Estimating Differential Equations

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

1 Introduction

Differential equations model change in a process by linking the behavior of a derivative to the behavior of the process itself and, possibly, to one or more exogenous inputs. Perhaps the grande dame of such dynamic models is $F = Ma$ connecting the rate of change of velocity a of a body in motion to an exogenous force F and mass M. In fact, functional covariates like F are often called *forcing functions*, perhaps out of deference to this equation and Newton.

Current methods for estimating differential equations (DIFE's) from noisy data are slow, uncertain to provide the best results, and do not lend themselves well to statistical techniques such as interval estimation and inference. As a consequence, one sees little evidence of the impact of statistics in fields routinely using DIFE models. This paper describes a 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.

1.1 Some notation and background

Let $\mathbf{x}$ be a function varying over time t, that is possibly vector-valued, and that has first derivative values $D\mathbf{x}(t)$. Let $\mathbf{u}$ be a vector containing one or more forcing functions and let $\boldsymbol{\theta}$ be a vector of parameters defining a differential equation. Then a general formulation for a differential equation is

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

In most cases the mapping $\mathbf{f}$ will only depend on t through $\mathbf{x}(t)$ and $\mathbf{u}(t)$, in which case the system is called *autonomous*. Systems involving higher order derivatives $D^m\mathbf{x}$ are reducible to this first order form by defining new variables, $\mathbf{x}_1 = \mathbf{x}, \mathbf{x}_2 = D\mathbf{x}_1$ and so on up to $\mathbf{x}_{m-1} = D^{m-1}\mathbf{x}$.

Let n denote the total number of equations in the system after expanding the system in this way. In many applications, n may be impressive; fifty is not unusual in modelling polymer production, for example. Of these variables, moreover, only a subset, and sometimes a small subset, may be measured or observed. Of those for which data are available, sampling rates and observational error variances can be diverse.

Most differential equation systems that are used in practice in fields such as biology, chemical engineering, pharmacokinetics, physics and physiology are not solvable analytically. The main exceptions are linear systems with constant coefficients, where the machinery of the Laplace transform and transform functions plays a role, and a statistical reference to this special case is Bates and Watts (1988). Discrete versions of such systems, that is systems of difference equations for equally spaced time points, are well treated in the classical time series ARIMA and Kalman filter literature, and will not be considered further here.

The numerical methods most often used to approximate solutions of DIFE's over a range $[t_0, t_1]$ use fixed initial values $\mathbf{x}_0 = \mathbf{x}(t_0)$ and adaptive discretization techniques. 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 $\partial f/\partial x$.

1.2 Current estimation strategies

Ignoring the simple linear constant coefficient DIFE, the standard approach for fitting general differential equations to data, often referred to by textbooks as nonlinear least squares or NLS method, goes as follows. A numerical method is used to approximate the solution given a trial set of parameter values and initial conditions, a procedure referred to in engineering circles as *simulation*. The fit, usually defined as the sum of squared differences between data and solution, is then input into an optimization algorithm to update parameter estimates and the initial conditions. That is, the initial state of the system must often be estimated along with the parameters defining the system. This approach is built into the widely used Simulink system in MATLAB, for example. 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, employed with gradient descent in the survey paper, Biegler et al. (1986), and with a Nelder-Mead simplex algorithm in Fussmann et al. (2000). Bock (1983) proposes a multiple shooting method tied to Gauss-Newton minimization. A similar approach is followed in Li et al. (2005). This has been extended to systems of partial differential equations in Müller and Timmer (2004).

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 size of the parameter set is increased by the set of initial conditions needed to solve the system, the inaccuracy of the numerical approximation is added to that of the data so as to further degrade parameter estimates. This approach also only produces point estimates of parameters, and where interval estimation is needed, a great deal more computation is required. As a consequence of all this, Marlin (2000) warns process control engineers to expect an error level of the order of 25% in parameter estimates.

A particular concern with such optimization procedures is the presence of

local minima in the fit surface. The existence of many such local minima has been commented on in Esposito and Floudas (2000) and a number of computationally demanding algorithms proposed to overcome this problem. A common approach in practise has been the use of simulated annealing methods to find global minima. This has been particularly computationally intensive; Jaeger et al. (2004) reported using weeks of computation to reach a point estimate.

An alternative, indirect, approach is available when all components $\mathbf{x}$ of the system are measured. This is to approximate $D\mathbf{x}$ using the data $\mathbf{y}$ and then choose $\boldsymbol{\theta}$ to minimize the discrepancy between $\hat{D}\mathbf{x}$ and its predicted value $\mathbf{f}(\mathbf{x}, \mathbf{u}, t|\boldsymbol{\theta})$. This has been suggested in Voss et al. (1998) using finite difference methods to approximate $\hat{D}\mathbf{x}$ and in Varah (1982) using smoothing spline methods. This approach, while computationally attractive, suffers from the drawback that approximations to $D\mathbf{x}$ are frequently both noisy and biassed. This can lead to biassed parameter estimates and the method still does not provide interval estimates.

A Bayesian approach which avoids the problems of local minima is suggested in chapter 20 of Gelman et al. (2004). The authors set up a model where observations y_t at times t conditional on the parameters and the initial conditions $\boldsymbol{\theta}$ are modelled with a density centered on the numerical solution to the differential equation, $g(\boldsymbol{\theta}, t)$. For example $y_t|\boldsymbol{\theta} \sim N(g(\boldsymbol{\theta}, t), \sigma^2)$. Since $g(\boldsymbol{\theta}, t)$ 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 and $g(\boldsymbol{\theta}, t)$ is numerically calculated conditional on the latest parameter estimates. Our method is similar to a Bayesian method in that we usually have prior information about the form of the DIFE and we wish to combine that information with the data to produce parameter estimates. However, we differ in that we use profiling instead of conditioning and avoid the need for a numerical solution to the DIFE.

Our approach may be thought of as a mid-way point between the approaches of simulation and indirect estimation. We will produce a spline fit to the data. However, we will use the differential equation as a smoothing operator. This will ensure better estimates of $D\mathbf{x}$, provide a computationally tractable procedure and allow us to model situations in which some components are not measured.

1.3 An overview of the paper

In the next Section, we extend this brief introduction by describing two differential equation systems that are of interest in chemical engineering and neuroscience. These systems were chosen with a view to pointing out a range of practical issues and also because fitting them to data has posed a range of important computational and statistical challenges.

Our approach to fitting differential equation models is developed in Section 3, where we develop the concepts of estimating functions and a generalization of profiled estimation. Section 4 follows up with some results on limiting behavior of estimates as the smoothing parameters increase, and discusses some heuristics.

Sections 5 and 6 show how the method performs in practice. Section 5 uses simulated data for the two test bed problems in Section 2, and Section 6 estimates differential equation models for data drawn from chemical engineering and medicine. Generalizations of the method are discussed in Section 7 and some open problems in fitting differential equations are given in Section 8.

2 Two test bed problems

2.1 The 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. The simplified equations reduce a description of four ion channels to a two-component system describing the reciprocal dependencies of the voltage V across an axon membrane and a recovery variable R reflecting outward currents, and the impact of a time-varying external input g Wilson (1999).

The equations are

$$\dot{V} = c\left(V - \frac{V^3}{3} + R\right) + g(t)$$
$$\dot{R} = -\frac{1}{c}\left(V - a + bR\right)$$

where $\boldsymbol{\theta} = \{a, b, c\}$ are unspecified parameters. Although not intended to provide a perfect fit to observable data, solutions to the FitzHugh-Nagumo exhibit behavior that is typical of many dynamical systems. The R equation is the simple constant coefficient system $\dot{R} = -(b/c)R$ linearly forced by V and a constant. However, V 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 $g(t) = 0$, exhibit periodic behavior that combines relatively linear motion with sharp changes in direction. Figure 1 gives a sample path from these equations. The existence of limit cycle, and how it varies as a function of parameters a, b and c, affords a useful study of the properties of these systems.

The FitzHugh-Nagumo system also strikingly demonstrates the difficulty of minimizing over a response surface defined by a differential equation. An example surface obtained by varying only the parameters a and b of the FitzHugh-Nagumo equations is provided in Figure 2. 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.

2.2 The tank reactor equations

A continuously stirred tank reactor, or a *CSTR*, consists of a tank surrounded by cooling jacket and an impeller which stirs the contents. A fluid is pumped

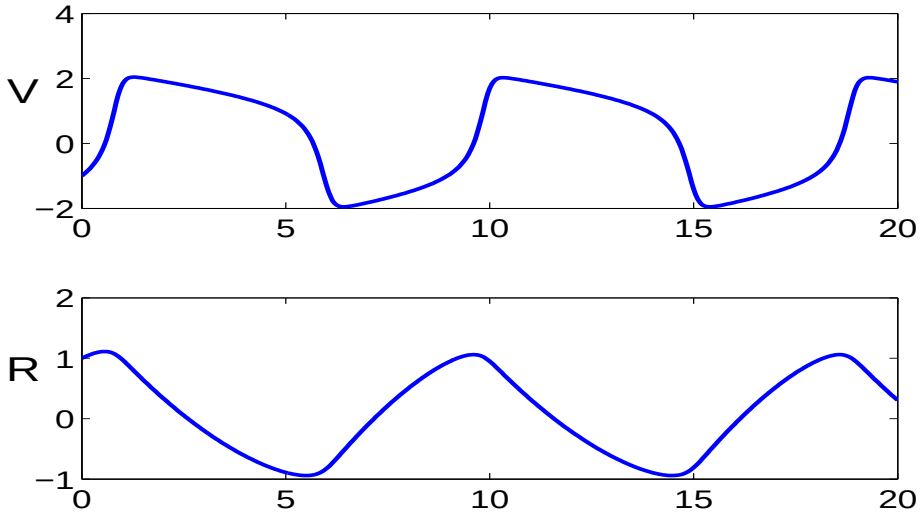

Figure 1: Sample paths from the unforced FitzHugh-Nagumo equations on the interval [0 20]. V represents voltage across an axon membrane and R gives outward currents in ion channels.

