temperature by 30 degrees Kelvin generates oscillations that come close to instability when the coolant temperature decreases, something that is undesirable in an actual industrial process. These perturbations are due to the double impact of a decrease in output temperature, which increases the size of both β_{CC} and β_{TC}. Increasing β_{TC} raises the forcing term in the T equation, thus increasing temperature. Increasing β_{CC} makes concentration more responsive to changes in temperature, but decreases the size of the response. This push–pull process has a resonant frequency that depends on the kinetic constants, and when the ambient operating temperature reaches a certain level, the resonance appears. For coolant temperatures either above or below this critical zone, the oscillations disappear.

The CSTR equations present two challenges that are not an issue for the Fitz-Hugh Nagumo equations. The step changes in inputs induce corresponding discontinuities in the output derivatives that complicate the estimation of solutions by numerical methods. Moreover, the engineer must estimate the reaction kinetics parameters in order to estimate the cooling temperature range to avoid, but a key question is whether all four parameters are actually estimable given a particular data configuration. Step changes in inputs and near over-parameterization are common problems in dynamic systems modeling.

1.3. Review of current ODE parameter estimation strategies

Procedures for estimating the parameters defining an ODE from noisy data tend to fall into three broad classes: linearization, discretization methods for initial value problems and basis function expansion or collocation methods for boundary and distributed data problems. Linearization involves replacing nonlinear structures by first order Taylor series expansions, and tends only to be useful over short time intervals combined with rather mild nonlinearities, and will not be considered further. There is a large literature on numerical methods for solving constrained optimization problems, under which parameter estimation usually falls; see Biegler and Grossman (2004) for an excellent overview.

1.3.1. Data fitting by numerical approximation of an initial value problem

The numerical methods most often used to approximate solutions of ODEs over a range $[t_0, t_1]$ use fixed initial values $\mathbf{x}_0 = \mathbf{x}(t_0)$ and adaptive discretization techniques (Biegler et al. (1986)). The data fitting process, often referred to by textbooks as the *nonlinear least squares* or *NLS* method, works as follows. A numerical method such as the Runge-Kutta algorithm is used to approximate the solution given a trial set of parameter values and initial conditions, a procedure referred to by engineers as *simulation*. The fit value is input into an optimization algorithm that updates parameter estimates. If the initial conditions $\mathbf{x}(0)$ are unavailable, they must be appended to the parameters $\boldsymbol{\theta}$ as quantities with respect to which the fit is optimized. The optimization process can proceed without using gradients, or these may also be approximated by solving the *sensitivity differential equations*

$$\frac{d}{dt}\left(\frac{d\mathbf{x}}{d\boldsymbol{\theta}}\right) = \frac{\partial \mathbf{f}}{\partial \boldsymbol{\theta}} + \frac{\partial \mathbf{f}}{\partial \mathbf{x}}\frac{d\mathbf{x}}{d\boldsymbol{\theta}}, \quad \text{with} \quad \left.\frac{d\mathbf{x}}{d\boldsymbol{\theta}}\right|_{t=0} = 0. \tag{5}$$

In the event that $\mathbf{x}(0) = \mathbf{x}_0$ must also be estimated, the corresponding sensitivity equations are

$$\frac{d}{dt}\left(\frac{d\mathbf{x}}{d\mathbf{x}_0}\right) = \frac{\partial \mathbf{f}}{\partial \mathbf{x}}\frac{d\mathbf{x}}{d\mathbf{x}_0}, \quad \text{with} \quad \left.\frac{d\mathbf{x}}{d\mathbf{x}_0}\right|_{t=0} = \mathbf{I}. \tag{6}$$

Systems for which solutions beginning at varying initial values tend to converge to a common trajectory are called *stiff*, and require special methods that make use of the Jacobian $\partial f / \partial x$.

The NLS procedure has many problems. It is computationally intensive since a numerical approximation to a possibly complex process is required for each update of parameters and initial conditions. The inaccuracy of the numerical approximation can be a problem, especially for stiff systems or for discontinuous inputs such as step functions or functions concentrating their masses at discrete points. The size of the parameter set may be increased by the set of initial conditions needed to solve the system, and the data may not provide much information for estimating them. NLS also only produces point estimates of parameters, and where interval estimation is needed, a great deal more computation can be required. As a consequence of all this, Marlin (2000) warns process control engineers to expect an error level of the order of 25% in parameter estimates.

A Bayesian approach which may escape minor ripples in the optimization surface is outlined in Gelman et al. (1996). This model uses a likelihood centered on the numerical solution to the differential equation $\hat{\mathbf{x}}(t_j|\hat{\boldsymbol{\theta}})$, such as $y_j \sim N[\hat{\mathbf{x}}(t_j|\boldsymbol{\theta}), \sigma^2]$. Since $\hat{\mathbf{x}}(t_j|\boldsymbol{\theta})$ has no closed form solution, the posterior density for $\boldsymbol{\theta} \mid \mathbf{y}$ has no closed form and inference must be based on simulation from a Metropolis-Hastings algorithm or other sampler. At each iteration of the sampler $\boldsymbol{\theta}$ is proposed and the numerical approximation $\hat{\mathbf{x}}(t_j|\boldsymbol{\theta})$ is used to compute the likelihood. Parallels between this approach and NLS mean that they share many of the same optimization problems. To fix this, the Bayesian model often requires strong finitely bounded priors. Extensions to this method are outlined in Campbell (2007).

1.3.2. Collocation methods or basis function expansions

Our own approach belongs in the family of *collocation* methods that express the approximation $\hat{x}_i$ of x_i in terms of a basis function expansion

$$\hat{x}_i(t) = \sum_k^{K_i} c_{ik}\phi_{ik}(t) = \mathbf{c}_i'\boldsymbol{\phi}_i(t), \tag{7}$$

where the number K_i of basis functions in vector $\boldsymbol{\phi}_i$ is chosen so as to ensure enough flexibility to capture the variation in the approximated function x_i and its derivatives. Typically, this will require substantially more flexibility than is required to fit the data, since $\hat{x}_i$ and $d\hat{x}/dt$ must also satisfy the differential equation to an extent considered acceptable. Although the original collocation methods used polynomial bases, spline basis systems are now preferred because they allow control over the smoothness of the solution at specific values of t, including discontinuities in $d\hat{x}/dt$ or higher order derivatives associated with step and point changes in the inputs $\mathbf{u}$. Using a spline basis to approximate an initial value problem is equivalent to the use of an implicit Runge-Kutta method for stepping points located at the knots defining the basis (Deuflhard and Bornemann (2000)). For solving boundary value problems, collocation tries to satisfy (1) at a discrete set of points; resulting in a large sparse system of nonlinear equations which must then be solved numerically.

Collocation with spline bases was applied to dynamic data fitting problems by Varah (1982), who suggested a two-stage procedure in which each x_i is first estimated by data smoothing methods without considering (1), followed by the minimization of a least squares measure of the fit of $d\hat{\mathbf{x}}/dt$ to $\mathbf{f}(\hat{\mathbf{x}}, \mathbf{u}, t|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. The method is attractive when $\mathbf{f}$ is nearly linear in $\boldsymbol{\theta}$, but nonlinear in $\mathbf{x}$. Varah's approach worked well for the simple equations that were considered, but considerable care was required in the smoothing step

to ensure a satisfactory estimate of $\dot{\mathbf{x}}$, and the technique also required that all variables in the system be measured.

Ramsay and Silverman (2005) and Poyton et al. (2006) took Varah's method further by iterating the two steps, and replacing the previous iteration's roughness penalty by a penalty on $\|d\hat{\mathbf{x}}/dt - f(\hat{\mathbf{x}}, \mathbf{u}, t|\boldsymbol{\theta})\|$ using the last minimizing value of $\boldsymbol{\theta}$. They found that this process, *iterated principal differential analysis* (iPDA), converged quickly to estimates of both $\mathbf{x}$ and $\boldsymbol{\theta}$ that had substantially improved bias and precision. However, iPDA is a joint estimation procedure in the sense that it optimizes a single roughness-penalized fitting criterion with respect to both $\mathbf{c}$ and $\boldsymbol{\theta}$, an aspect that will be discussed further in the next section.

A number of procedures have attempted to solve the parameter estimation problem at the same time as computing a numerical solution to (1). Tjoa and Biegler (1991) proposes to combine a numerical solution of the collocation equations with an optimization over parameters to obtain a single constrained optimization problem, see also Arora and Biegler (2004). Similar ideas can be found in Bock (1983), where the *multiple shooting method* is proposed that breaks the time domain into a series of smaller intervals, over each of which (1) is solved.

1.4. Overview of the paper

Our approach to fitting differential equation models is developed in Section 2, where we develop the concepts of estimating functions and a generalization of profiled estimation. Section 3 tests the method on simulated data for the FitzHugh-Nagumo and CSTR equations, and Section 4 estimates differential equation models for data drawn from chemical engineering and medicine. Generalizations of the method are discussed in Section 5.

2. Generalized profiling estimation procedure

We first give an overview of our estimation strategy, and then provide further details below. As we noted above, our method is a variant of the collocation method, and as such, represents each variable in terms of a basis function expansion (7). Let $\mathbf{c}$ indicate the composite vector of length $K = \sum_{i \in \mathcal{I}} K_i$ that results from concatenating the $\mathbf{c}_i$'s. Let $\boldsymbol{\Phi}_i$ be the N_i by K_i matrix of values $\phi_k(t_{ij})$, and let $\boldsymbol{\Phi}$ be the $N = \sum_{i \in \mathcal{I}} N_i$ by K super–matrix constructed by placing the matrices $\boldsymbol{\Phi}_i$ along the diagonals and zeros elsewhere. According to this notation, we have the composite basis expansion $\hat{\mathbf{x}} = \boldsymbol{\Phi}\mathbf{c}$.

2.1. Overview of the estimation procedure

Defining $\hat{\mathbf{x}}$ as a set of basis function expansions implies that there are two classes of parameters to estimate: the parameters $\boldsymbol{\theta}$ defining the equation, such as the four reaction kinetics parameters in the CSTR equations; and the coefficients in $\mathbf{c}_i$ defining each basis function expansion. The equation parameters are *structural* in the sense of being of primary interest, as are the error distribution parameters in $\boldsymbol{\sigma}_i, i \in \mathcal{I}$. But the coefficients $\mathbf{c}_i$ are considered as *nuisance* parameters that are essential for fitting the data, but usually not of direct concern. The sizes of these vectors are apt to vary with the length of the observation interval, density of observation, and other aspects of the structure of the data; and the number of these nuisance parameters can be orders of magnitude larger than the number of struc-

tural parameters, with a ratio of about 200 applying in the CSTR and FitzHugh-Nagumo problems.

In our profiling procedure, the nuisance parameter estimates are defined to be *implicit* functions $\hat{\mathbf{c}}_i(\boldsymbol{\theta}, \boldsymbol{\sigma}; \boldsymbol{\lambda})$ of the structural parameters, in the sense that each time $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}$ are changed, an *inner* fitting criterion $J(\hat{\mathbf{c}}|\boldsymbol{\theta}, \boldsymbol{\sigma}, \boldsymbol{\lambda})$ is re-optimized with respect to $\hat{\mathbf{c}}$ alone. The estimating function $\hat{\mathbf{c}}_i(\boldsymbol{\theta}, \boldsymbol{\sigma}; \boldsymbol{\lambda})$ is *regularized* by incorporating a penalty term in J that controls the size of the extent that $\hat{\mathbf{x}} = \hat{\mathbf{c}}'\boldsymbol{\phi}$ fails to satisfy the differential equation exactly, in a manner specified below. The amount of regularization is controlled by smoothing parameters in vector $\boldsymbol{\lambda}$. This process of eliminating the direct impact of nuisance parameters on the fit of the model to the data resembles the common practice of eliminating random effect parameters in mixed effect models by marginalizing over $\mathbf{c}$ with respect a prior density.

