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 non-linear differential equations. These models represent change in a system by linking the behavior of a derivative 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 statistical techniques such as inference and interval estimation. This paper describes a new method that uses noisy data 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 interval estimates and show that these have good coverage properties on data simulated from chemical engineering and neurobiology. The method is demonstrated using real-world data from chemistry and from the progress of the auto-immune disease lupus.

Keywords: Differential equations, profiled estimation, estimating equations, Gauss-Newton methods, functional data analysis

1. The challenges in dynamic systems estimation

We have in mind a process that transforms a set of m input functions, with values as functions of time $t \in [0, T]$ indicated by vector $\mathbf{u}(t)$, into a set of d output functions with values $\mathbf{x}(t)$. Examples are a single neuron whose response is determined by excitation from a number of other neurons, and a chemical reactor that transforms a set of chemical species into a product within the context of additional inputs such as temperature and flow of a coolant and additional outputs such as the temperature of the product. The number of outputs may be impressive; $d \approx 50$ is not unusual in modeling polymer production, for example, and Deuflhard and Bornemann (2000), in their nice introduction to the world of dynamic systems models, cite chemical reaction kinetic models where d is in the thousands.

It is routine that only some of the outputs will be measured. For example, temperatures in a chemical process can usually be obtained online cheaply and accurately, but concentrations of chemical species can involve expensive assays that can take months to complete and have relatively high error levels as well. The abundance of a predacious species may be estimable, but the subpopulation of reproducing individuals may be impossible to count. On the other hand, we take the values $\mathbf{u}(t)$ to be available with negligible error at all times t.

Ordinary differential equations (ODE's) model output change directly by linking the derivatives of the output to $\mathbf{x}$ itself and, possibly, to inputs $\mathbf{u}$. That is, using $\dot{\mathbf{x}}(t)$ to denote the value of the first derivative of $\mathbf{x}$ at time t,

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}, \mathbf{u}, t | \boldsymbol{\theta}). \tag{1}$$

Solutions of the ODE given initial values $\mathbf{x}(0)$ exist and are unique over a neighborhood of $(0, \mathbf{x}(0))$ if f is continuously differentiable with respect to $\mathbf{x}$ or, more generally, Lipschitz continuous with respect to $\mathbf{x}$. Vector $\boldsymbol{\theta}$ contains any parameters defining $\mathbf{f}$ whose values are not known from experimental data, theoretical considerations or other sources of information. Although (1) appears to cover only first order systems, systems with the highest order derivative $D^n x$ on the left side are reducible to a first order form by defining n new variables, $x_1 = x, x_2 = \dot{x}_1$ and so on up to $x_{n-1} = D^{n-1}x$, and (1) can easily be extended to include more general differential equation systems. Dependencies of $\mathbf{f}$ on t other than through $\mathbf{x}$ and $\mathbf{u}$ arise when, for example, certain parameters defining the system are themselves time-varying.

Most ODE systems are not solvable analytically, so that conventional data-fitting methodology is not directly applicable. Exceptions are linear systems with constant coefficients, where the machinery of the Laplace transform and transform functions plays a role, are solvable, and a statistical treatment of these is available in Bates and Watts (1988) and Seber and Wild (1989). Discrete versions of such 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, where we consider systems of nonlinear ordinary differential equations or ODE's. In fact, it is the capacity of relatively simple nonlinear differential equations to define functional relations of great complexity that explains why they are so useful.

We also set aside stochastic differential equation systems involving inputs or perturbations of parameters that are the derivative of a Wiener process. The mathematical complexities of working with such systems has implied that, in practice, the range of ODE structures considered has remained extremely restricted, and we focus on modeling situations where much more complex ODE structures are required and where inputs can be considered as being at least piecewise smooth.

The insolvability of most ODE's has meant that statistical science has had little impact on the fitting of such models to data. Current methods for estimating ODE's from noisy data are often slow, uncertain to provide satisfactory results, and do not lend themselves well collateral 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. Finally, although one would hope that the total number of measured values, along with its distribution over measured values, would have a healthy ratio to the dimension of the parameter vector $\boldsymbol{\theta}$, such is often not the case. Measurements in biological, medical and physiology, for example, may require invasive or destructive procedures that can strictly control the number of measurements that can realistically be obtained. These problems can be often be offset, however, by a high level of measurement precision.

This paper describes a method that is based 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 functions in $\mathbf{x}$, and the 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 parameter on the estimation of structural parameters is controlled through a multi-criterion optimization process rather than the more usual marginalization procedure.

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

If initial or boundary values are considered to be available without error, then the large collection of numerical methods for estimating these solutions, treated in texts such as Deuflhard and Bornemann (2000), may be brought into play. On the other hand, if either there are no observations at 0 and T or the observations supplied are subject to measurement error, than these initial or boundary values, if required, can be considered parameters that must be included in an augmented parameter vector $\boldsymbol{\theta}^* = (x(0)', \boldsymbol{\theta}')'$. Our

approach may be considered as an extension of methods for these two situations where the data over-determine the system, are distributed anywhere in $[0, T]$, and are subject to observational error. We may call such a situation a *distributed data ODE* problem.

1.1. The data and error model contexts

We assume that a subset $\mathcal{I}$ of the d output variables are 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 $e_{ij} = y_{ij} - x_i(t_{ij})$. 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 common to assume independently distributed Gaussian errors with mean 0 and standard deviation σ_i, but in fact autocorrelation structure and nonstationary variance are often evident in the data, and when these features are also modeled, these parameters are also incorporated into $\boldsymbol{\sigma}_i$. 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.

1.2. Two test-bed problems

Two elementary problems will be used in the paper to illustrate aspects of the data fitting problem.

1.2.1. The FitzHugh-Nagumo neural spike potential 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:

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

The system describes the reciprocal dependencies of the voltage V across an axon membrane and a recovery variable R summarizing outward currents, as well as the impact of a time-varying external input u. Although not intended to provide a close fit to actual neural spike potential data, solutions to the FitzHugh-Nagumo ODE's 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 $\dot{R} = -(b/c)R$ linearly forced by 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, unforced solutions, where $u(t) = 0$, quickly converge from a range of starting values to periodic behavior that alternates between the smooth evolution and the sharp changes in direction shown in Figure 1.

A particular concern in ODE 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)
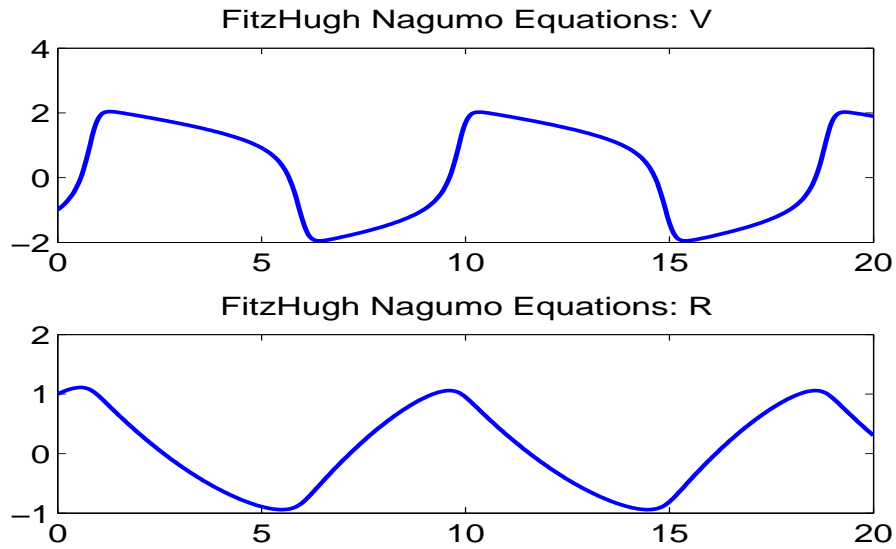
Fig. 1. The solid lines show the limiting behavior of voltage V and recovery R defined by the unforced FitzHugh-Nagumo equations (2) with parameter values $a = 0.2, b = 0.2$ and $c = 3.0$ and initial conditions $(V_0, R_0) = (-1, 1)$.
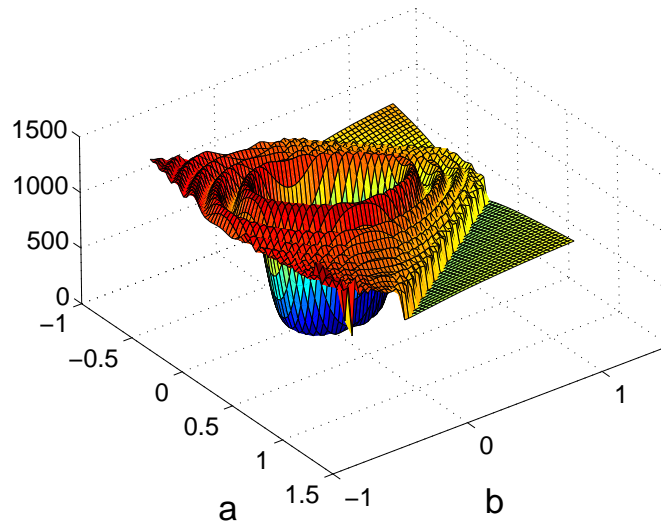

Fig. 2. The integrated squared difference between solutions of the FitzHugh-Nagumo equations for parameters (a, b) and $(0.2, 0.2)$ as a and b are varied about $(0.2, 0.2)$.

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 surface obtained by varying only the parameters a and b of the FitzHugh-Nagumo equations (2) in a fit to the errorless paths shown in Figure 1. 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. The tank reactor equations

The concept of a continuously stirred tank reactor, or a *CSTR*, in chemical engineering consists of a tank surrounded by cooling jacket and an impeller which stirs the contents. A fluid is pumped into the tank containing a reagent with concentration C_{in} at a flow rate F_{in} and temperature T_{in}. The reaction produces a product that leaves the tank with concentration C_{out} and temperature T_{out}. A coolant enters the cooling jacket with temperature T_{cool} and flow rate F_{cool}.

The differential equations used to model a CSTR, taken from Marlin (2000) and simplified by setting the volume of the tank to one, are

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

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_{cin}, respectively. These time-varying multipliers depend on four system parameters as follows:

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

where T_{ref} 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_{out} and T_{out} as determined by the five inputs $C_{in}, F_{in}, T_{in}, T_{cool}$ and F_{cool}. 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_{cool}.

