J. R. Statist. Soc. B (2007)
69, *Part* 5, *pp.* 741–796

Parameter estimation for differential equations: a generalized smoothing approach

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

McGill University, Montreal, Canada

[*Read before* The Royal Statistical Society *at a meeting organized by the* Research Section *on Wednesday, May 9th, 2007*, Professor T. J. Sweeting *in the Chair*]

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

Keywords: Differential equation; Dynamic system; Estimating equation; Functional data analysis; Gauss–Newton method; Parameter cascade; Profiled estimation

1. Challenges in dynamic systems estimation

1.1. Basic properties of dynamic systems

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

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

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

Differential equations as a rule do not define their solutions uniquely, but rather as a manifold of solutions of typical dimension d. For example, $\mathrm{d}^2x/\mathrm{d}t^2 = -\omega^2 x(t)$, reduced to $\dot{x}_1 = x_2$ and $\dot{x}_2 = -\omega^2 x_1$, implies solutions of the form $x_1(t) = c_1 \sin(\omega t) + c_2 \cos(\omega t)$, where coefficients c_1 and c_2 are arbitrary; and at least $d = 2$ observations are required to identify the solution that

Address for correspondence: J. O. Ramsay, 2748 Howe Street, Ottawa, Ontario, K2B 6W9, Canada.
E-mail: ramsay@psych.mcgill.ca

1369–7412/07/69741

best fits the data. *Initial value* problems supply $\mathbf{x}(0)$, whereas *boundary value* problems require d values selected from $\mathbf{x}(0)$ and $\mathbf{x}(T)$.

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

Solutions of the ordinary differential equation (ODE) system (1) given initial values $\mathbf{x}(0)$ exist and are unique over a neighbourhood of $(0, \mathbf{x}(0))$ if f is continuously differentiable or, more generally, Lipschitz continuous with respect to $\mathbf{x}$. However, most ODE systems are not solvable analytically, which typically increases the computational burden of data fitting methodology such as non-linear regression. Exceptions are linear systems with constant coefficients, where the machinery of the Laplace transform and transform functions plays a role, and a statistical treatment of these is available in Bates and Watts (1988) and Seber and Wild (1989). Discrete versions of linear constant coefficient systems, i.e. stationary systems of difference equations for equally spaced time points, are also well treated in the classical time series autoregressive integrated moving average and state space literature, and will not be considered further in this paper.

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

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

1.2. Two test bed problems

1.2.1. FitzHugh–Nagumo equations

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

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

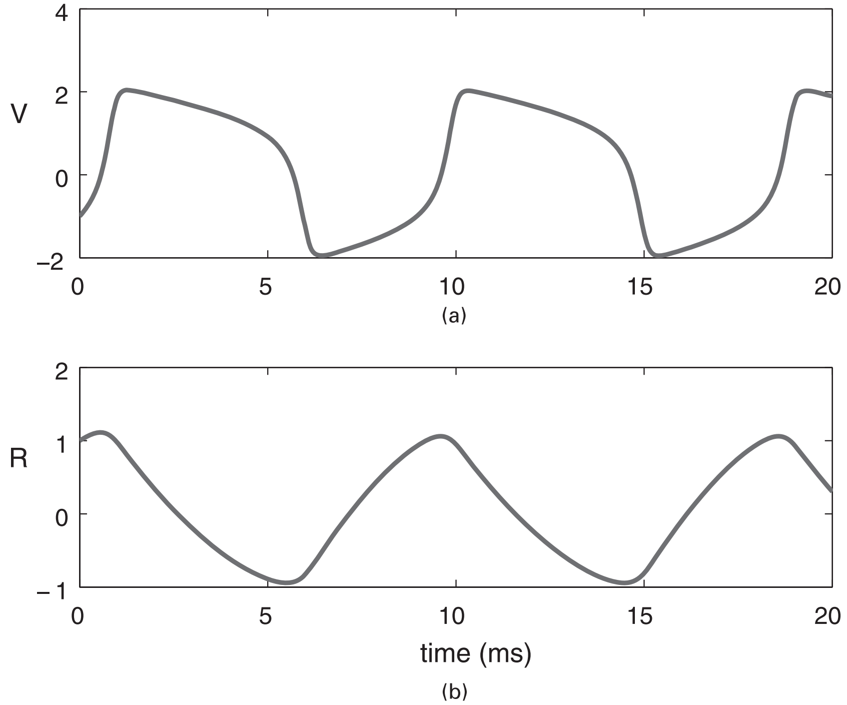


Fig. 1. Limiting behaviour of (a) voltage V and (b) recovery R variables defined by the FitzHugh–Nagumo equations (2) with parameter values $a=0.2$, $b=0.2$ and $c=3.0$ and initial conditions $(V_0,R_0)=(-1,1)$

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

The parameters are $\boldsymbol{\theta}=\{a,b,c\}$, to which we shall assign values $(0.2,0.2,3)$ respectively. The R-equation is the simple constant coefficient linear system $\dot{R}=-(b/c)R$ with linear inputs V and a. However, the V-equation is non-linear; when $V>0$ is small, $\dot{V}\approx cV$ and consequently exhibits nearly exponential increase but, as V passes $\pm\surd 3$, the influence of $-V^3/3$ takes over and turns V back towards 0. Consequently, solutions corresponding to a range of initial values quickly settle down to alternate between the smooth evolution and the sharp changes in direction that are shown in Fig. 1.

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

1.2.2. Tank reactor equations

The chemical engineering concept of a continuously stirred tank reactor (CSTR) consists of a tank surrounded by a cooling jacket containing an impeller which stirs its contents. A fluid

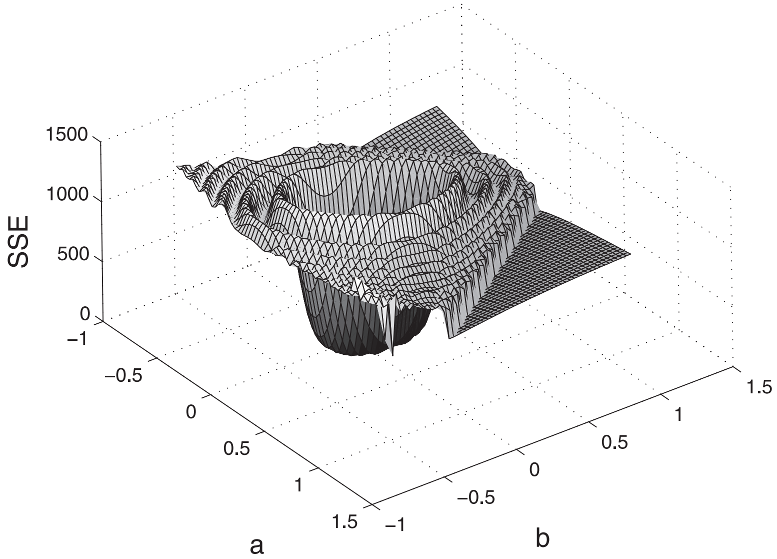


Fig. 2. Response surface for solutions of the FitzHugh–Nagumo equations (2) as parameters a and b are varied: surface values give the integrated squared difference between solutions at parameters $a=0.2$ and $b=0.2$ with solutions at the values of a and b given on the x- and y-axes respectively; $c=3$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

containing a reagent with concentration C_{in} enters the tank at a flow rate F_{in} and temperature T_{in}. A reaction produces a product that leaves the tank with concentration C and temperature T. A coolant in the cooling jacket has temperature T_{co} and flow rate F_{co}.

The differential equations that are used to model a CSTR, simplified by setting the volume of the tank to 1, are

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

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

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

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

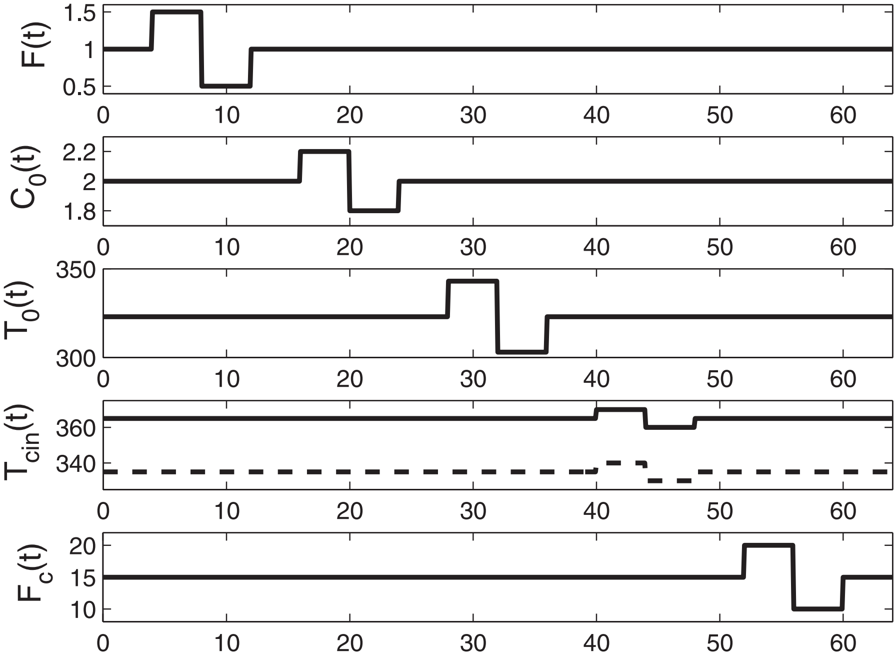


Fig. 3. The five inputs to the chemical reactor modelled by equations (3) and (4): flow rate $F(t)$, input concentrations $C_0(t)$, input temperature $T_0(t)$, coolant temperature $T_{co}(t)$ and coolant flow $F_0(t)$ ($T_{co}(t)$ was set at two base-line levels, cool (– – –) and hot (———))

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

The behaviours of output variables C and T under the two experimental regimes, given values 0.833, 0.461, 1.678 and 0.5 for parameters τ, κ, a and b respectively, are shown in Fig. 4. When the reactor runs in the cool mode, where the base-line coolant temperature is 335 K, the two outputs respond smoothly to the step changes in all inputs. However, an increase in base-line coolant temperature by 30 K generates oscillations that come close to instability when the coolant temperature decreases, something that is undesirable in an actual industrial process. These perturbations are due to the double effect of a decrease in output temperature, which increases the size of both β_{CC} and β_{TC}. Increasing β_{TC} raises the forcing term in the T-equation, thus increasing temperature. Increasing β_{CC} makes concentration more responsive to changes in temperature but decreases the size of the response. This push–pull process has a resonant frequency that depends on the kinetic constants and, when the ambient operating temperature reaches a certain level, the resonance appears. For coolant temperatures that are either above or below this critical zone, the oscillations disappear.

The CSTR equations present two challenges that are not an issue for the FitzHugh–Nagumo equations. The step changes in inputs induce corresponding discontinuities in the output derivatives that complicate the estimation of solutions by numerical methods. Moreover, the engineer must estimate the reaction kinetics parameters to estimate the cooling temperature range to avoid, but a key question is whether all four parameters are actually estimable given a particular

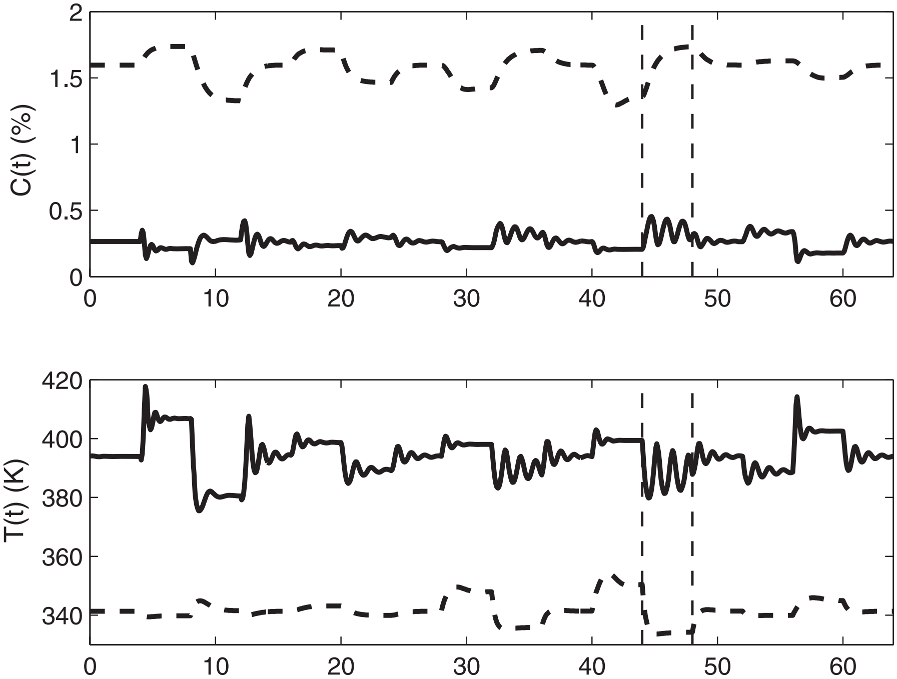


Fig. 4. The two outputs, for each of base-line coolant temperatures T_{co} of 335 K (– – –) and 365 K (———), from the chemical reactor modelled by the two equations (3): concentration $C(t)$ and temperature $T(t)$ (the input functions are shown in Fig. 3; ¦, times at which an input variable $T_{\text{co}}(t)$ was stepped down and then up)

data configuration. Step changes in inputs and near overparameterization are common problems in dynamic systems modelling.

1.3. Review of current ordinary differential equation parameter estimation strategies

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

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

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

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

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

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

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

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

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

1.3.2. Collocation methods or basis function expansions

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

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

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

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

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

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

1.4. Overview of the paper

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

2. Generalized profiling estimation procedure

We first give an overview of our estimation strategy and then provide further details below. As we noted above, our method is a variant of the collocation method and, as such, represents each variable in terms of a basis function expansion (7). Let $\mathbf{c}$ indicate the composite vector of length $K = \Sigma_{i \in \mathcal{I}} K_i$ that results from concatenating the $\mathbf{c}_i$s. Let $\boldsymbol{\Phi}_i$ be the $N_i \times K_i$ matrix of values $\phi_k(t_{ij})$, and let $\boldsymbol{\Phi}$ be the $N = \Sigma_{i \in \mathcal{I}} N_i \times K$ supermatrix that is constructed by placing the matrices $\boldsymbol{\Phi}_i$ along the diagonals and 0s elsewhere. According to this notation, we have the composite basis expansion $\hat{\mathbf{x}} = \boldsymbol{\Phi}\mathbf{c}$.

2.1. Overview of the estimation procedure

Defining $\hat{\mathbf{x}}$ as a set of basis function expansions implies that there are two classes of parameters to estimate: the parameters $\boldsymbol{\theta}$ defining the equation, such as the four reaction kinetics parameters in the CSTR equations, and the coefficients in $\mathbf{c}_i$ defining each basis function expansion. The equation parameters are *structural* in the sense of being of primary interest, as are the error distribution parameters in $\boldsymbol{\sigma}_i$, $i \in \mathcal{I}$. But the coefficients $\mathbf{c}_i$ are considered as *nuisance* parameters that are essential for fitting the data, but usually not of direct concern. The sizes of these vectors are apt to vary with the length of the observation interval, density of observation and other aspects of the structure of the data; and the number of these nuisance parameters can be orders of magnitude larger than the number of structural parameters, with a ratio of about 200 applying in the CSTR and FitzHugh–Nagumo problems.

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

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

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

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

2.2. Data fitting criterion

Let $\mathbf{e}_i$ indicate the vector of errors that is associated with observed variable $i \in \mathcal{I}$, and let $g_i(\mathbf{e}_i|\boldsymbol{\sigma}_i)$ indicate the joint density of these errors conditional on a parameter vector $\boldsymbol{\sigma}_i$. In practice it is usual to assume independently distributed Gaussian errors with mean 0 and standard deviation

σ_i. However, autocorrelation structure and non-stationary variance are often evident in the data and, when these features are also modelled, these parameters are also incorporated into error distribution parameters $\boldsymbol{\sigma}_i$. Let $\boldsymbol{\sigma}$ indicate the concatenation of the $\boldsymbol{\sigma}_i$-vectors. Although our notation is consistent with assuming that errors are independent across variables, intervariable error dependences, also, can be accommodated by the approach that is developed in this paper. In general, the data fitting criterion can be taken to be the negative log-likelihood

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

where

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

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

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

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

2.3. Assessing fidelity to the equations

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

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

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

$$\mathrm{PEN}_i(\hat{\mathbf{x}}) = \int L_{i,\boldsymbol{\theta}}\{\hat{x}_i(t)\}^2 \, \mathrm{d}t \tag{11}$$

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

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

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

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

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

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

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

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

or the least squares criterion

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

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

2.5. Optimizing with respect to θ

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

The gradient or total derivative with respect to $\boldsymbol{\theta}$ is

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

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

$$\begin{aligned}\frac{\mathrm{d}\hat{\mathbf{c}}}{\mathrm{d}\boldsymbol{\theta}} &= -\left(\frac{\partial^2 J}{\partial\hat{\mathbf{c}}^2}\right)^{-1}\frac{\partial^2 J}{\partial\hat{\mathbf{c}}\,\partial\boldsymbol{\theta}},\\ \frac{\mathrm{d}H}{\mathrm{d}\boldsymbol{\theta}} &= \frac{\partial H}{\partial\boldsymbol{\theta}} - \frac{\partial H}{\partial\hat{\mathbf{c}}}\left(\frac{\partial^2 J}{\partial\hat{\mathbf{c}}^2}\right)^{-1}\frac{\partial^2 J}{\partial\hat{\mathbf{c}}\,\partial\boldsymbol{\theta}}.\end{aligned} \tag{17}$$

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

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

Let $\boldsymbol{\Sigma}$ be the variance–covariance matrix for $\mathbf{y}$. Making explicit the dependence of H on the data $\mathbf{y}$ by using the notation $H(\boldsymbol{\theta}|\mathbf{y})$, the estimate $\hat{\boldsymbol{\theta}}(\mathbf{y})$ of $\boldsymbol{\theta}$ is the solution of the stationary equation $\partial H(\boldsymbol{\theta},|\mathbf{y})/\partial\boldsymbol{\theta} = 0$. Here and below, all partial derivatives as well as total derivatives are assumed to be evaluated at $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{c}}(\hat{\boldsymbol{\theta}})$, which are in turn evaluated at $\mathbf{y}$.

The usual δ-method that is employed in non-linear least squares produces a variance estimate of the form

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

by making use of the approximation

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

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

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

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

where

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

and

$$\frac{\mathrm{d}^2H}{\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\mathbf{y}}=\frac{\partial^2H}{\partial\boldsymbol{\theta}\,\partial\mathbf{y}}+\frac{\partial^2H}{\partial\hat{\mathbf{c}}\,\partial\mathbf{y}}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}}+\frac{\partial^2H}{\partial\boldsymbol{\theta}\,\partial\hat{\mathbf{c}}}\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}+\left(\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}}\right)'\frac{\partial^2H}{\partial\hat{\mathbf{c}}^2}\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}+\frac{\partial H}{\partial\hat{\mathbf{c}}}\frac{\partial^2\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}\,\partial\mathbf{y}}. \tag{21}$$

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

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

Let $\boldsymbol{\mu}=E(\mathbf{y})$; the first-order Taylor series expansion for $\mathrm{d}\hat{\boldsymbol{\theta}}/\mathrm{d}\mathbf{y}$ is

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

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

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

from which we derive

$$\operatorname{var}\{\hat{\boldsymbol{\theta}}(\mathbf{y})\}\approx\left(\frac{\mathrm{d}\hat{\boldsymbol{\theta}}}{\mathrm{d}\boldsymbol{\mu}}\right)\boldsymbol{\Sigma}\left(\frac{\mathrm{d}\hat{\boldsymbol{\theta}}}{\mathrm{d}\boldsymbol{\mu}}\right)'\approx\left(\frac{\mathrm{d}\hat{\boldsymbol{\theta}}}{\mathrm{d}\mathbf{y}}\right)\boldsymbol{\Sigma}\left(\frac{\mathrm{d}\hat{\boldsymbol{\theta}}}{\mathrm{d}\mathbf{y}}\right)', \tag{24}$$

since

$$E\left(\frac{\mathrm{d}\hat{\boldsymbol{\theta}}}{\mathrm{d}\boldsymbol{\mu}}\right)\approx E\left(\frac{\mathrm{d}\hat{\boldsymbol{\theta}}}{\mathrm{d}\mathbf{y}}\right).$$

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

$$\operatorname{var}[\hat{\mathbf{c}}\{\hat{\boldsymbol{\theta}}(\mathbf{y})\}]=\left(\frac{\mathrm{d}\hat{\mathbf{c}}}{\mathrm{d}\mathbf{y}}\right)\boldsymbol{\Sigma}\left(\frac{\mathrm{d}\hat{\mathbf{c}}}{\mathrm{d}\mathbf{y}}\right)', \tag{25}$$

where

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

2.7. Numerical integration in the inner optimization

The integrals in PEN_i will normally require approximation by the linear functional

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

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

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

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

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

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

2.8. Choosing the amount of smoothing

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

2.8.1. *Robustness with respect to initial parameter values*

Fig. 2 shows the severe non-convexity of least squares fitting criteria for $\boldsymbol{\theta}$ when using an exact solution of the FitzHugh–Nagumo ODE, implying a small neighbourhood of the optimal parameter values from which convergence is assured using the Gauss–Newton method. However, Fig. 5, displaying the much more regular surface corresponding to $\boldsymbol{\lambda} = 10^5$, suggests a much wider region of convergence; and our experience for other problems confirms this robustness with respect to poor initialization of parameters for smaller $\boldsymbol{\lambda}$-values. Because the criterion $H(\boldsymbol{\theta}, \boldsymbol{\sigma}|\boldsymbol{\lambda})$ is increasing in each λ_i, it underestimates the response surface for exact solutions to the differential equation. Moreover, results in Appendix A imply that $\|\mathrm{d}\mathbf{c}/\mathrm{d}\boldsymbol{\theta}\|$ increases in $\boldsymbol{\lambda}$, implying that relaxing the differential equation model regularizes the search for $\boldsymbol{\theta}$.