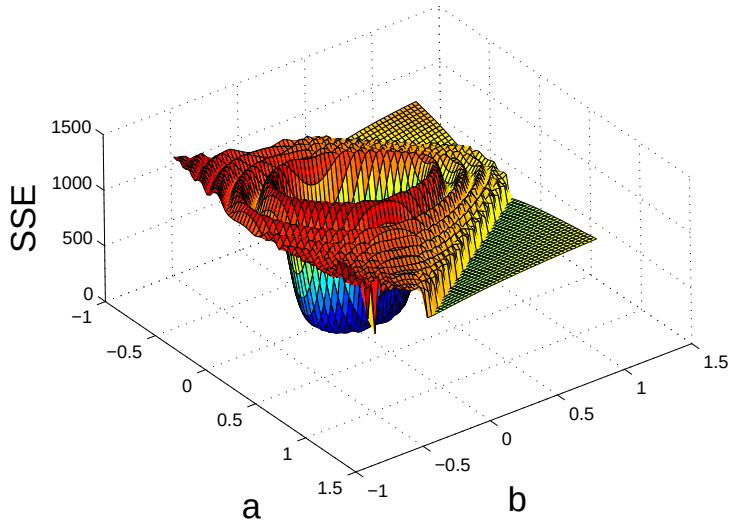

Figure 2: The squared difference between solutions of the FitzHugh-Nagumo equations as a and b are varied around "true" values of 0.2 and 0.2.

into the tank containing a reagent with concentration C_{in} at a flow rate F_{in} and temperature T_{in}. Inside the tank a reaction takes place, producing 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

$$\begin{aligned} DC_{out} &= -\beta_{CC}(T_{out})C_{out} + F_{in}C_{in} \\ DT_{out} &= -\beta_{TT}(F_{cool}, F_{in})T_{out} + \beta_{TC}(T_{out})C_{out} \\ &\quad + F_{in}T_{in} + \alpha(F_{cool})T_{cool}. \end{aligned} \tag{2}$$

The input variables play two roles in the right sides of these equations: Through added terms such as $F_{in}C_{in}$ in the concentration equation, 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 in the follow way:

$$\begin{aligned} \beta_{CC}(T_{out}, F_{in}) &= \kappa \exp[-10^4 \tau(1/T_{out} - 1/T_{ref})] + F_{in} \\ \beta_{TT}(F_{cool}, F_{in}) &= \alpha(F_{cool}) + F_{in} \\ \beta_{TC}(T_{out}) &= 130\beta_{CC}(T_{out}, F_{in}) \\ \alpha(F_{cool}) &= aF_{cool}^{b+1}/(F_{cool} + aF_{cool}^{b}/2), \end{aligned} \tag{3}$$

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 β_{CC} and (a, b) defining 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 certain values of the four parameters, 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, which would be 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

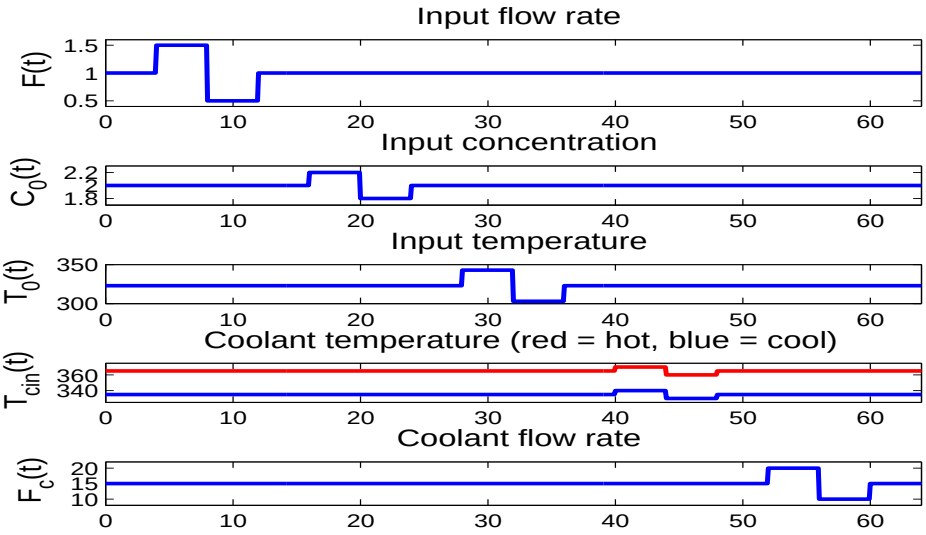

Figure 3: The five inputs to the chemical reactor modelled by the two equations (2): flow rate $F(t)$, input concentration $C_0(t)$, input temperature $T_0(t)$, coolant temperature $T_{cin}(t)$ and coolant flow $F_0(t)$.

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 engineer generally does not know the four reaction kinetic parameters, and therefore must estimate them from noisy data in order to estimate the cooling temperature range to avoid. Figure 5 indicates the nature of the data that may be available. These 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 0.2 times the standard deviations of each of the variable values across a fine mesh of time values. That is, the observational error was roughly 20% of the variable values, an error level that is considered typical for this process.

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. In any case, the actual behavior of the data may not match the behavior of an exact solution of the differential equations for even the best-fitting parameter values, and it will be of interest to explore the lack of fit in various ways with the possible aim of altering the equations themselves.

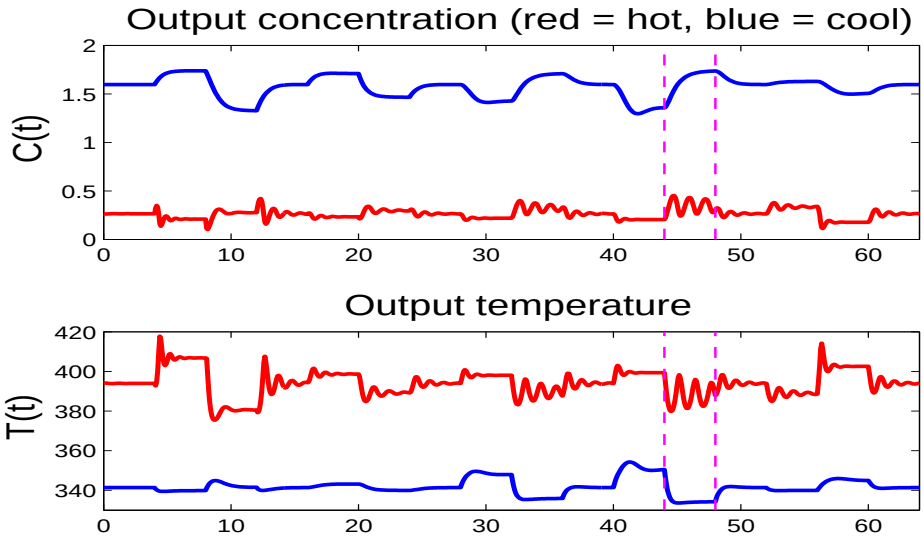

Figure 4: The two outputs from the chemical reactor modelled by the two equations (2): concentration $C(t)$ and temperature $T(t)$. Times at which an input variable is changed are shown as vertical dotted lines.

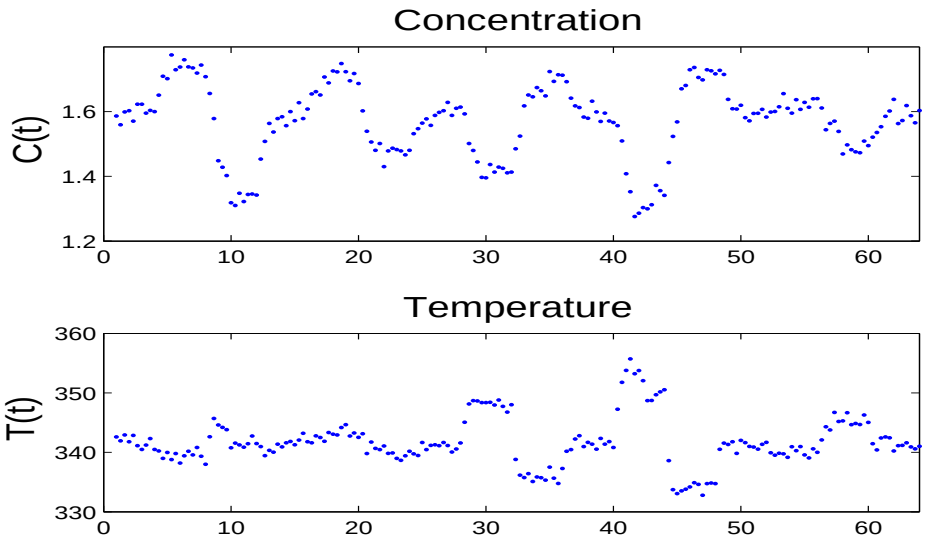

Figure 5: Simulated data for the two outputs from the chemical reactor modelled by the two equations (2): concentration $C(t)$ and temperature $T(t)$.

3 The estimation procedure

An overview is as follows. For each solution x_i in $\mathbf{x}$, we define a basis function expansion $\mathbf{c}_i'\boldsymbol{\phi}_i$, where $\mathbf{c}_i$ and $\boldsymbol{\phi}_i$ are a coefficient vector and a vector of basis functions, respectively. A data-fitting criterion $F(\mathbf{y}|\mathbf{x})$ is chosen that measures the fidelity of $\mathbf{x}$ to the data in vector $\mathbf{y}$. The extent to which $\mathbf{x}$ is a solution of the differential equation system is assessed by the use of additional penalty terms, and the relative balance between these two desiderata is controlled by a set of smoothing parameters.