The behaviors of output variables C_{out} and T_{out} under the experimental regime, 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. However, an increase in baseline coolant temperature by 30 degrees Kelvin generates oscillations that come close to instability when the coolant temperature decreases, and this
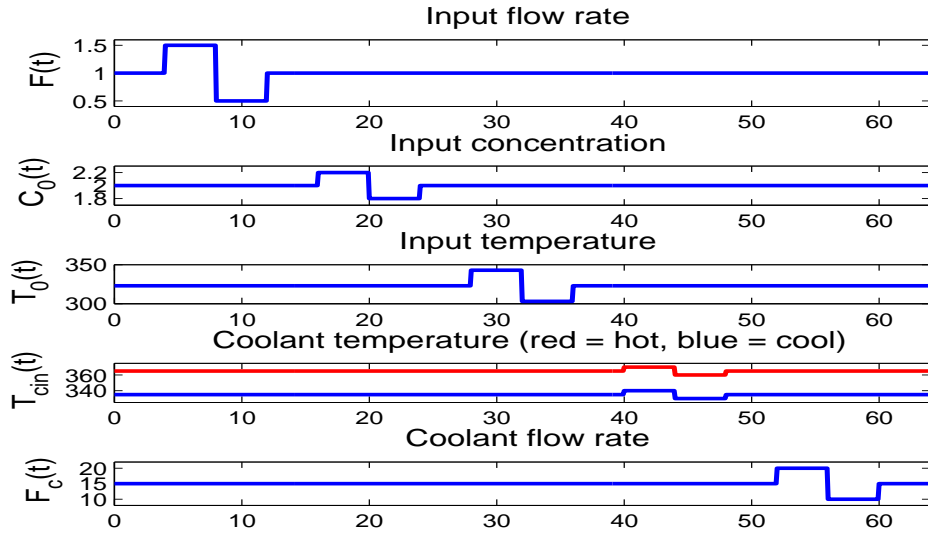
Fig. 3. The five inputs to the chemical reactor modeled by the two equations (3): flow rate $F(t)$, input concentration $C_0(t)$, input temperature $T_0(t)$, coolant temperature $T_{cin}(t)$ and coolant flow $F_0(t)$.
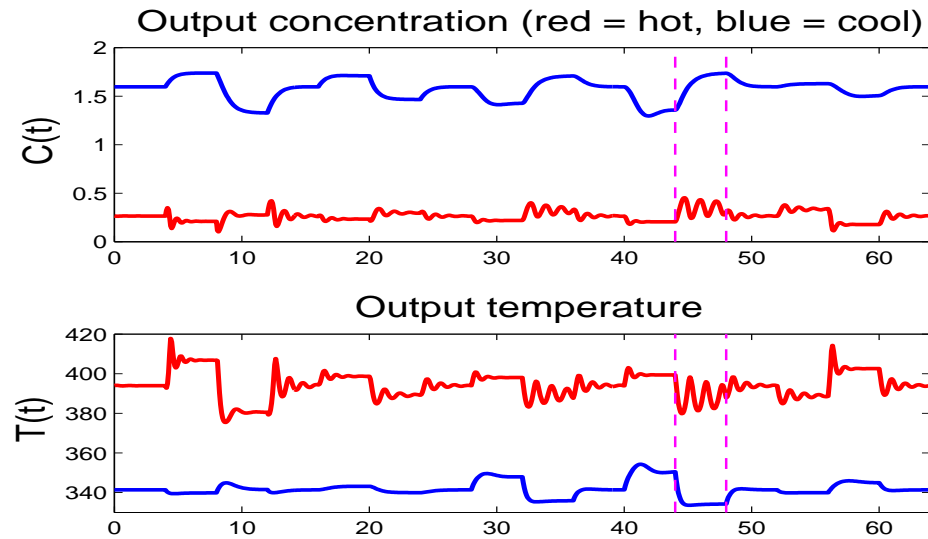

Fig. 4. The two outputs, for each of coolant temperatures T_{cool} of 335 and 365 deg. K, from the chemical reactor modeled by the two equations (3): concentration $C(t)$ and temperature $T(t)$. The input functions are shown in Figure 3. Times at which an input variable is changed are shown as vertical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

would be highly 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. We have noted that step changes in inputs and near over-parameterization are common problems in dynamic modeling.

1.3. A 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 and discretization methods for initial value 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.

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

The numerical methods most often used to approximate solutions of ODE's over a range $[t_0, t_1]$ use fixed initial values $\mathbf{x}_0 = \mathbf{x}(t_0)$ and adaptive discretization techniques. The data fitting process, often referred to by textbooks as the nonlinear least squares or NLS method, goes 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 added 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 be also be approximated by solving the *sensitivity differential equations*

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

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

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

There are a number of variants on this theme; any numerical method could conceivably be used with any optimization algorithm. The most conventional of these are Runge-Kutta integration methods, combined with gradient descent in the survey paper Biegler et al.

(1986), and with a Nelder-Mead simplex algorithm in Fussmann et al. (2000). 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, and especially for stiff systems or for discontinuous inputs such as step functions or functions concentrating their masses at discrete points. In any case, numerical solution noise is added to that of the data so as to further degrade parameter estimates. The size of the parameter set may be increased by the set of initial conditions needed to solve the system. 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. Nevertheless, the wide use of NLS testifies to the fact that, at least for simple smooth systems, it can meet the goals of the application.

A Bayesian approach which avoids the problems of local minima was suggested in Gelman et al. (2004). The authors set up a model where observations y_j at times t_j, conditional on $\boldsymbol{\theta}$, are modelled with a density 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 $\boldsymbol{\theta}$ 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 updated. Since $\hat{\mathbf{x}}(t_j|\boldsymbol{\theta})$ must be numerically approximated conditional on the latest parameter estimates, this approach has some of the problems of the NLS method.

1.3.2. Collocation methods using basis function expansions

Our own approach belongs in the family of collocation methods that express x_i in terms a basis function expansion

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

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 x_i and its derivatives that is required to satisfy the system equations (1). Although the original collocation methods used polynomial bases, spline systems tend to be used currently because of their computational efficiency, but also because they allow control over the smoothness of the solution at specific values of t. The latter property is especially useful for dealing with discontinuities in $\dot{\mathbf{x}}$ associated with step and point changes in inputs $\mathbf{u}$. The problem of estimating x_i is transformed into the problem of estimating the coefficients in $\mathbf{c}_i$. Collocation, of course, has its analogues everywhere in applied mathematics and statistics, and is especially close in spirit to *finite element methods* for approximating solutions to partial differential equations. Basis function approaches to data smoothing in statistics adopt the same approach, but in the approach that we propose, $x_i(t|\mathbf{c}_i)$ must come at least close to solving (1), the structure of $\mathbf{f}$ being a source of additional "data" that inform the fitting process.

Collocation methods were originally developed for boundary value problems, but the use of 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)). Collocation with spline bases was applied to data fitting problems

involving an ODE model by Varah (1982), who suggested a two-stage procedure in which each x_i is first estimated by data smoothing methods without considering satisfying (1), followed by the minimization of a least squares measure of the fit of $\dot{\mathbf{x}}$ to $\mathbf{f}(\mathbf{x}, \mathbf{u}, t|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. The method worked well for the simple equations that were considered in that paper, 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. Voss et al. (1998) suggested using finite difference methods to approximate $\dot{\mathbf{x}}$, but difference approximations are frequently too noisy and biassed to be useful.

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 the size of $\dot{\mathbf{x}} - f(\mathbf{x}, \mathbf{u}, t|\boldsymbol{\theta})$ 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.

Bock (1983) proposed a *multiple shooting method* for data fitting combined with Gauss-Newton minimization, and a similar approach is followed in Li et al. (2005). Multiple shooting has been extended to systems of partial differential equations in Müller and Timmer (2004). These methods incorporate parameter estimation into the numerical scheme for solving the differential equation; an approach also followed in Tjoa and Biegler (1991). They bear some similarity to our own methods in the sense that solutions to the differential equations are not achieved at intermediate steps. However, our method can be viewed as enforcing an soft-threshold that represents an interpretable compromise between data fitting and solving the ODE.

1.4. An 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 2.8 follows up with some results on limiting behavior of estimates as the smoothing parameters increase, and discusses some heuristics.

Sections 3 and 4 show how the method performs in practice. 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 and some open problems in fitting differential equations are given in Section 6. Some consistency results are provided in the Appendix.

2. The 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 $\mathbf{x} = \boldsymbol{\Phi}\mathbf{c}$.

2.1.   An overview of the estimation procedure

Defining $\mathbf{x}$ 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 structural parameters, with a ratio of about 200 applying in the CSTR problem.

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 marginalization.

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}.

While in some applications users may be happy to adjust the values in $\boldsymbol{\lambda}$ manually, we envisage also the data-driven estimation of $\boldsymbol{\lambda}$ 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. Our estimation procedure is, in effect, a multi-criterion optimization problem, and we can refer to J, H and F as *inner*, *middle* and *outer* criteria, respectively. We have applied this approach to semi-parametric regression in Cao and Ramsay (2006), and also note that Keilegom and Carroll (2006) use a similar approach, also in semiparametric 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. We find that this results in much faster computation than in our parallel experiments with MCMC methods, and is far easier to deploy to users in the form of flexible and extendable computer code.

2.2. The data fitting criterion

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

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

Because the use of least squares as a criterion is so common, some remarks are offered on the case e_{ij}'s are independently distributed as $N(0, \sigma_i^2)$. 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, each 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 $x_i(t_{ij})$ 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 x_i satisfies the ODE system can then be assessed by

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

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

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

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

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

where $\mathbf{L}$ is denotes the vector containing the d differential operators $L_{i,\boldsymbol{\theta}}$. 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 $\hat{\mathbf{c}}(\boldsymbol{\theta}; \boldsymbol{\lambda})$

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

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

In the least squares case, this reduces to

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

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

2.5. Outer optimization for θ

In this and the remainder of the section, we simplify the notation considerably by dropping the dependency of criterion H on $\boldsymbol{\sigma}$ and $\boldsymbol{\lambda}$; 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 $\boldsymbol{\theta}$ and error distribution parameters $\boldsymbol{\sigma}$ where required. It is assumed that H is twice continuously differentiable with respect to both $\boldsymbol{\theta}$ and $\mathbf{c}$, and that the second partial derivative or Hessian matrices

$$\frac{\partial^2 H}{\partial \boldsymbol{\theta}^2} \text{ and } \frac{\partial^2 H}{\partial \mathbf{c}^2}$$

are positive definite over a nonempty neighborhood $\mathcal{N}$ of $\mathbf{y}$ in data space.