However, as $\boldsymbol{\lambda}$ becomes smaller, the estimates that are obtained for $\boldsymbol{\theta}$ become both more biased and more variable. Theorem 2, in contrast, demonstrates that, ignoring error due to equation (7), parameter estimates must approximate those that would have been obtained from a straightforward maximum likelihood fit as $\boldsymbol{\lambda}$ increases. This suggests the following algorithm.

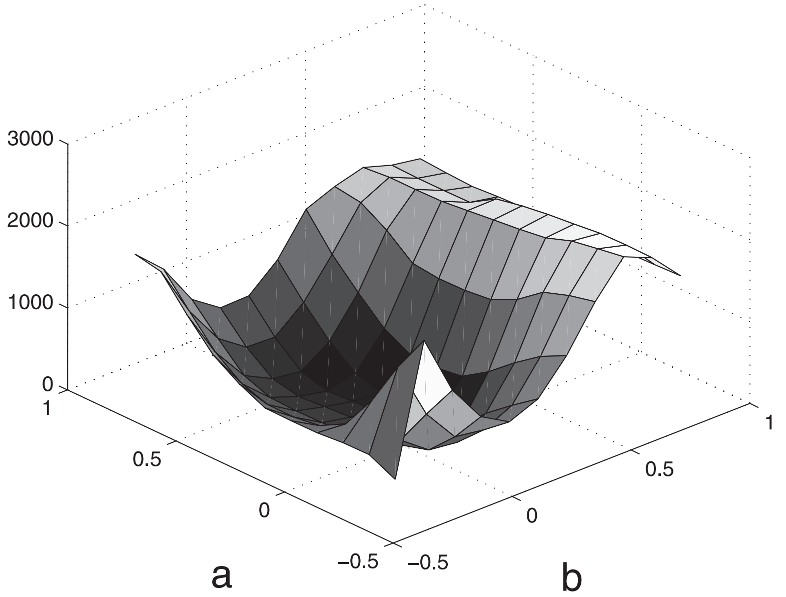


Fig. 5. Squared discrepancy between exact solutions to the FitzHugh–Nagumo equations and a model-based smooth that minimizes equation (14) with $\lambda = 10^5$: values of the surface are calculated by using the same data as in Fig. 2

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

To assess when $\boldsymbol{\lambda}$ has become too large:

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

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

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

2.8.2. Robustness with respect to model misspecification

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

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

$$F(\boldsymbol{\lambda}) = \frac{\sum_{\mathcal{I}} \|\mathbf{y}_i - \hat{x}_i(\mathbf{t}_i)\|^2}{\left[\sum_{\mathcal{I}} \left\{N_i - \sum_j \mathrm{d}\hat{x}_i(t_{ij})/\mathrm{d}y_{ij}\right\}\right]^2}, \tag{27}$$

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

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

where $\mathrm{d}\hat{x}_i(t_{ij})/\mathrm{d}\mathbf{c}$ is simply the value of the basis expansion (7) at t_{ij} and $\mathrm{d}\mathbf{c}/\mathrm{d}\mathbf{y}$ has been calculated in equation (25). Note that this explicitly takes the dependence of $\hat{\mathbf{y}}$ on $\hat{\boldsymbol{\theta}}$ into account. This construction is offered as speculation; it is well known that the first-order approximation that is used in $F(\boldsymbol{\lambda})$ can be biased (Friedman and Silverman, 1989). Furthermore, $F(\boldsymbol{\lambda})$ is only indirectly related to $\boldsymbol{\theta}$, and our experience suggests that, for misspecified models, estimators that are based on cross-validation tend to select $\boldsymbol{\lambda}$ at values that produce good estimates of $\mathbf{x}$, but which are smaller than optimal for estimating $\boldsymbol{\theta}$.

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

In this section, we consider the behaviour of our parameter estimate as λ becomes large. This analysis takes an idealized form in the sense that we assume that this optimization may be done globally and that the function being estimated can be expressed exactly and without the approximation error that would come from a basis expansion. We show that, as λ becomes large, the estimates that are defined through our profiling procedure converge to the estimates that we would obtain if we estimated $\boldsymbol{\theta}$ by minimizing the negative log-likelihood over both $\boldsymbol{\theta}$ and the initial conditions $\hat{\mathbf{x}}_0$. In other words, we treat $\hat{\mathbf{x}}_0$ as nuisance parameters and estimate $\boldsymbol{\theta}$ by profiling. When $\mathbf{f}$ is Lipschitz continuous in $\hat{\mathbf{x}}$ and continuous in $\boldsymbol{\theta}$, the likelihood is continuous in $\boldsymbol{\theta}$ and the usual consistency theorems (e.g. Cox and Hinkley (1974)) hold and, in particular, the estimate $\hat{\boldsymbol{\theta}}$ is asymptotically unbiased.

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

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

Secondly, we shall represent

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

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

We shall also assume that solutions to the data fitting problem exist and are well defined, and that there are objects $\mathbf{x}$ that satisfy $\mathrm{PEN}(\mathbf{x}|\boldsymbol{\theta}) = 0$. Such objects are guaranteed to exist *locally* whenever $\mathbf{f}$ is locally Lipschitz continuous, i.e. there is a time interval $[t_0, t_0 + h]$ on which $\mathbf{x}$

exists. On this interval $\mathbf{x}$ is uniquely determined by $\mathbf{x}(t_0)$; see Deuflhard and Bornemann (2000). Existence on the interval of the experiment is more difficult to show in general.

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

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

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

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

Theorem 1. Let $\lambda_k \to \infty$ and assume that

$$\mathbf{x}_k = \arg\min_{\mathbf{x}\in(W^1)^n}\{l(\mathbf{x})\} + \lambda_k\,\mathrm{PEN}(\mathbf{x}|\boldsymbol{\theta})$$

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

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

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

Theorem 2. Let $\mathcal{X} \subset (W^1)^n$ and $\Theta \subset \mathbb{R}^p$ be bounded. Assume that, for $\lambda > K$,

$$\mathbf{x}_{\boldsymbol{\theta},\lambda} = \arg\min_{\mathbf{x}\in\mathcal{X}}\{l(\mathbf{x})\} + \lambda\,\mathrm{PEN}(\mathbf{x}|\boldsymbol{\theta})$$

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

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

and let

$$\boldsymbol{\theta}(\lambda) = \arg\min_{\boldsymbol{\theta}\in\Theta}\{l(\mathbf{x}_{\boldsymbol{\theta},\lambda})\}$$

and

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}\in\Theta}\{l(\mathbf{x}^*_{\boldsymbol{\theta}})\}$$

also be well defined. Then

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

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

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

3. Simulated data examples

3.1. Fitting the FitzHugh–Nagumo equations

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

Fig. 6 gives quartiles of the parameter estimates for 60 simulations as λ is varied from 10^{-2} to 10^5. There is large bias for small values of λ, where smoothing is emphasized and $\boldsymbol{\theta}$ has little effect on $\hat{\mathbf{c}}$, but, as λ increases, parameter estimates become nearly unbiased. Table 1 provides bias and variance estimates from 500 simulations at $\lambda=10^4$, along with our estimate (24) and the Gauss–Newton standard error (18). We obtain good coverage properties for our estimates of variance whereas the Gauss–Newton estimates are somewhat less accurate. We note, however, that computing expression (24) increased computational effort by a factor of about 10 for this simulation. As a practical matter, using approximation (18) may be considered sufficient if expression (24) becomes too costly.

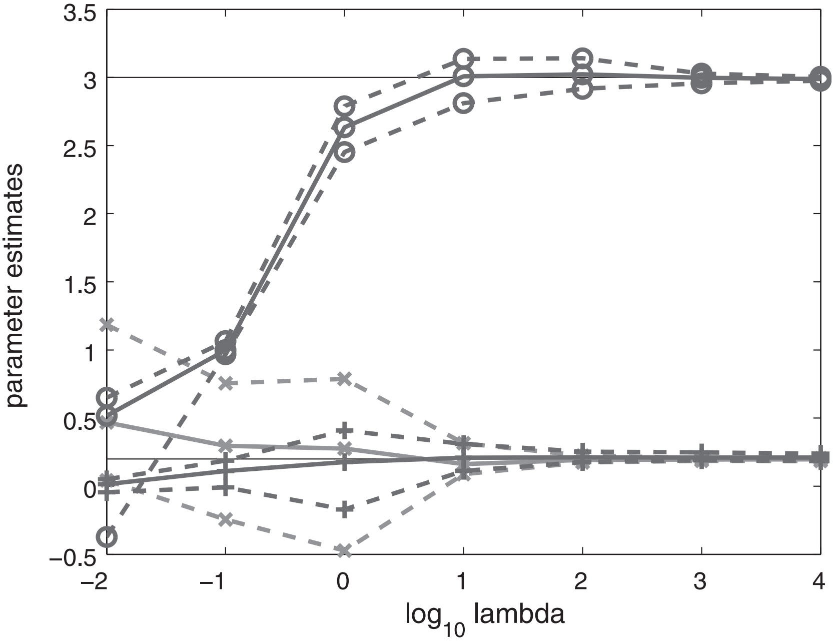


Fig. 6. 25%, 50% and 75% quantiles of parameter estimates for the FitzHugh–Nagumo equations as λ is varied: ———, true parameter values; ×, *a*; +, *b*; ○, *c*

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

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

3.2. *Fitting the tank reactor equations*

Data for concentration C and temperature T were simulated by adding zero-mean Gaussian noise with standard deviations 0.0223 and 0.79 respectively to the values for the cool mode experimental condition that is shown in Fig. 4. These error levels were about 20% of the variation of the respective outputs over the experimental conditions, an error level which is considered to be typical for many chemical engineering processes. We estimated only the parameters κ, τ and a, keeping b fixed at 0.5 because we had determined that the accurate estimation of all four parameters is impossible within the data design that was described above. Since the data are generated here from functions satisfying the differential equation system, we can expect the fit to improve with increasingly larger values for smoothing parameters λ_C and λ_T. Results are reported here for 100 and 10 respectively, which are sufficiently large that further increases were found to yield negligible improvement in parameter estimates.

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

Table 2 displays bias and sampling precision results for parameter estimates by our parameter cascade method for 1000 simulated samples for each of two measurement regimes: both variables measured, and only temperature measured. The first two rows of Table 2 compare the true parameter values with the mean estimates, and the last two rows compare the biases of the estimates with the standard errors of the mean estimates. We see that the estimation biases can be considered negligible for both measurement situations. The third and fourth rows compare the actual standard deviations of the parameter estimates with the values estimated with the Gauss–Newton method in approximation (18), and the two values seem sufficiently close for all three parameters to permit us to trust the Gauss–Newton estimates in this case. As we might expect, the main effect of having only temperature measurements is to increase the sampling error in the parameter estimates.

When the equations were solved by using the parameters estimated from measurements on both variables, the maximum absolute discrepancies between the fitted and true curves were 0.11% and 0.03% respectively and, when these parameter estimates were used for the hot mode of operation, the discrepancies became 1.72% and 0.05% respectively. Finally, when the param-

Table 2. Summary statistics for parameter estimates for 1000 simulated samples†

	Results for C- and T-data			*Results for only T-data*		
	κ	τ	a	κ	τ	a
True value	0.4610	0.8330	1.6780	0.4610	0.8330	1.6780
Mean value	0.4610	0.8349	1.6745	0.4613	0.8328	1.6795
Bias standard error	0.0002	0.0004	0.0012	0.0005	0.0005	0.0024
Acutual standard deviation	0.0034	0.0057	0.0188	0.0084	0.0085	0.0377
Estimate (18) standard deviation	0.0035	0.0056	0.0190	0.0088	0.0090	0.0386

†The results are for measurements on both concentration and temperature, and also for temperature measurements only. The estimate of the standard deviation of parameter values is by the δ-method that is usual in NLS analyses.

eters were estimated from only the temperature data, the concentration and temperature discrepancies in cool mode became 0.10% and 0.04% respectively, so using only the quickly and cheaply attainable temperature measurements is sufficient for identifying this system in either mode of operation.

4. Two real data examples

4.1. Modelling nylon production

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

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

where

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

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

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

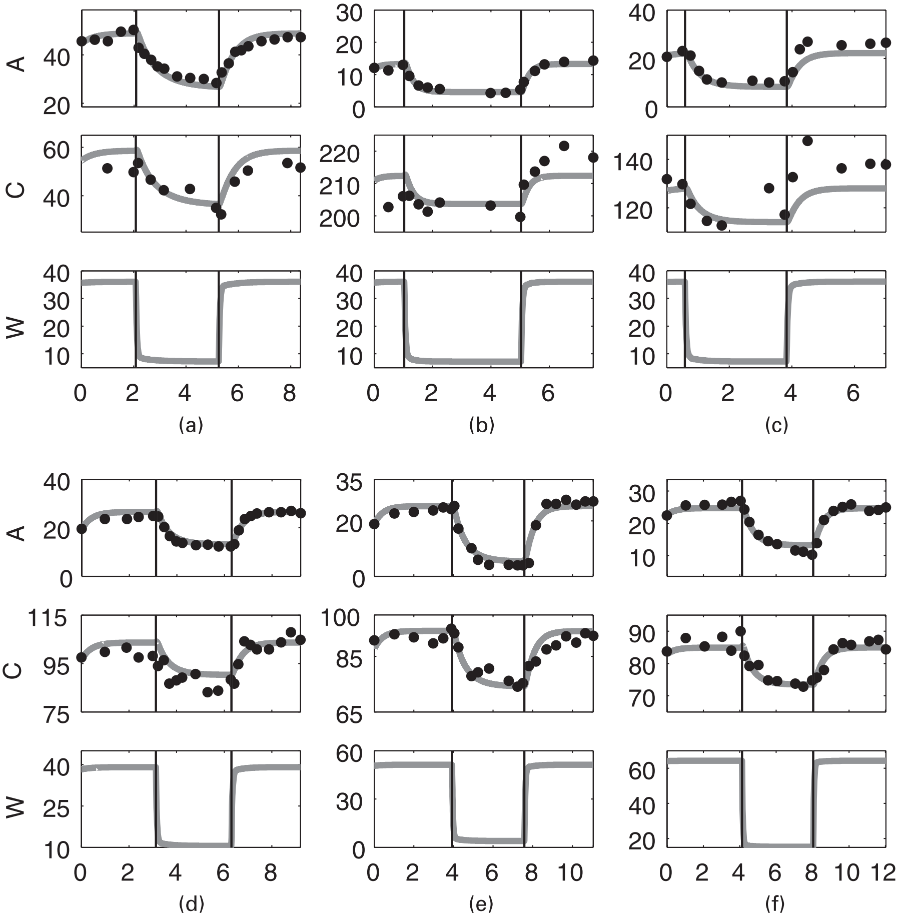


Fig. 7. Nylon components A, C and W along with the solution to the differential equations using initial values estimated by the smooth for each of six experiments (|, times of step change in input pressures; horizontal axes indicate time in hours; vertical axes indicated concentrations in moles): (a) $T=557$ K; (b) $T=557$ K; (c) $T=557$ K; (d) $T=554$ K; (e) $T=544$ K; (f) $T=536$ K

within columns. Since concentrations of A and C are expected to differ only by a vertical shift, their plots within an experimental run are shifted versions of the same vertical spread.

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

The parameter estimates along with 95% limits were $k_p=20.59\pm3.26$, $g=26.86\pm6.82$, $K_{a0}=50.22\pm6.34$ and $\Delta H=-36.46\pm7.57$. The solutions to the differential equations by using the final parameter estimates for $\hat{\boldsymbol{\theta}}$ and the initial system states estimated by the data smooth are shown in Fig. 7.

4.2. *Modelling flare dynamics in lupus*

Lupus is a disease which is 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

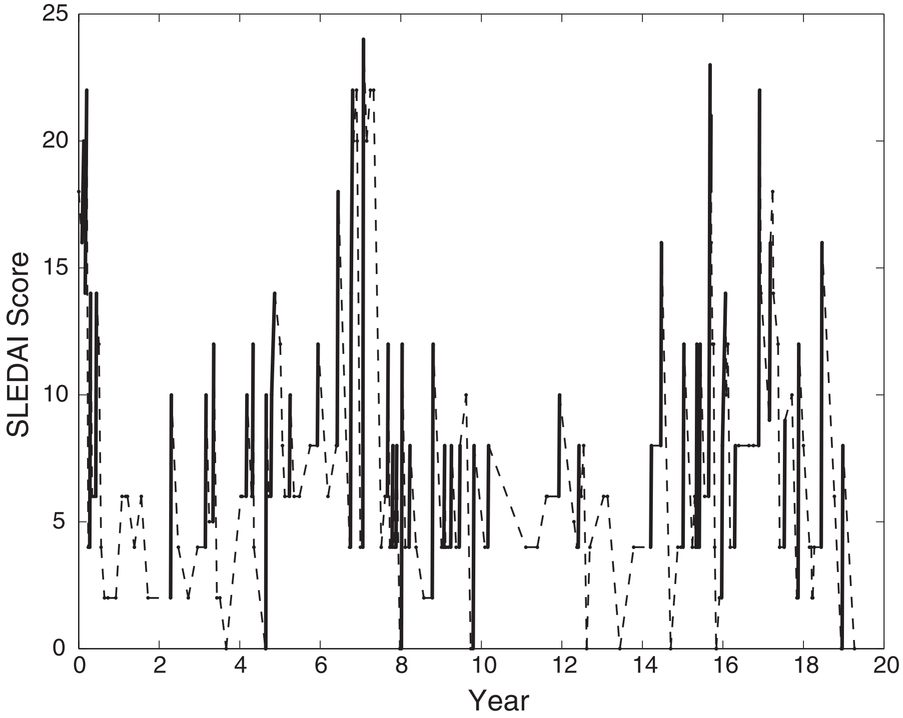


Fig. 8. Symptom level $s(t)$ for a patient suffering from lupus as assessed by the SLEDAI scale: |, changes in SLEDAI score corresponding to a flare; `\`, other changes in SLEDAI score

that is characteristic, but the most serious effects tend to be in the kidneys. The resulting nephritis and other symptoms can require immediate treatment, usually with the drug Prednisone, a corticosteroid that itself has serious long-term side-effects such as osteoporosis.

Various scales have been developed to measure the severity of symptoms, and Fig. 8 shows the course of one of the more popular measures, the systemic lupus erythematosus disease activity index (SLEDAI) scale, for a patient who experienced 48 flares over about 19 years before expiring. A definition of a flare event is commonly agreed to be a change in a scale value of at least 3 with a terminal value of at least 8, and Fig. 8 shows flare events as bold lines.

Because of the rapid onset of symptoms, and because the resulting treatment programme usually involves a SLEDAI assessment and a substantial increase in Prednisone dose, we can pin down the time of a flare with 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 period of onset and the period of recovery. We hope that this model will also show how these short-term flare dynamics interact with longer-term trends in symptom severity.

We postulated that the immune system goes on the attack for a fixed period of δ years, after which it returns to normal function because of treatment or normal recovery. For purposes of this illustration, we took $\delta = 0.02$ years, or about 2 weeks, and represented the time course of attacks as a box function $u(t)$ that is 0 during normal functioning and 1 during a flare.

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

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

and has the solution

$$s(t) = C\, s_0(t) + s_0(t) \int_0^t \alpha(z)\, u(z)/s_0(z)\, \mathrm{d}z$$

where

$$s_0(t) = \exp\left\{ -\int_0^t \beta(z)\, \mathrm{d}z \right\}.$$

Function $\alpha(t)$ tracks the long-term trend in the severity of the disease over the 19 years, and we represented this as a linear combination of eight cubic B-spline basis functions defined by equally spaced knots, with about 3 years between knots. We expected that a flare plays itself out over a much shorter time interval, so $\alpha(t)$ 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 0 with rate $\beta(t)$.

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

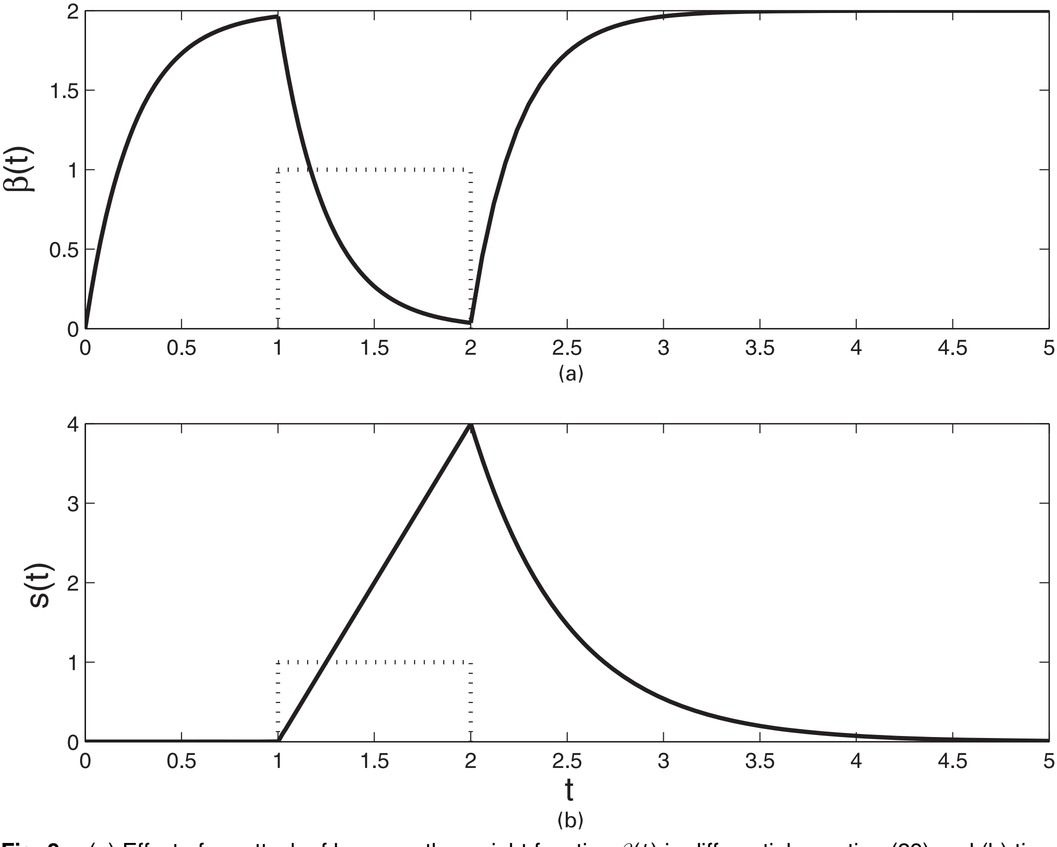


Fig. 9. (a) Effect of an attack of lupus on the weight function $\beta(t)$ in differential equation (29) and (b) time course of the symptom severity function $s(t)$

$\beta(t)$. Consequently, we proposed a second equation to model this mechanism:

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

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

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

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

Fig. 9(a) 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. Fig. 9(b) shows that the increase in symptoms is nearly linear during the period of attack but that, when the attack ceases, the symptom level declines exponentially and takes around 3 time units to return to 0.