This approach 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 $\mathbf{c}_i$ defining each basis function expansion. The equation parameters are structural in the sense of being of primary interest, but the coefficients $\mathbf{c}_i$ are considered as nuisance parameters that are essential for fitting the data because they are not of direct concern and because their numbers are apt to vary with the length of the observation interval, density of observation, and other factors. As a rule, the number of nuisance parameters can be orders of magnitude larger than the number of equation parameters, with a ratio of about 100 applying in the CSTR problem.

Nuisance parameters are removed from the problem by defining them as functions of the structural parameters using a modified profiling procedure, and the composite fitting criterion is then optimized with respect to the the structural parameters alone. An analytic expression for the gradient is developed using the Implicit Function Theorem. Each of these steps will now be described in more detail.

3.1 Basis function expansions for the solution functions

We assume that each output function $x_i, i = 1, \ldots, n$ is represented as a basis function expansion

$$x_i(t) = \mathbf{c}_i'\boldsymbol{\phi}_i(t) = \boldsymbol{\phi}_i(t)'\mathbf{c}_i. \tag{4}$$

The vector of basis functions $\boldsymbol{\phi}_i$ is of length K_i, as is the corresponding coefficient vector $\mathbf{c}_i$.

The number of basis functions as well as other aspects of the system must permit the model to accommodate any important variation in an actual solution. Moreover, since the equations involve one or more derivatives, the expansion must also have the capacity to be faithful to the behavior of the highest order derivative in the system. Although solutions to linear constant coefficient differential equations systems are sums of exponential functions and therefore smooth, nonlinear systems may give rise to solutions that have extremely sharp features, such as peaks, valleys, high frequency oscillations and near discontinuities in derivatives. Moreover, even linear systems are often forced in industrial settings by step changes in control systems, and these lead to sharp changes for first and higher derivatives. Splines are usually the basis of choice, but it may be advantageous to use multiple knots at certain critical locations, such as step

changes in inputs. A successful allocation of knots may require a preliminary exploration of the system initial estimates of parameters, possibly accompanied by updates of the knot sequence as the estimation process proceeds. If equally spaced knots are used, it may useful to begin with a very large number of these, followed by reducing knot density where appropriate.

We assume that the forcing functions in vector $\mathbf{u}$ are known exactly, although in practice these also may arise from smoothing noisy data, and the method that we use can easily be extended to accommodate this case. The step-wise definition of these functions that we see in the CSTR equations (2) is not at all unusual in actual experiments, where the dynamics of the response to a sudden change in input due, perhaps, to the failure of some component, can be of direct concern. Such discontinuities in input can imply corresponding discontinuities in a derivative of an output. For this reason, spline bases may be the best choice since multiple knots assigned to the time of a input step change can accommodate this feature nicely. For the data shown in Figure 5, where each variable was observed every 20 seconds, the resulting B-spline basis systems each had $K_i = 193$ basis functions.

If we need to refer to all n output functions simultaneously, we will use the notation $\mathbf{x}$ to refer to the vector of n output functions. Let $\mathbf{c}$ indicate the composite vector $(\mathbf{c}_1', \ldots, \mathbf{c}_n)'$ of length $\sum_i K_i$, and let $\mathbf{\Phi}$ be the $\sum_i N_i$ by $\sum_i K_i$ super–matrix constructed by placing the matrices $\mathbf{\Phi}_i$ along the diagonals and using zeros elsewhere. According to this notation, we have the composite basis expansion $\mathbf{x} = \mathbf{\Phi c}$.

3.2 The data fitting criterion

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 so that the normalized error sums of squares are of roughly comparable sizes. Suitable weights may be the reciprocals of initial values $w_i = x_i(0)$ or of the variance taken over values $x_i(t_{ij})$ for some trial or initial estimate of a solution of the equation.

Let $\mathbf{y}_i$ indicate the data available for variable i consisting of observations at time points $\mathbf{t}_i$, and let $\mathbf{y}$ indicate the total data available. The notation $x_i(\mathbf{t}_i)$ is used for the vector of fitted values corresponding to $\mathbf{y}_i$. The composite fit measure using error sum of squares is

$$\texttt{SSE}(\mathbf{y}|\mathbf{x}) = \sum_{i}^{n} w_i \|\mathbf{y}_i - x_i(\mathbf{t}_i)\|^2. \tag{5}$$

The summation over variables is only over the variables for which observations are available; it will be routine that only certain variables in the system are actually measured. If the weighted least squares criterion (5) is used, then the

normalization can be incorporated into the design of the weight matrix $\mathbf{W}_i$ in

$$\texttt{SSE}(\mathbf{y}|x) = \sum_i^n [\mathbf{y}_i - x_i(\mathbf{t}_i)]' \mathbf{W}_i [\mathbf{y}_i - x_i(\mathbf{t}_i)]. \tag{6}$$

Finally, many applications will require loss functions other than error sum of squares, and our approach can accommodate these situations without difficulty.

3.3 Assessing fidelity to the equations

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

$$L_i(x_i) = Dx_i - f_i(\mathbf{x}, \mathbf{u}, t|\boldsymbol{\theta}) = 0. \tag{7}$$

The extent to which an actual function x_i satisfies the DIFE system can then be assessed by

$$\texttt{PEN}_i(\mathbf{x}) = w_i \int [L_i(x_i)]^2 dt \tag{8}$$

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

$$\texttt{PEN}_i(\mathbf{x}) = w_i \int |L_i(x_i)| dt \tag{9}$$

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

$$\texttt{PEN}(\mathbf{x}|\mathbf{L}, \boldsymbol{\lambda}) = \sum_i^n \lambda_i \texttt{PEN}_i(\mathbf{x}) \tag{10}$$

where $\mathbf{L}$ is denotes the vector containing the n differential operators L_i. Note that in this case the summation will be over all n 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.

Finally, the data-fitting and equation-fidelity criteria are combined into the penalized least squares criterion

$$\texttt{PENSSE}(\mathbf{y}|\mathbf{x}, \mathbf{L}, \boldsymbol{\lambda}) = \texttt{SSE}(\mathbf{y}|\mathbf{x}) + \texttt{PEN}(\mathbf{x}|\boldsymbol{\lambda}) \tag{11}$$

Although this formulation may resemble the data smoothing methods based on roughness penalties or regularization such as those described in Ramsay and Silverman (2005), in fact the perspective is now much more symmetric in data fitting versus equation solving. In a sense, we may say that the data fitting criterion $\texttt{SSE}(\mathbf{y}|\mathbf{x})$ *data-regularizes* the differential solution that would be defined if we only paid attention to reducing each $\texttt{PEN}_i(\mathbf{x})$ to zero.

3.4 Estimating $\mathbf{c}(\boldsymbol{\theta})$

The parameters in $\mathbf{c}$ defining the fitting functions $x_i, i = 1, \ldots, n$ or $\mathbf{x}(\mathbf{c})$, may be regarded as *nuisance parameters*, meaning that they are essential to fit the data, but are not of direct interest. By contrast, the *structural parameters* $\boldsymbol{\theta}$ will usually be of more central interest, although, of course, confidence regions and other inferential issues concerning $\mathbf{x}$ may also prove useful. Vector $\mathbf{c}$, moreover, will tend to be far larger than $\boldsymbol{\theta}$, since DIFE solutions often exhibit sharp local behavior, as we saw with the CSTR equations, that can require large numbers of basis functions to capture.

Consequently, joint estimation of both $\boldsymbol{\theta}$ and $\mathbf{c}$ is likely to be unwise since the parameter space is apt to be of exceedingly high dimensionality, and the $\boldsymbol{\theta}$ estimates are likely to be unstable due to the limited number of degrees of freedom for error that will be left in the data. The usual marginalization strategy involving numerical integration over $\mathbf{c}$ with respect to some prior measure is also problematic since computational overhead involved is apt to be unacceptable.

Instead, we adopted a *generalized profiling* strategy, involving minimizing the penalized error sum of squares (11) each time any element of $\boldsymbol{\theta}$ is changed. That is, our approach is to embed an *inner optimization* where $\mathbf{c}$ alone is updated within an *outer optimization* loop optimizing $\boldsymbol{\theta}$. To keep the notation compact and to emphasize that $\mathbf{c}$ is optimized conditional on a value for $\boldsymbol{\theta}$, we now use the notation $H(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})$ for (11), and also $h = DH$, the gradient of H taken with respect to $\mathbf{c}$.

In effect, then, the inner optimization defines an *estimating function* $\mathbf{c}(\theta)$, if we can assume that the conditional minimum of $H(\mathbf{c}|\boldsymbol{\theta}, \mathbf{y})$ is unique for a neighborhood of the optimal $\boldsymbol{\theta}$ containing the values of $\boldsymbol{\theta}$ that will be encountered during the outer optimization. In this way, the dimension of the parameter space is now reduced to that of $\boldsymbol{\theta}$. For example, in our analysis of the CSTR equations, this amounts to reducing the dimensionality from $4 + 193 \times 2 = 390$ to 4.