The gradient or total derivative, $DH(\boldsymbol{\theta})$, with respect to $\boldsymbol{\theta}$ is

$$DH(\boldsymbol{\theta}) = \frac{\partial H}{\partial \boldsymbol{\theta}} + \frac{\partial H}{\partial \mathbf{c}} \frac{d\hat{\mathbf{c}}}{d\boldsymbol{\theta}}. \tag{16}$$

Since $\hat{\mathbf{c}}(\boldsymbol{\theta})$ is not available explicitly, we apply the *implicit function theorem* to obtain

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

and

$$DH(\boldsymbol{\theta}) = \frac{\partial H}{\partial \boldsymbol{\theta}} - \frac{\partial H}{\partial \mathbf{c}} \left(\frac{\partial^2 J}{\partial \mathbf{c}^2}\right)^{-1} \frac{\partial^2 J}{\partial \mathbf{c} \partial \boldsymbol{\theta}}. \tag{18}$$

The matrices used in these equations and those below have complex expressions in terms of the basis functions in $\boldsymbol{\Phi}$ 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 $\boldsymbol{\Sigma}$ be the variance–covariance matrix for $\mathbf{y}$. Making explicit the dependency of H on the data $\mathbf{y}$ by using the notation $H(\boldsymbol{\theta}|\mathbf{y})$, the estimate $\hat{\boldsymbol{\theta}}(\mathbf{y})$ of $\boldsymbol{\theta}$ is the solution of the stationary equation $\partial H(\boldsymbol{\theta}, |\mathbf{y})/\partial\boldsymbol{\theta} = 0$. 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

$$\left[(\frac{d\mathbf{x}}{d\boldsymbol{\theta}})' \boldsymbol{\Sigma} (\frac{d\mathbf{x}}{d\boldsymbol{\theta}}) \right]^{-1}$$

by making use of the approximation

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

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

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

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 $\boldsymbol{\mu} = \mathrm{E}(\mathbf{y})$, 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 $\mathrm{E}(d\hat{\boldsymbol{\theta}}/d\boldsymbol{\mu}) \approx \mathrm{E}(d\hat{\boldsymbol{\theta}}/d\mathbf{y})$. 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 side of (24), we derive

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

$$\approx \quad \left[\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}}\right]\boldsymbol{\Sigma}\left[\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}}\right]', \quad \text{since} \quad \mathrm{E}\left(\frac{d\hat{\boldsymbol{\theta}}}{d\boldsymbol{\mu}}\right) \approx \mathrm{E}\left(\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}}\right).\tag{25}$$

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

$$\mathtt{Var}[\hat{\mathbf{c}}(\hat{\boldsymbol{\theta}}(\mathbf{y}))] = \left(\frac{d\hat{\mathbf{c}}}{d\mathbf{y}}\right)\boldsymbol{\Sigma}\left(\frac{d\hat{\mathbf{c}}}{d\mathbf{y}}\right)',\tag{26}$$

where

$$\frac{d\hat{\mathbf{c}}}{d\mathbf{y}} = \frac{d\hat{\mathbf{c}}}{d\hat{\boldsymbol{\theta}}}\frac{d\hat{\boldsymbol{\theta}}}{d\mathbf{y}} + \frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}\,.\tag{27}$$

2.7. Numerical integration in the inner optimization

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

$$\mathtt{PEN}_i(\mathbf{x}) \approx \sum_q^Q v_q[L_i(x_i(t_q))]^2\tag{28}$$

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. It may be that there will be multiple knots at such a location in order to deal with step function inputs that will imply discontinuous derivatives. We have obtained satisfactory results by dividing 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, great improvements in speed of computation are 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 $\sum_i \lambda_i \mathtt{PEN}_i(\mathbf{x})$ as

$$\sum_i \sum_q [0 - (\lambda_i v_q)^{1/2}L_i(x_i(t_q))]^2\,.$$

These squared residuals can then be appended to those in H, and Gauss-Newton minimization can then be used. When the coefficients enter linearly into the expression for the fitting function, the inner optimization can be avoided entirely by using the explicit solution that is available in this case.

2.8. Choosing the amount of smoothing

Recall that the central goal of this paper is to estimate parameters, rather than to smooth the data. This means that traditional approaches to the choice of smoothing parameter, such as those based on cross validation, may no longer be appropriate. The theory derived in Section 2.9, suggests that when the data agree well with the ODE model, the λ_i should be chosen as large as possible, bounded only by the possibility of distortion from our choice of basis expansion (7).

In our experience, however, real world systems are rarely perfectly described by ODEs. In such situations, we may wish to choose a limited value for λ_i in order to be able to account for systematic discrepancies between ODE solutions and the data. In this sense, the amount of smoothing provides a continuum of solutions representing trade-offs between the problem of estimating $\boldsymbol{\theta}$ and fitting the data well. For each value of the λ_i, we are given two fits to the data; the smooth $\mathbf{x}$ at the estimated $\hat{\boldsymbol{\theta}}$ and the set of exact solutions to the ODE at $\hat{\boldsymbol{\theta}}$. The discrepancy between these two will decrease as λ_i increases and can be viewed as a diagnostic for lack of fit in the model, and therefore an additional benefit of this approach. The fit to the data defined by an exact solution to the equations can be obtained by computing solutions to the initial value problem corresponding to the estimated initial values $\hat{\mathbf{x}}(0)$. It may be helpful to try optimizing these initial conditions using the NLS method, where parameter values are kept fixed at their estimated values.

The degree of smoothing also affects the numerical properties of our estimation scheme. Typically, larger values of λ_i make the inner optimization harder, increasing the number of Gauss-Newton iterations required. Smaller values also appear to make the response surface for the outer optimization more convex, a point discussed further in Section 2.10. This suggests a scheme of estimating $\hat{\boldsymbol{\theta}}$ at increasing amounts of smoothness in order to overcome the local minima seen in Figure 2. Under this scheme an upper limit on λ_i is reached when the basis approximation begins to add too much numerical error to the estimation of $\mathbf{x}$. A simple diagnostic is therefore to solve the ODEs by a Runge-Kutta method and attempt to perform the smoothing in the inner optimization on the resulting data. λ_i should be kept below a level at which the smoothing process distorts these data.

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 $\mathbf{x}_0$. In other words, we treat $\mathbf{x}_0$ as nuisance parameters and estimate $\boldsymbol{\theta}$ by profiling. When $\mathbf{f}$ is Lipschitz continuous in $\mathbf{x}$ 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

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

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 therefore that there are objects $\mathbf{x}$ that satisfy $\text{PEN}(\mathbf{x}|\boldsymbol{\theta}) = 0$. This is guaranteed locally by the following theorem adapted from Bellman (1953):

THEOREM 2.1. *Let* $\mathbf{f}$ *be Lipschitz continuous and* $\mathbf{u}$ *differentiable almost everywhere, then the initial value problem:*

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}, \mathbf{u}, t|\theta), \ \mathbf{x}(t_0) = \mathbf{x}_0$$

has a unique solution.

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

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

Specifically, we will assume that the minimizers of these are well-defined and bounded uniformly over λ. Guarantees on boundedness may be given for whenever $\mathbf{x} \cdot \mathbf{f}(\mathbf{x}, \mathbf{u}, t|\boldsymbol{\theta}) < 0$ for $\|\mathbf{x}\|$ greater than some K. This 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. It seems reasonable, however, that they will hold for systems of practical interest.

We will assume that the 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 2.2. *Let* $\lambda_k \to \infty$ *and assume that*

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

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

Further, when $\text{PEN}(\mathbf{x}|\boldsymbol{\theta})$ 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 left to Appendix B.

Turning to the outer optimization, we obtain the following:

THEOREM 2.3. *Let* $\mathcal{X} \subset (W^1)^n$ *and* $\Theta \subset \mathbb{R}^p$ *be bounded. Let*

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

be well defined for each $\boldsymbol{\theta}$ and λ, 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

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

also be well defined for each λ. Then

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

This theorem requires fairly strong assumptions about the regularity of solutions to the inner optimization problem. Conditions on $\mathbf{f}$ that will provide this regularity is a matter of ongoing research. 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.

2.10. Heuristics for robust estimates

We believe that our method provides a computationally tractable parameter estimate that is numerically stable and easy to implement. It has also been our experience that these estimates are robust with respect to starting values for the optimization procedure. Figure 5 plots a similar to Figure 2 but providing the squared error of the spline fit as parameters a and b are varied. The plot shown is for $\lambda = 10^5$, experimentally, as λ becomes smaller, the surfaces become more regular.

We do not have a formal mathematical statement to indicate that these response surfaces become more regular. As a heuristic, we have already noted that

$$l(\mathbf{x}_{\lambda,\boldsymbol{\theta}}) \leq l(\mathbf{x}_{\boldsymbol{\theta}})$$

for any $\mathbf{x}_{\boldsymbol{\theta}}$ that satisfies $P(\mathbf{x}_{\boldsymbol{\theta}}|\boldsymbol{\theta}) = 0$. The squared error surface at λ is therefore an *underestimate* of the response surface for exact solutions to the differential equation. Moreover, Appendix A provides an expression for the derivative of $\mathbf{c}$ with respect to $\boldsymbol{\theta}$ that is of the form

$$\lambda \left[A + \lambda B\right]^{-1} C$$

whose norm increases with λ. Thus these surfaces must be less steep as λ becomes smaller. This, however, does not demonstrate the observation that they eventually become convex.