When we estimated this model with smoothing parameter value $\lambda = 1$, we obtained the results that are shown in Fig. 10. We found that parameter γ was indeed so high that the fitted symptom rise was effectively linear, so we deleted γ and used the simpler equation (31). This left only the constant θ to estimate for $\beta(t)$, which now controls the rate of decrease of symptoms after an attack ceases. This was estimated to be 1.54, corresponding to a recovery period of about $4/1.54 = 2.6$ years. Fig. 10 shows the variation in $\alpha(t)$ as a broken curve, indicating the long-term change in the intensity of the symptoms, which are especially severe around years 6 and 11, and in the patient's last 3 years.

The fitted function $s(t)$ is shown as a full curve and was defined by positioning three knots at each of the flare onset and offset times to accommodate the sudden break in $\dot{s}(t)$, and a single

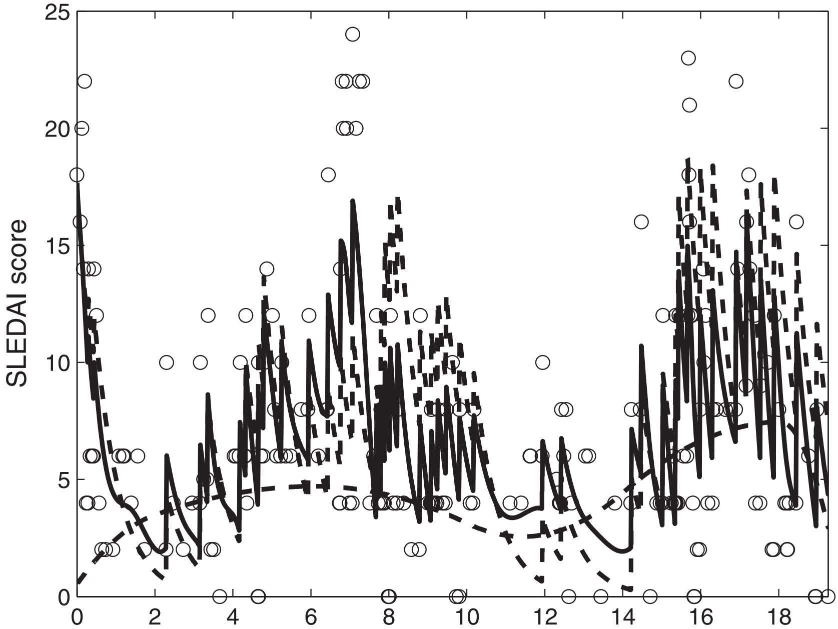


Fig. 10. SLEDAI scores (○), smoothing functions s(*t*) (∧), solution to the differential equation (∧) and smooth trend $\alpha(t)$ (– – –)

knot midway between two flare times. Order 4 B-splines were used, and this corresponded to 290 knot values and 292 basis functions in the expansion $\hat{s}(t)=\mathbf{c}'\,\phi(t)$. We see that the fitted function seems to do a reasonable job of tracking the SLEDAI scores, in both the period during and following an attack and also in terms of its long-term trend.

The model also defines the differential equation (29), and the solution to this equation is shown as a broken line. The discrepancy between the fit that is defined by the equation and the smoothing function $s(t)$ is important in years 8–11, where the equation solution overestimates the symptom level. In this region, new flares come too fast for recovery and thus build on each other. Nevertheless, the fit to the 208 SLEDAI scores that is achieved by an investment of nine structural parameters seems impressive for both the smoothing function $s(t)$ and the 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.

5. Generalizations and further problems

5.1. More general equations

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

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

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

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

(b) lagged equations,

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

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

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

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

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

5.2. Stochastic differential equations

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

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

where $\mathbf{W}(t)$ is d-dimensional Brownian motion. Thus for a fixed $\boldsymbol{\lambda}$, interpreted as the ratio of the variance of the Brownian motion to that of the observational error, the procedure may be thought of as profiling an estimate of the realized Brownian motion. This approach has been used for the problem of data assimilation in Apte *et al.* (2007), where they use criteria that are closely related to our own equation (14). This notion is appealing and suggests the use of alternative smoothing penalties based on the likelihood of other stochastic processes. The flares in the lupus data, for example, could be considered to be triggered by events in a Poisson process, and we expect this to be a fruitful area of future research. However, this interpretation relies on the representation of $\mathrm{d}\mathbf{W}(t)/\mathrm{d}t$ in terms of the discrepancy $\dot{\mathbf{x}}(t) - \mathbf{f}(\mathbf{x}, \mathbf{u}, t|\boldsymbol{\theta})$ where $\mathbf{x}$ is given by a basis expansion (7). For non-linear $\mathbf{f}$ the approximation properties of this discrepancy are not clear. Moreover, frequently a lack of fit in non-linear dynamics is due more to misspecification of the system under consideration than to stochastic inputs, and we are correspondingly wary of this interpretation.

5.3. *Further statistical problems*

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

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

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

6. 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. Our approach, a variety of collocation methods, combines the concepts of *smoothing* and *estimation*, providing a continuum of trade-offs between fitting the data well and fidelity to the hypothesized differential equations. This has been done by defining a fit through a penalized spline criterion for each value of $\boldsymbol{\theta}$ and then estimating $\boldsymbol{\theta}$ through a profiling scheme in which the fit is regarded as a nuisance parameter.

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

On the computational side, the algorithm is as fast as or faster than NLS and other approaches. Unlike Bayesian Markov chain Monte Carlo methods, the generalized profiling approach is relatively straightforward to deploy to a wide range of applications, and software in MATLAB described below merely requires that the user codes the various partial derivatives that are involved, and which are detailed in Appendix A. Finally, the method is also robust in the sense of converging over a wide range of starting parameter values. The possibility of beginning with smaller values of $\boldsymbol{\lambda}$ to work with a smooth criterion, and then stepping these values up towards those defining near approximations to the ODE, further adds to the method's robustness.

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

6.1. *Software*

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

Acknowledgements

The research was supported by grant 320 from the Natural Science and Engineering Research Council of Canada, grant 107553 from the Canadian Institute for Health Research and grant 208683 from Mathematics of Information Technology and Complex Systems to J. O. Ramsay. The authors thank Professor K. McAuley, Professor J. McLellan and Mr Saeed Varziri of the Department of Chemical Engineering at Queen's University for instruction in the language and principles of chemical engineering, many consultations and much useful advice. Appreciation is also due to the referees, whose comments on an earlier version of the paper were invaluable.

Appendix A: Matrix calculations for profiling

The calculations that are 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 straightforward 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:

$$\frac{\mathrm{d}x_i(t)}{\mathrm{d}\mathbf{c}_i} = \phi_i(t)$$

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

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

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

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

$$(\lambda_i v_q)^{1/2}[\dot{x}_i(t_q) - f_i\{\mathbf{x}(t_q), \mathbf{u}(t_q), t_q|\boldsymbol{\theta}\}]\, I(i=j)\, \dot{\phi}_i(t_q) - \left(\sum_{k=1}^{n} (\lambda_i v_q)^{1/2} \frac{\mathrm{d}f_k}{\mathrm{d}x_j} [D x_i(t_q) - f_i\{\mathbf{x}(t_q), \mathbf{u}(t_q), t_q|\boldsymbol{\theta}\}] \right) \phi_j(t_q)$$

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

A.2. Estimating structural parameters

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

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

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

$$\frac{\mathrm{d}\mathbf{c}}{\mathrm{d}\boldsymbol{\theta}}=-\left(\frac{\mathrm{d}^2H}{\mathrm{d}\mathbf{c}^2}\right)^{-1}\frac{\mathrm{d}^2H}{\mathrm{d}\mathbf{c}\,\mathrm{d}\boldsymbol{\theta}}$$

by the implicit function theorem.

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

$$\left\{\sum_{k=1}^{n_i}\phi_{ip}(t)\,\phi_{jq}(t)+\lambda\int\phi_{ip}(t)\,\phi_{jq}(t)\,\mathrm{d}t\right\}I(i=j)-\lambda_i\int\dot{\phi}_{ip}(t)\frac{\mathrm{d}f_i}{\mathrm{d}x_j}\phi_{jq}(t)\,\mathrm{d}t-\lambda_j\int\phi_{ip}(t)\frac{\mathrm{d}f_i}{\mathrm{d}x_j}\dot{\phi}_{jq}(t)\,\mathrm{d}t$$
$$+\int\phi_{ip}(t)\left\{\sum_{k=1}^{n}\lambda_k\left[\frac{\mathrm{d}^2f_k}{\mathrm{d}x_i\,\mathrm{d}x_j}\{f_k-\dot{x}_k(t)\}+\frac{\mathrm{d}f_k}{\mathrm{d}x_i}\frac{\mathrm{d}f_k}{\mathrm{d}x_j}\right]\right\}\phi_{jq}(t)\,\mathrm{d}t$$

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 $\mathrm{d}^2H/\mathrm{d}\mathbf{c}\,\mathrm{d}\boldsymbol{\theta}$ 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

$$\lambda_i\int\frac{\mathrm{d}f_i}{\mathrm{d}\boldsymbol{\theta}}\phi_{ip}(t)\,\mathrm{d}t-\int\left(\sum_{k=1}^{n}\lambda_k\left[\frac{\mathrm{d}^2f_k}{\mathrm{d}x_i\,\mathrm{d}\boldsymbol{\theta}}\{f_k-\dot{x}_k(t)\}+\frac{\mathrm{d}f_k}{\mathrm{d}x_i}\frac{\mathrm{d}f_k}{\mathrm{d}\boldsymbol{\theta}}\right]\right)\phi_{ip}(t)\,\mathrm{d}t.$$

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

The variance of the parameter estimates is calculated by using

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

where

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

and

$$\frac{\mathrm{d}^2H}{\mathrm{d}\boldsymbol{\theta}\,\mathrm{d}\mathbf{y}}=\frac{\partial^2H}{\partial\boldsymbol{\theta}\,\partial\mathbf{y}}+\frac{\partial^2H}{\partial\hat{\mathbf{c}}\,\partial\mathbf{y}}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}}+\frac{\partial^2H}{\partial\boldsymbol{\theta}\,\partial\hat{\mathbf{c}}}\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}+\frac{\partial^2H}{\partial\hat{\mathbf{c}}^2}\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}}+\frac{\partial H}{\partial\hat{\mathbf{c}}}\frac{\partial^2\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}\,\partial\mathbf{y}}.\tag{34}$$

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

(a) $\partial\hat{\mathbf{c}}/\partial\mathbf{y}$: similar to the deduction for $\mathrm{d}\hat{\mathbf{c}}/\mathrm{d}\boldsymbol{\theta}$, we obtain the formula for $\partial\hat{\mathbf{c}}/\partial\mathbf{y}$ by applying the implicit function theorem,

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

(b) $\partial \mathbf{c}^2/\partial \boldsymbol{\theta}\, \partial \mathbf{y}$: by taking the second derivative on both sides of the identity $\partial J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})/\partial \mathbf{c}|_{\hat{\mathbf{c}}}=0$ with respect to $\boldsymbol{\theta}$ and y_k, we derive

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

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

$$\frac{\partial^2 \hat{\mathbf{c}}}{\partial\boldsymbol{\theta}\,\partial y_k}=-\left\{\left.\frac{\partial^2 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2}\right|_{\hat{\mathbf{c}}}\right\}^{-1}\left\{\left.\frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}\,\partial\boldsymbol{\theta}\,\partial y_k}\right|_{\hat{\mathbf{c}}}+\left.\frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}\,\partial\boldsymbol{\theta}\,\partial c_i}\right|_{\hat{\mathbf{c}}}\frac{\partial \hat{c}_i}{\partial y_k}\right.$$
$$\left.+\left.\frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2\partial y_k}\right|_{\hat{\mathbf{c}}}\frac{\partial \hat{\mathbf{c}}}{\partial\boldsymbol{\theta}}+\left.\frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2\,\partial c_i}\right|_{\hat{\mathbf{c}}}\frac{\partial \hat{c}_i}{\partial y_k}\frac{\partial \hat{\mathbf{c}}}{\partial\boldsymbol{\theta}}\right\}. \qquad (37)$$

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

$$\frac{\partial^2 \hat{\mathbf{c}}}{\partial\boldsymbol{\theta}\,\partial \theta_j}=-\left\{\left.\frac{\partial^2 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2}\right|_{\hat{\mathbf{c}}}\right\}^{-1}\left\{\left.\frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}\,\partial\boldsymbol{\theta}\,\partial \theta_j}\right|_{\hat{\mathbf{c}}}+\left.\frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}\,\partial\boldsymbol{\theta}\,\partial c_i}\right|_{\hat{\mathbf{c}}}\frac{\partial \hat{c}_i}{\partial \theta_j}+\left.\frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2\,\partial \theta_j}\right|_{\hat{\mathbf{c}}}\frac{\partial \hat{\mathbf{c}}}{\partial\boldsymbol{\theta}}+\left.\frac{\partial^3 J(\mathbf{c}|\boldsymbol{\theta},\mathbf{y})}{\partial \mathbf{c}^2\,\partial c_i}\right|_{\hat{\mathbf{c}}}\frac{\partial \hat{c}_i}{\partial \theta_j}\frac{\partial \hat{\mathbf{c}}}{\partial\boldsymbol{\theta}}\right\}. \qquad (38)$$

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

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

$$\int\left\{\sum_{l=1}^{n}\lambda_l\left(\frac{\mathrm{d}^2 f_l}{\mathrm{d}x_i\,\mathrm{d}x_j}\frac{\mathrm{d}f_l}{\mathrm{d}x_k}+\frac{\mathrm{d}^2 f_l}{\mathrm{d}x_i\,\mathrm{d}x_k}\frac{\mathrm{d}f_l}{\mathrm{d}x_j}+\frac{\mathrm{d}^2 f_l}{\mathrm{d}x_j\,\mathrm{d}x_k}\frac{\mathrm{d}f_l}{\mathrm{d}x_i}\right)\right\}\phi_{ip}(t)\,\phi_{jq}(t)\,\phi_{ku}(t)\,\mathrm{d}t$$
$$+\int\sum_{l=1}^{n}\lambda_l\frac{\mathrm{d}^3 f_k}{\mathrm{d}x_i\,\mathrm{d}x_j\,\mathrm{d}x_k}\{f_l-\dot{x}_l(t)\}\,\phi_{ip}(t)\,\phi_{jq}(t)\,\phi_{ku}(t)\,\mathrm{d}t-\lambda_i\int\frac{\mathrm{d}^2 f_i}{\mathrm{d}x_j\,\mathrm{d}x_k}\dot{\phi}_{ip}(t)\,\phi_{jq}(t)\,\phi_{ku}(t)\,\mathrm{d}t$$
$$-\lambda_j\int\frac{\mathrm{d}^2 f_j}{\mathrm{d}x_i\,\mathrm{d}x_k}\phi_{ip}(t)\,\dot{\phi}_{jq}(t)\,\phi_{ku}(t)\,\mathrm{d}t-\lambda_k\int\frac{\mathrm{d}^2 f_k}{\mathrm{d}x_i\,\mathrm{d}x_j}\phi_{ip}(t)\,\phi_{jq}(t)\,\dot{\phi}_{ku}(t)\,\mathrm{d}t$$

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

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

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

$$\int\left\{\sum_{l=1}^{n}\lambda_l\left(\frac{\mathrm{d}^2 f_l}{\mathrm{d}\theta_i\,\mathrm{d}x_j}\frac{\mathrm{d}f_l}{\mathrm{d}x_k}+\frac{\mathrm{d}^2 f_l}{\mathrm{d}\theta_i\,\mathrm{d}x_k}\frac{\mathrm{d}f_l}{\mathrm{d}x_j}+\frac{\mathrm{d}^2 f_l}{\mathrm{d}x_j\,\mathrm{d}x_k}\frac{\mathrm{d}f_l}{\mathrm{d}\theta_i}\right)\right\}\phi_{kp}(t)\,\phi_{jq}(t)\,\mathrm{d}t$$
$$+\int\sum_{l=1}^{n}\lambda_l\frac{\mathrm{d}^3 f_k}{\mathrm{d}x_j\,\mathrm{d}x_k\,\mathrm{d}\theta_i}\{f_l-\dot{x}_l(t)\}\,\phi_{kp}(t)\phi_{jq}(t)\,\mathrm{d}t-\lambda_j\int\frac{\mathrm{d}^2 f_j}{\mathrm{d}\theta_i\,\mathrm{d}x_k}\dot{\phi}_{jq}(t)\,\phi_{kp}(t)\,\mathrm{d}t$$
$$-\lambda_k\int\frac{\mathrm{d}^2 f_k}{\mathrm{d}\theta_i\,\mathrm{d}x_j}\phi_{jq}(t)\,\dot{\phi}_{kp}(t)\,\mathrm{d}t$$

where c_p corresponds to the basis representation of x_k.

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

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

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

$$-\left(\frac{\partial^2 J}{\partial\mathbf{c}^2}\right)^{-1}\frac{\partial^2 J}{\partial\mathbf{c}\,\partial\mathbf{y}}$$

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

References

Apte, A., Hairer, M., Stuart, A. M. and Voss, J. (2007) Sampling the posterior: an approach to non-gaussian data assymilation. *Physica* D, **230**, 50–64.

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

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

Bauer, I., Bock, H. G., Körkel, S. and Schlöder, J. P. (2000) Numerical methods for optimum experimental design in DAE systsems. *J. Computnl Appl. Math.*, **120**, 1–25.

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

Biegler, L. and Grossman, I. (2004) Retrospective on optimization. *Comput. Chem. Engng*, **28**, 1169–1192.

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

Campbell, D. (2007) Bayesian collocation tempering and generalized profiling for estimation of parameters from differential equation models. *PhD Thesis*. McGill University, Montreal.

Cao, J. and Ramsay, J. O. (2006) Parameter cascades and profiling in functional data analysis. *Computnl Statist.*, to be published.

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

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

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

Esposito, W. R. and Floudas, C. (2000) Deterministic global optimization in nonlinear optimal control problems. *J. Glob. Optimizn*, **17**, 97–126.

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

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

Gelman, A., Bois, F. Y. and Jiang, J. (1996) Physiological pharamacokinetic analysis using population modeling and informative prior distributions. *J. Am. Statist. Ass.*, **91**, 1400–1412.

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

Hooker, G. (2007) Theorems and calculations for smoothing-based profiled estimation of differential equations. *Technical Report BU-1671-M*. Department of Biostatistics and Computational Biology, Cornell University, Ithaca.

Jaeger, J., Blagov, M., Kosman, D., Kolsov, K., Manu, Myasnikova, E., Surkova, S., Vanario-Alonso, C., Samsonova, M., Sharp, D. and Reinitz, J. (2004) Dynamical analysis of regulatory interactions in the gap gene system of drosophila melanogaster. *Genetics*, no. 167, 1721–1737.

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

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

Nagumo, J. S., Arimoto, S. and Yoshizawa, S. (1962) An active pulse transmission line simulating a nerve axon. *Proc. Inst. Radio Engrs*, **50**, 2061–2070.

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

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

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

Tjoa, I.-B. and Biegler, L. (1991) Simultaneous solution and optimization strategies for parameter estimation of differential-algebraic equation systems. *Industrl Engng Chem. Res.*, **30**, 376–385.

Van Keilegom, I. and Carroll, R. J. (2006) Backfitting versus profiling in general criterion functions. Submitted to *Statist. Sin.*

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

Wahba, G. (1990) *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.

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

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

Discussion on the paper by Ramsay, Hooker, Campbell and Cao

Arne Kovac (*University of Bristol*)

Estimation of parameters of an ordinary differential equation (ODE) from noisy data is an exciting area and we have to thank the authors for bringing this challenging problem to our attention. One reason why I think that this topic is so interesting is that many applications in science and engineering employ differential equations to model relationships between variables and one of the strengths of this paper is to share so many examples. Another reason is that it gives rise to a difficult optimization problem where the target function is usually not convex and can have many local minima. Finally given how natural the desire is to determine suitable values for the parameters of an ODE it is the more surprising that this topic is relatively unexplored.

Although the ‘discovery’ of this problem is certainly the highlight of this paper, the particular approach that is followed by the authors and the use of regularization in this context are another interesting contribution. Traditionally used to balance smoothness and closeness to data, regularization has recently also been used to estimate monotone functions (Ramsay, 1998), to obtain simple approximations without artificial local extrema (Davies and Kovac, 2001) and to select parameters in linear regression (Tibshirani, 1996). In this paper the authors use a new penalty that penalizes departure from solving the ODE to make an otherwise difficult optimization problem much easier to solve. We have to thank the authors for not only providing an explicit algorithm but also for making their implementation publicly available.