3.5 Outer optimization for $\boldsymbol{\theta}$

The outer optimization with respect to $\boldsymbol{\theta}$ minimizes the unpenalized composite error sum of squares (5), that is, the data-fitting criterion. There is no need to append a roughness criterion taking into account fidelity to the equations since this is already taken care of in the definition of $\mathbf{c}(\theta)$. We use the notation $F(\theta, \mathbf{c}(\theta)|\mathbf{y})$ to refer to $\mathtt{SSE}[\mathbf{y}|\mathbf{x} = \mathbf{c}'(\theta)\boldsymbol{\Phi}]$, the notation $f = DF$ for its gradient and $\hat{\boldsymbol{\theta}} = argmin\{F(\theta, \mathbf{c}(\boldsymbol{\theta})|\mathbf{y})\}$. It is assumed that F is twice continuously differentiable with respect to both $\boldsymbol{\theta}$ and $\mathbf{c}$, and that the second partial derivative or Hessian matrices

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

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

The gradient $f(\boldsymbol{\theta})$ is

$$f(\boldsymbol{\theta}) = \frac{\partial F}{\partial \boldsymbol{\theta}} + \frac{\partial F}{\partial \mathbf{c}} \frac{d\mathbf{c}}{d\boldsymbol{\theta}} \ . \tag{12}$$

Since $\mathbf{c}(\boldsymbol{\theta})$ is not available explicitly, we need the *implicit function theorem* to define $d\mathbf{c}/d\boldsymbol{\theta}$. To keep the notation compact, we will no longer include $\mathbf{y}$ explicitly as an argument for the functions F or H. The optimal value of $\mathbf{c}$ will satisfy $h(\mathbf{c}|\boldsymbol{\theta}) = 0$, and we have that

$$D_{\boldsymbol{\theta}} h = \frac{\partial h}{\partial \boldsymbol{\theta}} + \frac{\partial h}{\partial \mathbf{c}} \frac{d\mathbf{c}}{d\boldsymbol{\theta}} = 0.$$

Therefore

$$\frac{d\mathbf{c}}{d\boldsymbol{\theta}} = -\Big(\frac{\partial h}{\partial \mathbf{c}}\Big)^{-1} \frac{\partial h}{\partial \boldsymbol{\theta}} = -\Big(\frac{\partial^2 H}{\partial \mathbf{c}^2}\Big)^{-1} \frac{\partial^2 H}{\partial \mathbf{c} \partial \boldsymbol{\theta}}. \tag{13}$$

and

$$f(\boldsymbol{\theta}) = \frac{\partial F}{\partial \boldsymbol{\theta}} - \frac{\partial F}{\partial \mathbf{c}} \Big(\frac{\partial^2 H}{\partial \mathbf{c}^2}\Big)^{-1} \frac{\partial^2 H}{\partial \mathbf{c} \partial \boldsymbol{\theta}} \ . \tag{14}$$

The matrices used in these equations and those below have often 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.

3.6 Approximating the variation in $\hat{\theta}$ and $\hat{\mathbf{c}}$

Let $\boldsymbol{\Sigma}$ be the variance–covariance matrix for $\mathbf{y}$, The estimate $\hat{\boldsymbol{\theta}}(\mathbf{y})$ of $\boldsymbol{\theta}$ is the solution of the stationary equation $f(\hat{\boldsymbol{\theta}}, \hat{\mathbf{c}}|\mathbf{y}) = 0$ where $\hat{\mathbf{c}} = \mathbf{c}(\hat{\boldsymbol{\theta}})$. Here and below, all partial derivatives as well as total derivatives are assumed to be evaluated at $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{c}}$, which are in turn evaluated at $\mathbf{y}$.

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

$$(J^T \boldsymbol{\Sigma} J)^{-1}$$

where J is the Jacobian matrix given by $dx_k(t_i)/d\theta_j$, concatenated row-wise across the components of $\mathbf{x}$. This makes use of the approximation

$$\frac{d^2 F}{d\boldsymbol{\theta}^2} \approx J^T J.$$

Such an approximation may not be warranted when there is large curvature in parameter space. We will instead provide an exact estimation of the Hessian above and employ it with a pseudo δ-method. This is undertaken at the cost of considerably more computation. Our experiments in Section 5.1 suggest that this method provides more accurate results that the usual estimate, but that the latter will still provide a reasonable approximation.

By applying the Implicit Function Theorem to $D_{\boldsymbol{\theta}}F$ as a function of $\mathbf{y}$, we may say that for any $\mathbf{y}$ in $\mathcal{N}$ there exists a value $\boldsymbol{\theta}(\mathbf{y})$ such that $f[\boldsymbol{\theta}(\mathbf{y}), \mathbf{c}(\boldsymbol{\theta}(\mathbf{y}))|\mathbf{y}] = 0$. Consequently

$$D^2_{\boldsymbol{\theta},\mathbf{y}}F = \frac{\partial D_{\boldsymbol{\theta}}F}{\partial \mathbf{y}} + D^2_{\boldsymbol{\theta}}F\frac{d\boldsymbol{\theta}}{d\mathbf{y}} = 0, \tag{15}$$

where

$$D^2_{\boldsymbol{\theta}}F(\hat{\boldsymbol{\theta}}, \hat{\mathbf{c}}|\mathbf{y}) = \frac{\partial^2 F}{\partial \boldsymbol{\theta}^2} + \frac{\partial^2 F}{\partial \mathbf{c}\partial \boldsymbol{\theta}}\frac{\partial \mathbf{c}}{\partial \boldsymbol{\theta}} + \frac{\partial^2 F}{\partial \mathbf{c}^2}\big[\frac{\partial \mathbf{c}}{\partial \boldsymbol{\theta}}\big]^2 + \frac{\partial F}{\partial \mathbf{c}}\frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\theta}^2}$$

and

$$D^2_{\boldsymbol{\theta},\mathbf{y}}F(\hat{\boldsymbol{\theta}}, \hat{\mathbf{c}}|\mathbf{y}) = \frac{\partial^2 F}{\partial \boldsymbol{\theta}\partial \mathbf{y}} + \frac{\partial^2 F}{\partial \mathbf{c}\partial \mathbf{y}}\frac{\partial \mathbf{c}}{\partial \mathbf{y}}\frac{\partial \mathbf{c}}{\partial \boldsymbol{\theta}} + \frac{\partial F}{\partial \mathbf{c}}\frac{\partial^2 \mathbf{c}}{\partial \boldsymbol{\theta}\partial \mathbf{y}}.$$

Now to a first order of approximation over $\mathcal{N}$, we can approximate $\boldsymbol{\theta}(\mathbf{y}^*)$ evaluated at an alternative observation $\mathbf{y}^* \in \mathcal{N}$ by

$$\begin{aligned} \boldsymbol{\theta}(\mathbf{y}^*) - \boldsymbol{\theta}(\mathbf{y}) &\approx \frac{d\boldsymbol{\theta}}{d\mathbf{y}}(\mathbf{y}^* - \mathbf{y}) \\ &= [D^2_{\boldsymbol{\theta}}F(\hat{\boldsymbol{\theta}}, \hat{\mathbf{c}}|\mathbf{y})]^{-1}D^2_{\boldsymbol{\theta},\mathbf{y}}F(\hat{\boldsymbol{\theta}}, \hat{\mathbf{c}}|\mathbf{y})(\mathbf{y}^* - \mathbf{y}). \end{aligned} \tag{16}$$

Consequently, the variance of $\hat{\boldsymbol{\theta}}(\mathbf{y})$ can be estimated by

$$\mathtt{Var}[\hat{\boldsymbol{\theta}}(\mathbf{y})] = \mathbf{Z}(\hat{\boldsymbol{\theta}}, \hat{\mathbf{c}}|\mathbf{y})\boldsymbol{\Sigma}\mathbf{Z}'(\hat{\boldsymbol{\theta}}, \hat{\mathbf{c}}|\mathbf{y}),$$

where

$$\mathbf{Z}(\hat{\boldsymbol{\theta}}, \hat{\mathbf{c}}|\mathbf{y}) = [D^2_{\boldsymbol{\theta}}F(\hat{\boldsymbol{\theta}}, \hat{\mathbf{c}}|\mathbf{y})]^{-1}D^2_{\boldsymbol{\theta},\mathbf{y}}F(\hat{\boldsymbol{\theta}}, \hat{\mathbf{c}}|\mathbf{y}).$$

The sampling variance of $c(\hat{\boldsymbol{\theta}}(\mathbf{y})|\mathbf{y})$ is estimated by

$$\mathtt{Var}[\mathbf{c}(\hat{\boldsymbol{\theta}}(\mathbf{y})|\mathbf{y})] = \big(\frac{\partial \mathbf{c}}{\partial \boldsymbol{\theta}}\big)\mathtt{Var}[\hat{\boldsymbol{\theta}}]\big(\frac{\partial \mathbf{c}}{\partial \boldsymbol{\theta}}\big)' + \big(\frac{\partial \mathbf{c}}{\partial \mathbf{y}}\big)\boldsymbol{\Sigma}\big(\frac{\partial \mathbf{c}}{\partial \mathbf{y}}\big)'.$$

3.7 Numerical integration

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

$$\mathtt{PEN}_i(\mathbf{x}) \approx w_i \sum_q^Q v_q[L_i(x_i(t_q))]^2 \tag{17}$$

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 consider that normally these breakpoints will usually be at times t_{ij} of observation of output variable x_i.

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 the sparse matrix methods in Matlab.

Efficiency in the inner optimization is essential since this will be invoked far more often than the outer optimization. The minimization of penalized least squares criterion (11), can be expressed as a large nonlinear least squares approximation problem by observing that we can express the numerical quadrature approximation to $\sum_i \lambda_i \texttt{PEN}_i(\mathbf{x})$ as

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

These squared residuals can then be appended to those in $\texttt{SSE}(\mathbf{y})$, and Gauss-Newton minimization can then be used. In those situations where the coefficients enter linearly into the expression for the fitting function, the inner optimization can be avoided entirely by using the explicit solution.