Our experimental evidence suggests that for small values of λ, parameter estimates tend to be more variable and can become quite biassed. However, Theorem 2.3 demonstrates that as λ becomes large, the estimates become approximately unbiassed. This suggests that a scheme that uses a small values of λ to find a global optimum and then increases λ incrementally may be useful for particularly challenging surfaces.
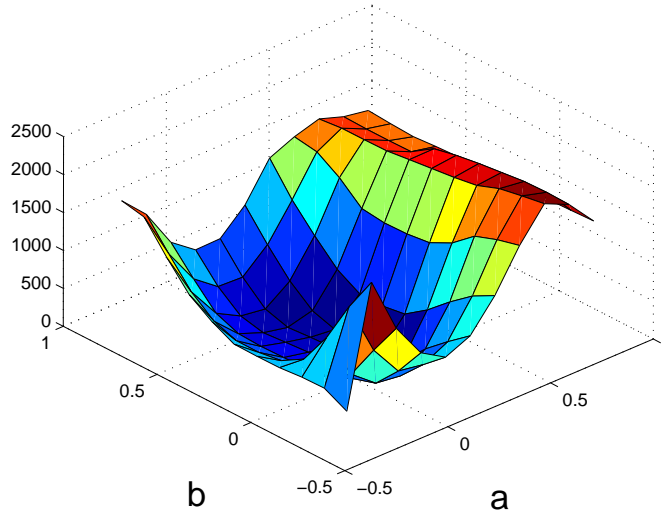
Fig. 5. FitzHugh-Nagumo response surfaces over a and b for $\lambda = 10^5$. Values of the surface are calculated using the same data as in Figure 2

3. Simulated data examples

3.1. Fitting the FitzHugh-Nagumo equations

We set up simulated data for V from the FitzHugh-Nagumo equations as a mathematical test-bed of our estimation procedure. Data were generated by taking solutions to the equations with parameters $\{a, b, c\} = \{0.2, 0.2, 3\}$ and initial conditions $\{V, R\} = \{-1, 1\}$ measured at 0.05 time units on the interval [0,20]. Noise was then added to the solution with standard deviation 0.5.

We estimated the smooths for each component using a third order B-spline basis with knots at each data point. A five-point quadrature rule was used for the numerical integration. Figure 6 gives quartiles of the parameter estimates for 60 simulations as λ is varied from 10^{-2} to 10^5. It is apparent that there is a large amount of bias for small values of λ. This is not surprising – the spline fit is affected very little by $\boldsymbol{\theta}$ and, in being very irregular, has high derivatives. Effectively, we select a fit that nearly interpolates the data and then choose $\boldsymbol{\theta}$ to try to mimic the fit as well as possible. However, as λ becomes large, parameter estimates become nearly unbiased and tightly centered on the true parameter values. Table 3.1 provides bias and variance estimates from 500 simulations at $\lambda = 10^4$. These are provided along with the estimate of standard error developed in Section 2.6 and the usual Gauss-Newton standard error. We obtain good coverage properties for our estimates of variance while the Gauss-Newton estimates are somewhat less accurate. However, the estimates based on Section 2.6 required 10 times the computer time than the standard estimates and we found that these could be unreliable for smaller sample sizes. Parameter estimates for a and c are very close to the true values. There appears to be a small amount of bias for the estimate of d, which we conjecture to be due to the use of a basis expansion.
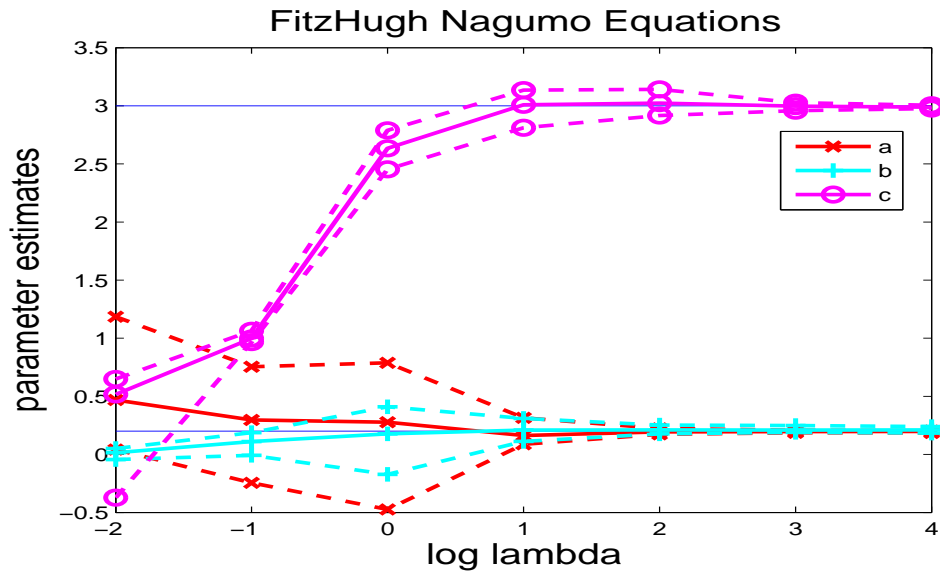
Fig. 6. Quartiles of parameter estimates for the FitzHugh-Nagumo Equations as λ is varied. Horizontal lines represent the true parameter values.

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

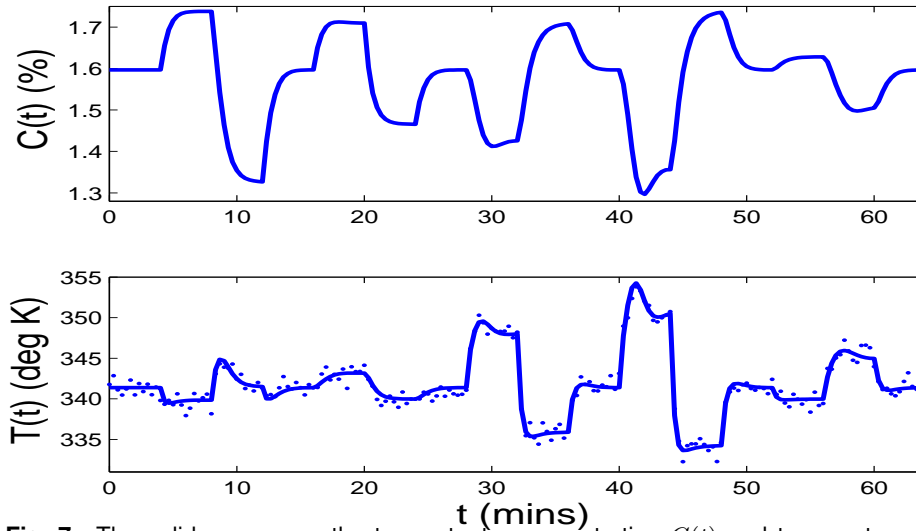|	a	b	c
True value	0.2000	0.2000	3.0000
Mean value	0.2005	0.1984	2.9949
Std. Dev.	0.0149	0.0643	0.0264
Est. Std. Dev.	0.0143	0.0684	0.0278
GN. Std. Dev.	0.0167	0.0595	0.0334
Bias	0.0005	-0.0016	-0.0051
Std. Err.	0.0007	0.0029	0.0012

Fig. 7. The solid curves are the two outputs, concentration $C(t)$ and temperature $T(t)$, defined by the chemical reactor model (3). The dots associated with the temperature curve are simulated measurements with a error level of about 20% of the variability in the smooth curve.

3.2. Fitting the tank reactor equations

The data in Figure 7 were simulated by adding zero mean Gaussian noise to numerical estimates of the solutions $C(t)$ and $T(t)$ of the equations for values of the parameters given in Marlin (2000): $\kappa = 0.461, \tau = 0.833, a = 1.678$ and $b = 0.5$. The standard deviations of the errors were 0.0223 for concentration and 0.79 for temperature, values which are about 20% of the standard deviations of the respective variable values, this being an error level that is considered typical for such processes.

Temperature measurements are relatively cheap and accurate relative to those for concentration, and the engineer may wish to base his estimates on these alone, in which case concentration effectively becomes a functional latent variable. Naturally, it would be wise to use data collected in the stable cool experimental regime in order to predict the response in the hot reaction mode.

We now consider how well the parameters κ, τ and a and the equation solutions can be estimated from the simulated data in Figure 7, keeping b fixed at 0.5 because we have determined that the accurate estimation of all four parameters is impossible within the data design described above.

We attempted to estimate these parameters using the nonlinear least squares or NLS method described in Section 1.3.1. At the times of step changes in inputs, the approximation to solutions using the Runge-Kutta algorithm with 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 to realize. 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

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
Std. Dev.	0.0034	0.0057	0.0188	0.0084	0.0085	0.0377
Est. Std. Dev.	0.0035	0.0056	0.0190	0.0088	0.0090	0.0386
Bias	0.0000	0.0000	-0.0001	0.0003	-0.0002	0.0015
Std. Err.	0.0002	0.0004	0.0012	0.0005	0.0005	0.0024

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 smoothing parameters λ_C and λ_T were 100 and 10, respectively. 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 usual Gauss-Newton method, using the Jacobian with respect to the parameters, and the two values seem sufficiently close for all three parameters to permit us to trust the Gauss-Newton estimates. As one might expect, the main impact of having only temperature measurements is to increase the sampling error in the parameter estimates.

The principal components of variation of the correlation matrix for the parameter estimates derived from both variables measured accounted for 85.0, 14.0 and 1.0 percent of the variance, respectively, indicating that, even after re-scaling the parameters, most of the sampling variation in these three parameters is in only two dimensions. Moreover, the scatter is essentially Gaussian in distribution, indicating that a further reduction the dimensionality of the parameter space using linear transformations might be worth considering. In particular, the correlation between parameters κ and a is 0.94, suggesting that these may be linked together without much loss in fitting power.

When the equations were solved using the parameters estimated from measurements on both variables, the maximum absolute discrepancy between the fitted concentration curve and the true curve was 0.11% of the true curve. The corresponding temperature discrepancy was 0.03%. When these parameter estimates were used to calculate the solutions in the hot mode of operation, the maximum concentration and temperature discrepancies became 1.72% and 0.05%, respectively. These error levels would be regarded as negligible by engineers interested in forecasting the consequences of running the reactor in hot mode. Finally, when the parameters were estimated from only the temperature data, the concentration and temperature discrepancies became 0.10% and 0.04%, respectively, so that only the quickly and cheaply attainable measurements of temperature seem sufficient for identifying this system in either mode of operation.

4. Working with real data

4.1. Modeling nylon production

This illustration concerns the decomposition of the polymer nylon into its constituents. 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 $A + C \rightleftharpoons L + W$. In an experiment described in (Zheng et al. (2005)), a mixture of steam and an inert gas was bubbled into a molten nylon to maintain an approximately constant amount of W in the system, thereby causing A, C, L and W to move towards equilibrium concentrations. Within each of six experimental runs the pressure of the steam was first stepped down from its initial level at times $\tau_{j1}, j = 1, \ldots, 6$, then back up at to it's initial pressure at time τ_{j2} until the end of the experiment. The temperature T_j was kept constant within a run, but varied over runs, as did the initial concentrations of A and C. The goal was to estimate the rate parameters governing the chemical reactions of nylon production.