One of many interesting questions is whether it is possible to assess the goodness of fit. If there were no noise at all we would just have to solve the ODE for the given set of parameters and to check whether the solution coincides with the data. With noise present and/or departures from the idealistic model this is more difficult. A set of parameters may be regarded as a good model if the residuals r_i look like noise and one way of checking this is to look at their sums on different scales and locations,

$$w_{j,k} = \frac{1}{\surd(k-j+1)} \sum_{i=j}^{k} r_i, \qquad 1 \leqslant j \leqslant k \leqslant n,$$

and to verify whether these are all sufficiently small, i.e. $|w_{j,k}| < \surd\{2\ \log(n)\}\hat{\sigma}$ where $\hat{\sigma}$ is some estimate of the noise level. Fig. 11 shows data from the FitzHugh–Nagumo ODE with an approximation from a slightly different model. Visual inspection shows hardly any lack of approximation; however, some $w_{j,k}$ exceeded the threshold. In contrast, the corresponding solution for the true value $c=3$ would have been accepted by the multiresolution criterion.

It is not quite clear to me whether we should calculate the residuals with respect to the solution to the ODE by using the parameter estimates or the aproximations $\hat{x}_i$ from the regularization problem. These functions may considerably differ if λ is small. How do we then interpret the parameter estimates given that the data do not follow the trajectory of the ODE? Do we estimate parameters at all?

Another challenging problem is how to deal with possible changes over time. Is there one global set of parameters that provides a good model for all of the data? And, if this is not possible, how would one estimate the parameters locally? A partial answer may be given again by the multiresolution criterion that was sketched above. We could try to devise an algorithm that aims to find parameter values such that as many of the coefficients $w_{j,k}$ as possible are below the threshold. If for any set of parameters coefficients $w_{j,k}$ exist which exceed the threshold, a local version needs to be determined. Fig. 12 shows another 401 data simulated from the FitzHugh–Nagumo ODE where the parameters changed after the first half,

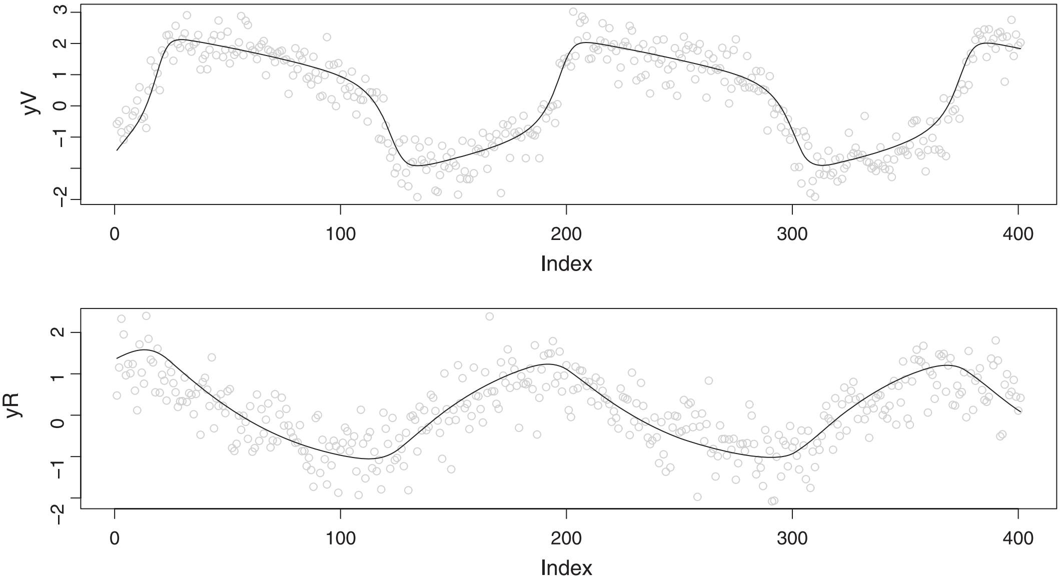


Fig. 11. FitzHugh–Nagumo data by using the values from Section 3 and the solution for $\lambda = 1000$ and $a = b = 0.2$ as in the true model, but $c = 2.2$ instead of 3

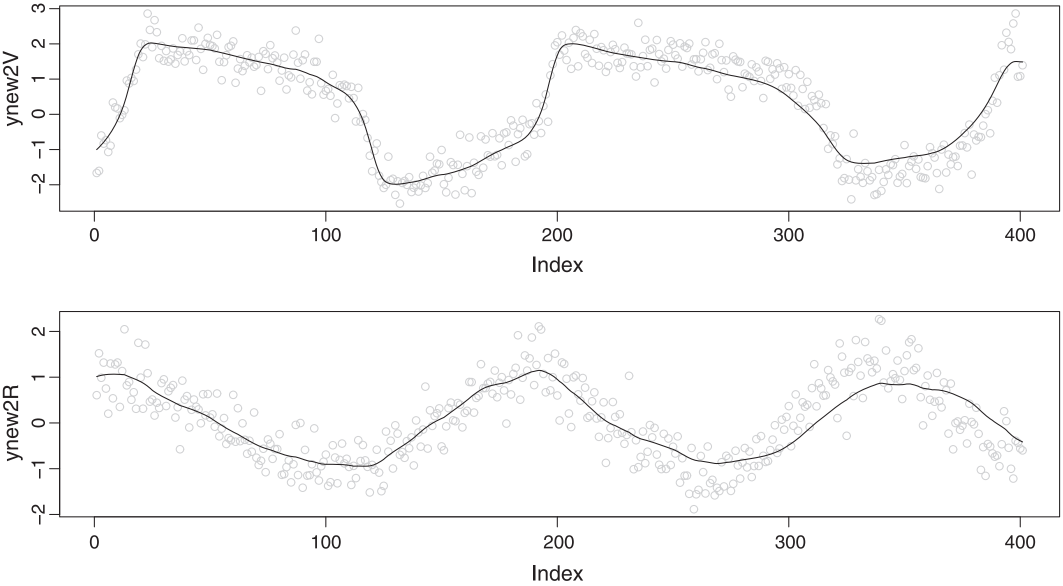


Fig. 12. FitzHugh–Nagumo data by using $\sigma = 0.4$ and the parameters as in Section 3 for the left-hand half, but using $a = 0.6$ and $c = 2$ in the right-hand half (———, solution for $a = b = 0.2$, $c = 3$ and $\lambda = 0.01$)

but where the solution was calculated globally by using the true parameters from the first half. For V all $w_{j,k}$ with $k \leqslant 152$ were below the threshold and for R even all $w_{j,k}$ with $k \leqslant 174$. Thus the multiresolution criterion clearly indicates that the approximation is adequate for at least the first 150 data points, but that a different approximation is needed for the second half.

Further questions include statements about rates of convergence, whether there is any use in a local choice of λ and whether L_1-penalties would offer any improvements when the functions have discontinuities. I am convinced that this paper will stimulate plenty of research and consequently I have great pleasure in proposing the vote of thanks.

S. Olhede (*Imperial College London*)
I congratulate the authors on their thought provoking contribution to the estimation of parameters of ordinary differential equations (ODEs). This is an important and currently much neglected area of statistics.

The main innovation of this paper is the attempt to combine various measures of misfit into a coherent likelihood framework, so that the parameters of a system of ODEs, denoted $\boldsymbol{\theta}$, can be estimated. For simplicity I in this discussion take $N=\min_i(N_i)$. As the ODEs cannot be solved numerically for each posited value of $\boldsymbol{\theta}$, the solutions are approximated by using B-spline bases $\{\phi_{ik}(t)\}_{k=1}^{K_i}$ (see Varah (1982)), by

$$\hat{x}_i(t)=\sum_{k=1}^{K_i}\hat{c}_{ik}(\boldsymbol{\theta},\mathbf{x}_0)\;\phi_{ik}(t).$$

I am concerned that the authors do not give an automated criterion for the selection of K_i. Once the number of measurements increases (Mendes *et al.* (2003) already have used eight coupled ODEs) choosing K_i on the basis of a qualitative assessment for each output variable of the data set will become infeasible.

Approximating the true solutions introduces two measures of disparity: the deviation of the B-spline approximation from the data, $e_{ij}=y_{ij}-\hat{x}_i(t_{ij})$ (data misfit), and the deviation of the spline approximation from solutions to the ODE, $x_i(t_{ij})-\hat{x}_i(t_{ij})$ (model misfit). The data misfit can be computed directly once some criterion has been determined for proposing $\hat{c}_{ik}(\boldsymbol{\theta},\mathbf{x}_0)$, but the model misfit is not available. Instead the deviation of $\hat{x}_i'(t_{ij})$ from $f_i(\mathbf{x},\mathbf{u},t_{ij}|\boldsymbol{\theta})$ is used to measure model misfit.

I would like to note that methods of combining the measures of misfit will vary in suitability dependent on the inference problem that is attacked. If we only seek to estimate $\boldsymbol{\theta}$ then it does not matter whether $\hat{\mathbf{x}}_N(t)\rightarrow\mathbf{x}(t)$ as $N\rightarrow\infty$ but only that

$$\frac{\partial}{\partial\theta_k}\|\hat{\mathbf{x}}_N(t)-\mathbf{x}(t)\|_\beta$$

becomes negligible with increasing N. If we are interested in *prediction* of the output variables, then this fact changes.

The most important component of the procedure is the choice of regularization parameters $\{\lambda_i\}_{i=1}^d$, and the norms that are chosen for the data and model misfits, which are denoted by α and γ respectively. For large N with $\alpha=2$ some further remarks can be made. Unless the data are very strongly correlated in time, $H(\boldsymbol{\theta}|\boldsymbol{\lambda})=O(N)$. I point out that $K_i=K_i(N)=O(N)$ and $\lambda_i=\lambda_i(N)$. To provide consistent large sample theory $\text{PEN}_i(\hat{\mathbf{x}})=\|\hat{x}_i'(t)-f_i(\mathbf{x},\mathbf{u},t|\boldsymbol{\theta})\|_\gamma=O(N^{-\delta})$ for some $\delta>0$ must be imposed. The choice of δ determines the rate of convergence of the approximation to the true solution. To ensure that the approximation of $\mathbf{x}(t)$ becomes exact for increasing sample sizes, a condition such as

$$\lim_{N\rightarrow\infty}\{N^{-1}\sum_i\lambda_i(N)\;\text{PEN}_i(\hat{x})\}=C<\infty \tag{38}$$

must be imposed. With smoothness assumptions on $x_i(t)$, or equivalently on $f_i(\mathbf{x},\mathbf{u},t|\boldsymbol{\theta})$, $\text{PEN}_i(\hat{x})$ can be bounded with $K_i(N)$ sufficiently large, in powers of N. This combines with a large sample argument for determining the optimal order of $\lambda_i(N)$. If we take K_i order $O(N)$ then $\lambda_i(N)=O(N^{1+\delta})$ for suitable δ will for large N ensure that $\hat{\mathbf{x}}_N(t)\rightarrow\mathbf{x}(t)$, in some suitable sense. Furthermore by taking $\lambda_i(N)\rightarrow\infty$ as $N\rightarrow\infty$ the results of theorems 1 and 2 gain additional interpretability. Clearly $\lambda_i(N)$ should be chosen to ensure the asymptotic efficiency of $\hat{\boldsymbol{\theta}}_N$. Decreasing $\lambda_i(N)$ appears to increase the variance of the estimators (see the curvature of Figs 2 and 5, and note the change of axes.)

To ensure the existence of a 'good' solution, we need to take $K_i(N)$ sufficiently large. Arguments to confirm the existence of the solution for a specific γ, $\lambda_i(N)$ and $K_i(N)$ can be made. $\lambda_i(N)$ might also be chosen to account for how informative (sensitive) $x_i(t)$ is to $\boldsymbol{\theta}$.

The specification of $\lambda_i(N)$ needs to be automated. For method 1 proposed on page 753,

(a) how do we know whether the first minimum is appropriate,
(b) can random variability due to the errors cause many minima and order mixing of minima and
(c) is there a strict theoretical justification for this procedure?

Method 2 proposed on page 754 is speculative and appears to underestimate the size of the regularization parameter. I think that, if a semiparametric model is appropriate, relevant assumptions must be made about the deviation of the derivatives of the sample paths from the ODE. Another possible approach to the problem is to combine the model misfit with the data misfit by using a Bayesian formulation of the problem; see Wahba (1978). In this case the variability of each sample path of $\mathbf{x}(t)$ needs to be modelled. Many alternatives to a Brownian coupled set of stochastic differential equations are available. Wahba and Wang (1990) have discussed issues with the usage of generalized cross-validation for the selection of regularization parameters. Certainly the choice of loss function should be approached with some care and should be linked to the inferential problem that is addressed. Neither proposed automated procedure was actually used for the simulated data or the nylon example.

Permitting a very weak norm for the measure of the deviation of $\hat{x}_i'(t)$ from $f_i(\mathbf{x}, \mathbf{u}, t_{ij}|\boldsymbol{\theta})$ will lead to large deviations in $\hat{x}_i(t)$ from $x_i(t)$, as large deviations will aggregate once $\hat{x}_i'(t)$ has been integrated. Thus, despite the recent popularity of the $\gamma=1$ norm in signal processing, I do *not* suggest usage of this norm for the penalty, but I would rather advocate $\gamma=\infty$. With an appropriate basis expansion an l_1-penalty on $\mathbf{c}$ may be appropriate; see the Danzig selector (Candès and Tao, 2005).

Another issue which is glossed over by the authors is model checking. The distribution of the error terms determines $H(\boldsymbol{\theta}|\boldsymbol{\lambda})$. The residuals should be checked for serial correlation, which appears to be present in the nylon data set residuals; see Fig. 13(a). Determining the second-order structure of e_{ij} is equally important to specifying an appropriate choice of regularization parameter. A simple autoregressive length 1 model seems to explain the serial correlation; see Fig. 13(b). The time sampling is not evenly spaced; hence models such as autoregressive processes may not always be appropriate. Non-parametric methods such as runs tests could be employed to test for serial correlation of the residuals; see for example Mood (1940).

There are issues with usage of profile likelihood: see the discussion in Berger *et al.* (1999), and note that unfortunate 'ridge maximization' may ensue. With the observed ripples in the likelihood such effects may lead to unfortunate properties of the procedure. Some care must therefore be taken with the profile maximization.

I have outlined some questions with regard to the performance of the methods proposed. A very adventurous step has been taken by the authors to construct a coherent likelihood framework for inference of systems of ODEs. Numerous modelling, consistency and fitting issues remain, as inevitably will be the case when boldly embarking on a new area of inference: I am very pleased to join Dr Arne Kovac in thanking the authors for their innovative and challenging paper.

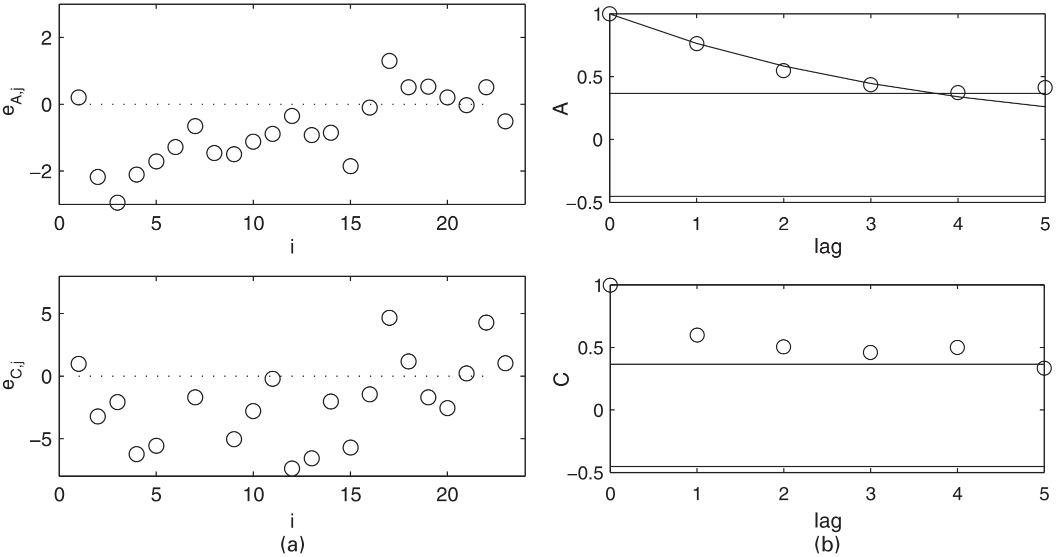


Fig. 13. (a) Residuals from the nylon data from run 2 (the serial correlation is apparent from the residuals) and (b) non-parametric estimated autocovariance of the residuals (assuming stationarity) for run 2 (———, critical region based on approximations to the distribution of the estimated autocorrelation under H_0 of no correlation with a pointwise level of 0.05 at each lag; an AR(1) fit to the first set of residuals is also included as the curve in the top panel)

The vote of thanks was passed by acclamation.

Steven M. Boker (*University of Virginia, Charlottesville*)
I congratulate Ramsay and his colleagues on a stimulating paper that addresses a problem that has been long considered important. 80 years ago Hotelling (1927) wrote of the difficulty of estimating differential equations in the presence of error. When data are sampled from real systems and a model is estimated, this error can be divided into at least three parts: a part that is associated with the measurement instrument itself, a part that is associated with exogenous influences which propagate in time and a part that is associated with inadequacy of the model to account for the relationships between the time derivatives of the system. Separation of these sources of error from signal while simultaneously estimating parameters of a system and providing goodness-of-fit estimates for the chosen model allows model comparison. These goals are particularly problematic when the system is non-linear and realizations of the system may diverge exponentially.

Some widely used methods for parameter estimation of differential equations are variants of Kalman filtering or Kalman smoothing (Kalman, 1960; Molenaar and Newell, 2003) and methods from stochastic differential equations (Itô, 1951; Bergstrom, 1966; Singer, 1993). These forward prediction methods operate on the integral form and thus require analytic solutions to the chosen system of differential equations. One advantage of the method that is outlined in this paper is that it estimates the parameters of the differential equations directly and thus does not require the analytic solution, which for non-linear systems may be unknown. A second interesting feature of the method is that it allows separate cost functions for the equation and error parameters.

There are three practical problems that I see arising when using the approach of Ramsay and his colleagues. The first concerns the choice of the smoothing complexity parameter λ, some potential solutions to which the paper covers. The second problem is that it is unclear how the separation of time-independent and time-dependent error is to be accomplished such that solution uniqueness is obtained given that the smoothness λ must also be chosen. Perhaps a latent variable form of the differential equation in question could be specified if multivariate indicators were available for each variable (Boker *et al.*, 2004). The third problem arises when the model structure is unknown: by what metric are we to perform model comparison given the flexibility of this method? Some penalty for lack of parsimony might unify solutions to these three problems. I do not see these problems as insurmountable and I hope that Ramsay and colleagues will consider them in hopes of widening the applicability of their interesting work.

Leonard Smith (*London School of Economics and Political Science*)
The paper is an important contribution to parameter estimation in non-linear systems of ordinary differential equations. We lack a general coherent theory here, despite important applications ranging from the small scale industrial processes that are discussed in the paper to informing decision support in climate change (Stainforth *et al.*, 2005). I thank the authors for the chance to suggest links between their work and approaches from non-linear dynamics, as the geometric–dynamics view provides a complementary perspective on parameter estimation which might allow

(a) better estimation when the model structure is exact and the non-linearities are non-trivial,
(b) improvement in model structure when it is known to be imperfect and
(c) clarification of the role of stochastic dynamics.

Imperfections in the model structure reopen the question of which parameter values should be used when elements of the parameter vector Θ *are* known from 'theoretical considerations of other sources of information'. Even in artificial cases where the model structure and the observational noise model are known exactly, traditional approaches like least squares are likely to prove unsatisfactory, as even normally distributed input uncertainties yield outputs under the model which are not normally distributed (Judd, 2007; McSharry and Smith, 2004).

In practice we are never in that perfect model scenario; the goal of parameter estimation, and indeed state estimation, is not only unclear but also unlikely to have a single well-posed definition (Smith, 2000; Judd and Smith, 2004). Focusing on information from the dynamics rather than focusing on the statistics abandons one notion of optimality for the goal of improved consistency. One simply asks whether the model admits trajectories that are consistent with the observations (Judd *et al.*, 2004). The distribution of the durations of shadowing orbits allows parameter estimation, provides a structured approach to estimating Ramsay's λ and locates regions of the model state space where the system dynamics are systematically inconsistent with those of the model (McSharry and Smith, 2004). When shadowing trajectories cannot be found, we can examine the mismatch 'errors' of pseudo-orbits that are consistent with the observations. This has the dual aims of model improvement and of developing stochastic models which are more likely to yield useful trajectories (Judd and Smith, 2004). These models are not, however, the traditional form of stochastic models: the innovations reflect the geometric failings of the model flow in model state space and aim to allow for the attracting manifolds that are common in non-linear dissipative models. Ideally the innovation distribution will be state dependent and perhaps path dependent. The clear formulation of such truly non-linear stochastic models which respect the geometrical dynamics of the model and observations of the underlying system would prove of great value in refining the models that Ramsay and his colleagues now provide us with.

Steven Gilmour (*Queen Mary, University of London*)
Parameter estimation for differential equations is a topic of enormous importance and applicability, which requires much more attention from statisticians. I welcome this paper which addresses the problem from one particular viewpoint, which seems to work rather well. My own interests are in the design of experiments which will enable the parameters to be estimated efficiently.

At a simple level, a design could be chosen which optimizes some function of var$\{\hat{\boldsymbol{\theta}}(\mathbf{y})\}$, as given in equation (24). Usually we would have to integrate over prior distributions for $\boldsymbol{\theta}$ and λ, so this is a far from trivial task.