4 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 4.1, suggests that when the data agree well with the DIFE model, the λ_i should be chosen as large as possible, bounded only by the possibility of distortion from our choice of basis expansion (4).

In our experience, however, real world systems are rarely perfectly described by DIFEs. In such situations, we may wish to make the λ_i smaller, in order to be able to account for systematic discrepancies between DIFE 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 DIFE 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.

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 4.2. 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.

4.1 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 exactly expressed as a spline basis function expansion, and therefore does not have approximation error. 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 squared error 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. This approach corresponds to a maximum likelihood estimate of $\boldsymbol{\theta}$ assuming Gaussian errors. 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:

$$\texttt{SSE}(\mathbf{x}) = \texttt{SSE}(\mathbf{y}|\mathbf{x}).$$

Secondly, we will represent

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

where the c_i are taken to be constants and the λ_i used in the definition (10) are given by λc_i for some λ.

We will also assume that solutions to the problem (1) exist and are well defined, and therefore that there are objects $\mathbf{x}$ that satisfy $\texttt{PEN}(\mathbf{x}|\boldsymbol{\theta}) = 0$. This is guaranteed locally by the following theorem adapted from Bellman (1953):

Theorem 4.1. *Let* $\mathbf{f}$ *be Lipschitz continuous and* $\mathbf{u}$ *differentiable almost everywhere, then the initial value problem:*

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

has a unique solution.

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

$$\texttt{SSE}(\mathbf{x}) + \lambda \texttt{PEN}(\mathbf{x}|\boldsymbol{\theta}).$$

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.

4.1.1 Preliminaries

The following theorem is a well-known consequence of the method of Lagrange mutipliers:

Theorem 4.2. *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 $\lambda' > \lambda$, $P(x_{\lambda'}) \leq P(x_\lambda)$.*

A two corollaries:

Corollary 4.1. *For $\lambda' > \lambda$, $F(x_{\lambda'}) \geq F(x_\lambda)$.*

Corollary 4.2. *If $\exists x$ such that $P(x) = 0$, then $P(x_\lambda) \to 0$ as $\lambda \to \infty$.*

follow immediately. The following theorem is proved in Appendix B:

Theorem 4.3. *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.

4.1.2 The inner optimization

The first step of the analysis is to show that as $\lambda \to \infty$ the solutions of the inner optimization criterion converge to an exact solution of the differential equation.

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 $[t_1, t_2]$ whose first derivatives are square integrable. This space is equipped with the inner product

$$(f, g) = \int_{t_1}^{t_2} f(t)g(t)dt + \int_{t_1}^{t_2} f'(t)g'(t)dt.$$

In particular, point-wise evaluation is a continuous operation on W^1 and hence on $\mathcal{H}$ (Gu (2002)).

Theorem 4.4. *Let $\lambda_k \to \infty$ and assume that*

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

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

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

$$\mathbf{x}_k = \operatorname*{argmin}_{\mathbf{x} \in (W^1)^n} (1 - \alpha_k)\mathtt{SSE}(\mathbf{x}) + \alpha_k \mathtt{PEN}(\mathbf{x}_k|\boldsymbol{\theta}) \tag{18}$$

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

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

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

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

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

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

By Theorem 4.3,

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

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 $\texttt{PEN}(\mathbf{x}|\boldsymbol{\theta})$, $\texttt{PEN}(\mathbf{x}^*|\boldsymbol{\theta}) = 0$. □

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 $g(\mathbf{x}, 1) = \texttt{PEN}(\mathbf{x}|\boldsymbol{\theta})$ 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 $\texttt{PEN}(\mathbf{x}|\boldsymbol{\theta})$ is given by (10), 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 4.1 ensures that

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

with

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

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

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

then

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

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

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

By the assumption that the solutions to (18) are well defined and bounded, this specifies a unique set of initial conditions $\mathbf{x}_0^*$ such that

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

with

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

4.1.3 The outer optimization

For the problem of estimating $\boldsymbol{\theta}$ we need to again assume that well-defined solutions to

$$\boldsymbol{\theta}(\lambda) = \operatorname*{argmin}_{\boldsymbol{\theta}\in\Theta} \mathtt{SSE}(\mathbf{x}_{\lambda,\boldsymbol{\theta}})$$

exist where $\mathbf{x}_{\lambda,\boldsymbol{\theta}}$ is now also indexed by candidate parameter values $\boldsymbol{\theta}$. In particular, we assume that

$$\boldsymbol{\theta}^* = \operatorname*{argmin}_{\boldsymbol{\theta}} \mathtt{SSE}(\mathbf{x}_{\infty,\boldsymbol{\theta}}),$$

the least-squares estimate of $\boldsymbol{\theta}$ obtained by directly solving the differential equation, is well defined.

In this section we will show that $\lim_{\lambda\to\infty} \boldsymbol{\theta}(\lambda) = \boldsymbol{\theta}^*$ provided we can bound $\boldsymbol{\theta}(\lambda)$ uniformly. As with the bounds required on $\mathbf{x}$, conditions on functions $\mathbf{f}(\mathbf{x}, \mathbf{u}, t|\boldsymbol{\theta})$ that guarantee this regularity are a matter of ongoing investigation.

Theorem 4.5. *Let $\mathcal{X} \subset (W^1)^n$ and $\Theta \subset \mathbb{R}^p$ be bounded. Let*

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

be well defined for each $\boldsymbol{\theta}$ and λ, define $\mathbf{x}^*_{\boldsymbol{\theta}}$ to be such that

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

and let

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

also be well defined for each λ. Then

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

Proof. The proof is very similar to that of Theorem 4.4. Setting $\alpha = \lambda/(1+\lambda)$

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

is uniformly continuous in α, $\boldsymbol{\theta}$ and $\mathbf{x}$. As observed in Theorem 4.4, $\mathbf{x}_{\boldsymbol{\theta},\lambda}$ 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 4.3, $\mathbf{x}_{\boldsymbol{\theta},\alpha}$ is continuous in $\boldsymbol{\theta}$ and α. On the set $\mathcal{X}$, therefore, $\mathtt{SSE}(\mathbf{x})$ is uniformly continuous in $\mathbf{x}$ and $\mathbf{x}_{\boldsymbol{\theta},\alpha}$ is uniformly continuous in $\boldsymbol{\theta}$ and α. $\mathtt{SSE}(\mathbf{x}_{\boldsymbol{\theta},\alpha})$ 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 4.3 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

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

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 $\texttt{SSE}(\mathbf{x}_{\boldsymbol{\theta},\alpha})$ in α and $\boldsymbol{\theta}$, we can choose $a_1 > a$ so that

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

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

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

giving

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

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

4.2 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 6 plots surfaces similar to Figure 2 but providing the squared error of the spline fit as parameters a and b are varied. These are given for three different values of λ and it can be seen that for smaller λ, the surfaces are 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

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

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.

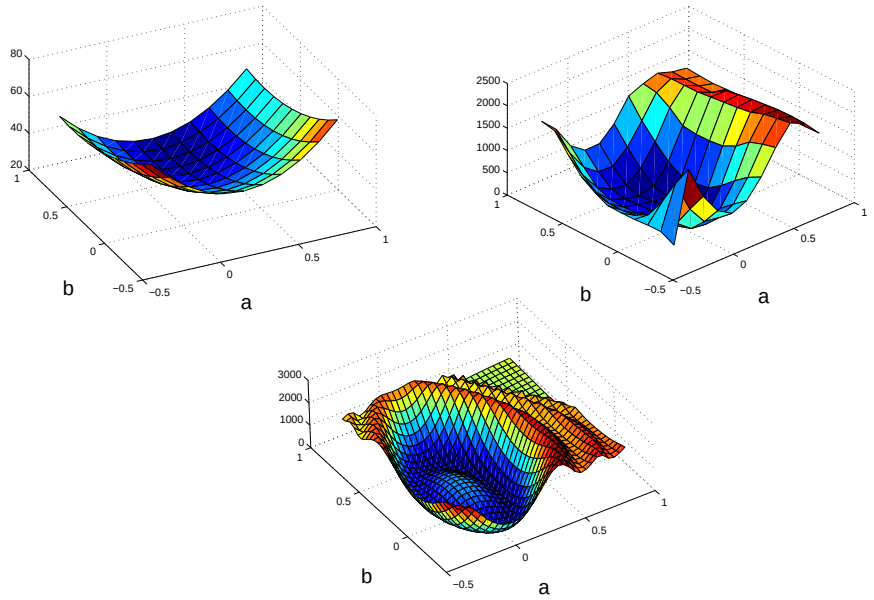

Figure 6: A comparison of the FitzHugh-Nagumo response surfaces over a and b for $\lambda = 10^2, 10^5$ and at exact solutions. Surfaces give the value of the squared difference between exact solutions at $\{a, b, c\} = \{0.2, 0.2, 3\}$ and the spline approximation to those solutions using exact data and perturbed values of a and b.

Our experimental evidence suggests that for small values of λ, parameter estimates tend to be more variable and can become quite biassed. However, Theorem 4.5 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.