A data fitting criterion $H(\boldsymbol{\theta}, \boldsymbol{\sigma}|\boldsymbol{\lambda})$ is then optimized with respect to the structural parameters alone. The dependency of H on $(\boldsymbol{\theta}, \boldsymbol{\sigma})$ is two-fold: directly, and implicitly through the involvement of $\hat{\mathbf{c}}_i(\boldsymbol{\theta}, \boldsymbol{\sigma}; \boldsymbol{\lambda})$ in defining the fit $\hat{x}_i$. Because $\hat{\mathbf{c}}_i(\boldsymbol{\theta}, \boldsymbol{\sigma}; \boldsymbol{\lambda})$ is already regularized, criterion H does not require further regularization, and is a straightforward measure of fit such as error sum of squares, log likelihood or some other measure that is appropriate given the distribution of the errors e_{ij}.

For the examples in this paper, $\boldsymbol{\lambda}$ has been adjusted manually using some numerical and visual heuristics. However, we also envisage that $\boldsymbol{\lambda}$ may be estimated automatically through the use of a measure $F(\boldsymbol{\lambda})$ of model complexity or mean squared error, such as the generalized cross-validation or GCV criterion often used in least squares spline smoothing. In this event, the vector $\boldsymbol{\lambda}$ defines a third level of parameters, and leads us to define a *parameter cascade* in which structural parameter estimates are in turn defined to be functions $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$ and $\hat{\boldsymbol{\sigma}}(\boldsymbol{\lambda})$ of regularization or complexity parameters, and nuisance parameters now also become functions of $\boldsymbol{\lambda}$ via their dependency on structural parameters. We have applied this notion to semi-parametric regression in Cao and Ramsay (2006) where the estimation procedure is a multi-criterion optimization problem, and we can refer to J, H and F as *inner*, *middle* and *outer* criteria, respectively. Keilegom and Carroll (2006) use a similar approach, also in semi-parametric regression.

We motivate this approach as follows. Fixing complexity parameters $\boldsymbol{\lambda}$ for the purposes of discussion, we appreciate here, as in random effects modeling and nonparametric regression, that it would be unwise to employ joint estimation using a fixed data-fitting criterion H with respect to all of $\boldsymbol{\theta}, \boldsymbol{\sigma}$ and $\mathbf{c}$ since the overwhelmingly larger number of nuisance parameters would tend to lead to over-fitting the data and consequently unacceptable bias and sampling variance in $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\sigma}}$. By assessing smoothness of the fit $\hat{\mathbf{x}}$ to the data in terms of departure from satisfying (1), we are, in effect, bringing additional "data" into the fitting process in the form of the roughness penalty in much the same way that a Bayesian brings prior information to parameter estimation in the form of the logarithm of a prior density. However, the Bayesian strategy suffers from the problem that the integration in the marginalization process is seldom available analytically, thus leading to computationally intensive MCMC technology. We show here that our parameter cascade approach leads to analytic derivatives required for efficient optimization, and also for linear approximation to interval estimates.

2.2. Data fitting criterion

Let $\mathbf{e}_i$ indicate the vector of errors associated with observed variable $i \in \mathcal{I}$, and let $g_i(\mathbf{e}_i|\boldsymbol{\sigma}_i)$ indicate the joint density of these errors conditional on a parameter vector $\boldsymbol{\sigma}_i$. In practice

it is usual to assume independently distributed Gaussian errors with mean 0 and standard deviation σ_i. However autocorrelation structure and non-stationary variance are often evident in the data, and when these features are also modeled, these parameters are also incorporated into error distribution parameters $\boldsymbol{\sigma}_i$. Let $\boldsymbol{\sigma}$ indicate the concatenation of the $\boldsymbol{\sigma}_i$ vectors. Although our notation is consistent with assuming that errors are independent across variables, inter-variable error dependencies, too, can be accommodated by the approach developed in this paper. In general, the data-fitting criterion can be taken to be the negative log likelihood

$$H(\boldsymbol{\theta}, \boldsymbol{\sigma}|\boldsymbol{\lambda}) = -\sum_{i \in \mathcal{I}} \ln g(\mathbf{e}_i|\boldsymbol{\sigma}_i, \boldsymbol{\theta}, \boldsymbol{\lambda}) \tag{8}$$

where

$$e_{ij} = y_{ij} - \hat{\mathbf{c}}_i(\boldsymbol{\sigma}_i, \boldsymbol{\theta}; \boldsymbol{\lambda})'\boldsymbol{\phi}(t_{ij}).$$

The output variables x_i will as a rule have different units; the concentration of the output in the CSTR equations is a percentage, while temperature is in degrees Kelvin. Consequently, fit measures such as error sum of squares must be multiplied by a normalizing weight w_i that, ideally, should be $1/\sigma_i^2$, so that the normalized error sums of squares are of roughly comparable sizes. However, given enough data per variable, it can suffice to use data-defined values, such as the squared reciprocals of initial values $w_i = x_i(0)$ or the variance taken over values $\hat{x}_i(t_{ij})$ for some trial or initial estimate of a solution of the equation. Letting $\mathbf{y}_i$ indicate the data available for variable i consisting of observations at time points $\mathbf{t}_i$, and $\hat{x}_i(\mathbf{t}_i)$ indicate the vector of fitted values corresponding to $\mathbf{y}_i$, the composite error sum of squares criterion is

$$H(\boldsymbol{\theta}|\boldsymbol{\lambda}) = \sum_{i \in \mathcal{I}} w_i \|\mathbf{y}_i - \hat{x}_i(\mathbf{t}_i)\|^2, \tag{9}$$

where the norm may allow for features like autocorrelation and heteroscedasticity.

2.3. Assessing fidelity to the equations

We may express each equation in (1) as the differential operator equation

$$L_{i, \boldsymbol{\theta}}(x_i) = \dot{x}_i - f_i(\mathbf{x}, \mathbf{u}, t|\boldsymbol{\theta}) = 0. \tag{10}$$

The extent to which an actual function $\hat{x}_i$ satisfies the ODE system can then be assessed by

$$\text{PEN}_i(\hat{\mathbf{x}}) = \int [L_{i, \boldsymbol{\theta}}(\hat{x}_i(t))]^2 dt \tag{11}$$

where the integration is over an interval which contains the times of measurement. The normalization constant w_i may be required here, too, to allow for different units of measurement. Other norms are also possible, and *total variation*, defined as

$$\text{PEN}_i(\hat{\mathbf{x}}) = \int |L_{i, \boldsymbol{\theta}}(\hat{x}_i(t))| dt \tag{12}$$

has turned out to be an important alternative in situations where there are sharp breaks in the function being estimated, such as in image analysis (Koenker and Mizera (2002)). A composite fidelity to equation measure is

$$\text{PEN}(\hat{\mathbf{x}}|\mathbf{L}_{\boldsymbol{\theta}}, \boldsymbol{\lambda}) = \sum_i^n \lambda_i \text{PEN}_i(\hat{\mathbf{x}}) \tag{13}$$

where $\mathbf{L_\theta}$ denotes the vector containing the d differential operators $L_{i,\theta}$. Note that in this case the summation will be over all d variables in the equation. The multipliers $\lambda_i \geq 0$ permit us to weight fidelities differently, and also control the relative emphasis on fitting the data and solving the equation for each variable.

2.4. Estimating $\hat{\mathbf{c}}(\theta; \lambda)$

Finally, the data-fitting and equation-fidelity criteria are combined into the penalized log likelihood criterion

$$J(\mathbf{c}|\theta, \sigma, \lambda) = -\sum_{i \in \mathcal{I}} \ln g(\mathbf{e}_i | \sigma_i, \theta, \lambda) + \text{PEN}(\hat{\mathbf{x}}|\lambda), \tag{14}$$

or the least squares criterion

$$J(\mathbf{c}|\theta, \sigma, \lambda) = \sum_{i \in \mathcal{I}} w_i \|\mathbf{y}_i - \hat{x}_i(\mathbf{t}_i)\|^2 + \text{PEN}_i(\hat{\mathbf{x}}|\lambda). \tag{15}$$

In general the minimization of J will require numerical optimization, but in the least squares case and linear ODEs, it is possible to express $\hat{\mathbf{c}}(\theta; \lambda)$ analytically (Ramsay and Silverman (2005)).

2.5. Optimizing with respect to θ

In this and the remainder of the section, we simplify the notation considerably by dropping the dependency of criterion H on σ and λ; and regarding the latter as a fixed parameter. These results can easily be extended to get the results for the joint estimation of system parameters θ and error distribution parameters σ where required. It is assumed that H is twice continuously differentiable with respect to both θ and $\mathbf{c}$, and that the second partial derivative or Hessian matrices $\partial^2 H / \partial \theta^2$ and $\partial^2 H / \partial \hat{\mathbf{c}}^2$ are positive definite over a nonempty neighborhood $\mathcal{N}$ of $\mathbf{y}$ in data space.

The gradient or total derivative with respect to θ is

$$\frac{dH}{d\theta} = \frac{\partial H}{\partial \theta} + \frac{\partial H}{\partial \hat{\mathbf{c}}} \frac{d\hat{\mathbf{c}}}{d\theta}. \tag{16}$$

Since $\hat{\mathbf{c}}(\theta)$ is not available explicitly, we apply the Implicit Function Theorem to obtain

$$\frac{d\hat{\mathbf{c}}}{d\theta} = -\left(\frac{\partial^2 J}{\partial \hat{\mathbf{c}}^2}\right)^{-1} \frac{\partial^2 J}{\partial \hat{\mathbf{c}} \partial \theta}. \text{ and } \frac{dH}{d\theta} = \frac{\partial H}{\partial \theta} - \frac{\partial H}{\partial \hat{\mathbf{c}}} \left(\frac{\partial^2 J}{\partial \hat{\mathbf{c}}^2}\right)^{-1} \frac{\partial^2 J}{\partial \hat{\mathbf{c}} \partial \theta}. \tag{17}$$

The matrices used in these equations and those below have complex expressions in terms of the basis functions in $\mathbf{\Phi}$ and the functions $\mathbf{f}$ on the right side of the differential equation. Appendix A provides explicit expressions for them for the case of least squares estimation.

2.6. Approximating the sampling variation of $\hat{\theta}$ and $\hat{\mathbf{c}}$

Let $\mathbf{\Sigma}$ be the variance–covariance matrix for $\mathbf{y}$. Making explicit the dependency of H on the data $\mathbf{y}$ by using the notation $H(\theta|\mathbf{y})$, the estimate $\hat{\theta}(\mathbf{y})$ of θ is the solution of the stationary equation $\partial H(\theta, |\mathbf{y})/\partial \theta = 0$. Here and below, all partial derivatives as well as

total derivatives are assumed to be evaluated at $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{c}}(\hat{\boldsymbol{\theta}})$, which are in turn evaluated at $\mathbf{y}$.

The usual δ-method employed in nonlinear least squares produces a variance estimate of the form

$$\text{Var}_{GN}[\hat{\boldsymbol{\theta}}(\mathbf{y})] \approx \sigma^2 \left[\left(\frac{d\hat{\mathbf{x}}}{d\boldsymbol{\theta}} \right)' \left(\frac{d\hat{\mathbf{x}}}{d\boldsymbol{\theta}} \right) \right]^{-1} \tag{18}$$

by making use of the approximation

$$\frac{d^2 H}{d\boldsymbol{\theta}^2} \approx \left(\frac{d\hat{\mathbf{x}}}{d\boldsymbol{\theta}} \right)' \left(\frac{d\hat{\mathbf{x}}}{d\boldsymbol{\theta}} \right).$$

We will instead provide an exact estimation of the Hessian above and employ it with a pseudo δ-method. Although this implies considerably more computation, our experiments in Section 3.1 suggest that this method provides more accurate results that the usual δ-method estimate.