However, it is important that we get the basics correct. The classical principles of good design have a role to play in complex experiments, which is at least as important as their role in simple text-book experiments. The tank reactor experiment that is described in Section 1.2.2 and illustrated in Fig. 3 is typical of many experiments on dynamical systems. It is not obvious even how to describe it in classical terms. We need to identify

(a) the treatments—combinations of the levels of six factors, F_{in}, C_{in}, T_{in}, T_{co}, F_{co} and the base-line T_{co},
(b) the experimental units—these seem to be runs of the process of length $t=4$—and
(c) the responses from each experimental unit–time series of C and T.

Then the design, using a standard coding, is shown in Table 3.

This design is poor on several counts: there is no randomization, no sensibly chosen replication, no blocking (so long-term drifts will have systematic effects), no protection of experimental units (it might be sensible to exclude the first part of the time series on each unit), no use of the factorial treatment structure (so interactions cannot be estimated) and a failure to recognize multiple strata (base-line T_{co} is a whole-plot factor). Such a poor design would be useless even for simple responses.

Also important are the implications of the design for the analysis. The concept of experimental units is always meaningful and implies a discreteness, even in dynamical systems. Each time that the system is

Table 3. Design of the tank reactor experiment

Base	F_{in}	C_{in}	T_{in}	T_{co}	F_{co}	*Base*	F_{in}	C_{in}	T_{in}	T_{co}	F_{co}
−1	0	0	0	0	0	1	0	0	0	0	0
−1	−1	0	0	0	0	1	−1	0	0	0	0
−1	1	0	0	0	0	1	1	0	0	0	0
−1	0	0	0	0	0	1	0	0	0	0	0
−1	0	−1	0	0	0	1	0	−1	0	0	0
−1	0	1	0	0	0	1	0	1	0	0	0
−1	0	0	0	0	0	1	0	0	0	0	0
−1	0	0	−1	0	0	1	0	0	−1	0	0
−1	0	0	1	0	0	1	0	0	1	0	0
−1	0	0	0	0	0	1	0	0	0	0	0
−1	0	0	0	−1	0	1	0	0	0	−1	0
−1	0	0	0	1	0	1	0	0	0	1	0
−1	0	0	0	0	0	1	0	0	0	0	0
−1	0	0	0	0	−1	1	0	0	0	0	−1
−1	0	0	0	0	1	1	0	0	0	0	1
−1	0	0	0	0	0	1	0	0	0	0	0

disturbed by changing the level of a factor a new, discrete, error is introduced (e.g. through small uncertainties in setting the levels) and so the model should contain random unit effects.

The following contributions were received in writing after the meeting.

Caroline Bampfylde (*University of Alberta, Edmonton*)
I thank Ramsay, Hooker, Campbell and Cao for their contribution to the practicalities of fitting dynamical models to data and estimating model parameters. This is a task which is commonly encountered by applied scientists and the rigorous solution technique that is provided by this manuscript is most welcome.

Although Ramsay and co-authors make efficient use of matrix algebra to simplify the calculation of the derivatives in Appendix A, the resulting formulae that are presented seem to be overly complicated. I am concerned that the implementation of their methods is non-trivial, especially for many applied scientists whose interest is in the results and their application rather than the details of the method. However, I do applaud the publishing of on-line materials providing open source software and numerical code to facilitate the implementation of their techniques. It appears that the Web site http://www.functionaldata.org needs to be updated to reflect the new statistical techniques and to present some examples that users can then modify to fit to their own problems.

I should like to end my discussion with thoughts about the wider applications of the authors' techniques. The methods have thus far been applied to systems of ordinary differential equations. Would it be possible to consider the extension to discrete time dynamical systems such as systems of difference equations? In my research I have to deal with dynamical systems both continuous and discrete in nature and a consistent technique for parameter estimation would be very useful. Have the authors considered the application to partial differential equations and integrodifferential equations which are regularly used for spatial problems? Any further extensions or generality that can be derived from their methods would be a great addition to the parameter estimation toolbox.

My thanks go to the Research Section of the Royal Statistical Society, for the opportunity to contribute to the discussion of this important paper.

Lorenz Biegler (*Carnegie Mellon University, Pittsburgh*)
It is a pleasure to comment on this paper. I found this paper very informative and useful and my comments are mostly from an optimization algorithm perspective. The approach that is mentioned in the paper complements strategies for dynamic optimization but specializes them with interesting statistical concepts and problem formulations.

On page 750, using the total variation penalty in equation (12) may be advantageous, although it is not used in the analysis or the examples. Although the necessary smoothness conditions are absent, the finite dimensional analogue to equation (12) is actually preferred over equation (11) because only a finite value of λ is needed to satisfy theorem 2. Also, equation (11) has the disadvantage that λ must approach ∞ to force PEN$(\cdot)$ to zero, thus leading to severe ill conditioning in the optimization. Some discussion on these numerical aspects (and possible improvements) can be found in chapter 17 of Nocedal and Wright (2006).

Sections 3 and 4 contain excellent examples that illustrate the benefits of the approach in Section 2 and also show how they apply to real world data. In the second paragraph of page 760, it should be mentioned that a Runge–Kutta *initial* value algorithm was used. The failure for this unstable system is due to this single shooting approach. Instead, if the instrumental variables were replaced with corresponding (dichotomous) boundary conditions, and the solver replaced by a corresponding boundary value solver (e.g. COLSYS or COLDAE; see Ascher and Petzold (1998)), the problem should also solve easily, just as the principal differential analysis method does. A method to do this along with a pathological parameter estimation problem is given in Tanartkit and Biegler (1995, 1996).

Section 5 is very useful in exploring future topics. More detail could be added in several areas. The exploration of differential algebraic equations (DAEs) has been done for some time and DAE systems have now been well studied and understood for parameter estimation. Ascher and Petzold (1998) provided a comprehensive discussion and summary of these systems. Many practical systems can be written as index 1 DAEs (or can be reformulated as index 1 DAEs). For these, much of the discussion in this paper could be extended directly.

Partial differential equation constrained optimization enjoys considerable current research attention and several approaches have been explored that are relevant to principal differential analysis. The authors might find Biegler *et al.* (2003) useful. Finally, for future work I think that the greatest potential of this approach is for stochastic systems. I look forward to further developments in this area with this approach.

Emery N. Brown (*Massachusetts Institute of Technology, Cambridge, and Harvard Medical School, Boston*)
It remains to be clearly established what the methods of Ramsay and colleagues add to current methods for differential equation model analyses.

The Ramsay analyses provide no comparisons with existing methods for analysing differential equation models. Therefore, the current work tells neither the dynamicist nor the statistician what if any improvements the new approach brings. For example, if a likelihood-based analysis had been applied to the nylon production problem as was done for the circadian data in Brown (1987) and Brown *et al.* (2000) what improvements would the new methods have provided? The likelihood-based analysis that was used in those references estimated non-linear dynamical systems models from observations with very strong serial dependence, computed confidence intervals for parameter estimates and demonstrated that estimation with approximate and exact solutions of the differential equation system gave similar answers.

Dynamical systems often have specific properties such as Hopf bifurcations, limit cycles and chaotic dynamics. Inferring these specific properties from experimental data is a fundamental question in dynamical systems analyses (Czeisler *et al.*, 1989, 1999; Diks, 1999). The authors give no evidence to show that their approach would allow dynamicists to determine whether particular types of dynamic behaviour can be more reliably determined from experimental data analyses by using their methods compared with current methods.

Another fundamental question in many dynamical systems analyses in neuroscience is how to model the stochastic features of a given neural system. The smoothness constraint should reflect specific hypotheses about the stochastic features of the dynamical system. The smoothness constraint in the Ramsay analyses represents an explicit (mathematically convenient) assumption about the stochastic features of the dynamical system. It does not relate to any specific hypothesis about the physical, chemical or biological origins of the stochastic features of the systems that are studied in their examples.

A dynamical system with noise in its observation process and/or its system equation falls naturally into the state space and the partially observed systems framework. The authors miss an important opportunity to relate their work to these established paradigms.

The FitzHugh–Nagumo example does not provide a true illustration of the issues that computational neuroscientists address in relating dynamical systems models to experimental data. Time courses of actual subthreshold membrane voltage potentials of single neurons (what the FitzHugh–Nagumo model is intended to characterize) are recorded by many neurophysiologists. Estimating the dynamic properties of these data is a challenging problem being investigated by many computational neuroscientists (Koch, 2001). Would the methods of Ramsay and colleagues outperform current methods in the study of this problem?

Sy-Miin Chow and Stacey S. Tiberio (*University of Notre Dame*)
The authors are to be congratulated for providing a comprehensive treatment of using smoothing methods to fit non-linear ordinary differential equation models. We particularly like the proposed approach's ease of use with irregularly spaced discrete time observations, and the authors' discussion on its diagnostic utility. We believe that the method proposed can be effectively integrated with recent advances in fitting non-linear, non-Gaussian state space models. In particular, we ask the authors to consider a non-linear continuous time state space model of the form

$$\mathrm{d}x_i(t) = f\{x_i(t), t|\theta\}\,\mathrm{d}t + \mathrm{d}w_i(t), \tag{39}$$

$$y_{ij} = h_i\{x_i(t_j), t_{ij}|\theta\} + e_{ij}, \tag{40}$$

where f is a non-linear drift function, h is a (possibly) non-linear measurement function, $w_i(t)$ is a Wiener (or possibly other dynamic noise) process and e_j is a vector of measurement errors.

If basis function expansion is used to obtain smoothed estimates of equation (39), the log-likelihood function $H(\theta, \sigma|\lambda)$ can then be written as a function of the innovations $e_{ij} = y_{ij} - h_i\{\hat{c}_i(\sigma_i, \theta; \lambda)'\,\phi(t_{ij})\}$. Along a similar line, $H(\theta, \sigma|\lambda)$ and the penalty function can then be used as the basis for assessing misfits stemming from the dynamic model and the measurement model (equations (39) and (40)) respectively. In addition, we do see some merits in incorporating process noise in the dynamic model in equation (39) in addition to allowing for non-Gaussian measurement processes (e.g. Poisson processes; Durbin and Koopman (2001) and Fahrmeir and Tutz (1994)) in equation (40). For instance, serially independent measurement errors can play a very different role from that of dynamic noises that do show continuity over time. More research along this line is certainly warranted. Some of the recent continuous time adaptations of Monte Carlo techniques (e.g. Beskos *et al.* (2006) and Särkkä (2006)) may also be a helpful alternative or addition to the generalized smoothing approach.

Our remaining comments are mainly questions to help to pave future extensions along this line. We wonder whether the authors can comment on the relationship between the complexity of the basis functions and the choice of the smoothing parameter λ. Specifically, how does the role of λ change if the numbers of knot points and basis functions that are used are overfitting compared with underfitting the data, especially when sample sizes are small to moderate? Furthermore, if an ordinary differential equation model is fitted to data with mild process noise, can the model misspecification be partially compensated by, for example, using more basis functions to construct $\hat{x}(t)$?

Sophie Donnet and Adeline Samson (*University Paris Descartes*)
When a biological or physical process is measured, the regression function of the statistical model describing the observed data often derives from dynamic systems based on ordinary differential equations (ODEs). The differential system often does not have any analytical solution, leaving only the combination of estimation procedures and discretization schemes to solve the ODE. As an alternative to addressing the various problems that are involved in these methods—computational time and stability—the authors suggest an original and efficient solution. Their method relies on a basis function expansion of the dynamic process and then consists of data fitting and an equation fidelity criterion combined in a penalized log-likelihood.

We shall now stress the numerous qualities of the method. First, the fact that no discretization scheme is used to solve the problem makes it possible to consider boundary or/and distributed data problems and, most of all, side-steps the instability problems that are involved with non-continuous input functions. These discontinuous functions are common in biology or physics and constitute a major limit to the use of classical discretization schemes such as Euler or Runge–Kutta schemes in estimation algorithms. Moreover, this method seems robust to the starting parameter values, which is often a concern with a non-linear least squares approach. Furthermore, the authors provide explicit expressions for the derivatives, allowing the use of an efficient Gauss–Newton algorithm. Finally, one of the major advances of this paper is the fact that it provides accurate estimations of the confidence intervals of the estimated parameters.

Obviously, this work opens many new perspectives in the active research field of the estimation in ODE models. As stressed by the authors, many extensions can be considered. Firstly, in biology, experimental studies often consist of repeated measurements of a biological criterion obtained from a population of subjects. The statistical parametric approach that is commonly used to analyse these data is mixed models. The extension of the estimation method that is proposed by Ramsay and his colleagues to mixed models would be an interesting alternative to classical methods that are based on discretization schemes. Secondly, it would be of considerable interest to develop such a method for stochastic differential equations, which are a natural extension of the models that are defined by ODEs, as it allows taking into account errors that are associated with misspecifications and approximations in the dynamic system.

Michael Dowd (*Dalhousie University, Halifax*)
My congratulations go to the authors for their interesting and topical study. Rigorous statistical examination of estimation problems for systems that are governed by differential equations (DEs) is important and timely. Such models are the theoretical foundation for many scientific fields and the synthesis of dynamical models and data is a pressing issue, e.g. for data assimilation (Lewis *et al.*, 2006).

This study offers a unique approach to parameter estimation for DEs by using a weak constraint formulation and exploiting the functional nature of the system state. The 'parameter cascade' appears an effective strategy for estimating different types of parameter. It offers a viable alternative to non-linear regression (Thompson *et al.*, 2000).

The paper also emphasizes the importance of identifying efficient and effective methods for parameter and state estimation for stochastic (and partial) DEs. An approach that supports these extensions directly is the state space model. It treats partially observable non-linear stochastic dynamics and multivariate non-Gaussian observations according to

$$x_t \sim p(x_t|x_{t-1}, \theta),$$
$$y_t \sim p(y_t|x_t, \phi).$$

The first equation describes the Markovian transition of the state x_t, with parameters θ. This corresponds to stochastic dynamic prediction using discretized DEs, i.e. $x_t = f(x_{t-1}, n_t, \theta)$. Observations y_t can be related to x_t through a non-linear measurement operator with ϕ being parameters of the measurement distribution. In Dowd (2006), I applied such a model for complex non-linear dynamical systems to recover state and dynamic parameters for a system which regularly transitioned across a bifurcation point.

The problem that is considered by this paper is the estimation of static parameters. Given the observation set as $y_{1:T}=(y_1',\ldots,y_T')'$ and using Bayes's theorem yield the target density for the state, $p(x_\tau|y_{1:T},\theta,\phi)$. This can be computed with sampling-based sequential Monte Carlo (MC) techniques (Künsch, 2005; Godsill *et al.*, 2004).

Unknown parameters θ and ϕ can be then be determined by maximizing the likelihood (see Kitagawa (1996)):

$$\begin{aligned} L(\theta,\phi|y_{1:T}) &= \prod_{t=1}^{T} \int p(y_t|x_t,\phi)\, p(x_t|y_{1:t-1},\theta)\,\mathrm{d}x_t \\ &\approx \prod_{t=1}^{T} \left\{ N_t^{-1} \sum_{i=1}^{N_t} p(y_t|x_{t|t-1}^{(i)},\theta,\phi) \right\} \end{aligned}$$

where the latter approximation relies on $\{x_{t|t-1}^{(i)}\}, i=1,\ldots,N_t$, which is a sample from the predictive density generated via sequential MC sampling. The resultant likelihood is affected by MC sampling variability and challenges optimizers; incorporation of kernel density estimation appears useful (de Valpine, 2004).

Computationally, application of these MC approaches to higher dimensional dynamic systems is a major challenge. Ideas based on dynamical analysis (Chorin and Krause, 2004) and effective approximations, e.g. the ensemble Kalman filter (Evensen, 2003), appear promising. I have compared some of these methods for non-linear dynamic systems (Dowd, 2007). It would be interesting to compare these further with the parameter estimation method of this paper, extended to the case of stochastic DEs.

David J. D. Earn (*McMaster University, Hamilton*)
Estimation of parameters of non-linear differential equations from noisy, observed time series is a problem that arises frequently in applied science. Unfortunately, anyone who has tried this is likely to be familiar with serious theoretical and computational challenges. The new method of Ramsay and colleagues is very welcome, and it will be interesting to see how it fares on a wide range of problems.

The method may prove particularly useful for the study of transmission dynamics of infectious diseases (Anderson and May, 1991). The state variables in the basic *susceptible–infectious–recovered* (*SIR*) *model* are the numbers of individuals who are susceptible (S), infectious (I) and recovered or immune (R), and the parameters are the rates of birth (ν), death (μ), transmission (β) and recovery (γ):

$$\dot{S}=\nu-(\beta I+\mu)S, \tag{41a}$$

$$\dot{I}=\beta IS-(\gamma+\mu)I, \tag{41b}$$

$$\dot{R}=\gamma I-\mu R. \tag{41c}$$

Note that I records the *prevalence* of the disease, i.e. the number of individuals who are currently infected. We typically observe *incidence*, i.e. $\int \beta SI\,\mathrm{d}t$, where the integral is over the reporting interval (typically weekly or monthly, but sometimes daily).

For human diseases that have been present in the population for years, we typically have estimates of all the parameters from data other than time series of reported cases or deaths. Moreover, the SIR model as formulated in equation (41) has a globally asympotically stable equilibrium, so we can easily compare the predicted equilibrium with the observed times series (without the aid of Ramsay and colleagues).

The catch is that the transmission rate β is rarely constant in practice. Instead, β often varies seasonally, either because of seasonally changing aggregation patterns (London and Yorke, 1973) or other seasonal factors that may be difficult to pin down (Dushoff *et al.*, 2004). Seasonal forcing drastically changes the dynamics of the SIR model, often leading to co-existing stable cycles (Schwartz and Smith, 1983) or chaos (Schaffer, 1985). Since the conclusions that we draw depend strongly on the estimated amplitude of seasonal forcing (Earn *et al.*, 2000; Bauch and Earn, 2003), we need a credible way of estimating time variation in β and we rarely have useful data to work with other than incidence time series. The method of Ramsay and colleagues is begging to be applied to this problem and I look forward to comparing the results that are inferred from it and from previous methods (Fine and Clarkson, 1982; Ellner *et al.*, 1998; Finkenstädt and Grenfell, 2000; Bjornstad *et al.*, 2002; Wallinga and Teunis, 2004).

Finally, it is worth mentioning a vexing issue that has the potential to undermine parameter estimation for differential equation models of disease spread. The process of infectious disease transmission is fundamentally stochastic. Solutions of the SIR model (41) can be thought of as ensemble means of the true stochastic process (Kurtz, 1980), but any incidence time series represents only one realization of that stochastic process and may not accurately reflect the mean. In the specific context that I have highlighted—estimating a seasonal forcing function—this problem may not be serious if we have data covering many seasons, but it is worth bearing in mind.

Stephen P. Ellner (*Cornell University, Ithaca*)
I congratulate the authors for two important contributions: the profiling method and for highlighting to the statistical community the problems of fitting dynamical systems models. Non-linear differential equations are core models in many sciences, including my own discipline of ecology, but are sorely neglected in statistical research (Ellner and Guckenheimer, 2006). In ecology, low dimensional non-linear dynamic models that would have been called a caricature or metaphor 20 years ago have proved remarkably successful in confrontations with real data (e.g. Zimmer (1999) and Turchin (2003)). These models necessarily leave out many 'inessentials' (rare species, spatial variability, etc.) and are often deterministic even though we know better. Omitted inessentials are problematic for non-linear systems because they can have a large effect on long-term model trajectories even if their effect at any instant is small. Wanting $f\{\hat{x}(t),\theta\}$ to be near $\mathrm{d}\hat{x}(t)/\mathrm{d}t$, where $\hat{x}(t)$ is near the data, is a more reasonable hope for a model with the right 'essentials'.

But for practical acceptance I believe that selection of the smoothing parameter λ must be automated on a defensible basis. The profiling criterion immediately suggests cross-validation. Straight leave-one-out methods are computationally infeasible for end-users (though computer and algorithmic improvements may change this situation), but we can still use the principle of predicting something that was not used in fitting. Dynamic models predict the future, so we can evaluate them on the basis of forecasting accuracy. Let $\phi_t(x_0;\theta)$ be the model solution at time t starting from $x(0)=x_0$. A measure of prediction error at time interval τ is

$$\mathrm{PE}(\lambda;\tau)=\sum_j \|y(t_j)-\varphi_\tau\{\hat{x}_\lambda(t_j-\tau);\hat{\theta}_\lambda\}\|^2. \tag{42}$$

PE should be large if $\hat{x}_\lambda$ undersmooths or oversmooths the data, either way throwing off parameter estimates. I tried this criterion on the FitzHugh–Nagumo system (modifying MATLAB code that was provided by Hooker), with the omitted inessential being an additive perturbation to $\mathrm{d}V/\mathrm{d}t$ (Fig. 14(a)) that changes the period of the oscillations (Fig. 14(b) *versus* Fig. 14(c)). Fitting five artificial data sets by profiling with a range of λ-values, PE selects a range of λ-values that is good for parameter estimation (Figs 14(d) and 14(e)). Profiling with a 'good' λ performs comparably with two-step methods in which the data are smoothed without regard to the model, and the ordinary differential equation is then fitted to the smooth or its time derivative; with a 'bad' λ profiling is less successful. Profiling's big advantage over two-step methods is that it does not need data on all state variables but, as this small example indicates, success may depend on choosing λ well.

Chong Gu (*Purdue University, West Lafayette*)
The authors are to be congratulated for a fine paper on a challenging problem. As shown in the paper, fitting data to models derived from ordinary differential equations (ODEs) involves numerous issues such as the numerical strategies and the methodological framework, and it is the methodological aspects that we shall comment on.