5 Simulated data examples

5.1 Fitting the FitzHugh-Nagumo equations

We set up simulated data 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 these data 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 numeric integration. Figure 7 gives quartiles of the parameter estimates for 50 simulations as λ is varied from 10^{-2} to 10^5. It is apparent that there is a

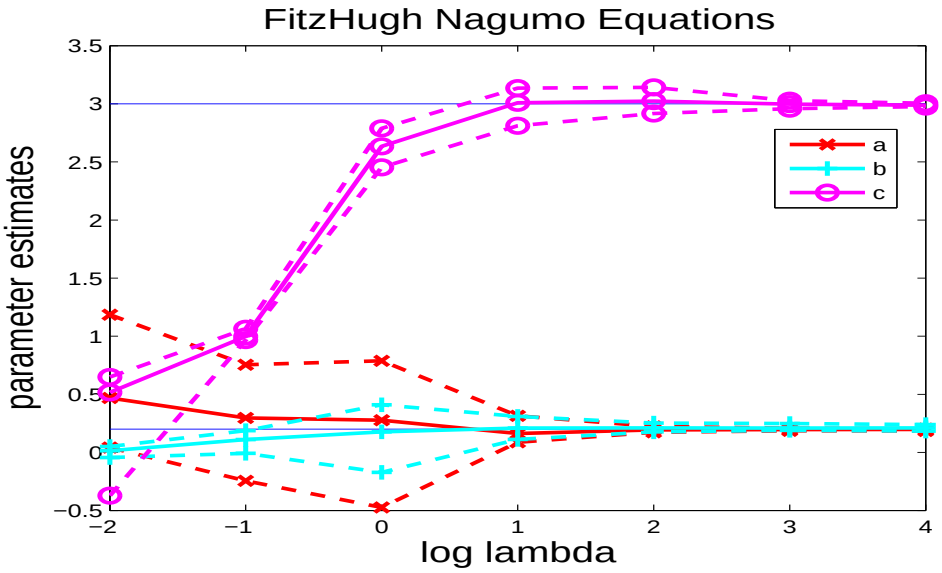

Figure 7: Quartiles of parameter estimates for the FitzHugh-Nagumo Equations as λ is varied. Horizontal lines represent the true parameter values.

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 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 3.6 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. The estimates based on Section 3.6 required 10 times the computer time than the standard estimates. Parameter estimates for a and b are very close to the true values. There appears to be a small amount of bias for the estimate of c, which we conjecture to be due to the use of a basis expansion.

5.2 Fitting the tank reactor equations

We now consider how well the parameters and the equation solutions can be estimated from the simulated data in Figure 5. The smoothing parameters λ_C and λ_T were 100 and 10, respectively.

Table 2 displays bias and sampling precision results for parameter estimates for 1000 simulated samples. The first two lines of the table compare the true parameter values with the mean estimates, and the last two lines compare the

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

Table 2: Summary statistics for parameter estimates for 1000 simulated samples from the same population illustrated in Figure 5. The estimate of the standard deviation of parameter values is by the delta method usual in nonlinear least squares analyses.

	κ	τ	a
True value	0.4610	0.8330	1.6780
Mean value	0.4610	0.8349	1.6745
Std. Dev.	0.0034	0.0057	0.0188
Est. Std. Dev.	0.0035	0.0056	0.0190
Bias	0.0000	0.0000	-0.0001
Std. Err.	0.0002	0.0004	0.0012

biases of the estimates with the standard errors of the mean estimates. We see that the biases can be considered negligible. The third and fourth lines compare the actual standard deviations of the 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 the usual estimate of sampling variance.

The principal components of variation of the correlation matrix for the parameter estimates account 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 rotation of the parameters might be worth considering in order to reduce the dimensionality of the parameter space. 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 estimated parameters, the maxi-

mum 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 concentration and temperature discrepancies became 1.72% and 0.05%, respectively. 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.

6 Working with real data

6.1 Thermal decomposition of α-Pinene

The compound α-pinene is a component of tea-tree oils and is used in pharmaceutical and aroma-chemical products. When heated to between 189.5^o and 285^oC it undergoes a thermal reaction, yielding dipentene and allo-ocimene. Allo-ocimene further decomposes into α-pyronene and β-pyronene and a dimer. Fuguitt and Hawkins (1947) provide the results of an experiment in which pure α-pinene was heated to 189.5^oC in an oil bath and the proportions of α-pinene and the resulting chemicals were measured at eight time intervals. Since mass must be conserved in the experiment, the data are given as *percentages* of total mass. While Fuguitt and Hawkins (1947) reported concentrations for the sum of the pyronenes, these were calculated from the other measured quantities and mass balance considerations. We have therefore taken this to be an unmeasured component.

A number of models and estimation procedures have been attempted for these data. Stewart and Sorensen (1981) proposed a non-linear model, fit using a Bayesian procedure. Bates and Watts (1988) chose a linear model but noted that residual plots indicated systematic deviation from the model. Here we have chosen to combine the two. Let x_i, $i = 1, \cdots, 4$ represent the weight percentage of α-pinene, dipentene, allo-ocimene and the dimer. We employ the following nonlinear differential equations:

$$\frac{dx_1}{dt} = -(\theta_1 + \theta_2)x_1 - 2\theta_3 x_1^2 \tag{19}$$

$$\begin{aligned} \frac{dx_2}{dt} &= \theta_1 x_1 \\ \frac{dx_3}{dt} &= \theta_2 x_1 - \theta_4 x_3 - 2\theta_5 x_3^2 - \theta_7 x_3 + 2\theta_6 x_4 \\ \frac{dx_4}{dt} &= \theta_3 x_1^2 + \theta_5 x_3^2 - \theta_7 x_3 + \theta_6 x_4 \end{aligned} \tag{20}$$

Compounds α- and β- pyronene appear only on the left hand side of their equation and do not influence any of the right hand side functions. Since they

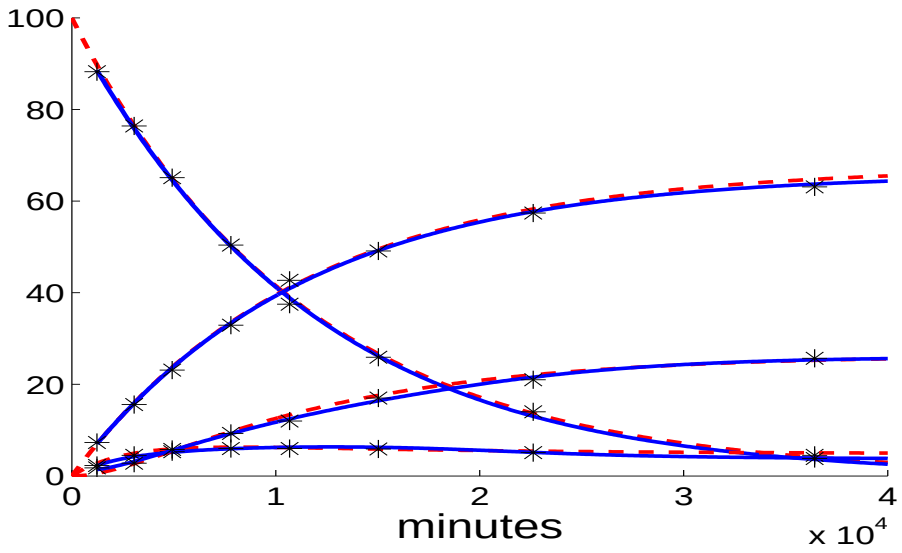

Figure 8: Solutions to the differential equation (19) using initial parameter and initial condition estimates (dashed) and with re-estimated parameters (solid). The data are given by asterisks.

are unmeasured, their inclusion only makes their initial conditions undetermined and does not affect the rest of the estimates.

The equations given in Bates and Watts (1988) are equivalent to setting $\theta_3 = \theta_4 = \theta_5 = 0$. Stewart and Sorensen (1981) corresponds to $\theta_7 = 0$. That paper also includes terms for pyronene in the derivative of x_3. We have found that, without measurements for pyronene, this system was effectively unidentifiable and have therefore dropped the extra terms from our model. An approximate initial condition may be given at time zero as having 100% α-pinene concentration and listing the others as 0.

Using these data, we estimated the seven parameters using the estimates in Bates and Watts (1988) as starting values. We used 160 equally spaced knots in each component and set $\lambda = 10^8$. Solving the differential equation using the theoretical initial conditions, our parameters correspond to a 14% decrease in total squared error. Allowing the initial conditions to vary and using the value of the smoothing spline at the first data point resulted in a 35% decrease in total squared error. Figure 8 compares the exact solution fit to the differential equation provided in Bates and Watts (1988) with the new fit using new parameters and new initial values chosen at the first measurement time.

Table 3 presents our parameter estimates and confidence intervals. Here all confidence intervals apart from that for θ_4 do not contain zero. However, the negative signs for θ_3 and θ_5 suggests, that the system may remain miss-specified. An examination of the residuals plotted in Figure 9 indicates systematic bias

Table 3: Parameter estimates and 95% confidence intervals for the α-pinene data.

	estimate	lower bound	upper bound
θ_1	$5.930 * 10^{-5}$	$5.606 * 10^{-5}$	$6.255 * 10^{-5}$
θ_2	$3.488 * 10^{-5}$	$3.002 * 10^{-5}$	$3.974 * 10^{-5}$
θ_3	$-1.140 * 10^{-7}$	$-1.698 * 10^{-7}$	$0.582 * 10^{-7}$
θ_4	$1.145 * 10^{-5}$	$-0.665 * 10^{-5}$	$3.574 * 10^{-5}$
θ_5	$-7.775 * 10^{-5}$	$-8.590 * 10^{-5}$	$-6.959 * 10^{-5}$
θ_6	$7.780 * 10^{-5}$	$6.099 * 10^{-5}$	$9.461 * 10^{-5}$
θ_7	$8.332 * 10^{-4}$	$8.156 * 10^{-4}$	$8.508 * 10^{-4}$

for individual components, suggesting that the differential equation is still not exact.

6.2 Modelling 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, 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.