Samples of the molten mixture were extracted at irregularly spaced intervals, and the concentrations of A and C were measured, all though the more expensive measurements of C were not made at all A measurement times. Figure 8 shows the data for the runs aligned by experiment within columns. Vertical lines correspond to τ_{j1} and τ_{j2}. 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 model for the reaction dynamics was

$$
\begin{aligned}
-DL = DA = DC &= -k_p * 10^{-3}(CA - LW/K_a) \\
DW &= k_p * 10^{-3}(CA - LW/K_a) - k_m(W - W_{eq})
\end{aligned}
\tag{29}
$$

The constant $k_m = 24.3$ was estimated in previous studies. The two step changes in input W_{eq} induces two discontinuities in the derivatives given in (29). Due to the mass balance of the reactions, if A, C and W are known then L can be algebraically removed from the equations. Consequently, we will only estimate those three components. The reaction rate parameter K_a, which depends on the temperature T, is

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

where the ideal gas constant $R = 8.3145 * 10^{-3}$, $C_T = 20.97 \exp[-9.624 + 3613/T]$ and a reference temperature $T_0 = 549.15$ was chosen to be in the middle of the range of experimentally manipulated temperatures. The parameter vector to estimate is $\boldsymbol{\theta} = [k_p, g, K_{a0}, \Delta H]$. The scaling factor of 1000 selected to scale all initial parameter absolute values into the range $[17, 78.1]$. Further details concerning the experiment, and these and other analyses, can be found in Zheng et al. (2005) and Campbell et al. (2006).

Since the input W_{eq} is a step function of time, it induces a discontinuity in the derivative of the smooth for all three system outputs. This means that the linear differential operator in (10) is not defined at the times $\{\tau_{j1}, \tau_{j2}\}$, and consequently we removed a small neighborhood $[\tau - \delta, \tau + \delta]$ around these points before computing the integral in PEN, δ being 10^{-6} times the smallest interval between unique neighboring knots. We used a fifth order
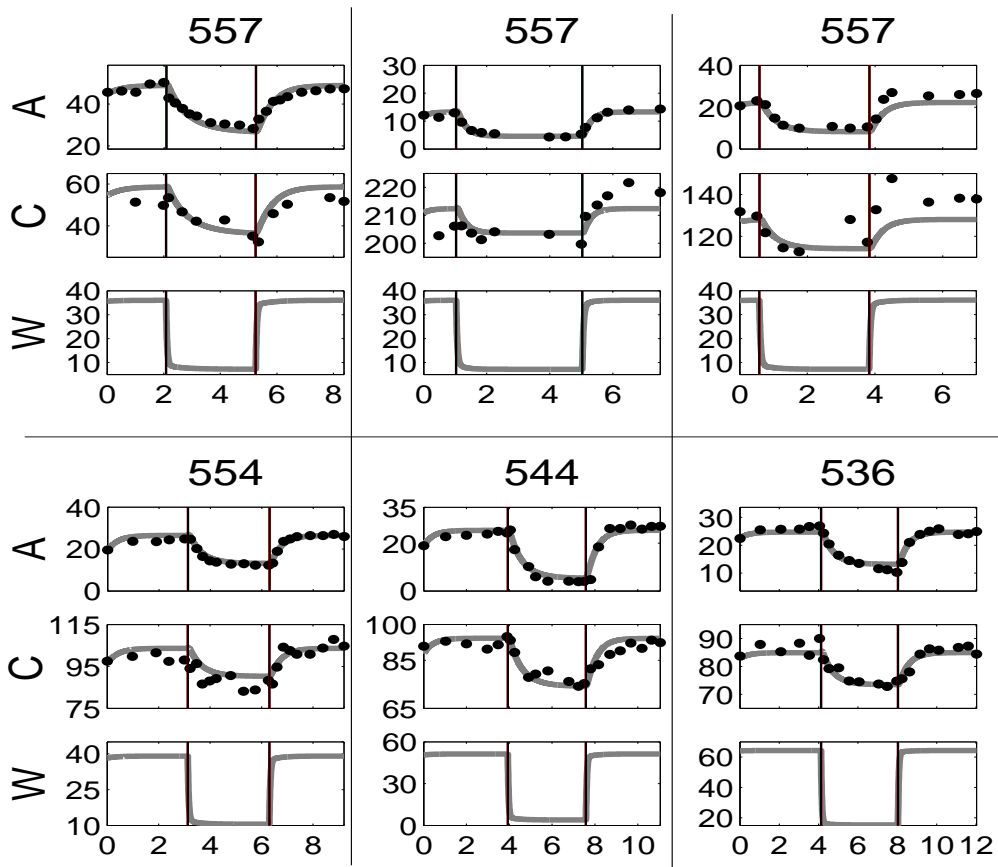
Fig. 8. Nylon components A, C and W along with the solution to the differential equations using initial values estimated by the smooth for each of six experiments. The times of step change in input pressures are marked by thin vertical lines. Horizontal axes indicate time in hours, and vertical axes are concentrations in moles. The labels above each experiment indicate the constant temperature in degrees Kelvin.

b-spline basis with knots at each observation of A and included additional knots in order to assure a knot rate of at least five per hour. Multiple knots were included at times τ_{j1} and τ_{j2} to allow a discontinuity in the smoothing function's first derivative. The same basis was used for the all three components within an experimental run. We use weights $w_A = 1/.6$ and $w_C = 1/2.4$, these being the reciprocals of the measurement standard deviations.

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.

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 8. While the fit to the data is quite good overall, there does seem to be a positive autocorrelation of residuals within a run.

4.2. Modeling flare dynamics in lupus

Lupus is an auto-immune 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.

Various scales have been developed to measure the severity of symptoms, and Figure 9 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 postulate 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 take $\delta = 0.02$ years, or about two weeks. We represent the time course of attacks as a box function $u(t)$ that is 0 during normal functioning and 1 during a flare.

We begin with the following simple linear differential equation for symptom severity $s(t)$ at time t

$$Ds(t) = -\beta(t)s(t) + \alpha(t)u(t). \tag{30}$$

This equation has the solution

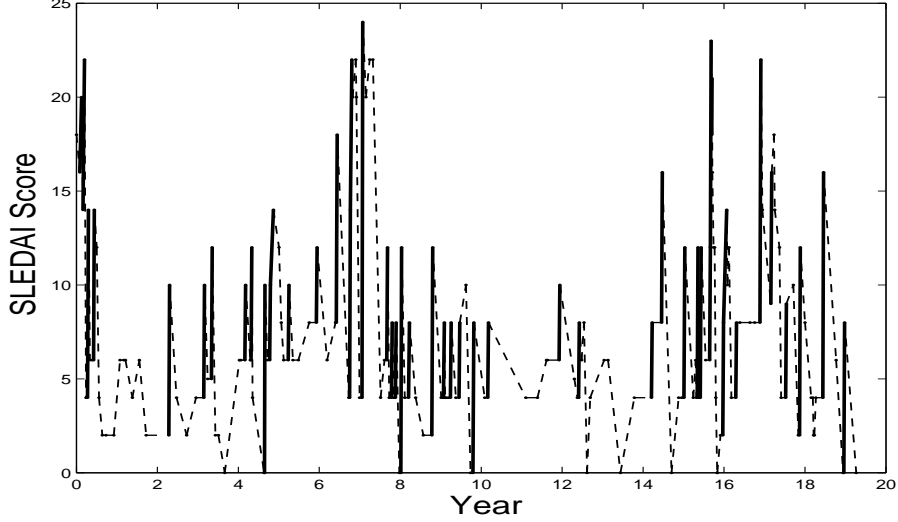$$s(t) = Cs_0(t) + s_0(t) \int_0^t \alpha(z)u(z)/s_0(z) \, dz$$

Fig. 9. 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.

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 will represent this as a linear combination of 8 cubic B-spline basis functions defined by equally spaced knots and with about three years between knots. We expect 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 seems 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 propose a second simple linear equation to model this mechanism:

$$D\beta(t) = -\gamma\beta(t) + \theta[1 - u(t)]. \tag{31}$$

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

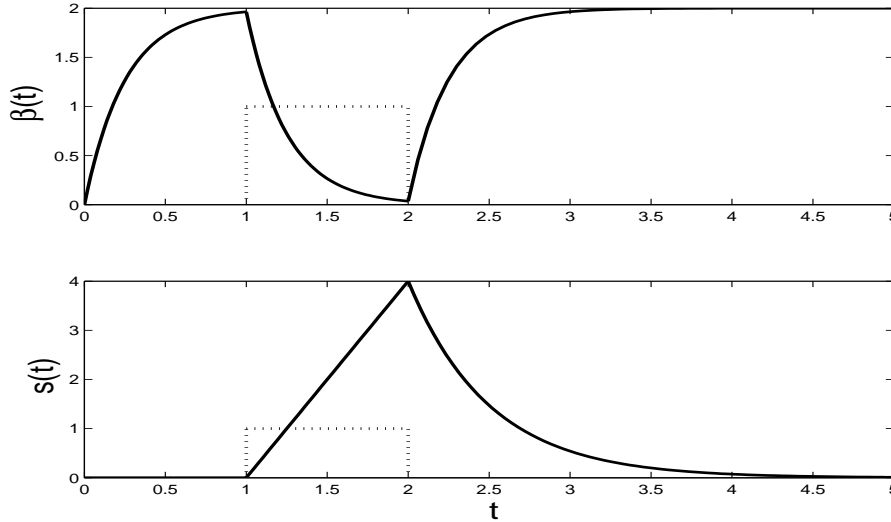$$D\beta(t) = \theta[1 - u(t)], \tag{32}$$

Fig. 10. The top panel shows the effect of a lupus attack on the weight function $\beta(t)$ in differential equation (30). The bottom panel shows the time course of the symptom severity function $s(t)$. These results are for parameters $\gamma = \theta 4$.

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

The top panel in Figure 10 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 11. 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 (32). 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 11 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.