First let us attempt a crude parallel between the setting of the paper and a standard cubic spline as the minimizer of

$$\sum_{i=1}^n \{y_i-x(u_i)\}^2+\lambda\int\left(\frac{\mathrm{d}^2x}{\mathrm{d}u^2}\right)^2\mathrm{d}u. \tag{43}$$

Setting $\lambda=\infty$ in expression (43) forces $\mathrm{d}^2x/\mathrm{d}u^2=0$ that characterizes a *static system*, with the solution of the form $x(u)=c_1+c_2u$, where (c_1,c_2) are to be determined by the data (y_i,u_i) through the least squares; if precise readings of (u,x) are available from $x(u)=c_1+c_2u$, we need only two pairs of 'initial values' to pin down (c_1,c_2). Likewise, replacing $\mathrm{pen}(x)=\int(\mathrm{d}^2x/\mathrm{d}u^2)^2\,\mathrm{d}u$ in expression (43) by $\mathrm{pen}(x)=\int(\mathrm{d}^2x/\mathrm{d}u^2+\omega^2x)^2\,\mathrm{d}u$ yields an L-spline, and setting $\lambda=\infty$ then forces $\mathrm{d}^2x/\mathrm{d}u^2+\omega^2=0$ with the solution of the form $x(u)=c_1\sin(\omega u)+c_2\cos(\omega u)$. Compare these with

$$\sum_{i=1}^n \{y_i-x(t_i)\}^2+\lambda\int\left\{\frac{\mathrm{d}x}{\mathrm{d}t}-f(x,\mathbf{u},t|\boldsymbol{\theta})\right\}^2\mathrm{d}t, \tag{44}$$

where for simplicity we consider only a single ODE. The main difference between expressions (43) and (44) is the time variable t in expression (44) and the implicit dependence of x on u. The system parameter

$\boldsymbol{\theta}$ is absent for the cubic spline and is the period ω for the L-spline. Setting $\lambda=\infty$ in expression (44) forces $\mathrm{d}x/\mathrm{d}t - f(x,\mathbf{u},t|\boldsymbol{\theta})=0$ with the solution of the form $x_u(t;\boldsymbol{\theta},\mathbf{c})$, say, and the parameters $\boldsymbol{\theta}$ and $\mathbf{c}$ may be fixed via least squares as in expression (44) or through alternative 'initial values'.

As crude as the parallel is, it sheds light into the roles of various components in the proposed setting. For data smoothing via expression (43) or the like, the stochastic structure of the data is typically well

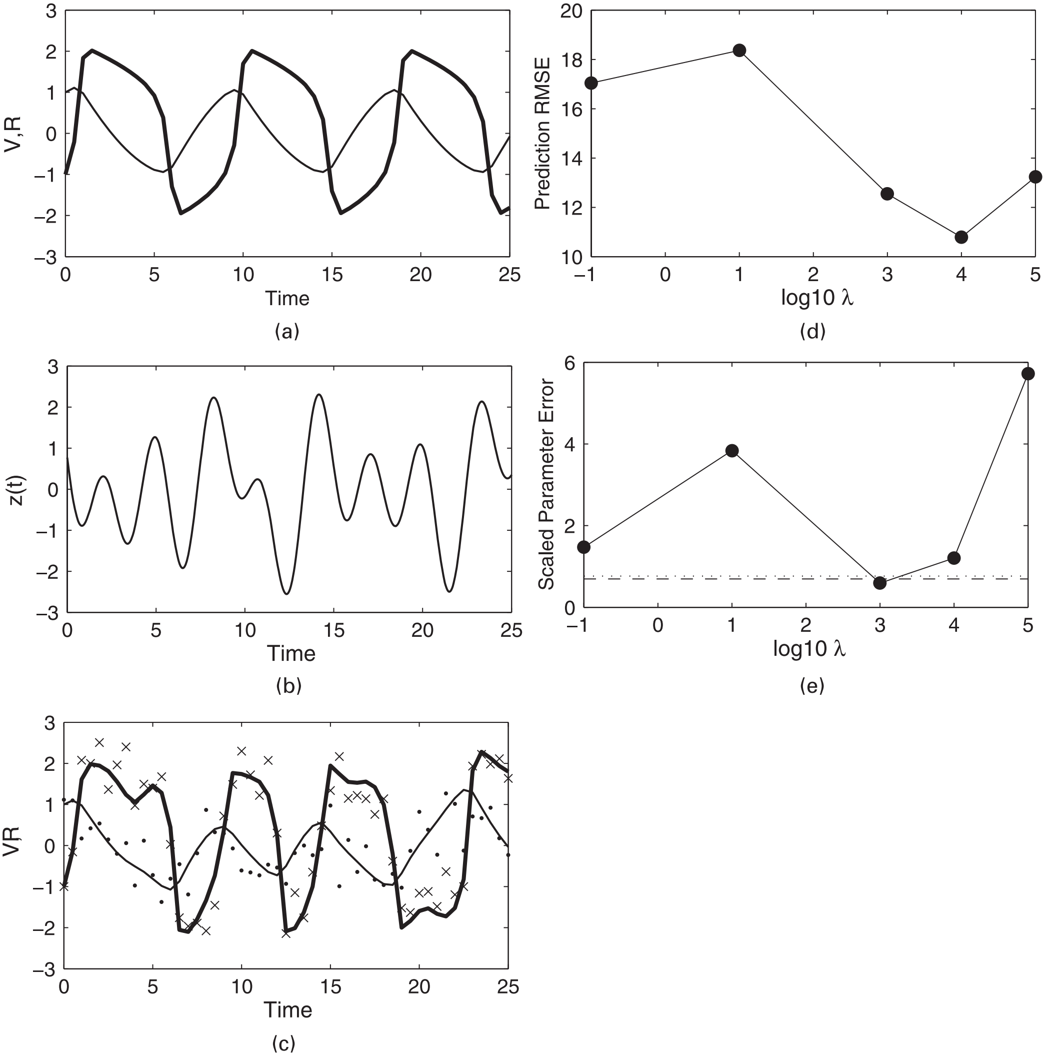


Fig. 14. Parameter estimation by profiling for the Fitzhugh–Nagumo model $\dot{V}=c(V-V^3/3-R)$, $\dot{R}=-(V-a+bR)/c$: (a) solution trajectories with parameters $(a,\ b,\ c)=(0.2,\ 0.2,\ 3)$ (the thicker curve is $\dot{V}(t)$); (b) time varying perturbation $z(t)$ added to $\dot{V}$; (c) solution trajectories and one of the artificial 'data sets' generated by the perturbed model $\dot{V}=c(V-V^3/3-R)+z(t)$, $\dot{R}=-(V-a+bR)/c$; (d) root-mean-square prediction error (averaging over five artificial data sets) as a function of the value of λ used when fitting the data by profiling, for prediction horizon $\tau=5$ time units; (e) mean scaled parameter error, averaging across the same five artificial data sets (scaled parameter error = |estimated value – true value|/(true value); ———, mean scaled parameter error from two-step methods; each data set was first smoothed without reference to the differential equation, using gam from the mgcv package in R (R Core Development Team, 2006; Wood, 2006); parameters were estimated by 'gradient matching' (·······, Ellner *et al.* (2002)), i.e. fitting the right-hand side of the model to the time derivative of the smooth by non-linear least squares, or by minimizing equation (42) with $y(t_j)$ and $\hat{x}(t_j-\tau)$ both given by the smooth and $\tau=5$ (-------))

specified, whereas the roughness penalty pen(x) is virtually an afterthought mainly to provide 'stability' to the end results, and one is more than willing to 'warp' the function away from the 'null model' characterized by pen$(x)=0$ to fit the data. For solutions to the dynamic systems, however, the roles of goodness of fit and 'roughness penalty' seems more likely reversed, with fidelity to the ODE the major concern and the 'error distribution' of the data an afterthought. With such an understanding, automatic λ-selection via cross-validation may not be the most appropriate for expression (44); cross-validation was designed to minimize the estimation error for data smoothing, $\Sigma_{i=1}^{n}\{\hat{x}(u_i)-x(u_i)\}^2$ in the setting of expression (43) with $y_i=x(u_i)+\varepsilon_i$. Instead, a manual selection of λ that keeps $\int\{\mathrm{d}x/\mathrm{d}t-f(x,\mathbf{u},t|\boldsymbol{\theta})\}^2\,\mathrm{d}t\leqslant\rho$, say, for some prespecified tolerance level ρ, might be more appropriate.

John Guckenheimer (*Cornell University, Ithaca*) **and Joseph Tien** (*Fred Hutchinson Cancer Research Center, Seattle*)

A key issue in parameter estimation problems for differential equations is minimizing residual functions with optimization algorithms. As illustrated in Fig. 2, the graph of the residual as a function of the parameters may be so convoluted that smooth optimization algorithms that are based on quadratic models require initial parameter values that are very close to the optimal values. Ramsay and his colleagues smooth the residual by a spline fit, together with a penalty on discrepancies between the fitted curve to solutions of the differential equations.

Our work also introduces residual functions which involve penalties, but we focus on the relationships between qualitative properties of the differential equation solutions to the geometry of the response surface. Those relationships prompt us to propose new residual functions that incorporate geometric features of the dynamical system and simplify the landscape. Examples of these geometric features include periodic orbits, bifurcation boundaries and fast–slow decompositions of multiple-timescale solution trajectories.

This paper represents solutions of differential equations through their initial values. When these solutions depend sensitively on initial values or system parameters, the residual function has large gradients. This is evident in Fig. 2, showing a residual function for solutions to the FitzHugh–Nagumo equation fitted over approximately 2.5 periods of an oscillatory solution. Since the oscillation period varies with the system parameters, the residual is more sensitive when evaluated for longer time intervals. If the data to be fitted are at its periodic asymptotic state, we suggest fitting the periodic orbit of the model to the data instead of the solution of an initial value problem. This approach was developed by Casey (2004). Matching the period of a periodic orbit to its measured period is a step towards solving the parameter estimation problem. Furthermore, the 'cliff' in the response surface of Fig. 2 suggests that there is a bifurcation of the model at these parameter values. Bifurcation boundaries in the model form natural constraints of the 'reasonable' parameter region for fitting attractors to stationary data. We advocate using computations of bifurcation boundaries in this context.

In multiple-timescale systems, abrupt changes in solutions occur due to changes in the transitions between slow and fast segments of solutions. The geometry of fast–slow decompositions of solution trajectories can be used to define residual functions for both non-periodic and periodic solutions (Tien and Guckenheimer, 2007; Tien, 2007).

Serge Guillas (*Georgia Institute of Technology, Atlanta*)

I congratulate the authors for their paper. They have introduced a technique for the estimation of parameters for differential equations that is fast and precise. Unlike many smoothing situations, the large range (e.g. several orders of magnitude in the FitzHugh–Nagumo equations) of good λ is quite surprising. The analysis for which the authors examine the asymptotic behaviour of the estimates when $\lambda\rightarrow\infty$ is very helpful and rarely done in traditional smoothing settings. It would be interesting to study further the range of values of λ that give accurate estimates.

The authors mention Bayesian methods as an alternative to their method. In this framework, the numerical solution to the differential equation at each sample time point is assumed to be normally distributed, with the use of the Metropolis–Hastings algorithm. In the more general context of complex computer models, two approaches have been recently developed to take into account the functional form of the output better. For well-chosen designs for the parameters, and sufficient computing power, these methods are efficient and robust, in particular if there is no complete knowledge of the set of differential equations. Higdon *et al.* (2007) represent a functional output through a principal components analysis. Bayarri *et al.* (2007) considered a decomposition of the time series of outputs in a wavelet basis. Wavelets can easily model abrupt changes in the outputs. This could be helpful for a better understanding of certain types of solutions to differential equations. Calibration can then be directly carried out on the coefficients themselves following a traditional approach (Kennedy and O'Hagan, 2001). These formulations may improve the estimation of the parameters in the case where complicated noise and biases are present. The additional discrepancy term can accommodate biases that depend on the initial conditions. Also the Bayesian approach naturally leads to an assessment of the uncertainties. Combining Gaussian processes and information from derivatives is also possible (O'Hagan, 1992; Morris *et al.*, 1993; Mitchell *et al.*, 1994; Solak *et al.*, 2003).

Jianhua Huang (*Texas A&M University, College Station*) **and Yanyuan Ma** (*Université de Neuchâtel*)
We are glad to have the opportunity to discuss this stimulating and exciting paper. We tried to approach the problem from the viewpoint of familiar M-estimation. To simplify the notation and to focus on the main idea, consider the case of only one equation. The penalized least squares criterion function is

$$L(\mathbf{c},\boldsymbol{\theta})=\sum_j\{y_j-\mathbf{c}'\phi(t_j)\}^2+\lambda\int[\mathbf{c}'\dot{\phi}(t)-f\{\mathbf{c}'\phi(t),t|\boldsymbol{\theta}\}]^2\,\mathrm{d}t,$$

minimization of which for fixed λ gives a joint estimation of $\mathbf{c}$ and $\boldsymbol{\theta}$. Potential overfitting of the data that is caused by a high dimensional parameter $\mathbf{c}$ is avoided owing to the second term in the criterion function, where a large λ can reduce significantly the effective dimension of $\mathbf{c}$.

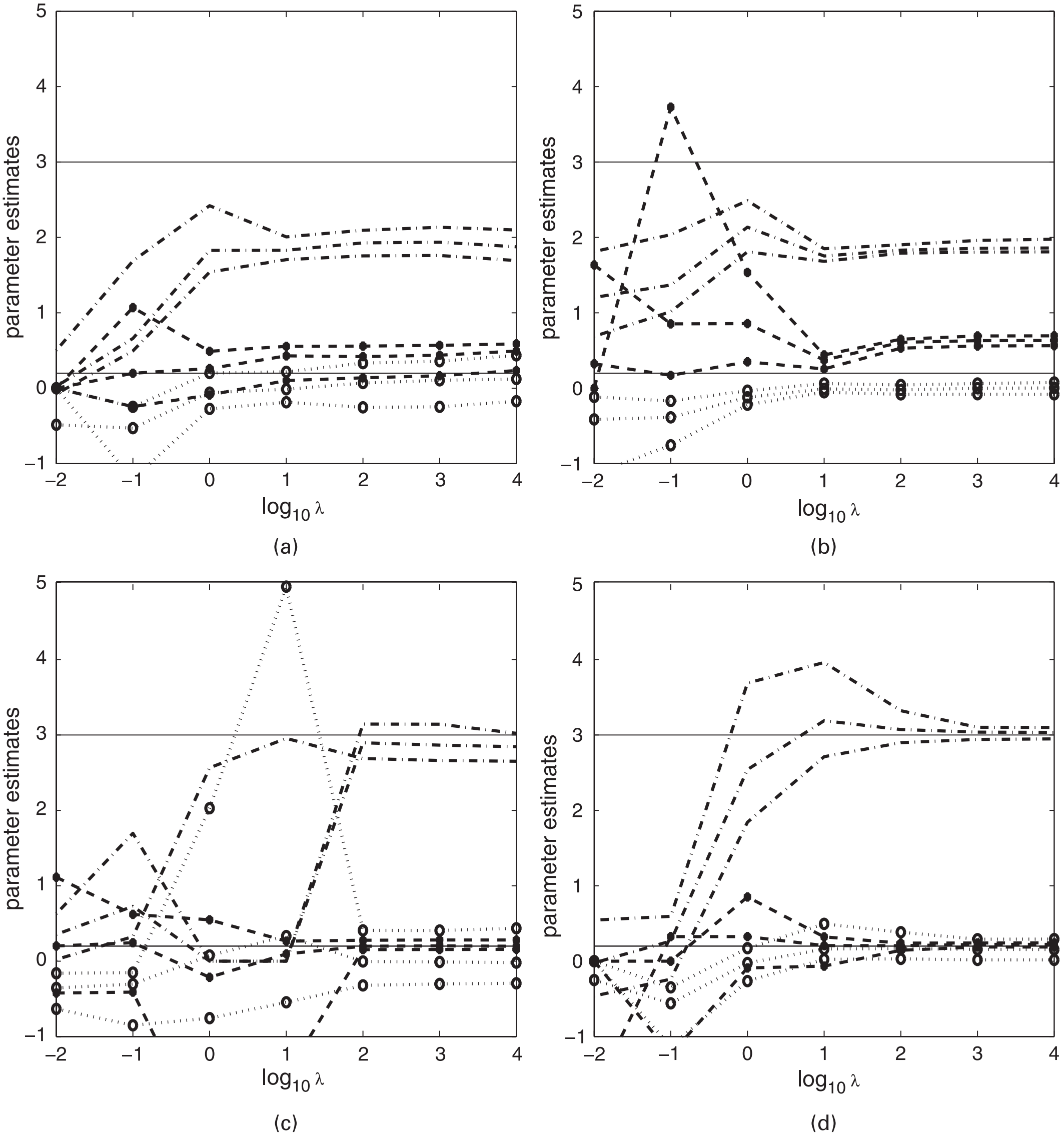
There are several computational approaches to solving the optimization problem—joint optimization of $\mathbf{c}$ and $\boldsymbol{\theta}$, backfitting or profiling. The method that is proposed corresponds to a variation of the profiling method that drops the second term in $L(\mathbf{c},\boldsymbol{\theta})$ when optimizing $L\{\hat{\mathbf{c}}(\boldsymbol{\theta}),\boldsymbol{\theta}\}$ in the second step. Dropping the second term simplifies computation and can be justified when the differential equation is satisfied or approximately so, which is the case when a very large penalty parameter λ is used. The M-estimation formulation also enables us to approximate the sampling variation of $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{c}}$ easily, without relying on the implicit function theorem. The authors could probably give us some insight on advantages of the proposed multicriterion approach in comparison with the more direct approach here.

Next we report some results from an experiment on fitting the FitzHugh–Nagumo equations in example 3.1 by using the software that was kindly provided by the authors. Motivated by many real biomedical data sets where only a sparse sample is available, we considered a sparse sampling of the profiles of V with only 21 observations. Fig. 15(a) shows that the parameter estimate can be seriously biased. However, when we reran the program but increased the number of bases in the collocation method to 10 times the sample size, we obtained reasonable estimation, as shown in Fig. 15(c). This prompted us to believe that the number of bases that are used in collocation should be decided by the essential nature of the ordinary differential equation instead of just the number of observations. Our belief is reinforced by the results from using 21 and 201 basis functions to fit data sampled at 201 time points as given in Figs 15(b) and 15(d). Our finding indicates an important difference between pure data smoothing and smoothing in parameter estimation for ordinary differential equations.

Edward L. Ionides (*University of Michigan, Ann Arbor*)
The authors are to be congratulated for their elegant approach to reconciling mechanistic dynamic models with time series data. Their methodology appears to be readily applicable to a range of challenging inference problems. I would like to compare and contrast the deterministic dynamic modelling approach, which was adopted by the authors, with a stochastic dynamic modelling approach. For the sake of discussion, ordinary differential equations (ODEs) can be compared with stochastic differential equations (SDEs), though similar considerations will apply to other models, such as Markov chains.

A drawback of the authors' method is that the fitted model is not readily apparent. One may be led to interpret the fitted model as an ODE with parameter vector $\hat{\theta}$, but of course the trajectories that are fitted to the data do not perfectly follow this ODE. There is allowance for some deviation, which is controlled by the parameter λ, and this deviation may be important for both the qualitative and the quantitative behaviour of the system. The differences between stochastic dynamic models and their approximating ODEs, which is termed the 'deterministic skeleton' of the model, have been found to be relevant in ecological systems (Coulson *et al.*, 2004). One related issue is, how should trajectories be simulated from the fitted model? In the context of the tank reactor, for example, it would seem desirable if the variability between simulated trajectories were comparable with variability between replications of the experiment. Additionally, such simulated trajectories should be available to a researcher who is aware of only the reported values of $\hat{\theta}$ and λ.

One way around these difficulties is to consider the equivalent SDE, which is given by the authors in Section 5.2, as the fitted model. The authors are reluctant to do this since 'lack of fit in non-linear dynamics is due more to misspecification of the system under consideration than to stochastic inputs'. I would argue



(a)

(b)

(c)

(d)

Fig. 15. 25%, 50% and 75% quantiles of the parameter estimates for the FitzHugh–Nagumo equations (the experiment is conducted by using equally spaced observations of V in the time range from 0 to 10, with different numbers of knots and observations; the parameter values, the initial conditions and the level of noise of the observed data are the same as those in example 3.1 of the paper; the starting values of the parameters are the true values plus random noise with the standard deviation equal to 20% of the parameter values; ———, truth; -◆-, *a* (0.2); ··○··, *b* (0.2); ·-·-·-·, *c* (3)): (a) 21 observations, 21 knots; (b) 201 observations, 21 knots; (c) 21 observations, 201 knots; (d) 201 obervations, 201 knots

that it should be acceptable to interpret the noise as model misspecification combined with random variation; such interpretations are certainly routine in linear regression, for example. Quite general methods exist for carrying out inference in the context of partially observed non-linear SDE systems (Ionides *et al.*, 2006). However, the authors' penalized spline approach has considerable computational advantages that should motivate future work into clarifying the relationship between the penalized splines and comparable SDE models.

Satish Iyengar (*University of Pittsburgh*)
I congratulate the authors for bringing to the attention of the statistics community methods of inference for differential equations.

Early in their paper, the authors mention the case 'when only a subset of variables of a system is actually measured...'. I suspect that this case is quite common in many areas. It typically leads to the non-identifiability of parameters of the model. We encountered this problem in our studies of varying spike rates in certain monkey interneurons (Czanner, 2004; Czanner *et al.*, 2007). We fit a leaky integrate-and-fire model (Liu and Wang, 2001) for the (observed) membrane potential V and the (latent) intracellular calcium concentration X. The model has the form

$$\begin{aligned} \mathrm{d}V_t &= (\alpha + \beta V_t - \gamma X_t V_t)\,\mathrm{d}t + \mathrm{d}W_t, \\ \mathrm{d}X_t &= -\delta X_t\,\mathrm{d}t + \mathrm{d}\tilde{W}, \end{aligned}$$

where W and $\tilde{W}$ are independent Brownian motions. On firing V returns to its reset potential and X is increased by a constant to model the resulting calcium influx. In the discretized version, there are about a dozen parameters, with the number of identifiable functions of the parameters depending on the details of the experiment. A careful study of what those identifiable functions are can be used to suggest auxiliary experiments that are needed to estimate the original parameters. However, determining the identifiable parameters can be a rather involved task. Widely applicable approaches to do that would be useful.