Various scales have been developed to measure the severity of symptoms, and Figure 10 shows the course of one of the more popular measures, the SLEDAI scale, for a patient that experienced 41 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. 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 begin by postulating 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

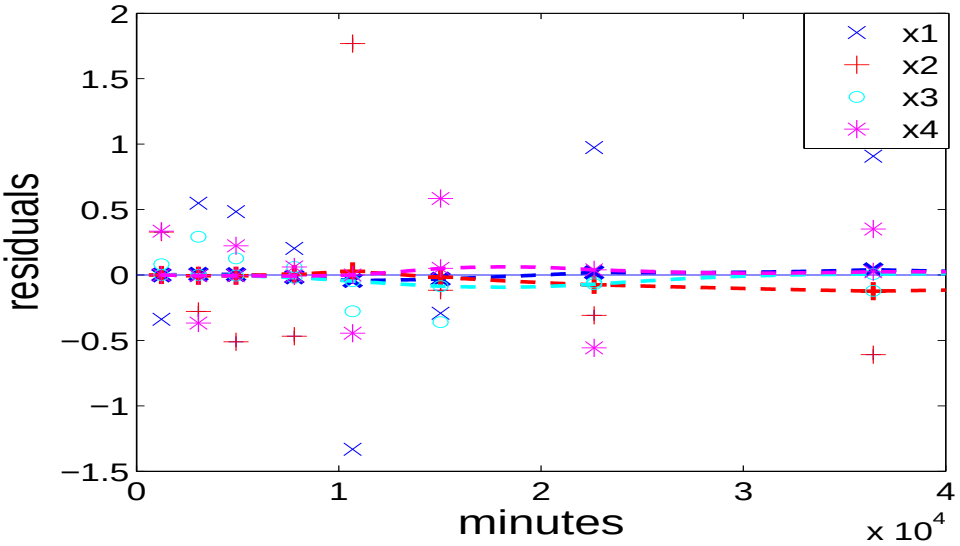

Figure 9: Residuals from a fit to the α-pinene data given as points. Lines correspond to the discrepancy between the smoothing spline and the exact differential equation.

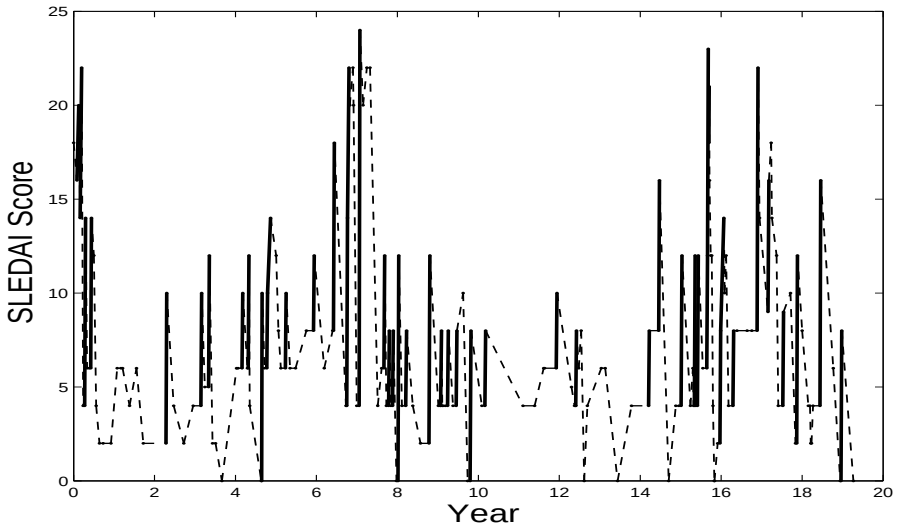

Figure 10: Symptom level $s(t)$ for a patient suffering from lupus as assessed by the SLEDAI scale. Changes in SLEDAI score corresponding to a flare are shown as heavy solid lines, and other the remaining changes are shown as dashed lines.

severity $s(t)$ at time t:

$$Ds(t) = -\beta(t)s(t) + \alpha(t)u(t). \tag{21}$$

This equation has the solution

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

where

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

Function $\alpha(t)$ tracks the long-term trend in the severity of the disease over the 19 years, and we 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 indirectly is likely to reduce $\beta(t)$. Consequently, we propose a second simple linear equation to model this mechanism:

$$D\beta(t) = -\gamma\beta(t) + \theta[1 - u(t)]. \tag{22}$$

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

$$D\beta(t) = \theta[1 - u(t)], \tag{23}$$

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

The top panel in Figure 11 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 12. We found that parameter γ was indeed

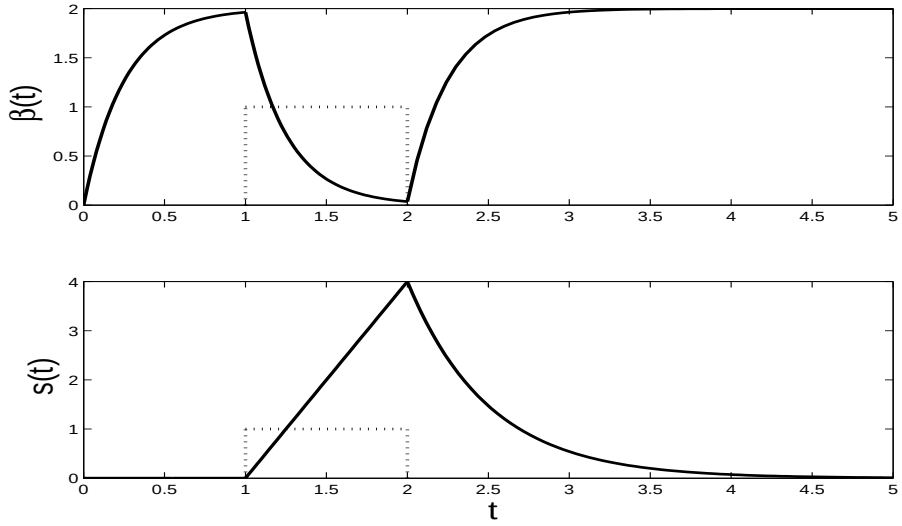

Figure 11: The top panel shows the effect of a lupus attack on the weight function $\beta(t)$ in differential equation (21). The bottom panel shows the time course of the symptom severity function $s(t)$. These results are for parameters $\gamma = \theta = 4$.

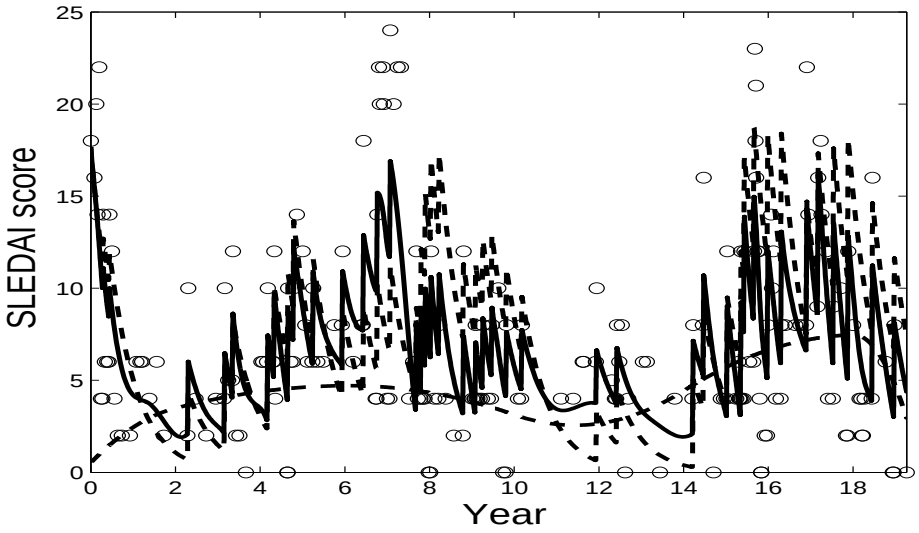

Figure 12: The circles indicate SLEDAI scores, the jagged solid line is the smoothing functions $s(t)$, the dashed jagged line is the solution to the differential equation and the smooth dashed line is the smooth trend $\alpha(t)$.

so high that the fitted symptom rise was effectively linear, so we deleted γ and use the simpler equation (23). 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 12 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 (21), 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-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 13, and we see there that the DIFE solution is less able than the smooth to track the data when

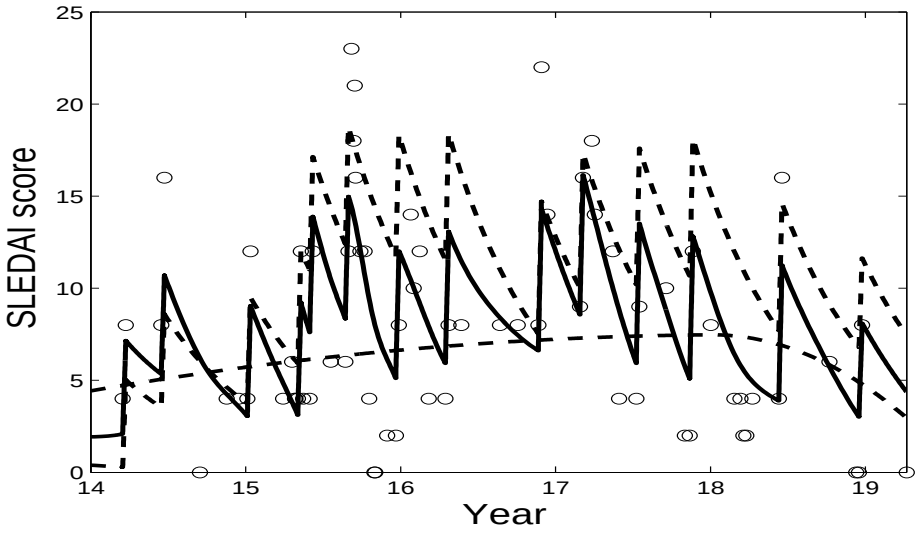

Figure 13: The data in Figure 12 plotted over the last five years of the record.