Our model provides two estimates of the symptom levels. The fitted function $s(t)$ is shown as a solid line. It was defined by positioning three knots at each of the flare onset and offset times in order to accommodate the sudden break in the first derivative of $s(t)$, 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 $s(t) = \mathbf{c}'\boldsymbol{\phi}(t)$. 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 (30), and the solution to this equation is shown as a dashed line. The discrepancy between the fit defined by the equation and the smoothing function $s(t)$ is important in years 8 to 11, where the equation solution over-
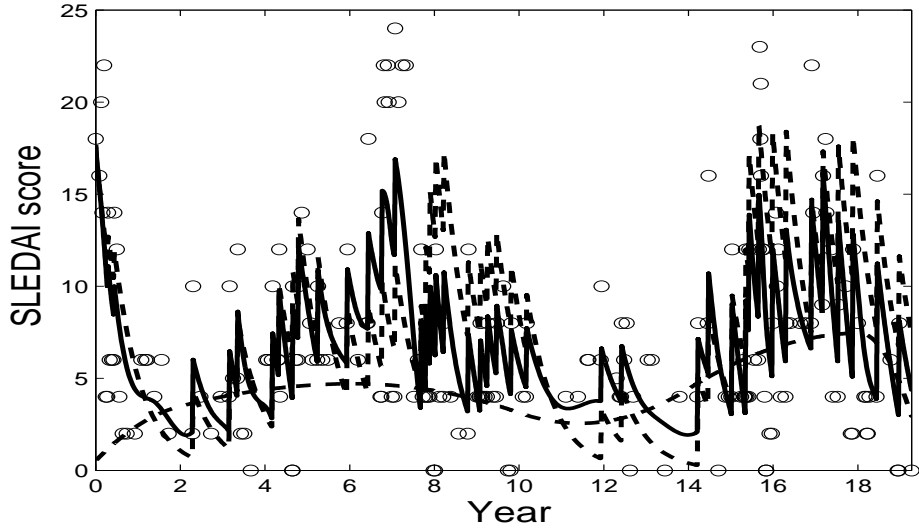
Fig. 11. The circles indicate SLEDAI scores, the jagged solid line is the smoothing functions $s(t)$, the dashed jagged line is the solution to the differential equation and the smooth dashed line is the smooth trend $\alpha(t)$.

estimates symptom level. In this region, new flares come too fast for recovery, and thus build on each other. A more detailed view over the years 14 to the end of the record is in Figure 12, and we see there that the ODE solution is less able than the smooth to track the data when flares come close together.

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

The methodology presented here has been described for systems of ordinary differential equations. However, the idea is much more general. In any parametric situation, if we can define a $\mathtt{PEN}(\mathbf{x}|\theta)$ whose zero set is indexed by nuisance parameters and the estimation of θ is of interest, then similar methods may be applied. The generalization of Theorems 2.2 and 2.3 are immediate.

In dynamical systems, we have already noted that an mth order system of the form:

$$D^n\mathbf{x}(t) = \mathbf{f}(\mathbf{x}, \dot{\mathbf{x}}, \dots, D^{n-1}\mathbf{x}, \mathbf{u}, t|\boldsymbol{\theta}) \tag{33}$$

may be reduced to a larger first-order system by defining the derivatives $\dot{\mathbf{x}}$ up to $D^{n-1}\mathbf{x}$ as new variables. Initial conditions need to be given for each of these new variables in
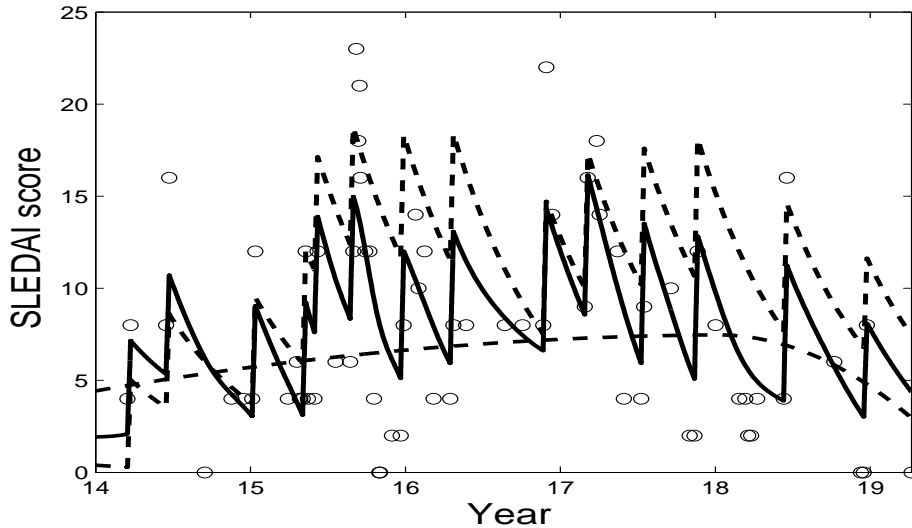
Fig. 12. The data in Figure 11 plotted over the last five years of the record.

order to define a unique solution. Equation (33), however, can be used directly to define a differential operator as in (10), saving the estimation of the derivative terms and all the initial conditions. There is, of course, no need for n in (33) to be constant across components of $\mathbf{x}$, or to restrict to equations that may be written in the form (33).

A slight generalization of (33) is to allow n to be zero for some components, that is define

$$x_i(t) = f_i(\mathbf{x}, \mathbf{u}, t|\boldsymbol{\theta}) \tag{34}$$

some some components i. Such a system is labelled a *Differential-Algebraic System* and these have been used in chemical engineering (Biegler et al. (1986)). In general, a numerical solution of such equations requires (34) to be solved numerically given the other values of $\mathbf{x}$. Our approach also allows (34) to appear as a term in $\text{PEN}(\mathbf{x}|\boldsymbol{\theta})$, providing an easier implementation of such systems.

A further generalization allows $\mathbf{f}$ to include lags. That is

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t - \delta_1), \mathbf{x}(t - \delta_2), \dots, \mathbf{x}(t - \delta_3), \mathbf{u}(t - \delta_4), t|\boldsymbol{\theta}) \tag{35}$$

in which case $\mathbf{x}(t)$ needs to be specified for all values in $[t_0 - \max \delta_i, t_0]$ as initial conditions. Again, in its generality, our methodology can include such systems without knowing initial conditions. We can also estimate the δ_i; an example of doing so in a simple system is given in Koulis et al. (2006).

Although we have only considered ordinary differential equations in this paper, the methodology extends naturally to partial differential equations in which a system $\mathbf{x}(s, t)$ is described over spatial variables s as well as time t. In this case, the system may be described

in terms of both time and space derivatives

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

The smooth $\mathbf{x}(s, t)$ now requires a multi-dimensional basis expansion, but the same estimation and variance estimation schemes already discussed can be carried out in a straightforward manner.

Finally, we note that the data criterion (14) may be interpreted as the log likelihood for an observation from the stochastic differential equation:

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

where $W(t)$ 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 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 $dW(t)/dt$ 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 immediately clear. Moreover, it is frequently the case that lack of fit in nonlinear dynamics is due more to miss-specification of the system under consideration than to stochastic inputs. We have therefore restricted the discussion in this paper solely to deterministic systems.

6. Further issues in fitting differential equations

Although we have emphasized situations where initial and/or boundary values for a system are not known, in fact these can be incorporated into the method as constraints on the optimization of inner criterion (14). These constraints can be incorporated explicitly by the use of constrained optimization methods, or implicitly as data that receive large weights or high prior probability through the specification of density $g_i(\mathbf{e}_i | \boldsymbol{\sigma}_i)$ used in fitting criterion (8). Integral constraints arise in statistical contexts such as the nonparametric estimation of density functions, and these, too, can be applied without much additional effort.

Our experiences with real-world data suggest that differential equation models are often not well specified. This is particularly true in biological sciences where the first principles from which they are commonly deduced tend to be less exact than those derived from physics and chemistry. These models are commonly selected only to provide the right *qualitative* behavior and may take values orders of magnitude different from the observed data.

There is therefore a great need for diagnostic tools for such systems. Both to determine the appropriateness of the model and, where it is inappropriate, to suggest ways in which it may be modified. One approach to this is to estimate additional components of $\mathbf{u}$ that will provide good fits. These may then be correlated with observed values of the system, or external factors, to suggest new model formulae.

A typical industrial process involves many outputs and many inputs, with at least some of each varying over time. Engineers plan experiments in which inputs are varied under

various regimes, including randomly or systematically timed changes; and step, ramp, curvilinear and harmonic perturbations. Often the effects of input perturbations are localized and also interactive. These considerations point to a wide spectrum of experimental design problems that statisticians need to address with the help of the system estimation technology proposed here.

We can add to these design issues the choice of sampling rate and accuracy for measurements taken on both input and output variables. For example, in stable systems minor changes in initial values of variables wash out quickly, but for systems that are close to instability, estimating the initial state of the system requires considerable high quality data at start-up. Certain parameters may also affect system behavior only locally, and therefore also require more information where it counts.

7. 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 θ and then estimating θ through a profiling scheme in which the fit is regarded as a nuisance parameter.

We have found that our approaches 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 NLS that add solution approximation noise to data noise. 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 parameter cascade algorithm is as fast or faster than NLS and other approaches, and much faster than the Bayesian-MCMC method, which has comparable estimation efficiency. Unlike MCMC, the parameter cascade or 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 up 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, and the possibility of beginning with a smaller range of smoothing parameters λ 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.

The methodology that we have presented can be adapted to a large number of problems that extend beyond ordinary differential equations; an area that we have yet to explore. For example, it seems fairly straightforward to apply the parameter cascade approach to the solution of partial or distributed differential equation systems, where a finite element basis system would be more practical than a spline basis. Differential-algebraic, integral-differential equations and other more general systems also seem approachable in this way.

Finally, we hope that this method will prove to open wide a door to the statistical community that leads to an exciting range of data analysis opportunities in the burgeoning world of dynamic systems modeling.

7.1. *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

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

Bellman, R. (1953). *Stability Theory of Differential Equations.* New York: Dover.

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

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. A., G. Hooker, J. Ramsay, K. McAuley, J. McLellan, and S. Varziri (2006). Parameter estimation in differential equation models: An application to dynamic systems. McGill University unpublished manuscript.

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

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

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.

Fussmann, G. F., S. P. Ellner, K. W. Shertzer, and N. G. J. Hairston (2000). Crossing the hopf bifurcation in a live predator-prey system. *Science 290*, 1358–1360.

Gelman, A., J. B. C. andH.S. Stern, and D. B. Rubin (2004). *Bayesian Data Analysis*. New York: Chapman and Hall/CRC.

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

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.

Koulis, T., J. . Ramsay, and D. Levitin (2006). Input-output systems in psychoacoustics. submitted to Psychometrika.

Li, Z., M. Osborne, and T. Prvan (2005). Parameter estimation in ordinary differential equations. *IMA Journal of Numerical Analysis 25*, 264–285.