By applying the Implicit Function Theorem to $\partial H / \partial \boldsymbol{\theta}$ as a function of $\mathbf{y}$, we may say that for any $\mathbf{y}$ in $\mathcal{N}$ there exists a value $\hat{\boldsymbol{\theta}}(\mathbf{y})$ satisfying $\partial H / \partial \boldsymbol{\theta} = 0$. By taking the $\mathbf{y}$-derivative of this relation, we obtain:

$$\frac{d}{d\mathbf{y}} \left(\frac{dH}{d\boldsymbol{\theta}} \Big|_{\hat{\boldsymbol{\theta}}(\mathbf{y})} \right) = \frac{d^2 H}{d\boldsymbol{\theta} d\mathbf{y}} \Big|_{\hat{\boldsymbol{\theta}}(\mathbf{y})} + \frac{d^2 H}{d\boldsymbol{\theta}^2} \Big|_{\hat{\boldsymbol{\theta}}(\mathbf{y})} \frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}} = 0 \,, \tag{19}$$

where

$$\frac{d^2 H}{d\boldsymbol{\theta}^2} = \frac{\partial^2 H}{\partial \boldsymbol{\theta}^2} + \frac{\partial^2 H}{\partial \boldsymbol{\theta} \partial \hat{\mathbf{c}}} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \left(\frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} \right)' \frac{\partial^2 H}{\partial \hat{\mathbf{c}} \partial \boldsymbol{\theta}} + \left(\frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} \right)' \frac{\partial^2 H}{\partial \hat{\mathbf{c}}^2} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \frac{\partial H}{\partial \hat{\mathbf{c}}} \frac{\partial^2 \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}^2} \,, \tag{20}$$

and

$$\frac{d^2 H}{d\boldsymbol{\theta} d\mathbf{y}} = \frac{\partial^2 H}{\partial \boldsymbol{\theta} \partial \mathbf{y}} + \frac{\partial^2 H}{\partial \hat{\mathbf{c}} \partial \mathbf{y}} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \frac{\partial^2 H}{\partial \boldsymbol{\theta} \partial \hat{\mathbf{c}}} \frac{\partial \hat{\mathbf{c}}}{\partial \mathbf{y}} + \frac{\partial^2 H}{\partial \hat{\mathbf{c}}^2} \frac{\partial \hat{\mathbf{c}}}{\partial \mathbf{y}} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \frac{\partial H}{\partial \hat{\mathbf{c}}} \frac{\partial^2 \hat{\mathbf{c}}}{\partial \boldsymbol{\theta} \partial \mathbf{y}} \,. \tag{21}$$

The formulas (20) and (21) involve the terms $\partial \hat{\mathbf{c}} / \partial \mathbf{y}$, $\partial^2 \hat{\mathbf{c}} / \partial \boldsymbol{\theta}^2$ and $\partial^2 \hat{\mathbf{c}} / \partial \boldsymbol{\theta} \partial \mathbf{y}$, which can also be derived by the Implicit Function Theorem and are given in Appendix A. Solving (19), we obtain the first derivative of $\hat{\boldsymbol{\theta}}$ with respect to $\mathbf{y}$:

$$\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}} = - \left(\frac{\partial^2 H}{\partial \boldsymbol{\theta}^2} \Big|_{\hat{\boldsymbol{\theta}}(\mathbf{y})} \right)^{-1} \left(\frac{\partial^2 H}{\partial \boldsymbol{\theta} \partial \mathbf{y}} \Big|_{\hat{\boldsymbol{\theta}}(\mathbf{y})} \right). \tag{22}$$

Let $\boldsymbol{\mu} = \text{E}(\mathbf{y})$, the first order Taylor expansion for $d\hat{\boldsymbol{\theta}} / d\mathbf{y}$ is:

$$\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}} \approx \frac{d\hat{\boldsymbol{\theta}}}{d\boldsymbol{\mu}} + \frac{d^2 \hat{\boldsymbol{\theta}}}{d^2 \boldsymbol{\mu}} (\mathbf{y} - \boldsymbol{\mu}) \,. \tag{23}$$

When $d^2 \hat{\boldsymbol{\theta}} / d^2 \boldsymbol{\mu}$ is uniformly bounded, we can take the expectation on both sides of (23) and derive $\text{E}(d\hat{\boldsymbol{\theta}} / d\boldsymbol{\mu}) \approx \text{E}(d\hat{\boldsymbol{\theta}} / d\mathbf{y})$. We can also approximate $\hat{\boldsymbol{\theta}}(\mathbf{y})$ by using the first order Taylor expansion:

$$\hat{\boldsymbol{\theta}}(\mathbf{y}) \quad \approx \quad \hat{\boldsymbol{\theta}}(\boldsymbol{\mu}) + \frac{d\hat{\boldsymbol{\theta}}}{d\boldsymbol{\mu}} (\mathbf{y} - \boldsymbol{\mu}) \,.$$

Taking variance on both sides of (24), we derive

$$\mathtt{Var}[\hat{\boldsymbol{\theta}}(\mathbf{y})] \approx \left[\frac{d\hat{\boldsymbol{\theta}}}{d\boldsymbol{\mu}}\right]\boldsymbol{\Sigma}\left[\frac{d\hat{\boldsymbol{\theta}}}{d\boldsymbol{\mu}}\right]' \approx \left[\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}}\right]\boldsymbol{\Sigma}\left[\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}}\right]', \quad \text{since} \quad \mathrm{E}\left(\frac{d\hat{\boldsymbol{\theta}}}{d\boldsymbol{\mu}}\right) \approx \mathrm{E}\left(\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}}\right). \tag{24}$$

Similarly, the sampling variance of $\hat{\mathbf{c}}[\hat{\boldsymbol{\theta}}(\mathbf{y})]$ is estimated by

$$\mathtt{Var}[\hat{\mathbf{c}}(\hat{\boldsymbol{\theta}}(\mathbf{y}))] = (\frac{d\hat{\mathbf{c}}}{d\mathbf{y}})\boldsymbol{\Sigma}(\frac{d\hat{\mathbf{c}}}{d\mathbf{y}})', \quad \text{where} \quad \frac{d\hat{\mathbf{c}}}{d\mathbf{y}} = \frac{d\hat{\mathbf{c}}}{d\hat{\boldsymbol{\theta}}}\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}} + \frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}. \tag{25}$$

2.7. Numerical integration in the inner optimization

The integrals in $\mathtt{PEN}_i$ will normally require approximation by the linear functional

$$\mathtt{PEN}_i(\hat{\mathbf{x}}) \approx \sum_q^Q v_q[L_i(\hat{x}_i(t_q))]^2 \tag{26}$$

where Q, the evaluation points t_q, and the weights v_q are chosen so as to yield a reasonable approximation to the integrals involved.

Let ξ_ℓ indicate a knot location or a breakpoint, and recall that there will be multiple knots at such a location in order to deal with step function inputs that will imply discontinuous derivatives. We divide each interval $[\xi_\ell, \xi_{\ell+1}]$ into four equal-sized intervals, and using Simpson's rule weights $[1, 4, 2, 4, 1](\xi_{\ell+1} - \xi_\ell)/5$. The total set of these quadrature points and weights along with basis function values may be saved at the beginning of the computation so as to save time. If a B-spline basis is used, improvements in speed of computation may be achieved by using sparse matrix methods.

Efficiency in the inner optimization is essential since this will be invoked far more often than the outer optimization. In the case of least squares fitting, the minimization of (14) can be expressed as a large nonlinear least squares approximation problem by observing that we can express the numerical quadrature approximation to $\sum_i \lambda_i \mathtt{PEN}_i(\hat{\mathbf{x}})$ as

$$\sum_i \sum_q [(\lambda_i v_q)^{1/2} L_i(\hat{x}_i(t_q))]^2.$$

These squared residuals can then be appended to those in H, and Gauss-Newton minimization can then be used.

2.8. Choosing the amount of smoothing

We now consider two rationales for choosing $\boldsymbol{\lambda}$, corresponding to the need for robustness with respect to poor initial parameter values or model mis-specification, respectively. Although $\boldsymbol{\lambda}$ was chosen manually for our examples, this choice can be automated under either paradigm, and we suggest some ways of doing so.

2.8.1. Robustness with respect to initial parameter values

Figure 2 shows the severe non-convexity of least-squares fitting criteria for $\boldsymbol{\theta}$ when using an exact solution of the FitzHugh-Nagumo ODE, implying a small neighborhood of the optimal parameter values from which convergence is assured using the Gauss-Newton method.

However, Figure 5, displaying the much more regular surface corresponding to $\boldsymbol{\lambda} = 10^5$, suggests a much wider region of convergence; and our experience for other problems confirms this robustness with respect to poor initialization of parameters for smaller $\boldsymbol{\lambda}$ values. Because the criterion $H(\boldsymbol{\theta}, \boldsymbol{\sigma}|\boldsymbol{\lambda})$ is increasing in each λ_i, it underestimates the response surface for exact solutions to the differential equation. Moreover, results in Appendix A imply that $\|d\mathbf{c}/d\boldsymbol{\theta}\|$ increases in $\boldsymbol{\lambda}$, implying that relaxing the differential equation model regularizes the search for $\boldsymbol{\theta}$.

However, as $\boldsymbol{\lambda}$ becomes smaller, the estimates obtained for $\boldsymbol{\theta}$ become both more biased and more variable. Theorem 2.2, on the other hand, demonstrates that, ignoring error due to (7), parameter estimates must approximate those that would have been obtained from a straightforward maximum-likelihood fit as $\boldsymbol{\lambda}$ increases. This suggests the following algorithm:

(a) Choose initial value $\boldsymbol{\lambda}_0$ so that $H(\boldsymbol{\theta}|\boldsymbol{\sigma}, \boldsymbol{\lambda}_0)$ dominates $\texttt{PEN}(\hat{\mathbf{x}}|\mathbf{L}_{\boldsymbol{\theta}}, \boldsymbol{\lambda}_0)$.
(b) Increase $\boldsymbol{\lambda}_i$ iteratively, and estimate $\boldsymbol{\theta}_i$, initializing the Gauss-Newton algorithm with parameter estimates $\boldsymbol{\theta}_{i-1}$. We typically choose $\lambda_i = 10^{i-k}$ where k represents a starting value.
(c) Stop when $\boldsymbol{\lambda}_0$ becomes so large that the collocation approximation (7) starts to distort the estimate of $\mathbf{x}$.

In order to assess when $\boldsymbol{\lambda}$ has become too large:

(a) Calculate solutions $\tilde{\mathbf{x}}(t)$ to (1) with the current estimate of $\boldsymbol{\theta}$ and $\mathbf{x}_0$.
(b) Smooth $\tilde{\mathbf{x}}(\mathbf{t})$, the solution at the observation times, using the model-based criterion (14) to get an estimate $\tilde{\mathbf{x}}^*$.
(c) Stop when $\|\tilde{\mathbf{x}} - \tilde{\mathbf{x}}^*\|$ begins to increase after attaining a minimum.

We have observed that there is usually a large range of $\boldsymbol{\lambda}$ values that provide stable and accurate estimates for $\boldsymbol{\theta}$.

For the simulated examples in Section 3 and for the Nylon production data, we chose $\boldsymbol{\lambda}$ large enough to guarantee that we could reproduce solutions to (1) to a visually high degree of accuracy without suffering distortion from the use of a basis expansion.

2.8.2. *Robustness with respect to model mis-specification*

For the Lupus data in Section 4.2, the ODE model provides only a partially adequate fit to the data, and consequently the optimal value of $\boldsymbol{\lambda}$ is not infinite. In such situations, a practical method of choosing $\boldsymbol{\lambda}$ is by visual inspection of the fit to the observed data, aided by examining the corresponding ODE solution at the estimated parameters. Initial conditions $\mathbf{x}(0)$ may be taken from the smooth $\hat{\mathbf{x}}(0)$, or may be separately optimized.