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 modelling the within-flare dynamics, the general trend in the data, and the interaction between flare dynamics and trend.

7 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 $\texttt{PEN}(\mathbf{x}|\theta)$ 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 4.4 and 4.5 are immediate.

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

$$D^m\mathbf{x}(t) = \mathbf{f}(\mathbf{x}, D\mathbf{x}, \ldots, D^{m-1}\mathbf{x}, \mathbf{u}, t|\boldsymbol{\theta}) \tag{24}$$

may be reduced to a larger first-order system by defining the derivatives $D\mathbf{x}$ up to $D^{m-1}\mathbf{x}$ as new variables. Initial conditions need to be given for each of these new variables in order to define a unique solution. Equation (24), however, can be used directly to define a differential operator as in (7), saving the estimation

of the derivative terms and all the initial conditions. There is, of course, no need for m in (24) to be constant across components of $\mathbf{x}$.

A slight generalization of (24) is to allow m to be zero for some components, that is define

$$x_i(t) = f_i(\mathbf{x}, \mathbf{u}, t|\boldsymbol{\theta}) \tag{25}$$

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 (25) to be solved numerically given the other values of $\mathbf{x}$. Our approach also allows (25) to appear as a term in $\texttt{PEN}(\mathbf{x}|\boldsymbol{\theta})$, providing an easier implementation of such systems.

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

$$D\mathbf{x}(t) = \mathbf{f}(\mathbf{x}(t-\delta_1), \mathbf{x}(t-\delta_2), \ldots, \mathbf{x}(t-\delta_3), \mathbf{u}(t-\delta_4), t|\boldsymbol{\theta}) \tag{26}$$

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, in theory, estimate the δ_i, although we have yet to experiment with this possibility.

Finally, 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.

8 Further issues in fitting differential equations

8.1 Assessing goodness of fit and model building

From our experiences with real-world data, 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.

8.2 Experimental design

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.

9 Conclusions

Differential equations have a long and illustrious history in mathematical modelling. However, there has been little development of statistical theory for estimating such models or assessing their agreement with observational data. We have proposed a novel method for estimating parameters from data derived from systems governed by differential equations. Our approach 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 penalize spline criterion for each value of $\boldsymbol{\theta}$ and then estimating $\boldsymbol{\theta}$ through a profiling scheme in which the fit is regarded as a nuisance parameter. We have found this scheme to have good numerical properties. We have also produced variance estimates that we show to have good coverage properties.

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. The theoretical properties of our proposed estimates are not well understood for λ not close to ∞. In particular, the extent to which using a smaller amount of smoothing provides robustness against mild mis-specification of the equations is not clear. We have mentioned experimental design and goodness of fit among the many standard statistical problems that have yet to be addressed, making the *statistics of dynamical systems* a source of important and challenging problems.

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.

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

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.

Fuguitt, R. and J. E. Hawkins (1947). Rate of the thermal isomerization of α-pinene in the liquid phase. *Journal of the American Chemical Society 69*, 319–322.

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.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. New York: Springer.

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.

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.

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.

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

Stewart, W. and J. Sorensen (1981). Bayesian estimation of common parameters from multiresponse data with missing observations. *Technometrics 23*, 131–141.

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.

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:

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

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

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

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

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

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:

$$\left(\sum_{k=1}^{n_i} w_i \phi_{i,p}(t_{i,k})\phi_{j,q}(t_{i,k}) + \lambda \int \phi_{i,p}(t)\phi_{j,q}(t)dt\right) I(i=j)$$
$$- \lambda_i \int D\phi_{i,p}(t)\frac{df_i}{dx_j}\phi_{j,q}(t)dt - \lambda_j \int \phi_{i,p}(t)\frac{df_i}{dx_j}D\phi_{j,q}(t)dt$$
$$+ \int \phi_{i,p}(t)\left[\sum_{k=1}^{n}\lambda_k\left(\frac{d^2f_k}{dx_i dx_j}(f_k - Dx_k(t)) + \frac{df_k}{dx_i}\frac{df_k}{dx_j}\right)\right]\phi_{j,q}(t)dt$$

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:

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

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

The variance of the parameter estimates is calculated using

$$\frac{d\theta}{dY} = -\left[\frac{d^2F}{d\boldsymbol{\theta}^2}\right]^{-1}\frac{d^2F}{d\boldsymbol{\theta}dY}.$$

Neither of the terms on the right hand side are trivial. The first may be expanded by differentiating:

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

Using the matrix identity

$$\frac{d}{dt}\left(A(t)^{-1}\right) = -\left[A(t)\right]^{-1}\frac{dA(t)}{dt}\left[A(t)\right]^{-1}$$

this expression expands (and simplifies) to:

$$-\frac{dF}{d\mathbf{c}}^T\left[\frac{d^2H}{d\mathbf{c}^2}\right]^{-1}\left\{\sum_{p=1}^{N}\left(\frac{d^3H}{d\mathbf{c}d\theta_i d\theta_j}\frac{d^3H}{d\mathbf{c}dc_p d\theta_i}\frac{dc_p}{d\theta_j} + \frac{d^3H}{d\mathbf{c}dc_p d\theta_j}\frac{dc_p}{d\theta_i}\right) + \right\}$$
$$-\frac{dF}{d\mathbf{c}}^T\left[\frac{d^2H}{d\mathbf{c}^2}\right]^{-1}\left\{\sum_{p,q=1}^{N}\frac{dc_p}{d\theta_i}\frac{d^3H}{d\mathbf{c}dc_p dc_q}\frac{dc_q}{d\theta_j}\right\} + \frac{d\mathbf{c}}{d\theta_i}^T\frac{d^2F}{d\mathbf{c}^2}\frac{d\mathbf{c}}{d\theta_j}$$

for the (i,j)th entry where the summations for p and q cover the coefficients from all components of $\mathbf{x}$. Here $d^2F/d\mathbf{c}^2$ 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 $-w_i\Phi_i(\mathbf{t}_i)^T(\mathbf{y}_i - x_i(\mathbf{t}_i))$.

Three-way array $d^3H/d\mathbf{c}dc_pdc_q$ can be written in the same block vector form as $d^2H/d\mathbf{c}d\boldsymbol{\theta}$ with the uth entry of the kth block given by

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

assuming c_p is a coefficient in the basis representation of x_i and c_q a corresponds to x_j. Three-way array $d^3H/d\mathbf{c}d\theta_i d\theta_j$ is also expressed in the same block form with entry p in the kth block being:

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

Three-way array $d^3H/d\mathbf{c}dc_pd\theta_i$ is in the same block from, with the qth entry of the jth block being:

$$\int \left(\sum_{l=1}^{n} \lambda_l \left[\frac{d^2 f_l}{d\theta_i dx_j} \frac{df_l}{dx_k} + \frac{d^2 f_l}{d\theta_i dx_k} \frac{df_l}{dx_j} + \frac{d^2 f_l}{dx_j dx_k} \frac{df_l}{d\theta_i} \right] \right) \phi_{k,p}(t)\phi_{j,q}(t)dt$$
$$+ \int \sum_{l=1}^{n} \lambda_l \left(\frac{d^3 f_k}{dx_j dx_k d\theta_i} (f_l - Dx_l(t)) \right) \phi_{k,p}(t)\phi_{j,q}(t)dt$$
$$- \lambda_j \int \frac{d^2 f_j}{d\theta_i dx_k} D\phi_{j,q}(t)\phi_{k,p}(t)dt - \lambda_k \int \frac{d^2 f_k}{d\theta_i dx_j} \phi_{j,q}(t)D\phi_{k,p}(t)dt$$

where c_p corresponds to the basis representation of x_k.

Similar calculations give matrix $d^2F/d\boldsymbol{\theta}dY$ explicitly as:

$$\frac{d\mathbf{c}}{d\boldsymbol{\theta}}^T \left[\frac{d^2F}{d\mathbf{c}dY} + \frac{d^2F}{d\mathbf{c}^2} \frac{d\mathbf{c}}{dY} \right]$$
$$- \frac{dF}{d\mathbf{c}} \left[\frac{d^2F}{d\mathbf{c}^2} \right]^{-1} \left\{ \sum_{p,q=1}^{N} \frac{dc_p}{d\boldsymbol{\theta}}^T \frac{d^3H}{d\mathbf{c}dc_pdc_q} \frac{dc_q}{dY} + \sum_{p=1}^{N} \frac{d^3H}{d\mathbf{c}dc_pd\boldsymbol{\theta}} \frac{dc_p}{dY} \right\}$$

with $d\mathbf{c}/dY$ given by

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

and $d^2H/d\mathbf{c}dY$ being block diagonal with the ith block containing $w_i\Phi_i(\mathbf{t}_i)$.

B Proofs of theorems in section 4.1

The central theorem of Section 4.1 concerns the behavior of a function $x(\alpha)$ defined by $x(\alpha) = \min_x g(x, \alpha)$. We begin with two lemmas:

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

$$x^* = \operatorname*{argmin}_{x\in\mathcal{X}} g(x) \tag{27}$$

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

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

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 (27) 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 $j(\alpha') = \min_{x\in\mathcal{X}} g(x, \alpha)$. □

Theorem 4.3: *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.

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

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

giving

$$\begin{aligned} |g(x(\alpha), \alpha) - g(x(\alpha'), \alpha)| &< |g(x(\alpha), \alpha) - g(x(\alpha'), \alpha')| + |g(x(\alpha'), \alpha') - g(x(\alpha'), \alpha)| \\ &= |j(\alpha) - j(\alpha')| + |g(x(\alpha'), \alpha') - g(x(\alpha'), \alpha)| \\ &< \delta/3 + \delta/3 \\ &< \delta \end{aligned}$$

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