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

Müller, T. G. and J. Timmer (2004). Parameter identification techniques for partial differential equations. *International Journal of Bifurcation and Chaos 14*, 2053–2060.

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.

Voss, H., M. M. Bünner, and M. Abel (1998). Indentification of continuous spaciotemporal systems. *Physical Review E 57*, 2820–2823.

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:

$$\frac{dx_i(t_{i,k})}{d\mathbf{c}_i} = \Phi_i(t_{i,k})$$

where $\Phi_i(t_{i,k})$ is the vector corresponding to the evaluation of all the basis functions used to represent x_i evaluated at $t_{i,k}$. 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:

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

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

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

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

A.2. Outer optimization

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:

$$\frac{d\mathbf{x}(t_{i,k})}{d\boldsymbol{\theta}} = \frac{d\mathbf{x}(t_{i,k})}{d\mathbf{c}} \frac{d\mathbf{c}}{d\boldsymbol{\theta}}$$

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.

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:

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

$$- \lambda_i \int D\phi_{i,p}(t)\frac{df_i}{dx_j}\phi_{jq}(t)dt - \lambda_j \int \phi_{i,p}(t)\frac{df_i}{dx_j}D\phi_{jq}(t)dt$$

$$+ \int \phi_{i,p}(t)\left[\sum_{k=1}^{n} \lambda_k \left(\frac{d^2 f_k}{dx_i dx_j}(f_k - Dx_k(t)) + \frac{df_k}{dx_i}\frac{df_k}{dx_j} \right) \right] \phi_{jq}(t)dt$$

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:

$$\lambda_i \int \frac{df_i}{d\boldsymbol{\theta}}\phi_{i,p}(t)dt - \int \left(\sum_{k=1}^{n} \lambda_k \left[\frac{d^2 f_k}{dx_i d\boldsymbol{\theta}}(f_k - Dx_k(t)) + \frac{df_k}{dx_i}\frac{df_k}{d\boldsymbol{\theta}} \right] \right) \phi_{i,p}(t)dt$$

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

The variance of the parameter estimates is calculated using

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

where

$$\frac{d^2 H}{d\boldsymbol{\theta}^2}\frac{\partial^2 H}{\partial\boldsymbol{\theta}^2} + 2\frac{\partial^2 H}{\partial\hat{\mathbf{c}}\partial\boldsymbol{\theta}}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}} + \left(\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}} \right)'\frac{\partial^2 H}{\partial\hat{\mathbf{c}}^2}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}} + \frac{\partial H}{\partial\hat{\mathbf{c}}}\frac{\partial^2\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}^2}, \tag{36}$$

and

$$\frac{d^2 H}{d\boldsymbol{\theta}d\mathbf{y}}\frac{\partial^2 H}{\partial\boldsymbol{\theta}\partial\mathbf{y}} + \frac{\partial^2 H}{\partial\hat{\mathbf{c}}\partial\mathbf{y}}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}} + \frac{\partial^2 H}{\partial\boldsymbol{\theta}\partial\hat{\mathbf{c}}}\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}} + \frac{\partial^2 H}{\partial\hat{\mathbf{c}}^2}\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}} + \frac{\partial H}{\partial\hat{\mathbf{c}}}\frac{\partial^2\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}\partial\mathbf{y}}. \tag{37}$$

The formulas (36) and (37) 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 caller *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 $\sum_i a_i x_i$, we merely write $a_i x_i$.

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

$$\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}} - \left[\frac{\partial^2 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}^2}\Big|_{\hat{\mathbf{c}}} \right]^{-1}\left[\frac{\partial^2 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial\mathbf{c}\partial\mathbf{y}}\Big|_{\hat{\mathbf{c}}} \right]. \tag{38}$$

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

  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:

  $$\frac{d^2}{d\boldsymbol{\theta} dy_k}\left(\frac{\partial J(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c}}\bigg|_{\hat{\mathbf{c}}}\right) \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c} \partial \boldsymbol{\theta} \partial y_k}\bigg|_{\hat{\mathbf{c}}} + \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c} \partial \boldsymbol{\theta} \partial c_i}\bigg|_{\hat{\mathbf{c}}} \frac{\partial \hat{c}_i}{\partial y_k}$$
  $$+ \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c}^2 \partial y_k}\bigg|_{\hat{\mathbf{c}}} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c}^2 \partial c_i}\bigg|_{\hat{\mathbf{c}}} \frac{\partial \hat{c}_i}{\partial y_k} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \frac{\partial^2 J(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c}^2}\bigg|_{\hat{\mathbf{c}}} \frac{\partial^2 \hat{\mathbf{c}}}{\partial \boldsymbol{\theta} \partial y_k}$$
  $$= 0 \tag{39}$$

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

  $$\frac{\partial^2 \hat{\mathbf{c}}}{\partial \boldsymbol{\theta} \partial y_k} = -\left[\frac{\partial^2 J(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c}^2}\bigg|_{\hat{\mathbf{c}}}\right]^{-1}\left[\frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c} \partial \boldsymbol{\theta} \partial y_k}\bigg|_{\hat{\mathbf{c}}} + \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c} \partial \boldsymbol{\theta} \partial c_i}\bigg|_{\hat{\mathbf{c}}} \frac{\partial \hat{c}_i}{\partial y_k}\right.$$
  $$\left. + \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c}^2 \partial y_k}\bigg|_{\hat{\mathbf{c}}} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c}^2 \partial c_i}\bigg|_{\hat{\mathbf{c}}} \frac{\partial \hat{c}_i}{\partial y_k} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}}\right] \tag{40}$$

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

  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:

  $$\frac{\partial^2 \hat{\mathbf{c}}}{\partial \boldsymbol{\theta} \partial \theta_j} = -\left[\frac{\partial^2 J(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c}^2}\bigg|_{\hat{\mathbf{c}}}\right]^{-1}\left[\frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c} \partial \boldsymbol{\theta} \partial \theta_j}\bigg|_{\hat{\mathbf{c}}} + \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c} \partial \boldsymbol{\theta} \partial c_i}\bigg|_{\hat{\mathbf{c}}} \frac{\partial \hat{c}_i}{\partial \theta_j}\right.$$
  $$\left. + \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c}^2 \partial \theta_j}\bigg|_{\hat{\mathbf{c}}} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}} + \frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})}{\partial \mathbf{c}^2 \partial c_i}\bigg|_{\hat{\mathbf{c}}} \frac{\partial \hat{c}_i}{\partial \theta_j} \frac{\partial \hat{\mathbf{c}}}{\partial \boldsymbol{\theta}}\right] \tag{41}$$

When estimating ODE's, 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, $d^2 H/d\mathbf{c}^2$ is a block-diagonal matrix with the ith block being $w_i \Phi_i(\mathbf{t}_i)^T \Phi_i(\mathbf{t}_i)$ and $dF/d\mathbf{c}$ is a block vector containing blocs $-w_i \Phi_i(\mathbf{t}_i)^T (\mathbf{y}_i - x_i(\mathbf{t}_i))$.

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

$$\int \left(\sum_{l=1}^{n} \lambda_l \left[\frac{d^2 f_l}{dx_i dx_j} \frac{df_l}{dx_k} + \frac{d^2 f_l}{dx_i dx_k} \frac{df_l}{dx_j} + \frac{d^2 f_l}{dx_j dx_k} \frac{df_l}{dx_i}\right]\right) \phi_{i,p}(t) \phi_{jq}(t) \phi_{ku}(t) dt$$
$$+ \int \sum_{l=1}^{n} \lambda_l \left(\frac{d^3 f_k}{dx_i dx_j dx_k}(f_l - Dx_l(t))\right) \phi_{i,p}(t) \phi_{jq}(t) \phi_{ku}(t) dt$$
$$- \lambda_i \int \frac{d^2 f_i}{dx_j dx_k} D\phi_{i,p}(t) \phi_{jq}(t) \phi_{ku}(t) dt - \lambda_j \int \frac{d^2 f_j}{dx_i dx_k} \phi_{i,p}(t) D\phi_{jq}(t) \phi_{ku}(t) dt$$
$$- \lambda_k \int \frac{d^2 f_k}{dx_i dx_j} \phi_{i,p}(t) \phi_{jq}(t) D\phi_{ku}(t) dt$$

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:

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

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

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

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

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

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

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 $w_i \Phi_i(\mathbf{t}_i)$.

B.   Proofs of theorems in section 2.9

B.0.1.   Preliminaries
The following theorem is a well-known consequence of the method of Lagrange multipliers:

THEOREM B.1. *Suppose that x_λ minimizes $F(x) + \lambda P(x)$, then x_λ minimizes $F(z)$ for $z \in \{x : P(x) < P(x_\lambda)\}$. Moreover, for $\lambda' > \lambda$, $P(x_{\lambda'}) \leq P(x_\lambda)$.*

Two corollaries:

COROLLARY B.1. *For $\lambda' > \lambda$, $F(x_{\lambda'}) \geq F(x_\lambda)$.*

COROLLARY B.2. *If $\exists x$ such that $P(x) = 0$, then $P(x_\lambda) \to 0$ as $\lambda \to \infty$.*

follow immediately.

The proofs of Theorems 2.2 and 2.3 rely heavily on the following:

THEOREM B.2. *Let $\mathcal{X}$ and $\mathcal{Y}$ be metric spaces with $\mathcal{X}$ closed and bounded. Let $g(x, \alpha)$: $\mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be uniformly continuous in x and α, such that*

$$x(\alpha) = \operatorname*{argmin}_{x \in \mathcal{X}} g(x, \alpha)$$

is well defined for each α. Then $x(\alpha) : \mathcal{Y} \to \mathcal{X}$ is continuous.

We begin with two lemmas:

LEMMA B.1. *Let $\mathcal{X}$ be a closed and bounded metric space. Suppose that*

$$x^* = \operatorname*{argmin}_{x \in \mathcal{X}} g(x) \tag{42}$$

is well defined and $g(x)$ is continuous. Then

$$\forall \epsilon > 0, \ \exists \delta > 0 \ such \ that \ \|x - x^*\| > \epsilon \Rightarrow f(x) - f(x^*) > \delta.$$

holds for all $x \in \mathcal{X}$.

PROOF. Assume that the the statement is not true. That is, for some $\epsilon > 0$ we can find a sequence $x_n \in \mathcal{X}$ such that $\|x_n - x^*\| > \epsilon$ but $\|g(x_n) - g(x^*)\| < 1/n$. Since $\mathcal{X}$ is closed and bounded, it is compact and there exists a subsequence $x_{n'} \to x^{**} \neq x^*$ for some x^{**}. By the continuity of g, we have $g(x^{**}) = g(x^*)$ violating the assumption that (42) is well defined.