When the objective is filtering the data, a GCV-type approach may be appropriate. The estimation of $\hat{\mathbf{x}}$ given $\boldsymbol{\lambda}$ is in general a nonlinear problem, so standard cross-validation measures are not available. Instead, the following GCV-like criterion has been adapted from Wahba (1990):

$$F(\boldsymbol{\lambda}) = \frac{\sum_{\mathcal{I}} \|\mathbf{y}_i - \hat{x}_i(\mathbf{t}_i)\|^2}{\left[\sum_{\mathcal{I}} \left(N_i - \sum_j \frac{d\hat{x}_i(t_{ij})}{dy_{ij}}\right)\right]^2}, \tag{27}$$
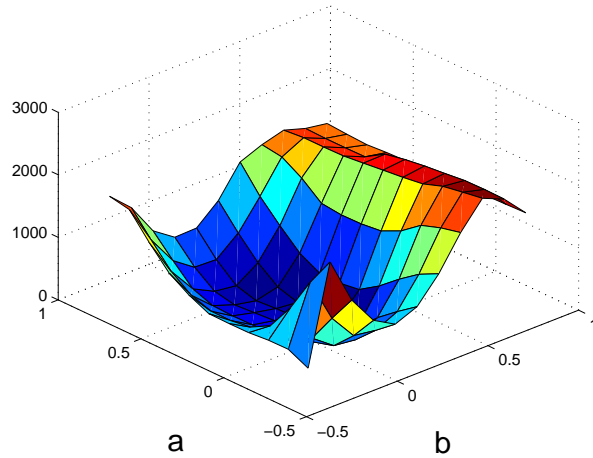
Fig. 5. The squared discrepancy between exact solutions to the FitzHugh-Nagumo equations and a model based smooth that minimizes (14) with $\lambda = 10^5$. Values of the surface are calculated using the same data as in Figure 2.

where the derivatives in the denominator are exactly the diagonal elements of the smoothing matrix in a linear smoothing problem. For the profiling procedure outlined above we have

$$\frac{d\hat{x}_i(t_{ij})}{dy_{ij}} = \frac{\partial \hat{x}_i(t_{ij})}{\partial \mathbf{c}} \frac{d\mathbf{c}}{dy_{ij}}$$

where $d\hat{x}_i(t_{ij})/d\mathbf{c}$ is simply the value of the basis expansion (7) at t_{ij} and $d\mathbf{c}/d\mathbf{y}$ has been calculated in (24). Note that this explicitly takes the dependence of $\hat{\mathbf{y}}$ on $\hat{\boldsymbol{\theta}}$ into account. This construction is offered as speculation; it is well known that the first order approximation used in $F(\boldsymbol{\lambda})$ can be biased (Friedman and Silverman (1989)). Furthermore, $F(\boldsymbol{\lambda})$ is only indirectly related to $\boldsymbol{\theta}$, and our experience suggests that, for mis-specified models, estimators based on cross-validation tend select $\boldsymbol{\lambda}$ at values that produce good estimates of $\mathbf{x}$, but which are smaller than optimal for estimating $\boldsymbol{\theta}$.

2.9. Parameter estimate behavior as $\lambda \to \infty$

In this section, we consider the behavior of our parameter estimate as λ becomes large. This analysis takes an idealized form in the sense that we assume that this optimization may be done globally and that the function being estimated can be expressed exactly and without the approximation error that would come from a basis expansion. We show that as λ becomes large, the estimates defined through our profiling procedure converge to the estimates that we would obtain if we estimated $\boldsymbol{\theta}$ by minimizing negative log likelihood over both $\boldsymbol{\theta}$ and the initial conditions $\hat{\mathbf{x}}_0$. In other words, we treat $\hat{\mathbf{x}}_0$ as nuisance parameters and estimate $\boldsymbol{\theta}$ by profiling. When $\mathbf{f}$ is Lipschitz continuous in $\hat{\mathbf{x}}$ and continuous in $\boldsymbol{\theta}$, the

likelihood is continuous in $\boldsymbol{\theta}$ and the usual consistency theorems (e.g. Cox and Hinkley (1974)) hold and in particular, the estimate $\hat{\boldsymbol{\theta}}$ is asymptotically unbiassed.

For the purposes of this section, we will make a few simplifying conventions. Firstly, we will take:

$$l(\mathbf{x}) = -\sum_{i \in \mathcal{I}} \ln g(\mathbf{e}_i | \boldsymbol{\sigma}_i, \boldsymbol{\theta}, \boldsymbol{\lambda}).$$

Secondly, we will represent

$$\mathtt{PEN}(\mathbf{x}|\boldsymbol{\theta}) = \sum_{i=1}^{n} c_i w_i \int \left(\dot{x}_i(t) - f_i(\mathbf{x}, \mathbf{u}, t|\boldsymbol{\theta}) \right)^2 dt$$

where the c_i are taken to be constants and the λ_i used in the definition (13) are given by λc_i for some λ.

We will also assume that solutions to the data fitting problem exist and are well defined, and that there are objects $\mathbf{x}$ that satisfy $\mathtt{PEN}(\mathbf{x}|\boldsymbol{\theta}) = 0$. Such objects are guaranteed to exist *locally* whenever $\mathbf{f}$ is locally Lipschitz continuous. That is, there is a time interval $[t_0, t_0 + h]$ on which $\mathbf{x}$ exists. On this interval $\mathbf{x}$ is uniquely determined by $\mathbf{x}(t_0)$; see Deuflhard and Bornemann (2000). Existence on the interval of the experiment is more difficult to show in general.

Finally, we will need to make some assumptions about the spline smooths minimizing

$$l(\mathbf{x}) + \lambda \mathtt{PEN}(\mathbf{x}|\boldsymbol{\theta}).$$

Specifically, we will assume that the minimizers of these are well-defined and bounded uniformly over λ. Guarantees on boundedness may be given whenever $\mathbf{x} \cdot \mathbf{f}(\mathbf{x}, \mathbf{u}, t|\boldsymbol{\theta}) < 0$ for $\|\mathbf{x}\|$ greater than some K (see Hooker (2007)). This condition is also sufficient for the global uniqueness of solutions to (1). It is true for reasonable parameter values in all systems presented in this paper. More general characteristics of functions $\mathbf{f}$ for which these properties hold is a matter of continued research.

Solutions of interest lie in the Hilbert space $\mathcal{H} = (W^1)^n$; the direct sum of n copies of W^1 where W^1 is the Sobolev space of functions on the the time-observation interval $[t_1, t_2]$ whose first derivatives are square integrable. The analysis will examine both inner and outer optimization problems as $\lambda \to \infty$. For the inner optimization, we can show

THEOREM 2.1. *Let $\lambda_k \to \infty$ and assume that*

$$\mathbf{x}_k = \operatorname*{argmin}_{\mathbf{x} \in (W^1)^n} l(\mathbf{x}) + \lambda_k \mathtt{PEN}(\mathbf{x}|\boldsymbol{\theta})$$

is well defined and uniformly bounded over λ. Then $\mathbf{x}_k$ converges to $\mathbf{x}^*$ with $\mathtt{PEN}(\mathbf{x}^*|\boldsymbol{\theta}) = 0$.

Further, when $\mathtt{PEN}(\mathbf{x}|\boldsymbol{\theta})$ is given by (13), $\mathbf{x}^*$ is the solution of the differential equations (1) that is obtained by minimizing squared error over the choice of initial conditions. The proof of this, and of the theorem below, is given in Hooker (2007).

Turning to the estimation of $\boldsymbol{\theta}$, we obtain the following:

THEOREM 2.2. *Let $\mathcal{X} \subset (W^1)^n$ and $\Theta \subset \mathbb{R}^p$ be bounded. Assume that for $\lambda > K$,*

$$\mathbf{x}_{\boldsymbol{\theta}, \lambda} = \operatorname*{argmin}_{\mathbf{x} \in \mathcal{X}} l(\mathbf{x}) + \lambda \mathtt{PEN}(\mathbf{x}|\boldsymbol{\theta})$$

is well defined for each $\boldsymbol{\theta}$. *Define* $\mathbf{x}_{\boldsymbol{\theta}}^{*}$ *to be such that*

$$l(\mathbf{x}_{\boldsymbol{\theta}}^{*}) = \min_{\mathbf{x}:P(\mathbf{x}|\boldsymbol{\theta})=0} l(\mathbf{x})$$

and let

$$\boldsymbol{\theta}(\lambda) = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}}\, l(\mathbf{x}_{\boldsymbol{\theta},\lambda})\ and\ \boldsymbol{\theta}^{*} = \underset{\boldsymbol{\theta}\in\Theta}{\operatorname{argmin}}\, l(\mathbf{x}_{\boldsymbol{\theta}}^{*})$$

also be well defined. Then

$$\lim_{\lambda\to\infty} \boldsymbol{\theta}(\lambda) = \boldsymbol{\theta}^{*}.$$

The conditions listed in this theorem are natural, in the sense that we merely require that the smoothing, parameter estimation and NLS optimization problems to have unique solutions. However, verifying that this is the case, even for the NLS problem, many not be straightforward for any given $\mathbf{f}$. We note a substantial literature on system identifiability: for example Denis-Vidal et al. (2003). We conjecture that it will hold for any $\mathbf{f}$ such that the parameter estimation problem is well defined for exact solutions to (1).

Taken together, these theorems state that as λ is increased, the solutions obtained from this scheme tend to those that would be obtained by estimating the parameters directly while profiling out the initial conditions. In particular, the path of parameter values as λ changes is continuous, motivating a successive approximation scheme. This analysis also highlights the distinction between these methods and traditional smoothing; our penalties are highly informative and it is, in fact, the data which plays the minor role in finding a solution.

3. Simulated data examples

3.1. *Fitting the FitzHugh-Nagumo equations*
We set up simulated data for V alone by adding Gaussian error with standard deviation 0.5 to the solution for parameters $\{a, b, c\} = \{0.2, 0.2, 3\}$ and initial conditions $\{V, R\} = \{-1, 1\}$ at times $0.0, 0.05, \ldots, 20.0$. Collocation fit $\hat{\mathbf{x}}$ was a third order B-spline with knots at each data point.

Figure 6 gives quartiles of the parameter estimates for 60 simulations as λ is varied from 10^{-2} to 10^{5}. There is large bias for small values of λ, where smoothing is emphasized and $\boldsymbol{\theta}$ has little impact on $\hat{\mathbf{c}}$; but, as λ increases, parameter estimates become nearly unbiased. Table 3.1 provides bias and variance estimates from 500 simulations at $\lambda = 10^{4}$, along with our estimate (24) and the Gauss-Newton standard error (18). We obtain good coverage properties for our estimates of variance while the Gauss-Newton estimates are somewhat less accurate. We note, however, that computing (24) increased computational effort by a factor of about 10 for this simulation. As a practical matter, using (18) may be considered sufficient if (24) becomes too costly.

3.2. *Fitting the tank reactor equations*
Data for concentration C and temperature T were simulated by adding zero mean Gaussian noise with standard deviations 0.0223 and 0.79, respectively to the values for the cool mode experimental condition shown in Figure (4). These error levels were about 20% of the variation of the respective outputs over the experimental conditions, an error level considered
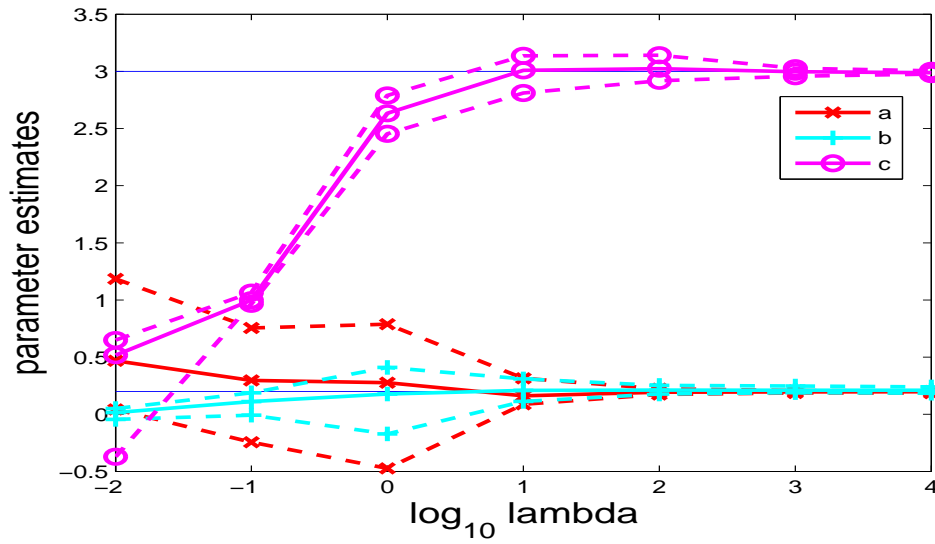
Fig. 6. 25%, 50% and 75% quantiles of parameter estimates for the FitzHugh-Nagumo Equations as λ is varied. Horizontal lines represent the true parameter values.