Robert E. Kass (*Carnegie Mellon University, Pittsburgh*) **and Jonathan E. Rubin and Sven Zenker** (*University of Pittsburgh*)
The fitting of differential equations to data has an illustrious history in neuroscience, but further progress requires solutions to several important problems. For example, in their pioneering work, Hodgkin and Huxley (1952) modelled action potential generation in the space-clamped squid giant axon by fitting parameters in a system coupling the voltage equation

$$\frac{\mathrm{d}V}{\mathrm{d}t} = \frac{I - I_{Na}(V, m, h) - I_K(V, n) - I_L(V)}{C} \tag{45}$$

to equations for auxiliary variables m, h and n each of the form

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \phi\{\alpha_x(V)(1-x) - \beta_x(V)x\}. \tag{46}$$

Each pair $(\alpha_x(V), \beta_x(V))$ incorporates five parameters, whereas the voltage-dependent currents in equation (45) include four (I_{Na}), three (I_K) and two (I_L) parameters respectively. Together, equations (45) and (46) contain 27 parameters.

An immediate issue is that the parameter values in equations (45) and (46) are not uniquely determined from readily available data, i.e. a näive statistical model will be non-identifiable. The best solution is to obtain additional data (as Hodgkin and Huxley (1952) did), but this is often impractical. Methodologically, two things are needed:

(a) a method for checking whether the statistical model is identifiable and,
(b) when it is not, a constructive method for proceeding.

Item (a) has been discussed in the optimization literature (e.g. Nocedal and Wright (2006)). Item (b) is generally more difficult. One possibility is to simplify models to reduce the number of parameters. This may be disadvantageous in situations where there is a direct correspondence between model structure and physiological interpretation, as inference about physiological parameters is often the objective, whereas non-uniqueness of parameter vectors may reflect physiological reality (Prinz *et al.*, 2004). For such scenarios, local optimization methods like that presented by Ramsay, Hooker, Cambell and Cao are of limited use. An alternative is to apply simulation-based Bayesian inference to compute a (potentially multimodal) posterior density on the parameter vector and thereby to quantify uncertainty about lower dimensional parameter subsets of interest (Zenker *et al.*, 2006).

We hope that the interesting overview by Ramsay, Hooker, Cambell and Cao will succeed in drawing attention to this important class of problems. The large body of literature on collocation methods should be considered carefully. This may be a case in which the field of statistics will advance most rapidly by incorporating results from other mathematical disciplines, via collaborative research that delves deeply into particular scientific problems.

Stefan Körkel (*Humboldt-Universität, Berlin*)
In this paper, the authors present an approach for the estimation of parameters in non-linear differential equation models.

For the parameterization of the differential equations, a collocation method is applied with an expansion of the state solution in terms of basis functions introducing the collocation coefficients as additional *nuisance* parameters.

The data fitting criterion, a negative log-likelihood of the observation error distribution, is augmented by a regularization term, the *equation fidelity*, which is a norm of the differential equation residual, which is numerically approximated by a quadrature formula. The two parts of this objective function are weighted by a *smoothing multiplier* λ to control the relative emphasis on fitting the data and solving the model equations.

The authors propose to solve the optimization problem for parameter estimation in a hierarchical way: an outer optimization with respect to the *structural* equation parameters is performed subject to an underlying inner optimization with respect to the collocation coefficients for fixed equation parameters.

The choice of the smoothing parameter λ is crucial for the robustness of the method. In the numerical examples presented, the authors could find suitable values by manual adjustment. Alternatively, they suggest an automatic iterative strategy based on the idea of preventing that the regularization distorts the estimate. The behaviour for $\lambda \rightarrow \infty$ is studied and shows a natural behaviour of the approach.

The method that is presented by the authors exhibits robustness and flexibility. This is demonstrated for four examples: two academic test problems, the FitzHugh–Nagumo equation system which leads to a very non-convex least squares estimation problem and the tank reactor equation system which, for particular experimental settings, has a behaviour which is close to instability. Moreover, the method is applied to two real data examples: nylon production and flare dynamics in lupus.

For all these examples, appropriate smoothing can be found and parameter estimates can be obtained from quite noisy data and in situations where not all model states can be observed. The choice of the initial guesses for the parameters is not critical at all. For comparision, for such problems with high non-convexity of the least squares fitting criteria, Gauss–Newton methods often are not usable because of small convergence regions.

The hierarchical optimization approach presented requires higher computational effort compared with an all-at-once approach, but it provides a very robust method for the estimation of parameters in intricate non-linear situations.

Reg Kulperger (*University of Western Ontario, London*)
I congratulate the authors on their proposal of a very useful and practical method. Their idea of projecting the differential equation (DE) solution to a linear space through expression (7) and then not having to find the coefficients $c(\theta)$ explicitly in terms of θ are the key elements. It is impressive that their method works amazingly well, and in some cases with data on only a subset of the components. The real example in Section 4.1 shows a very good fit of the data and estimated DE solutions.

In Section 3.1 you have chosen the standard deviation to be 0.5. How stable is the estimation over different noise levels? It is reasonable to hope for good estimates with small noise, but at what level does the estimation break down?

Fig. 6, and the discussion around it, suggests a practical way of choosing λ, the penalty tuning parameter. In an Akaike or Bayes information criterion the penalty is a function of the sample size n. The penalty in this paper is more in the spirit of spline regression but does not explicitly involve the sample size n or the level of noise. Are these implicitly reflected in the tuning parameter λ?

The estimator variances that are approximated by expressions (18) or (24) are compared in a simulation study in Section 3.2. They are first-order delta method approximations, and they perform very well in these examples compared with the actual sample standard deviation in the simulation experiment for the two models in Section 1.2. How do you expect this approximation to behave in other model applications and different noise levels?

Section 5 raises some other interesting questions. Does a lagged equation model also require data at offsets δ, or is it possible for the data still to be irregularly spaced or at least not depending on δ? If the former is needed then δ is a number that must be known and not estimated. Equivalently, is δ identifiable?

The stochastic DE (SDE) that is described in Section 5.2 is considered with diffusion term $\lambda \, \mathrm{d}W(t)$ (where λ is not the same as the penalty parameter). In general $X(t) = E\{x(t)\}$ does not satisfy the noise-free DE $X(t) = f\{X(t), u, t|\theta\}$. These SDE processes have quite different dynamics from those of the regression form that is described here. Is there some analogous method for an SDE setting, or are these a different class of estimation problems?

Subhash Lele (*University of Alberta, Edmonton*)
This paper proposes a method to confront non-linear dynamical models with real data so that they provide not just pretty pictures and qualitative understanding, but also quantitative predictions and model adequacy measures.

The method that is developed in this paper is intuitive and appealing but somewhat *ad hoc*.

(a) How does the choice of the number and the form of basis functions affect the estimates?
(b) Do and how do the standard errors and resultant confidence intervals reflect the amount of approximation that is involved in equation (7)?
(c) The method is based on estimating functions but it is unclear whether the resultant estimating functions are, in fact, zero unbiased or not. Are they information unbiased? If not, the asymptotic variances should be based on Godambe information rather than Fisher information.
(d) What kind of asymptotics are appropriate: infill asymptotics, or increasing domain asymptotics or both (Cressie, 1991)?
(e) Can we use resampling techniques to obtain robust standard errors?
(f) In population dynamics models, there is demographic stochasticity and environmental stochasticity (Lande *et al.*, 2003). Can the methodology that is developed in this paper be useful for such models?
(g) With hidden layers in the model, how would you know that the parameters that you are trying to estimate are, in fact, identifiable?

Recently, extending the work of Robert and Titterington (1998), I, jointly with my colleagues, have developed a technique, which is called data cloning, to conduct likelihood inference for hierarchical models (Lele *et al.*, 2007). Data cloning is based on the simple idea that, as the sample size increases, posterior distributions converge to a Gaussian distribution with mean equal to the maximum likelihood estimate and variance equal to the inverse of the Fisher information. One can artificially increase the sample size by cloning the data several times. Then, a standard application of Markov chain Monte Carlo methods provides the maximum likelihood estimate along with its standard error. We are currently using the data cloning method to conduct inference for stochastic population dynamics models for single or multiple populations such as the Lotka–Volterra model. We are also using data cloning to conduct inference for epidemiological models such as the susceptible–infected–recovered model. One of the major advantages of the data cloning method is that it provides a simple check for the identifiability of the parameters. We have found that the initial conditions are, in general, very difficult to estimate (if identifiable). But, otherwise, the data cloning method is computationally quite fast.

Lang Li (*Indiana University, Indianapolis*)
I congratulate the authors for their breakthrough in parameter estimation problems for differential equations. I would also like to express my appreciation for the effort of the Royal Statistical Society. This pioneer paper advocates the integration of cutting edge statistics and traditional mathematics.

As a statistician working exclusively in the pharmacology area, I can see an immediate application of this smoothing approach to pharmacokinetics models. Besides the work by Gelman (1996) that is referred to in the text, more comprehensive reviews of statistical and computational work in pharmacokinetics models can be found in Davidian and Giltinan (2003) and Pillai *et al.* (2005). It is worthwhile to mention that in Li (2002, 2004) the non-linear relationships between pharmacokinetics parameters and covariates were modelled by cubic splines. These works were probably the earliest integration of smoothing techniques and differential equations in pharmacokinetics models. So far, all pharmacokinetics model fittings are based on the numerical solution of a differential equation, when the analytical solution is not available.

Now, the generalized smoothing approach totally changed the paradigm of parameter estimation for differential equations. It transformed a fragmented numerical procedure into a uniformed non-linear regression. As the authors claimed in the paper, the computational stability is much improved. I think that this is a major improvement.

Computational speed is obviously a critical factor for its more general usage. When not all the response variables in the ordinary differential equations are measurable, the unmeasured variables still need to be solved from ordinary differential equations, and they will be used in the penalty term. According to current smoothing parameter selection strategy, the model may need to be fitted to the data multiple times. Hence, it is not clear whether or not its computational expense is lower than that of the other approaches. In all pharmacokinetics models, only blood samples can be assessed; the drug concentrations in all the other organs or peripheral compartments cannot be directly measured. Therefore, an evaluation of the speed for various approaches is absolutely necessary.

One important application of the pharmacokinetics model is its ability in prediction. It will be interesting to see whether the generalized smoothing approach can improve the prediction or not.

Sylvain Sardy (*Université de Genève*)
The backbone of the methodology proposed is the expansion representation of output functions so that both $x_i(t)=\Sigma_{k=1}^{K_i} c_{ik}\,\phi_i(t)$ and its derivative $\dot{x}_i(t)=\Sigma_{k=1}^{K_i} c_{ik}\,\dot{\phi}_i(t)$ are a linear form of the same coefficients $\mathbf{c}_i$. Besides solving the non-trivial optimization, providing variance estimation is also an achievement. The authors' substantial work is the source of many research directions for statisticians, like non-parametric estimation.

The authors essentially solve the least squares problem for systems of differential equations by letting λ_i become large in equation (13). At the limit, no regularization is performed: they solve the constrained problem, as in linear regression one could solve $\min_{\boldsymbol{\theta}}(\|\mathbf{y}-X\boldsymbol{\theta}\|_2^2)$ by successively solving $\min(_{\boldsymbol{\theta},\mathbf{x}}\|\mathbf{y}-\mathbf{x}\|_2^2)+\lambda\|\mathbf{x}-X\boldsymbol{\theta}\|_2^2$ and letting $\lambda\rightarrow\infty$. This observation leads to two points. First the constrained optimization could be solved efficiently by handling constraints directly. Second if the true parametric equations are not completely known, the practice is to do model selection. Take the FitzHugh–Nagumo equations (2) for instance: we could start with the richer model

$$\dot{V}=c\left(V-\frac{V^3}{3}+dR+eR^2\right),$$
$$\dot{R}=-\frac{1}{c}\{fV+g\log(V)-a+bR+hR^2\}$$

and estimate a sparse vector of coefficients $\boldsymbol{\theta}=(a,c,d,e,f,g,h)$ while satisfying the constraints that are imposed by the differential equations. A possible model selection strategy consists in solving a lasso-type

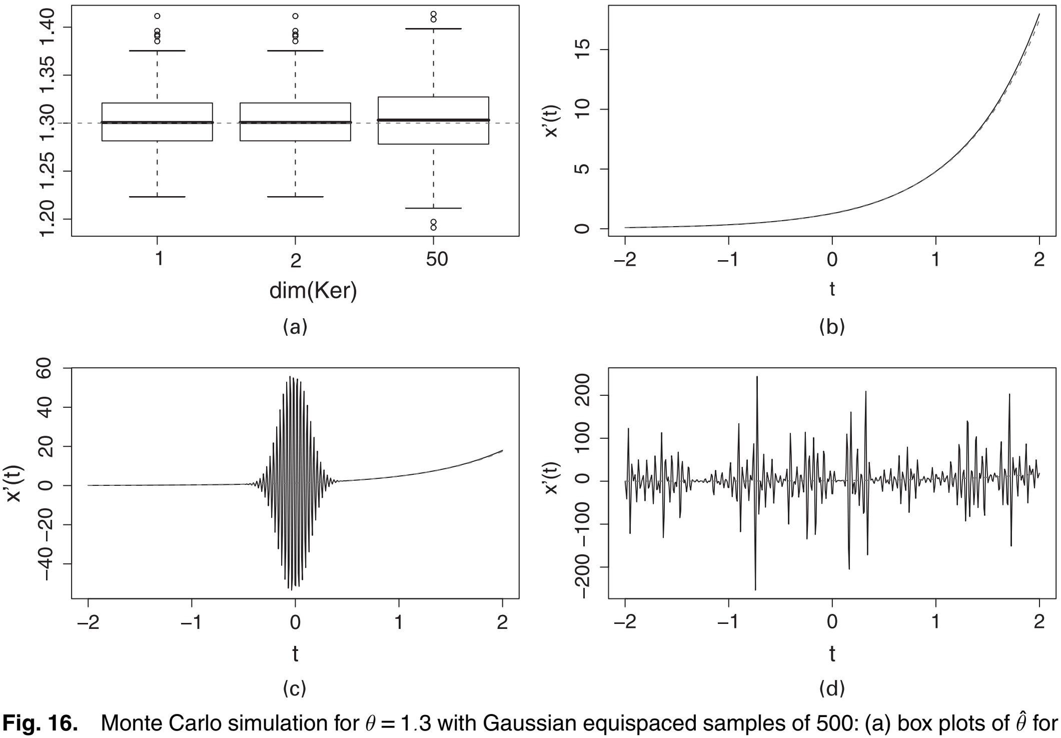


Fig. 16. Monte Carlo simulation for $\theta=1.3$ with Gaussian equispaced samples of 500: (a) box plots of $\hat{\theta}$ for $Q\in\{K-1, K-2, K-50\}$ and one typical estimated derivative of $\dot{x}(t)$, when (b) dim(Ker)=1, (c) dim(Ker) = 2 and (d) dim(Ker) = 50

l_1 penalized least squares. The convex l_1-penalty on $\boldsymbol{\theta}$ may also have the advantage of removing some of the ripples of Fig. 2. Solving the constrained l_1 penalized least squares is a worthy challenge to achieve model selection for systems of non-linear differential equations. Finally, increasing the dimension of $\boldsymbol{\theta}$ with the sample size, non-parametric estimation becomes possible.

The number of terms K_i that are used in each spline expansion relative to the number Q_i of collocation points is important. We illustrate with a toy differential equation: $\dot{x}(t) = f(t, x; \theta) = \theta\, x(t)$ defined on $[-2, 2]$ with $\theta = 1.3$. The data consist of $N = 500$ equispaced measurements $\mathbf{y} = \mathbf{x} + \varepsilon$, where $\varepsilon \sim N(\mathbf{0}, I)$. We solve

$$\min_{\theta, \mathbf{c}} (\tfrac{1}{2} \|\mathbf{y} - \mathbf{x}\|_2^2)$$

subject to

$$\mathbf{L}(\mathbf{c}, \theta) := \dot{B}\mathbf{c} - \theta B\mathbf{c} = \mathbf{0},$$
$$\mathbf{x} = X\mathbf{c}$$

where X is the $N \times K$ spline matrix ($K = 250$), and B and $\dot{B}$ are the $Q \times K$ matrices of splines and their derivatives. For the residuals $\mathbf{L}(\mathbf{c}, \theta) = \mathbf{0}$ to have a solution other than $\mathbf{c} = 0$ for all θ, the kernel of $\dot{B} - \theta B$ must be different from $\{\mathbf{0}\}$. A sufficient condition is $Q < K$. Here the correct choice seems to be $Q = K - 1$ to have $\dim\{\mathrm{Ker}(\dot{B} - \theta B)\} = 1$ and to fit the solution $x(t; \theta) = \exp(\theta t)$. Choosing a smaller value of Q has adverse effects. For each choice of $Q \in \{K - 1, K - 2, K - 50\}$ we estimate 500 times θ and $\mathbf{c}$. Looking at Fig. 16, we see that both bias and variance increase with $K - Q$ and that the estimation of $\dot{x}(t)$ becomes bad when $Q < K - 1$. This Monte Carlo experiment shows that the choice of K_i and Q_i calls for particular attention.

Hulin Wu (*University of Rochester*)

I congratulate Professor Ramsay and his collegues on their stimulating paper that introduces the inverse problem of ordinary differential equations (ODEs) to the statistical research community. The problem of predicting the results of measurements for a given ODE model is called the *forward problem*. The *inverse problem* is to use the measurements of state variables to estimate the parameters in the ODE model. This paper reflects the important effort to promote more statistical research to address the statistical inverse problem for differential equation models. The inverse problem for ODE models is a long-standing problem in the mathematical modelling research community, but it is less familiar in the statistical research community. However, this is an area in which statisticians can make significant contributions. Mathematicians and engineers have made great progress in addressing the ODE inverse problem, but mostly from theoretical perspectives and on the basis of the standard least squares principle (Anger, 1990; Lawson and Hanson, 1995; Englezos and Kalogerakis, 2001; Tarantola, 2005; Aster *et al.*, 2005; Li *et al.*, 2005). Modern statistical techniques have not been widely used in this field.

Ramsay and his colleagues introduced an interesting smoothing-based profiling estimation procedure to estimate parameters in ODE models. This method avoids numerically solving the ODEs, which is a good feature compared with the least squares method. The proposed penalized log-likelihood and least squares criteria (14) and (15) are weighted 'goodness-of-fit' measures to the observed data and to the ODE model. This indicates that both observations and ODE model have errors, and the criteria proposed are an attempt to trade off these two errors. Thus, the optimal weight (λ) should depend on the relative magnitudes of the observation error and the ODE model error. Thus, there is a need to introduce the model error into the specification of the ODE model which is similar to the Kalman filtering in the state space model (Stengel, 1994).

It is worthwhile to point out that there are a few publications on ODE parameter estimation in the statistical literature. For example, Li *et al.* (2002) proposed a spline-based estimation method to estimate the time varying parameters in a pharmacokinetic (ODE) model for longitudinal data, whereas Chen and Wu (2007) proposed a local kernel smoothing-based two-step estimation method to estimate time varying parameters in ODE models. Huang and Wu (2006) and Huang *et al.* (2006) employed a hierarchical Bayesian approach to estimate kinetic parameters in ODE models for longitudinal data.

The **authors** replied later, in writing, as follows.

We thank the Royal Statistical Society for providing the venue for this paper and its discussion, and the discussants for their many insightful comments from so many different backgrounds. There seems to be near universal agreement on the lack of good statistical methodology for estimation and inference in non-linear dynamics and on the need for greater involvement from the statistical community in these problems.

The generation of interest may be the most important contribution of our paper. We are, of course, not the only statisticians to have worked in this field, and we thank the discussants for adding to our references to previous work. The range of ideas in the commentaries indicates the breadth of research problems that remain open, and we look forward to exciting times. From among the many issues raised, we have selected a few that especially require further comment.

Choosing λ for inference and prediction

Smoothing parameter choice is clearly the most vexing aspect of our method. We do not have ready answers, and in fact we think that interesting answers will have to wait for a tighter specification of the questions. For example, Gu points out that there are two distinct and often contradictory goals here. The smoothing objective of representing the observed trajectories well will often require somewhat smaller values of λ than will the problem of estimation of the parameters $\boldsymbol{\theta}$.

Criteria such as cross-validation, generalized or not, are too tightly tied to data smoothing to be reliable routes to optimal parameter estimation. More generally, a data smoother is only one example of a function $g(\boldsymbol{\theta})$ that may be the actual target of the experiment and subsequent data analysis, and where we judge the quality of $\hat{\boldsymbol{\theta}}$ by the usefulness of $g(\hat{\boldsymbol{\theta}})$. Ellner raises the important question of *extrapolation*, either further forward in time or for new runs, given that our smooth is not a direct solution to the ordinary differential equation (ODE). We are intrigued by his ideas, and we note their resemblance to the path following techniques that are described by Smith. We look forward to seeing further developments; a particular question would be how far ahead we should look.

When the smooth and the ODE do not coincide, we suggest that it is the smooth that should be taken to represent the actual trajectory of the system. However, the discrepancy between the two can be used as a diagnostic for potentially misspecified measurement processes, such as in the autoregressive integrated moving average structure that is highlighted by Olhede.

Stochastic differential equations

We warmly agree with the many commentators who insist that no experiment is completely deterministic and free of external influences. Numerous of them have pointed out the resemblance of our methods to stochastic differential equations (SDEs), where the usual notation is

$$\mathrm{d}X_t = f(X_t, t)\,\mathrm{d}t + \sigma(X_t, t)\,\mathrm{d}B_t$$

with $\mathrm{d}B_t$ being the innovation distribution and $\sigma(X_t, t)$ specifying its standard deviation, possibly varying with the process level X_t and otherwise with respect to time t.