LEMMA B.2. *Let $\mathcal{X}$ and $\mathcal{Y}$ be metric spaces and $g(x, \alpha) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ be bounded below and uniformly continuous in α and x, then $j(\alpha) = \min_{x \in \mathcal{X}} g(x, \alpha)$ is a continuous function.*

PROOF. Assume $j(\alpha)$ is not continuous: that is, for some $\alpha \in \mathcal{Y}$, $\exists \epsilon > 0$ such that $\forall \delta > 0$, $\exists \alpha'$ with $|\alpha' - \alpha| < \delta$ and $|j(\alpha) - j(\alpha')| > \epsilon$.

By the uniformity of g in α across x, we can choose $\delta' > 0$ so that $|g(x, \alpha) - g(x, \alpha')| < \epsilon/3$ for all x when $|\alpha - \alpha'| < \delta'$. By assumption, we can find some such α' so that $|j(\alpha) - j(\alpha')| > \epsilon$. Without loss of generality, let $j(\alpha) < j(\alpha')$.

Now, choose $x \in \mathcal{X}$ so that $g(x, \alpha) < j(\alpha) + \epsilon/3$. Then $g(x, \alpha') < j(\alpha) + 2\epsilon/3 < j(\alpha')$, contradicting $j(\alpha') = \min_{x \in \mathcal{X}} g(x, \alpha)$.

Using these, we can now prove Theorem B.2:

PROOF. Let $\epsilon > 0$, by Lemma B.1 there exists $\delta' > 0$ such that

$$g(x, \alpha) - g(x(\alpha), \alpha) < \delta' \Rightarrow \|x - x(\alpha)\| < \epsilon.$$

By Lemma B.2, $j(\alpha)$ is continuous. Since $g(x, \alpha)$ is uniformly continuous , we can choose δ so that

$$|\alpha - \alpha'| < \delta \to |j(\alpha) - j(\alpha')| < \delta'/3 \text{ and } \forall x, |g(x, \alpha) - g(x, \alpha')| < \delta'/3$$

giving

$$|g(x(\alpha), \alpha) - g(x(\alpha'), \alpha)| < |g(x(\alpha), \alpha) - g(x(\alpha'), \alpha')| + |g(x(\alpha'), \alpha') - g(x(\alpha'), \alpha)|$$
$$= |j(\alpha) - j(\alpha')| + |g(x(\alpha'), \alpha') - g(x(\alpha'), \alpha)|$$
$$< \delta/3 + \delta/3$$
$$< \delta$$

from which we conclude $\|x(\alpha) - x(\alpha')\| < \epsilon$.

B.0.2. The inner optimization
Proof of Theorem 2.2:

PROOF. We first note that we can re-express $\mathbf{x}_k$ as

$$\mathbf{x}_k = \underset{\mathbf{x} \in (W^1)^n}{\operatorname{argmin}} (1 - \alpha_k)l(\mathbf{x}) + \alpha_k \text{PEN}(\mathbf{x}_k|\boldsymbol{\theta}) \tag{43}$$

where $\alpha_k = \lambda_k/(1 + \lambda_k) \to 1$.

By the continuity of point-wise evaluation in $(W^1)^n$, $l(\mathbf{x})$ is a continuous functional of $\mathbf{x}$ and $\text{PEN}(\mathbf{x}|\boldsymbol{\theta})$ is similarly continuous. Since the x_k lie in a bounded set $\mathcal{X}$, we have that

$$l(\mathbf{x}) < F^* \text{ and } \text{PEN}(\mathbf{x}|\boldsymbol{\theta}) < P^*$$

for all $\mathbf{x} \in \mathcal{X}$. Both $l(\mathbf{x})$ and $\text{PEN}(\mathbf{x}|\boldsymbol{\theta})$ are bounded below by 0 and we note that

$$g(\mathbf{x}, \alpha) = (1 - \alpha)l(\mathbf{x}) + \alpha \text{PEN}(\mathbf{x}|\boldsymbol{\theta})$$

is uniformly bounded on $\mathcal{C}$ by 0 and $F^* + P^*$ and is therefore uniformly continuous in α and $\mathbf{x}$.

By Theorem B.2,

$$\mathbf{x}(\alpha) = \underset{\mathbf{x} \in \mathcal{C}}{\operatorname{argmin}} g(\mathbf{x}, \alpha)$$

is a continuous function from $(0, 1)$ to $(W^1)^n$. Since $\|x(\alpha)\|$ is bounded by assumption, it is uniformly continuous. Since $\alpha_n \to 1$ is convergent, we must have that $\mathbf{x}_n = \mathbf{x}(\alpha_n) \to \mathbf{x}*$. By the continuity of $\text{PEN}(\mathbf{x}|\boldsymbol{\theta})$, $\text{PEN}(\mathbf{x}^*|\boldsymbol{\theta}) = 0$.

Note that if it were possible to define $\mathbf{x}(\alpha)$ as a continuous function on $[0, 1]$, the need for a bound on $\|\mathbf{x}(\alpha)\|$ would be removed. However, since we do not expect $g(\mathbf{x}, 1)\text{PEN}(\mathbf{x}|\boldsymbol{\theta})$ to have a well-defined minimum, boundedness is required to ensure that $\mathbf{x}(\alpha)$ has a limit as $\alpha \to 1$.

We can now go further when $\text{PEN}(\mathbf{x}|\boldsymbol{\theta})$ is given by (13), by specifying that $\mathbf{x}^*$ is the solution of the differential equations (1) that is obtained by minimizing squared error over the choice of initial conditions. To see this, we observe that Theorem 2.1 ensures that

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}, \mathbf{u}, t|\boldsymbol{\theta}).$$

with

$$\mathbf{x}(t_0) = \mathbf{x}_0$$

specifies a unique element of $(W^1)^n$. Let

$$\mathcal{F} = \{\mathbf{x}, \mathtt{PEN}(\mathbf{x}|\boldsymbol{\theta}) = 0\},$$

then

$$\lim_{k \to \infty} l(\mathbf{x}_n) \leq \min_{\mathbf{X} \in \mathcal{F}} l(\mathbf{x}).$$

Since l is a continuous functional on $(W^1)^n$, and $\mathtt{PEN}(\mathbf{x}^*|\boldsymbol{\theta})0$, we must have

$$l(\mathbf{x}^*) = \min_{\mathbf{X} \in \mathcal{F}} l(\mathbf{x}).$$

By the assumption that the solutions to (43) are well defined and bounded, this specifies a unique set of initial conditions $\mathbf{x}_0^*$ such that

$$\dot{\mathbf{x}}^*(t) = \mathbf{f}(\mathbf{x}^*, \mathbf{u}, t|\boldsymbol{\theta}).$$

with

$$\mathbf{x}^*(t_0) = \mathbf{x}_0^*.$$

B.0.3. The outer optimization
Proof of Theorem 2.3:

PROOF. The proof is very similar to that of Theorem 2.2. Setting $\alpha = \lambda/(1+\lambda)$

$$g(\mathbf{x}, \alpha, \boldsymbol{\theta}) = (1-\alpha)l(\mathbf{x}) + \alpha\mathtt{PEN}(\mathbf{x}|\boldsymbol{\theta})$$

is uniformly continuous in α, $\boldsymbol{\theta}$ and $\mathbf{x}$. As observed in Theorem 2.2, $\mathbf{x}_{\boldsymbol{\theta},\lambda}$ can be equivalently written as

$$\mathbf{x}_{\boldsymbol{\theta},\alpha} = \operatorname*{argmin}_{\mathbf{X} \in (W^1)^k} g(\mathbf{x}, \alpha, \boldsymbol{\theta}).$$

with $\alpha\lambda/(1+\lambda)$. By Theorem B.2, $\mathbf{x}_{\boldsymbol{\theta},\alpha}$ is continuous in $\boldsymbol{\theta}$ and α. On the set $\mathcal{X}$, therefore, $l(\mathbf{x})$ is uniformly continuous in $\mathbf{x}$ and $\mathbf{x}_{\boldsymbol{\theta},\alpha}$ is uniformly continuous in $\boldsymbol{\theta}$ and α. $l(\mathbf{x}_{\boldsymbol{\theta},\alpha})$ is therefore uniformly continuous in $\boldsymbol{\theta}$ and α. Under the assumption that $\boldsymbol{\theta}(\alpha)$ is well defined for each α, we can now employ Theorem B.2 again to give us that $\theta(\alpha)$ is continuous in α and the boundedness of Θ provides uniform continuity.
Assume that

$$\tilde{\boldsymbol{\theta}} = \lim_{\alpha \to 1} \boldsymbol{\theta}(\alpha) \neq \boldsymbol{\theta}^*$$

and in particular $\|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\| > \epsilon$. From Lemma B.1 there must exist a $\delta > 0$ such that

$$l(\mathbf{x}_{\boldsymbol{\theta}^*}^*) < l(\mathbf{x}_{\boldsymbol{\theta}}^*) - \delta.$$

for all $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| > \epsilon/2$. Since $\boldsymbol{\theta}(\alpha)$ is uniformly continuous in α, there is some a such that $\|\boldsymbol{\theta}(\alpha) - \boldsymbol{\theta}^*\| > \epsilon/2$ for all $\alpha > a$. Now by the uniform continuity of $l(\mathbf{x}_{\boldsymbol{\theta},\alpha})$ in α and $\boldsymbol{\theta}$, we can choose $a_1 > a$ so that

$$\left| l(\mathbf{x}_{\boldsymbol{\theta}(\alpha),\alpha}) - l(\mathbf{x}_{\boldsymbol{\theta}}^*) \right| < \delta/3$$

for all $\alpha > a_1$. By the same uniform continuity, we can choose $\alpha > a_1$ so that

$$|l(\mathbf{x}_{\boldsymbol{\theta}^*,\alpha}) - l(\mathbf{x}_{\boldsymbol{\theta}^*}^*)| < \delta/2$$

giving

$$l(\mathbf{x}_{\boldsymbol{\theta}^*,\alpha}) < l(\mathbf{x}_{\boldsymbol{\theta}(\alpha),\alpha})$$

contradicting the definition of $\boldsymbol{\theta}(\alpha)$. Finally, note that α is also uniformly continuous in λ and $\lim_{\lambda\to\infty} \alpha(\lambda) = 1$.