Table 1. Summary statistics for parameter estimates for 500 simulated samples of data generated from the FitzHugh-Nagumo equations.

	a	b	c
True value	0.2000	0.2000	3.0000
Mean value	0.2005	0.1984	2.9949
Bias Std. Err.	0.0007	0.0029	0.0012
Actual Std. Dev.	0.0149	0.0643	0.0264
Estimate (24) Std. Dev.	0.0143	0.0684	0.0278
Estimate (18) Std. Dev.	0.0167	0.0595	0.0334

Table 2. Summary statistics for parameter estimates for 1000 simulated samples. Results are for measurements on both concentration and temperature, and also for temperature measurements only. The estimate of the standard deviation of parameter values is by the delta method usual in nonlinear least squares analyses.

	C and T data			Only T data		
	κ	τ	a	κ	τ	a
True value	0.4610	0.8330	1.6780	0.4610	0.8330	1.6780
Mean value	0.4610	0.8349	1.6745	0.4613	0.8328	1.6795
Bias Std. Err.	0.0002	0.0004	0.0012	0.0005	0.0005	0.0024
Actual Std. Dev.	0.0034	0.0057	0.0188	0.0084	0.0085	0.0377
Estimate (18) Std. Dev.	0.0035	0.0056	0.0190	0.0088	0.0090	0.0386

typical for many chemical engineering processes. We estimated only the parameters κ, τ and a, keeping b fixed at 0.5 because we had determined that the accurate estimation of all four parameters is impossible within the data design described above. Since the data are generated here from functions satisfying the differential equation system, we can expect the fit to improve with larger and larger values for smoothing parameters λ_C and λ_T. Results are reported here for 100 and 10, respectively, which are sufficiently large that further increases were found to yield negligible improvement in parameter estimates.

We found, in applying the NLS method described in Section 1.3.1, that the approximation to $T(t)$ at the times of input step changes using the Runge-Kutta algorithm were inaccurate and unstable with respect to small changes in parameters. As a consequence, the estimation of the gradient of fit (9) by differencing was so unstable that gradient-free optimization was impossible. When we estimated the gradient by solving the sensitivity equations (5) and (6), we could only achieve optimization when starting values for parameters and initial values were much closer to the optimal values than could be realized in practice. By contrast, our approach was able to converge reliably from random starting values far removed from the optimal estimates.

Table 3.2 displays bias and sampling precision results for parameter estimates by our parameter cascade method for 1000 simulated samples for each of two measurement regimes: both variables measured, and only temperature measured. The first two lines of the table compare the true parameter values with the mean estimates, and the last two lines compare the biases of the estimates with the standard errors of the mean estimates. We see that the estimation biases can be considered negligible for both measurement situations. The third and fourth lines compare the actual standard deviations of the parameter estimates with the values estimated with the Gauss-Newton method in (18), and the two values seem sufficiently close for all three parameters to permit us to trust the Gauss-Newton estimates in this case. As one might expect, the main impact of having only temperature measurements is to increase the sampling error in the parameter estimates.

When the equations were solved using the parameters estimated from measurements on both variables, the maximum absolute discrepancy between the fitted and true curves was 0.11% and 0.03%, respectively, and when these parameter estimates were used for the hot mode of operation, the the discrepancies became 1.72% and 0.05%, respectively. Finally, when the parameters were estimated from only the temperature data, the concentration and temperature discrepancies in cool mode became 0.10% and 0.04%, respectively, so that using only the quickly and cheaply attainable temperature measurements is sufficient for identifying this system in either mode of operation.

4. Two real data examples

4.1. Modeling nylon production

If water (W) in the form of steam is bubbled through molten nylon (L) under high temperatures, W will split L into amine (A) and carboxyl (C) groups. To produce nylon, on the other hand, A and C are mixed together under high temperatures, and their reaction produces L and W, water then escaping as steam. These competing reactions are depicted symbolically by $A + C \rightleftharpoons L + W$. The reaction dynamic equations are

$$
\begin{aligned}
-\dot{L} = \dot{A} = \dot{C} &= -k_p * 10^{-3}(CA - LW/K_a) \\
\dot{W} &= k_p * 10^{-3}(CA - LW/K_a) - k_m(W - W_{eq})
\end{aligned} \tag{28}
$$

where

$$
K_a = \left[\left(1 + \frac{g}{1000}W_{eq}\right)C_T\right]K_{a0}\exp\left[-\frac{\Delta H}{R}\left(\frac{1}{T} - \frac{1}{T_0}\right)\right]
$$

and $R = 8.3145 * 10^{-3}$, $C_T = 20.97\exp[-9.624 + 3613/T]$ and a reference temperature $T_0 = 549.15$ was chosen to be in the middle of the range of experimentally manipulated temperatures. Rate parameter $k_m = 24.3$ was estimated in previous studies. Due to the reaction mass balance, if A, C and W are known then L can be algebraically removed from the equations, so that we will only estimate those three components.

In an experiment described in Zheng et al. (2005), a mixture of steam and an inert gas was bubbled into molten nylon to maintain a constant W, causing A, C, L and W to move towards equilibrium concentrations. Within each of six experimental runs the steam pressure was stepped down from its initial level at times $\tau_{i1}, i = 1, \ldots, 6$, and then returned to its initial pressure at times τ_{i2}. The temperature T_i and concentration difference $A_i(t) - C_i(t)$ varied over runs but were constant within a run. Samples of the molten mixture were extracted at irregularly spaced intervals, and the A and C concentrations measured. The goal was to estimate the rate parameters $\boldsymbol{\theta} = [k_p, g, K_{a0}, \Delta H]$. Figure 7 shows the data for the runs aligned by experiment within columns. Since concentrations of A and C are expected to differ only by a vertical shift, their plots within an experimental run are shifted versions of the same vertical spread. The temperature of each run is given above the plots for each set of components.

The profile estimation process was run initially with $\lambda = 10^{-4}$. Upon convergence of $\hat{\boldsymbol{\theta}}$, λ was increased by a factor of ten and the estimation process rerun using the most recent estimates as the latest set of initial parameter guesses, increasing λ up to 10^3. Beginning with such a small value of λ made the results robust to choice of initial parameter guesses. Further details concerning the data analysis are available in Campbell et al. (2007).

The parameter estimates along with 95% limits were: $k_p = 20.59 \pm 3.26$, $g = 26.86 \pm 6.82$, $K_{a0} = 50.22 \pm 6.34$ and $\Delta H = -36.46 \pm 7.57$. The solutions to the differential equations using the final parameter estimates for $\hat{\boldsymbol{\theta}}$ and the initial system states estimated by the data smooth are shown in Figure 7.

4.2. Modeling flare dynamics in lupus

Lupus is a disease characterized by sudden flares of symptoms caused by the body's immune system attacking various organs. The name derives from a rash on the face and chest that is characteristic, but the most serious effects tend to be in the kidneys. The resulting nephritis and other symptoms can require immediate treatment, usually with the drug Prednisone, a corticosteroid that itself has serious long-term side effects such as osteoporosis.
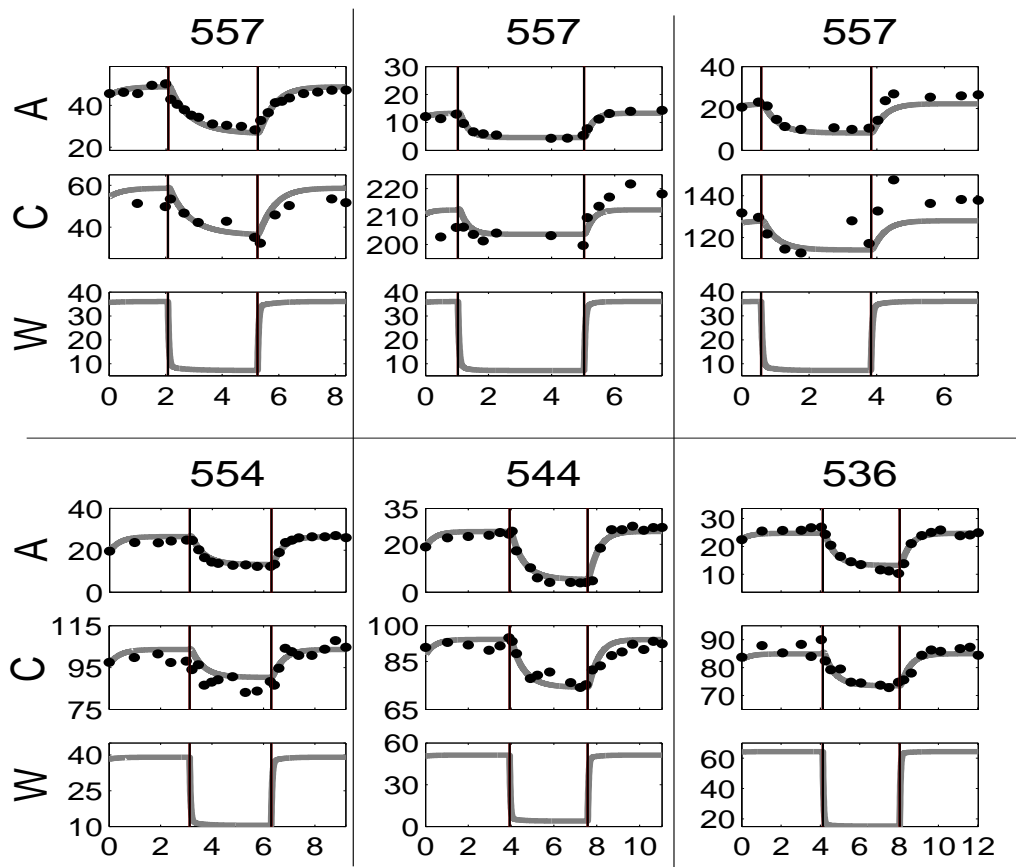
Fig. 7. Nylon components A, C and W along with the solution to the differential equations using initial values estimated by the smooth for each of six experiments. The times of step change in input pressures are marked by thin vertical lines. Horizontal axes indicate time in hours, and vertical axes are concentrations in moles. The labels above each experiment indicate the constant temperature in degrees Kelvin.
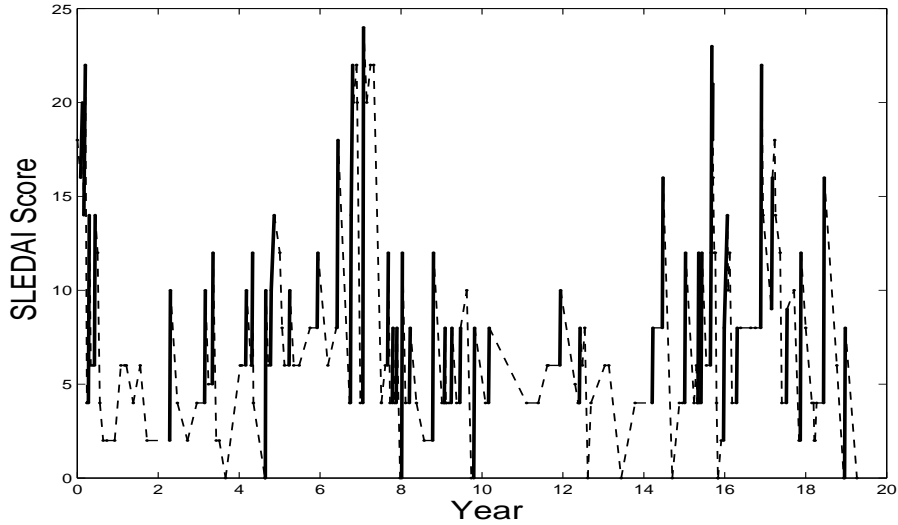
Fig. 8. Symptom level $s(t)$ for a patient suffering from lupus as assessed by the SLEDAI scale. Changes in SLEDAI score corresponding to a flare are shown as heavy solid lines, and other the remaining changes are shown as dashed lines.

Various scales have been developed to measure the severity of symptoms, and Figure 8 shows the course of one of the more popular measures, the SLEDAI scale, for a patient that experienced 48 flares over about 19 years before expiring. A definition of a flare event is commonly agreed to be a change in a scale value of at least 3 with a terminal value of at least 8, and the figure shows flare events as heavy solid lines.

Because of the rapid onset of symptoms, and because the resulting treatment program usually involves a SLEDAI assessment and a substantial increase in Prednisone dose, we can pin down the time of a flare with some confidence. Thus, the set of flare times combined with the accompanying SLEDAI score constitute a marked point process. Our goal here is to illustrate a simple model for flare dynamics, or the time course of symptoms over the onset period and the period of recovery. We hope that this model will also show how these short-term flare dynamics interact with longer term trends in symptom severity.

We postulated that the immune system goes on the attack for a fixed period of δ years, after which it returns to normal function due to treatment or normal recovery. For purposes of this illustration, we took $\delta = 0.02$ years, or about two weeks, and represented the time course of attacks as a box function $u(t)$ that is 0 during normal functioning and 1 during a flare.

This first order linear differential equation was proposed for symptom severity $s(t)$ at time t:

$$\dot{s}(t) = -\beta(t)s(t) + \alpha(t)u(t), \tag{29}$$

and has the solution

$$s(t) = Cs_0(t) + s_0(t) \int_0^t \alpha(z)u(z)/s_0(z)\, dz$$

where
$$s_0(t) = \exp[-\int_0^t \beta(z)\,dz].$$

Function $\alpha(t)$ tracks the long-term trend in the severity of the disease over the 19 years, and we represented this as a linear combination of 8 cubic B-spline basis functions defined by equally spaced knots, with about three years between knots. We expected that a flare plays itself out over a much shorter time interval, so that $\alpha(t)$ cannot capture any aspect of flare dynamics.

The flare dynamics depend directly on weight function $\beta(t)$. At the point where an attack begins, a flare increases in intensity with a slope that is proportional to β, and rises to a new level in roughly $4/\beta(t)$ time units if $\beta(t)$ is approximately constant. Likewise, when an attack ceases, $s(t)$ decays exponentially to zero with rate $\beta(t)$.

It seemed reasonable to propose that $\beta(t)$ is affected by an attack as well as $s(t)$. This is because $\beta(t)$ reflects to some extent the health of the individual in the sense that responding to an attack in various ways requires the body's resources, and these are normally at their optimum level just before an attack. The response drains these resources, and thus the attack is likely to reduce $\beta(t)$. Consequently, we proposed a second equation to model this mechanism:

$$\dot{\beta}(t) = -\gamma\beta(t) + \theta[1 - u(t)]. \tag{30}$$

This model suggests that an attack results in an exponential decay in β with rate γ, and that the cessation of the attack results in $\beta(t)$ returning to its normal level in about $4/\gamma$ time units. This normal level is defined by the gain $K = \theta/\gamma$. However, if γ is large, the model behaves like

$$\dot{\beta}(t) = \theta[1 - u(t)], \tag{31}$$

which is to say that $\beta(t)$ increases and decreases linearly.

The top panel in Figure 9 shows how $\beta(t)$ responds to an attack indicated by the box function $u(t)$ when $\gamma = \theta = 4$, corresponding to a time to reach a new level of about 1 time unit. The initial value $\beta(0) = 0$ in this plot. The bottom panel shows that the increase in symptoms is nearly linear during the period of attack, but that when the attack ceases, symptom level declines exponentially and takes around 3 time units to return to zero.

When we estimated this model with smoothing parameter value $\lambda = 1$, we obtained the results shown in Figure 10. We found that parameter γ was indeed so high that the fitted symptom rise was effectively linear, so we deleted γ and used the simpler equation (31). This left only the constant θ to estimate for $\beta(t)$, which now controls the rate of decrease of symptoms after an attack ceases. This was estimated to be 1.54, corresponding to a recovery period of about $4/1.54 = 2.6$ years. Figure 10 shows the variation in $\alpha(t)$ as a dashed line, indicating the long-term change in the intensity of the symptoms, which are especially severe around year 6, 11, and in the patient's last three years.

The fitted function $s(t)$ is shown as a solid line, and was defined by positioning three knots at each of the flare onset and offset times in order to accommodate the sudden break in $\dot{s}(t)$, and a single knot midway between two flare times. Order 4 B-splines were used, and this corresponded to 290 knot values and 292 basis functions in the expansion $\hat{s}(t) = \mathbf{c}'\boldsymbol{\phi}(t)$. We see that the fitted function seems to do a reasonable job of tracking the SLEDAI scores, both in the period during and following an attack and also in terms of its long-term trend.

The model also defines the differential equation (29), and the solution to this equation is shown as a dashed line. The discrepancy between the fit defined by the equation and the
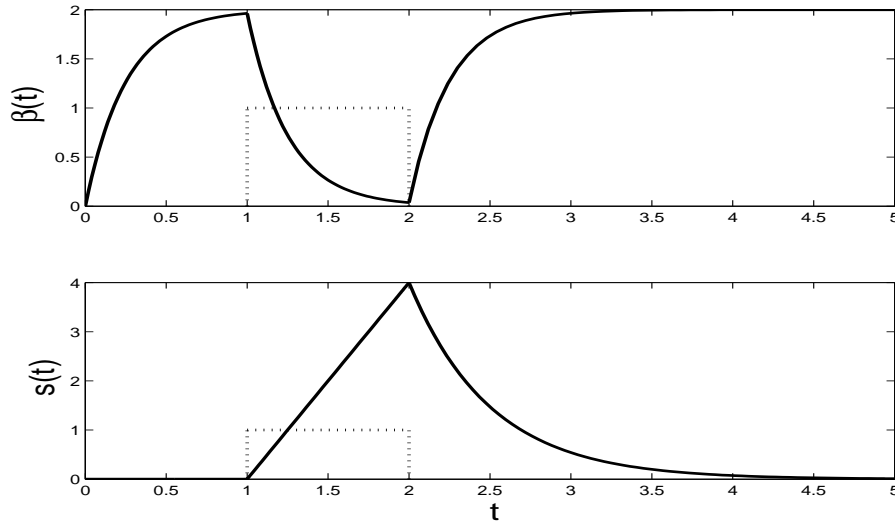
Fig. 9. The top panel shows the effect of a lupus attack on the weight function $\beta(t)$ in differential equation (29). The bottom panel shows the time course of the symptom severity function $s(t)$.
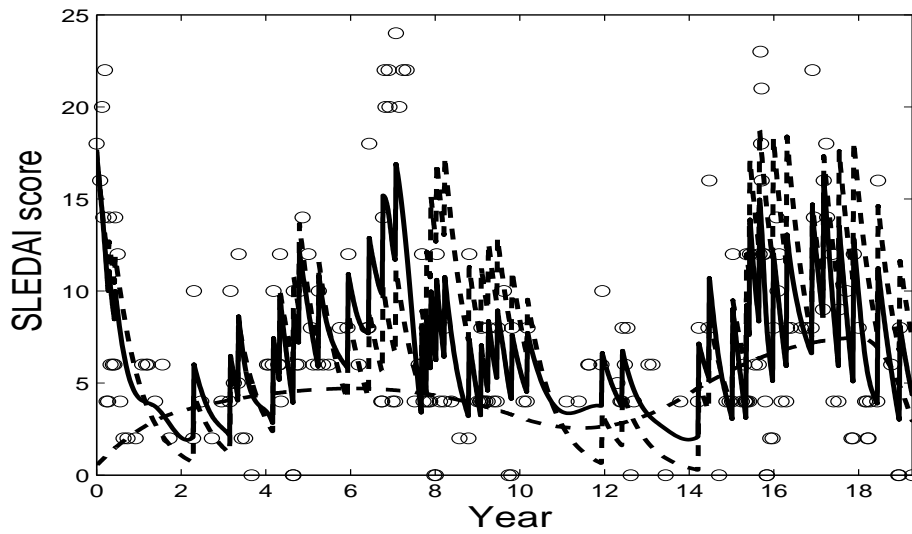

Fig. 10. The circles indicate SLEDAI scores, the jagged solid line is the smoothing functions $s(t)$, the dashed jagged line is the solution to the differential equation and the smooth dashed line is the smooth trend $\alpha(t)$.

smoothing function $s(t)$ is important in years 8 to 11, where the equation solution overestimates symptom level. In this region, new flares come too fast for recovery, and thus build on each other. Nevertheless, the fit to the 208 SLEDAI scores achieved by an investment of 9 structural parameters seems impressive for both the smoothing function $s(t)$ and equation solution, taking into consideration that the SLEDAI score is a rather imprecise measure. Moreover, the model goes a long way to modeling the within-flare dynamics, the general trend in the data, and the interaction between flare dynamics and trend.

5. Generalizations and further problems

5.1. More general equations

We have discussed the methods presented here with respect to systems of ODEs. However, these methods can be applied to the following situations in a direct manner:

- Differential-algebraic equations (DAEs), in which some components of $\mathbf{x}$ are specified directly rather than on the derivative scale:

$$x_i(t) = f_i(\mathbf{x}, \mathbf{u}, t|\boldsymbol{\theta}). \tag{32}$$

  Such systems are common in chemical engineering; see (Biegler et al. (1986)) for a classical example.

- Lagged equations:
$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t - \boldsymbol{\delta}_1), \mathbf{u}(t - \boldsymbol{\delta}_2), t|\boldsymbol{\theta}),$$

  where $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ are vectors of time lags for state and forcing functions, respectively.

- Partial differential equations (PDEs)in which a system $\mathbf{x}(s, t)$ is described over spatial variables s as well as time t:

$$\frac{\partial \mathbf{x}}{\partial t} = \mathbf{f}\left(\mathbf{x}, \frac{\partial \mathbf{x}}{\partial s}, \mathbf{u}, t|\boldsymbol{\theta}\right).$$

Both lagged and partial differential equations require the specification of an infinite dimensional boundary condition, rather than a finite set of initial conditions.

5.2. Stochastic differential equations

Criterion (14) may be interpreted as the log likelihood for an observation from the stochastic differential equation:

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}, \mathbf{u}, t|\boldsymbol{\theta}) + \boldsymbol{\lambda}\frac{d\mathbf{W}(t)}{dt}$$

where $\mathbf{W}(t)$ is a d-dimensional Brownian motion. Thus for a fixed $\boldsymbol{\lambda}$, interpreted as the ratio of the Brownian motion variance to that of the observational error, the procedure may be thought of as profiling an estimate of the realized Brownian motion. This approach has been used for the problem of data assimilation in Apte et al. (2007), where they use criteria closely related to our own (14). This notion is appealing and suggests the use of alternative smoothing penalties based on the likelihood of other stochastic processes. The flares in the Lupus data, for example, could be considered to be triggered by events in a Poisson process, and we expect this to be a fruitful area of future research. However, this interpretation

relies on the representation of $d\mathbf{W}(t)/dt$ in terms of the discrepancy $\dot{\mathbf{x}}(t) - \mathbf{f}(\mathbf{x}, \mathbf{u}, t|\boldsymbol{\theta})$ where $\mathbf{x}$ is given by a basis expansion (7). For nonlinear $\mathbf{f}$ the approximation properties of this discrepancy are not clear. Moreover, it is frequently the case that lack of fit in nonlinear dynamics is due more to mis-specification of the system under consideration than to stochastic inputs, and we are correspondingly wary of this interpretation.

5.3. Further statistical problems

Diagnostic tools are needed for differential equation models. Particularly in biological applications, these models often provide the right *qualitative* behavior and may take values orders of magnitude different from the observed data. Diagnostic analyses can estimate additional components of $\mathbf{u}$ that will provide good fits. These may be correlated with observed values of the system, or external factors, to suggest new model formulae.

Experimental design is a relatively unexplored area of research for nonlinear dynamical systems. Engineers plan experiments in which inputs are varied under various regimes; including step, ramp, periodic and other perturbations. These inputs are then continuous functions which join sampling rates for each component and replicated experiments as design variables. See Bauer et al. (2000) for an approach to these problems.

Finally, there are a large class of theoretical and inferential problems in fitting nonlinear differential equations to data, including inference near bifurcation boundaries, about system stability and on the relationship between statistical information and chaotic behavior.

6. Conclusions

Differential equations have a long and illustrious history in mathematical modeling. However, there has been little development of statistical theory for estimating such models or assessing their agreement with observational data. Our approach, a variety of collocation method, combines the concepts of *smoothing* and *estimation*, providing a continuum of trade-offs between fitting the data well and fidelity to the hypothesized differential equations. This has been done by defining a fit through a penalized spline criterion for each value of $\boldsymbol{\theta}$ and then estimating $\boldsymbol{\theta}$ through a profiling scheme in which the fit is regarded as a nuisance parameter.

We have found that this procedure has a number of important advantages relative to older methods such as nonlinear least squares. Parameter estimates can be obtained from data on partially measured systems, a common situation where certain variables are expensive to measure or are intrinsically latent. Comparisons with other approaches suggest that the bias and sampling variance of these estimates is at least as good as for other approaches, and rather better relative to methods such as NLS. The sampling variation in the estimates is easily estimable, and our simulation experiments and experience indicate that there is good agreement between these estimation precision indicators and the actual estimation accuracies. Our approach also gains from not requiring a formulation of the dynamic model as an initial value problem in situations where initial values are not available or not required.

On the computational side, the algorithm is as fast or faster than NLS and other approaches. Unlike Bayesian MCMC, the generalized profiling approach is relatively straightforward to deploy to a wide range of applications, and software in Matlab described below merely requires that the user to code the various partial derivatives that are involved, and which are detailed in the Appendix. Finally, the method is also robust in the sense of converging over a wide range of starting parameter values. The possibility of beginning with

smaller values of $\boldsymbol{\lambda}$ so as to work with a smooth criterion, and then stepping these values up toward those defining near approximations to the ODE further adds to the method's robustness.

Finally the fitting of a compromise between an actual ODE solution and a simple smooth of the data adds a great deal of flexibility that should prove useful to users wishing to explore variation in the data not representable in the ODE model. By comparing fits with smaller values of λ with fits that are near or exact ODE solutions, the approach offers a diagnostic capability that can guide further extensions and elaborations of the model.

6.1. Software

All the results in this paper have been generated in the MATLAB computing language, making use of functional data analysis software intended to compliment Ramsay and Silverman (2005). A set of software routines that may be applied to any differential equation is available from the URL: `http://www.functionaldata.org`.

References

Apte, A., M. Hairer, A. M. Stuart, and J. Voss (2007). Sampling the posterior: An approach to non-gaussian data assymilation. *Physica D, to appear.*

Arora, N. and L. T. Biegler (2004). A trust region SQP algorithm for equality constrained parameter estimation with simple parametric bounds. *Computational Optimization and Applications 28*, 51–86.

Bates, D. M. and D. B. Watts (1988). *Nonlinear Regression Analysis and Its Applications.* New York: Wiley.

Bauer, I., H. G. Bock, S. Körkel, and J. P. Schlöder (2000). Numerical methods for optimum experimental design in DAE systems. *Journal of Computational and Applied Mathematics 120*, 1–25.

Biegler, L., J. J. Damiano, and G. E. Blau (1986). Nonlinear parameter estimation: a case study comparison. *AIChE Journal 32*(1), 29–45.

Biegler, L. and I. Grossman (2004). Retrospective on optimization. *Computers and Chemical Engineering 28*, 1169–1192.

Bock, H. G. (1983). Recent advances in parameter identification techniques for ODE. In P. Deuflhard and E. Harrier (Eds.), *Numerical Treatment of Inverse Problems in Differential and Integral Equations*, pp. 95–121. Basel: Birkhäuser.

Campbell, D. (2007). *Bayesian Collocation Tempering and Generalized Profiling for Estimation of Parameters From Differential Equation Models.* Ph. D. thesis, McGill University.

Campbell, D., G. Hooker, J. O. Ramsay, K. McAuley, and J. McLellan (2007). Generalized profiling parameter estimation in differential equation models with constrained variables. unpublished manuscript.

Cao, J. and J. O. Ramsay (2006). Parameter cascades and profiling in functional data analysis. *Computational Statistics*, In press.

Cox, D. R. and D. V. Hinkley (1974). *Theoretical Statistics*. London: Chapman & Hall.

Denis-Vidal, L., G. Joly-Blanchard, and C. Noiret (2003). System identifiability (symbolic computation) and parameter estimation (numerical computation). *Numerical Algorithms 34*, 283–292.

Deuflhard, P. and F. Bornemann (2000). *Scientific Compuitng with Ordinary Differential Equations*. New York: Springer-Verlag.

Esposito, W. R. and C. Floudas (2000). Deterministic global optimization in nonlinear optimal control problems. *Journal of Global Optimization 17*, 97–126.

FitzHugh, R. (1961). Impulses and physiological states in models of nerve membrane. *Biophysical Journal 1*, 445–466.

Friedman, J. and B. W. Silverman (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics 3*, 3–21.

Gelman, A., F. Y. Bois, and J. Jiang (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *Journal of the American Statistical Association 91*(436), 1400–1412.

Hodgkin, A. L. and A. F. Huxley (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol. 133*, 444–479.

Hooker, G. (2007). Theorems and calculations for smoothing-based profiled estimation of differential equations. Technical Report BU-1671-M, Dept. Bio. Stat. and Comp. Bio., Cornell University.

Jaeger, J., M. Blagov, D. Kosman, K. Kolsov, Manu, E. Myasnikova, S. Surkova, C. Vanario-Alonso, M. Samsonova, D. Sharp, and J. Reinitz (2004). Dynamical analysis of regulatory interactions in the gap gene system of *drosophila melanogaster*. *Genetics* (167), 1721–1737.

Keilegom, I. V. and R. J. Carroll (2006). Backfitting versus profiling in general criterion functions. Submitted to Statistica Sinica.

Koenker, R. and I. Mizera (2002). Elastic and plastic splines: Some experimental comparisons. In Y. Dodge (Ed.), *Statistical Data Analysis based on the L1-norm and Related Methods*, pp. 405–414. Basel: Birkhäuser.

Marlin, T. E. (2000). *Process Control*. New York: McGraw-Hill.

Nagumo, J. S., S. Arimoto, and S. Yoshizawa (1962). An active pulse transmission line simulating a nerve axon. *Proceedings of the IRE 50*, 2061–2070.

Poyton, A. A., M. S. Varziri, K. B. McAuley, P. J. McLellan, and J. O. Ramsay (2006). Parameter estimation in continuous dynamic models using principal differential analysis. *Computational Chemical Engineering 30*, 698–708.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis*. New York: Springer.

Seber, G. A. F. and C. J. Wild (1989). *Nonlinear Regression*. New York: Wiley.

Tjoa, I.-B. and L. Biegler (1991). Simultaneous solution and optimization strategies for parameter estimation of differential-algebraic equation systems. *Industrial Engineering and Chemical Research 30*, 376–385.

Varah, J. M. (1982). A spline least squares method for numerical parameter estimation in differential equations. *SIAM Journal on Scientific Computing 3*, 28–46.

Wahba, G. (1990). *Spline Models for Observational Data.* Philadelphia: SIAM CBMS-NSF Regional Conference Series in Applied Mathematics.

Wilson, H. R. (1999). *Spikes, Decisions and Actions: The Dynamical Foundations of Neuroscience.* Oxford: Oxford University Press.

Zheng, W., K. McAuley, K. Marchildon, and K. Z. Yao (2005). Effects of end-group balance on melt-phase nylon 612 polycondensation: Experimental study and mathematical model. *Ind. Eng. Chem. Res. 44*, 2675–2686.

Appendices

A. Matrix calculations for profiling

The calculations used throughout this paper have been based on matrices defined in terms of derivatives of F and H with respect to $\boldsymbol{\theta}$ and $\mathbf{c}$. In many cases, these matrices are non-trivial to calculate and expressions for their entries are derived here. For these calculations, we have assumed that the outer criterion, F is a straight-forward weighted sum of squared errors and only depends on $\boldsymbol{\theta}$ through $\mathbf{x}$.

A.1. Inner optimization

Using a Gauss-Newton method, we require the derivative of the fit at each observation point:

$$\frac{dx_i(t)}{d\mathbf{c}_i} = \boldsymbol{\phi}_i(t)$$

where matrix $\boldsymbol{\phi}_i$ is the vector corresponding to the evaluation of all the basis functions used to represent x_i evaluated at t. This gradient of x_i with respect to $\mathbf{c}_j$ is zero.

A numerical quadrature rule allows the set of errors to be augmented with the evaluation of the penalty at the quadrature points and weighted by the quadrature rule:

$$(\lambda_i v_q)^{1/2} \left(\dot{x}_i(t_q) - f_i(\mathbf{x}(t_q), \mathbf{u}(t_q), t_q | \boldsymbol{\theta}) \right).$$

Each of these then has derivative with respect to $\mathbf{c}_j$:

$$(\lambda_i v_q)^{1/2} \left(\dot{x}_i(t_q) - f_i(\mathbf{x}(t_q), \mathbf{u}(t_q), t_q | \boldsymbol{\theta}) \right) I(i = j) \dot{\boldsymbol{\phi}}_i(t_q)$$
$$- \left(\sum_{k=1}^{n} (\lambda_i v_q)^{1/2} \frac{df_k}{dx_j} \left(Dx_i(t_q) - f_i(\mathbf{x}(t_q), \mathbf{u}(t_q), t_q | \boldsymbol{\theta}) \right) \right) \boldsymbol{\phi}_j(t_q)$$

and the augmented errors and gradients can be used in a Gauss-Newton scheme. $I(\cdot)$ is used as the indicator function of its argument.

A.2. Estimating structural parameters

As in the inner optimization, in employing a Gauss-Newton scheme, we merely need to write a gradient for the point-wise fit with respect to the parameters:

$$\frac{d\mathbf{x}(t)}{d\boldsymbol{\theta}} = \frac{d\mathbf{x}(t)}{d\mathbf{c}} \frac{d\mathbf{c}}{d\boldsymbol{\theta}}$$

where $d\mathbf{x}(t_i)/d\mathbf{c}$ has already be calculated and

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

by the implicit function theorem.

The Hessian matrix $d^2H/d\mathbf{c}^2$ may be expressed as a block form, the (i,j)th block corresponding to the cross-derivatives of the coefficients in the ith and jth components of $\mathbf{x}$. This block's (p,q)th entry is given by:

$$\left(\sum_{k=1}^{n_i}\phi_{ip}(t)\phi_{jq}(t)+\lambda\int\phi_{ip}(t)\phi_{jq}(t)dt\right)I(i=j)$$

$$-\lambda_i\int\dot{\phi}_{ip}(t)\frac{df_i}{dx_j}\phi_{jq}(t)dt-\lambda_j\int\phi_{ip}(t)\frac{df_i}{dx_j}\dot{\phi}_{jq}(t)dt$$

$$+\int\phi_{ip}(t)\left[\sum_{k=1}^{n}\lambda_k\left(\frac{d^2f_k}{dx_idx_j}(f_k-\dot{x}_k(t))+\frac{df_k}{dx_i}\frac{df_k}{dx_j}\right)\right]\phi_{jq}(t)dt$$

with the integrals evaluated by numeric integration. The arguments to $f_k(\mathbf{x},\mathbf{u},t|\boldsymbol{\theta})$ have been dropped in the interests of notational legibility.

We can similarly express the cross-derivatives $d^2H/d\mathbf{c}d\boldsymbol{\theta}$ as a block vector, the ith block corresponding to the coefficients in the basis expansion for the ith component of $\mathbf{x}$. The pth entry of this block can now be expressed as:

$$\lambda_i\int\frac{df_i}{d\boldsymbol{\theta}}\phi_{ip}(t)dt-\int\left(\sum_{k=1}^{n}\lambda_k\left[\frac{d^2f_k}{dx_id\boldsymbol{\theta}}(f_k-\dot{x}_k(t))+\frac{df_k}{dx_i}\frac{df_k}{d\boldsymbol{\theta}}\right]\right)\phi_{ip}(t)dt.$$

A.3. Estimating the variance of $\hat{\boldsymbol{\theta}}$

The variance of the parameter estimates is calculated using

$$\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}}=-\left[\frac{d^2H}{d\boldsymbol{\theta}^2}\right]^{-1}\frac{d^2H}{d\boldsymbol{\theta}d\mathbf{y}},$$

where

$$\frac{d^2H}{d\boldsymbol{\theta}^2}=\frac{\partial^2H}{\partial\boldsymbol{\theta}^2}+\left(\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}}\right)'\frac{\partial^2H}{\partial\hat{\mathbf{c}}\partial\boldsymbol{\theta}}+\frac{\partial^2H}{\partial\boldsymbol{\theta}\partial\hat{\mathbf{c}}}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}}+\left(\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}}\right)'\frac{\partial^2H}{\partial\hat{\mathbf{c}}^2}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}}+\frac{\partial H}{\partial\hat{\mathbf{c}}}\frac{\partial^2\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}^2},\qquad(33)$$

and

$$\frac{d^2H}{d\boldsymbol{\theta}d\mathbf{y}}=\frac{\partial^2H}{\partial\boldsymbol{\theta}\partial\mathbf{y}}+\frac{\partial^2H}{\partial\hat{\mathbf{c}}\partial\mathbf{y}}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}}+\frac{\partial^2H}{\partial\boldsymbol{\theta}\partial\hat{\mathbf{c}}}\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}+\frac{\partial^2H}{\partial\hat{\mathbf{c}}^2}\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}}+\frac{\partial H}{\partial\hat{\mathbf{c}}}\frac{\partial^2\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}\partial\mathbf{y}}.\qquad(34)$$

The formulas (33) and (34) for $d^2H/d\boldsymbol{\theta}^2$ and $d^2H/d\boldsymbol{\theta}d\mathbf{y}$ involve the terms $\partial\hat{\mathbf{c}}/\partial\mathbf{y}$, $\partial^2\hat{\mathbf{c}}/\partial\boldsymbol{\theta}^2$ and $\partial^2\hat{\mathbf{c}}/\partial\boldsymbol{\theta}\partial\mathbf{y}$. In the following, we derive their analytical formulas by the Implicit Function Theorem. We introduce the following convention, which is called *Einstein Summation Notation*. If a Latin index is repeated in a term, then it is understood as a summation with respect to that index. For instance, instead of the expression $\sum_i a_ix_i$, we merely write a_ix_i.

- $\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}$

  Similar as the deduction for $d\hat{\mathbf{c}}/d\boldsymbol{\theta}$, we obtain the formula for $\partial\hat{\mathbf{c}}/\partial\mathbf{y}$ by applying the Implicit Function Theorem:

$$\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}=-\left[\frac{\partial^2J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}^2}\bigg|_{\hat{\mathbf{c}}}\right]^{-1}\left[\frac{\partial^2J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}\partial\mathbf{y}}\bigg|_{\hat{\mathbf{c}}}\right].\qquad(35)$$

- $\dfrac{\partial \mathbf{c}^2}{\partial \boldsymbol{\theta} \partial \mathbf{y}}$

By taking the second derivative on both sides of the identity $\partial J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})/\partial \mathbf{c}|_{\hat{\mathbf{c}}} = 0$ with respect to $\boldsymbol{\theta}$ and y_k, we derive:

$$\frac{d^2}{d\boldsymbol{\theta} dy_k}\left(\frac{\partial J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}}\bigg|_{\hat{\mathbf{c}}}\right) \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}\partial\boldsymbol{\theta}\partial y_k}\bigg|_{\hat{\mathbf{c}}} + \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}\partial\boldsymbol{\theta}\partial c_i}\bigg|_{\hat{\mathbf{c}}} \frac{\partial \hat{c}_i}{\partial y_k}$$
$$+ \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2\partial y_k}\bigg|_{\hat{\mathbf{c}}} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2\partial c_i}\bigg|_{\hat{\mathbf{c}}} \frac{\partial \hat{c}_i}{\partial y_k}\frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \frac{\partial^2 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2}\bigg|_{\hat{\mathbf{c}}} \frac{\partial^2 \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}\partial y_k}. \quad (36)$$

Solving for $\dfrac{\partial^2 \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}\partial y_k}$, we obtain the second derivative of $\hat{\mathbf{c}}$ with respect to $\boldsymbol{\theta}$ and y_k:

$$\begin{aligned}
\frac{\partial^2 \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}\partial y_k} =& -\left[\frac{\partial^2 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2}\bigg|_{\hat{\mathbf{c}}}\right]^{-1}\left[\frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}\partial\boldsymbol{\theta}\partial y_k}\bigg|_{\hat{\mathbf{c}}} + \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}\partial\boldsymbol{\theta}\partial c_i}\bigg|_{\hat{\mathbf{c}}} \frac{\partial \hat{c}_i}{\partial y_k}\right. \\
&+ \left. \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2\partial y_k}\bigg|_{\hat{\mathbf{c}}} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2\partial c_i}\bigg|_{\hat{\mathbf{c}}} \frac{\partial \hat{c}_i}{\partial y_k}\frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}}\right]. \quad (37)
\end{aligned}$$

- $\dfrac{\partial^2 \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}^2}$

Similar to the deduction of $\partial^2 \hat{\mathbf{c}}/\partial \boldsymbol{\theta}\partial y_k$, the second partial derivative of $\mathbf{c}$ with respect to $\boldsymbol{\theta}$ and θ_j is:

$$\begin{aligned}
\frac{\partial^2 \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}\partial \theta_j} =& -\left[\frac{\partial^2 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2}\bigg|_{\hat{\mathbf{c}}}\right]^{-1}\left[\frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}\partial\boldsymbol{\theta}\partial \theta_j}\bigg|_{\hat{\mathbf{c}}} + \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}\partial\boldsymbol{\theta}\partial c_i}\bigg|_{\hat{\mathbf{c}}} \frac{\partial \hat{c}_i}{\partial \theta_j}\right. \\
&+ \left. \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2\partial \theta_j}\bigg|_{\hat{\mathbf{c}}} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2\partial c_i}\bigg|_{\hat{\mathbf{c}}} \frac{\partial \hat{c}_i}{\partial \theta_j}\frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}}\right]. \quad (38)
\end{aligned}$$

When estimating ODEs, we define $J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})$ as (14) and $H(\boldsymbol{\theta},\hat{\mathbf{c}}(\boldsymbol{\theta})|\mathbf{y})$ as (8), and further write the above formulas in terms of the basis functions in $\boldsymbol{\phi}$ and the functions $\mathbf{f}$ on the right side of the differential equation. For instance, $d^2H/d\mathbf{c}^2$ is a block-diagonal matrix with the ith block being $w_i\boldsymbol{\phi}_i(\mathbf{t}_i)^T\boldsymbol{\phi}_i(\mathbf{t}_i)$ and $dF/d\mathbf{c}$ is a block vector containing blocks $-w_i\boldsymbol{\phi}_i(\mathbf{t}_i)^T(\mathbf{y}_i - x_i(\mathbf{t}_i))$.

The three-dimensional array $\partial^3 J/\partial\mathbf{c}\partial c_p\partial c_q$ can be written in the same block vector form as $\partial^2 J/\partial\mathbf{c}\partial\boldsymbol{\theta}$ with the uth entry of the kth block given by

$$\int\left(\sum_{l=1}^{n}\lambda_l\left[\frac{d^2 f_l}{dx_i dx_j}\frac{df_l}{dx_k} + \frac{d^2 f_l}{dx_i dx_k}\frac{df_l}{dx_j} + \frac{d^2 f_l}{dx_j dx_k}\frac{df_l}{dx_i}\right]\right)\phi_{ip}(t)\phi_{jq}(t)\phi_{ku}(t)dt$$
$$+ \int\sum_{l=1}^{n}\lambda_l\left(\frac{d^3 f_k}{dx_i dx_j dx_k}(f_l - \dot{x}_l(t))\right)\phi_{ip}(t)\phi_{jq}(t)\phi_{ku}(t)dt$$
$$- \lambda_i\int\frac{d^2 f_i}{dx_j dx_k}\dot{\phi}_{ip}(t)\phi_{jq}(t)\phi_{ku}(t)dt - \lambda_j\int\frac{d^2 f_j}{dx_i dx_k}\phi_{ip}(t)\dot{\phi}_{jq}(t)\phi_{ku}(t)dt$$
$$- \lambda_k\int\frac{d^2 f_k}{dx_i dx_j}\phi_{ip}(t)\phi_{jq}(t)\dot{\phi}_{ku}(t)dt$$

assuming c_p is a coefficient in the basis representation of x_i and c_q a corresponds to x_j. The array $\partial^3 J/\partial\mathbf{c}\partial\theta_i\partial\theta_j$ is also expressed in the same block form with entry p in the kth block

being:

$$\int \left(\sum_{l=1}^{n} \lambda_l \left[\frac{d^2 f_l}{d\theta_i d\theta_j} \frac{df_l}{dx_k} + \frac{d^2 f_l}{d\theta_i dx_k} \frac{df_l}{d\theta_j} + \frac{d^2 f_l}{d\theta_j dx_k} \frac{df_l}{d\theta_i} \right] \right) \phi_{kp}(t) dt$$

$$+ \int \sum_{l=1}^{n} \lambda_l \left(\frac{d^3 f_k}{dx_k d\theta_i d\theta_j} (f_l - \dot{x}_l(t)) \right) \phi_{kp}(t) dt - \lambda_k \int \frac{d^2 f_k}{d\theta_i d\theta_k} \phi_{kp}(t) dt.$$

The term $\partial^3 J / \partial \mathbf{c} \partial c_p \partial \theta_i$ is in the same block from, with the qth entry of the jth block being:

$$\int \left(\sum_{l=1}^{n} \lambda_l \left[\frac{d^2 f_l}{d\theta_i dx_j} \frac{df_l}{dx_k} + \frac{d^2 f_l}{d\theta_i dx_k} \frac{df_l}{dx_j} + \frac{d^2 f_l}{dx_j dx_k} \frac{df_l}{d\theta_i} \right] \right) \phi_{kp}(t) \phi_{jq}(t) dt$$

$$+ \int \sum_{l=1}^{n} \lambda_l \left(\frac{d^3 f_k}{dx_j dx_k d\theta_i} (f_l - \dot{x}_l(t)) \right) \phi_{kp}(t) \phi_{jq}(t) dt$$

$$- \lambda_j \int \frac{d^2 f_j}{d\theta_i dx_k} \dot{\phi}_{jq}(t) \phi_{kp}(t) dt - \lambda_k \int \frac{d^2 f_k}{d\theta_i dx_j} \phi_{jq}(t) \dot{\phi}_{kp}(t) dt$$

where c_p corresponds to the basis representation of x_k.

Similar calculations give matrix $d^2 H / d\boldsymbol{\theta} d\mathbf{y}$ explicitly as:

$$\frac{d\hat{\mathbf{c}}}{d\boldsymbol{\theta}}^T \left[\frac{\partial^2 H}{\partial \hat{\mathbf{c}} \partial \mathbf{y}} + \frac{\partial^2 H}{\partial \mathbf{c}^2} \frac{d\hat{\mathbf{c}}}{d\mathbf{y}} \right]$$

$$- \frac{\partial H}{\partial \mathbf{c}} \left[\frac{\partial^2 H}{\partial \mathbf{c}^2} \right]^{-1} \left\{ \sum_{p,q=1}^{N} \frac{d\hat{c}_p}{d\boldsymbol{\theta}}^T \frac{\partial^3 J}{\partial \mathbf{c} \partial c_p \partial c_q} \frac{d\hat{c}_q}{d\mathbf{y}} + \sum_{p=1}^{N} \frac{\partial^3 J}{\partial \mathbf{c} \partial c_p \partial \boldsymbol{\theta}} \frac{d\hat{c}_p}{d\mathbf{y}} \right\}$$

with $d\hat{\mathbf{c}}/d\mathbf{y}$ given by

$$- \left[\frac{\partial^2 J}{\partial \mathbf{c}^2} \right]^{-1} \frac{\partial^2 J}{\partial \mathbf{c} \partial \mathbf{y}}$$

and $\partial^2 J / \partial \mathbf{c} d\mathbf{y}$ being block diagonal with the ith block containing $w_i \phi_i(\mathbf{t}_i)$.