The innovation distribution in SDEs is intended to account for random variation and unobserved influences. As Ionides points out, it may also be used to account for model misspecification, although this needs to be taken in conjunction with diagnostics for systematic lack of fit, As with serial correlation in linear regression, some care needs to be taken in evaluating the appropriateness of the innovation distribution. Here, Smith's methods have some interesting diagnostic ideas and we would like to see whether they could also be used to suggest some form of serial correlation as, for example, in an integrated Gaussian process. Allowing the innovation distribution to vary over state space (which is another intriguing idea) could be incorporated in the smoothing methods that we describe, but we are cautious about overcomplicating models without good reason.

The connection between penalized splines and Gaussian processes has long been recognized, and formalized, for example, in Wahba (1990). We are working on extending these results to non-linear penalties of the type that we use, and we thank Ionides for his encouragement. An alternative approach to estimating SDEs by using smoothing could be to represent the innovation distribution as

$$\mathrm{d}B(t) = \phi(t)'\mathbf{c}$$

where the $\mathbf{c}$ are random effects in the spirit of Ruppert *et al.* (2005). Our estimation procedure would then look like conditional inference in a non-linear mixed model. This opens the door, for example, to restricted maximum likelihood type estimates of λ. Such an approach reintroduces some of the numerical difficulties that we sought to avoid, but we are exploring how mixed model ideas could be translated into our methods.

Diagnostics

No models are perfect and Kovac, Olhede, Smith, Boker, and Chow and Tiberio have all pointed out the need for good methods to suggest model improvements. Kovac and Olhede both suggest methods for finding serial correlation that may need to be accounted for via a change in the likelihood, or for finding regime changes which would motivate a change in parameter values. Chow and Tiberio would use the penalty as a way of checking for misspecification on the *derivative* scale, which would give us direct access to where the model may be wrong structurally. We have developed this idea in Hooker (2007), including examining some identifiability issues. One problem is the vast range of model modifications that are possible in a non-linear dynamic model, and we advise particular caution and consultation with domain experts.

Extensions

A large range of further models to which our methods could be applied have been mentioned by commentators. We would like to point out that some of the desired functionality is already available in the publicly provided software, although these have not been directly addressed. In particular, we allow for $\boldsymbol{\theta}$ to be penalized by a twice differentiable function. This provides a way to include a Bayesian prior (Kass and Guillas), parameters that vary smoothly over time (Earn, Kovac and Smith) and mixed models over experiments (Li, and Donnet and Samson). We also allow for mixtures of derivatives, including zero order.

Unfortunately, the choice of norms that is desired by Biegler and Olhede is not available in current software, but we agree that this represents an important area of software development. The partial differential equations that are desired by Bampfylde are more problematic, both in terms of implementation and in terms of theory. Unlike ODEs, partial differential equation boundary conditions are infinite dimensional and must be constrained in some way to ensure that the problem is identifiable. This seems like a fascinating area for future work.

Bayesian methods and identifiability

Bayesian analysis (Dowd, Guillas, Kass and Wu) has been used in some of the most successful applications of statistical methods to non-linear dynamic systems. This is at least partly due to the ease of implementation of Markov chain Monte Carlo computation and its ability to side-step the issue of parameter identifiability. However, our own experience is that the local minima that plague non-linear least squares methods is also a problem for a Bayesian approach, so one must be cautious in concluding that the Markov chain has converged.

In recent work, Campbell (2007) has adapted our relaxed fit smoothing to a collocation tempering approach with Markov chain Monte Carlo methods. In this parallel chain Markov chain Monte Carlo algorithm, one chain uses the solution to the ODE $\mathbf{x}_{\boldsymbol{\theta}}(t)$, as the location parameter in the likelihood. The remaining parallel chains are constructed by substituting $\mathbf{x}_{\boldsymbol{\theta}}(t)$ with smooth approximations to the ODE solution $\mathbf{x}_{\boldsymbol{\theta},\lambda}(t) = \mathbf{c}(\boldsymbol{\theta},\lambda)'\,\boldsymbol{\phi}(t)$, where λ is fixed within each chain. Parameters are allowed to swap between parallel chains, similarly to parallel tempering (Geyer, 1991), leading to improving convergence and stability. Furthermore, the combination of chains using $\mathbf{x}_{\boldsymbol{\theta}}(t)$ and $\mathbf{x}_{\boldsymbol{\theta},\lambda}(t)$ allows inference on $\boldsymbol{\theta}$ and the fits to the data from the deterministic model and a relaxed smooth.

There is a substantial literature on identifiability in ODEs, as picked up by Kass and Iyengar, and it is not difficult to find systems which are unidentifiable. A simple diagnostic is to examine the Fisher information matrix at the current parameters, as do Wu *et al.* (2007) for the dynamics of human immunodeficiency virus.

Bases

Huang and Ma, and Olhede note that, unlike traditional smoothing, a large number of basis functions may be required by our collocation approach. Deuflhard and Bornemann (2000) reviewed the literature in numerical analysis on the size of basis. If we intend to let $\lambda \to \infty$, then it is sufficient to select a basis that is sufficiently rich to represent a solution to the ODE. Could stochastic differential equations require even richer bases? This may be possible, but we know of no work in the area.

The choice of quadrature technique is an issue, and we know that our implementation may not be optimal. Biegler, Kovacs and Olhede argue for penalty norms that would allow the penalty to be explicitly set to zero for a finite λ, and Sardy observes that this is only possible if the quadrature rule contains no more points than basis functions. In fact, collocation methods are usually based on Gauss–Radau quadrature between knots with the same number of Legendre polynomial terms as quadrature points.

We reran the FitzHugh–Nagumo simulations using Huang and Ma's 201 observations and 201 knots but placed equally weighted quadrature points only at each knot. At $\lambda = 10^4$, there was no observable difference over 200 simulations between this quadrature and the Simpson's rule that we initially employed in terms of parameter estimation bias and standard error. However, the new quadrature rule was about 100 times faster to provide answers.

However, we encountered a new issue when we smoothed a sample of simulated date at the higher smoothing level $\lambda = 10^7$ using the true parameters. Fig. 17 shows that Simpson's rule quadrature produces a substantial distortion in the initial shape of the path, whereas the simpler collocation regime remains indistinguishable from the true trajectory. In contrast, the FitzHugh–Nagumo dynamics are comparatively

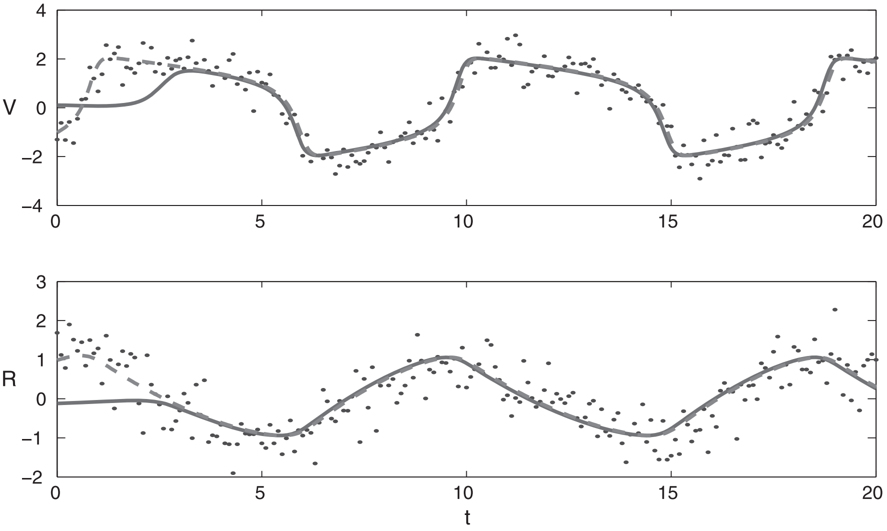


Fig. 17. Smooths of the data (·) with true parameters and by using Simpson's quadrature rule (———) and equally weighted quadrature points at the knots (– – –): the simpler quadrature rule is visually indistinguishable from the true trajectory

mild, and our experience is that the choices of bases and quadrature methods for systems with sharper dynamics and discontinuous inputs require considerable care.

Dynamical features

Dynamical features such as limit cycles, fixed points, bifurcations and chaos are central areas of interest in non-linear dynamics and, as both Brown, and Guckenheimer and Tien observe, they have played very little role in traditional parameter estimation techniques, including our own. Too little attention has been given to problems of inference about dynamical features. However, along with Kulperger, we note that dynamic behaviour can be quite different in stochastic differential equations and the analysis that is required to understand it is not necessarily easy.

Guckenheimer and Tien suggest only searching the parameter space where limit cycles exist. In general, dynamical features, when they can be readily analysed, can be incorporated in Bayesian priors. Using estimated features such as periods and peaks as data is also interesting, but methods for understanding uncertainty from this feature perspective remain to be developed.

Response surfaces

We are pleased to see so many commentaries on our Fig. 5: the nature of response surfaces that must be minimized has been one of the factors retarding progress in the area. In common with our approach, several methods have been developed over the years that rely on relaxing the solution to the differential equation, at least at intermediate steps. The idea of fitting cycles independently, as advocated by Guckenheimer and Tien, may be viewed as using different initial conditions for each cycle. This is similar to the methods in Bock (1983), in which the ODE is solved over adjoining small intervals, and where discontinuities at interval boundaries are successively reduced. Tjoa and Biegler (1991) also provided methods that do not explicitly solve the ODE until the final set of parameter estimates.

An explanation for why the approach provides better-conditioned minimization problems could be that, if the approximate trajectory is different from the trajectory that is given by the parameters, the response surface will be partly affected by $\|\dot{\mathbf{x}} - \mathbf{f}(\mathbf{x}, \boldsymbol{\theta})\|$, which is frequently more convex than the original likelihood criterion. We believe that firming up this conjecture may be useful for other difficult optimization problems.

Asymptotics

No statistical paper is complete without an asymptotic analysis, and we thank Olhede for providing ours. This is given in the context of a deterministic model in which the essential point is to ensure that $\boldsymbol{\theta}$ continues to affect $\mathbf{x}_{\theta,\lambda}$. With infill asymptotics, we are now back to maximum likelihood theory. In the expanding domain case, the situation is somewhat more complicated, since we need to ensure that λ increases at a rate that is sufficiently fast to force the convergence of $\mathbf{x}_{\theta,\lambda}$ to an exact solution; this again implies an $N^{1+\delta}$-rate. To answer Lele's question, infill asymptotics with independent and identically distributed residuals about a deterministic system does not appear to be reasonable and suggests either a Gaussian process for the errors or a stochastic differential equation, or both. In such cases, neither infill nor expanding domain asymptotics alone may be sufficient to provide consistency.

Conclusion

We have been impressed and stimulated by the range of ideas and perspectives in the commentaries and thank the Royal Statistical Society for making this discussion possible. There appear to be several independent suggestions that might benefit from collaboration between our discussants. Many other comments will require a paper or more to address adequately. It is clear that we still have much work to do, both for this method and for inference in non-linear dynamics generally. We hope that this paper has demonstrated both the challenge and the interest in these problems, and that it inspires more statisticians to help us to solve them.

References in the discussion

Anderson, R. M. and May R. M. (1991) *Infectious Diseases of Humans: Dynamics and Control*. Oxford: Oxford University Press.

Anger, G. (1990) *Inverse Problems in Differential Equations*. Berlin: Kluwer.

Ascher, U. M. and Petzold, L. R. (1998) *Computer Methods for Ordinary Differential Equations and Differential-algebraic Equations*. Philadelphia: Society for Industrial and Applied Mathematics.

Aster, R. C., Borchers, B. and Thurber, C. H. (2005) *Parameter Estimation and Inverse Problems*. Boston: Elsevier.

Bauch, C. T. and Earn, D. J. D. (2003) Transients and attractors in epidemics. *Proc. R. Soc. Lond.* B, **270**, 1573–1578.

Bayarri, M., Berger, J., Cafeo, J., Garcia-Donato, G., Liu, F., Palomo, J., Parthasarathy, R., Paulo, R., Sacks, J. and Walsh, D. (2007) Computer model validation with functional output. *Ann. Statist.*, to be published.

Berger, J. O., Liseo, B. and Wolpert, R. L. (1999) Integrated likelihood methods for eliminating nuisance parameters. *Statist. Sci.*, **14**, 1–28.

Bergstrom, A. R. (1966) Nonrecursive models as discrete approximations to systems of stochastic differential equations. *Econometrica*, **34**, 173–182.

Beskos, A., Papaspiliopoulos, O., Roberts, G. O. and Fearnhead, P. (2006) Exact and computationally efficient likelihood-based estimation for discretely observed diffusion processes (with discussion). *J. R. Statist. Soc.* B, **68**, 333–382.

Biegler, L. T., Ghattas, O., Heinkenschloss, M. and van Bloemen Waanders, B. (eds) (2003) Large-scale PDE-constrained optimization. *Lect. Notes Computnl Sci. Engng*, **30**.

Bjornstad, O. N., Finkenstadt, B. F. and Grenfell, B. T. (2002) Dynamics of measles epidemics: estimating scaling of transmission rates using a time series SIR model. *Ecol. Monogr.*, **72**, 169–184.

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

Boker, S. M., Neale, M. C. and Rausch, J. (2004) Latent differential equation modeling with multivariate multi-occasion indicators. In *Recent Developments on Structural Equation Models: Theory and Applications* (eds K. van Montfort, H. Oud and A. Satorra), pp. 151–174. Dordrecht: Kluwer.

Brown, E. N. (1987) Identification and estimation of differential equation models for circadian data. *PhD Dissertation*. Department of Statistics, Harvard University, Cambridge.

Brown, E. N., Choe, Y., Luithardt, H. and Czeisler, C. A. (2000) A statistical model of the human core-temperature circadian rhythm. *Am. J. Physiol.*, **279**, E669–E683.

Campbell, D. (2007) Bayesian collocation tempering and generalized profiling for estimation of parameters from differential equation models. *PhD Thesis*. McGill University, Montreal.

Candès, E. J. and Tao, T. (2005) The Dantzig selector: statistical estimation when *p* is much larger than *n*. *Technical Report*. Caltech, Pasadena.

Casey, R. (2004) Periodic orbits in neural models: sensitivity analysis and algorithms for parameter estimation. *PhD Thesis*. Cornell University, Ithaca.

Chen, J. and Wu, H. (2007) Estimation of time-varying parameters in deterministic dynamic models with application to HIV infections. *Statist. Sin.*, to be published.

Chorin, A. J. and Krause, P. (2004) Dimensional reduction for a Bayesian filter. *Proc. Natn. Acad. Sci. USA*, **101**, 15013–15017.

Coulson, T., Rohani, P. and Pascual, M. (2004) Skeletons, noise and population growth: the end of an old debate? *Trends Ecol. Evoln*, **19**, 359–364.

Cressie, N. A. C. (1991) *Statistics for Spatial Data*. New York: Wiley.

Czanner, G. (2004) Applications of statistics in neuroscience. *PhD Dissertation*. Department of Statistics, University of Pittsburgh, Pittsburgh.

Czanner, G., Iyengar, S., Zajtsev, A. and Krimer, L. (2007) Maximum likelihood estimation of state-space integrate-and-fire model of adapting neurons. *Technical Report*. Department of Statistics, University of Pittsburgh, Pittsburgh.

Czeisler, C. A., Duffy, J. F., Shanahan, T. L., Brown, E. N., Mitchell, J. F., Rimmer, D. W., Ronda, J. M., Silva, E., Allan, J. S., Emens, J. S., Dijk, D. J. and Kronauer, R. E. (1999) Age-independent stability, precision, and near 24 hour period of the human circadian pacemaker. *Science*, **284**, 2177–2181.

Czeisler, C. A., Kronauer, R. E., Allan, J. S., Duffy, J. F., Jewett, M. E., Brown, E. N. and Ronda, J. M. (1989) Bright light induction of strong (Type 0) resetting of the human circadian pacemaker. *Science*, **244**, 1328–1333.

Davidian, M. and Giltinan, D. M. (2003) Nonlinear models for repeated measurement data: an overview and update. *J. Agric. Biol. Environ. Statist.*, **8**, 387–419.

Davies, P. L. and Kovac, A. (2001) Local extremes, runs, strings and multiresolution (with discussion). *Ann. Statist.*, **29**, 1–65.

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

Diks, C. (1999) *Nonlinear Time-series Analysis: Methods and Applications*. Singapore: World Scientific Publishing.

Dowd, M. (2006) A sequential Monte Carlo approach to marine ecological prediction. *Environmetrics*, **17**, 435–455.

Dowd, M. (2007) Bayesian statistical data assimilation for ecosystem models using Markov Chain Monte Carlo. *J. Mar. Syst.*, doi 10.1016/j.jmarsys.2007.01.007, to be published.

Durbin, J. and Koopman, S. J. (2001) *Times Series Analysis by State-space Methods*. New York: Oxford University Press.

Dushoff, J., Plotkin, J. B., Levin, S. A. and Earn, D. J. D. (2004) Dynamical resonance can account for seasonality of influenza epidemics. *Proc. Natn. Acad. Sci. USA*, **101**, 16915–16916.

Earn, D. J. D., Rohani, P., Bolker, B. M. and Grenfell, B. T. (2000) A simple model for complex dynamical transitions in epidemics. *Science*, **287**, 667–670.

Ellner, S. P., Bailey, B. A., Bobashev, G. V., Gallant, A. R., Grenfell, B. T. and Nychka, D. W. (1998) Noise and nonlinearity in measles epidemics: combining mechanistic and statistical approaches to population modeling. *Am. Naturlst*, **151**, 425–440.

Ellner, S. P. and Guckenheimer, J. (2006) *Dynamic Models in Biology*. Princeton: Princeton University Press.

Ellner, S. P., Seifu, Y. and Smith, R. H. (2002) Fitting population dynamic models to time-series data by gradient matching. *Ecology*, **83**, 2256–2270.

Englezos, P. and Kalogerakis, N. (2001) *Applied Parameter Estimation for Chemical Engineers*. New York: Dekker.

Evensen, G. (2003) The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dyn.*, **53**, 343–367.

Fahrmeir, L. and Tutz, G. (eds) (1994) *Multivariate Statistical Modelling based on Generalized Linear Models*. Berlin: Springer.

Fine, P. E. M. and Clarkson, J. A. (1982) Measles in England and Wales—I: an analysis of factors underlying seasonal patterns. *Int. J. Epidem.*, **11**, 5–14.

Finkenstädt, B. F. and Grenfell, B. T. (2000) Time series modelling of childhood diseases: a dynamical systems approach. *Appl. Statist.*, **49**, 187–205.

Gelman, A., Bois, F. and Jiang, J. (1996) Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *J. Am. Statist. Ass.*, **91**, 1400–1412.

Geyer, C. (1991) Markov chain monte carlo maximum likelihood. In *Computing Science and Statistics: Proc. 23rd Symp. Interface* (ed. E. M. Keramidas), pp. 156–163. Fairfax Station: Interface Foundation.

Godsill, S. J., Doucet, A. and West, M. (2004) Monte Carlo smoothing for nonlinear time series. *J. Am. Statist. Ass.*, **99**, 156–168.

Higdon, D., Gattiker, J. and Williams, B. (2007) Computer model calibration using high dimensional output. *J. Am. Statist. Ass.*, to be published.

Hodgkin, A. and Huxley, A. (1952) A quantitative description of membrane current and its application to conduction and excitation in nerve. *J. Physiol.*, **117**, 500–544.

Hooker, G. (2007) Forcing function diagnostics for nonlinear dynamics. To be published.

Hotelling, H. (1927) Differential equations subject to error, and population estimates. *J. Am. Statist. Ass.*, **22**, 283–314.

Huang, Y., Liu, D. and Wu, H. (2006) Hierarchical Bayesian methods for estimation of parameters in a longitudinal HIV dynamic system. *Biometrics*, **62**, 413–423.

Huang, Y. and Wu, H. (2006) A Bayesian approach for estimating antiviral efficacy in HIV dynamic models. *J. Appl. Statist.*, **33**, 155–174.

Ionides, E. L., Bretó, C. and King, A. A. (2006) Inference for nonlinear dynamical systems. *Proc. Natn. Acad. Sci. USA*, **103**, 18438–18443.

Itô, K. (1951) On stochastic differential equations. In *American Mathematical Society Memoirs*, no. 4. New York: American Mathematical Society.

Judd, K. (2007) Failure of maximum likelihood methods for chaotic dynamical systems. *Phys. Rev.* E, **75**.

Judd, K. and Smith, L. A. (2004) Indistinguishable states II. *Physica* D, **196**, 224–242.

Judd, K., Smith, L. and Weisheimer, A. (2004) Gradient free descent: shadowing, and state estimation using limited derivative information. *Physica* D, **190**, 153–166.

Kalman, R. E. (1960) A new approach to linear filtering and prediction problems. *Trans. ASME J. Basic Engng*, **82**, 35–45.

Kennedy, M. C. and O'Hagan, A. (2001) Bayesian calibration of computer models (with discussion). *J. R. Statist. Soc.* B, **63**, 425–464.

Kitagawa, G. (1996) Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. *J. Computnl Graph. Statist.*, **5**, 1–25.

Koch, C. (1999) *Biophysics of Computation: Information Processing in Single Neurons*. New York: Oxford University Press.

Künsch, H. R. (2005) Recursive Monte Carlo filters: algorithms and theoretical analysis. *Ann. Statist.*, **33**, 1983–2021.

Kurtz, T. G. (1980) Relationships between stochastic and deterministic population models. *Lect. Notes Biomath.*, **38**, 449–467.

Lande, R., Engen, S. and Saether, B. (2003) *Stochastic Population Dynamics in Ecology and Conservation*. Oxford: Oxford University Press.

Lawson, C. L. and Hanson, R. J. (1995) *Solving Least Squares Problems*. Philadelphia: Society for Industrial and Applied Mathematics.

Lele, S. R., Dennis, B. and Lutscher, F. (2007) Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov Chain Monte Carlo Methods. *Ecol. Lett.*, **10**, 551–563.

Lewis, J. M., Lakshmivarahan, S. and Dhall, S. K. (2006) *Dynamic Data Assimilation: a Least Squares Approach*. Cambridge: Cambridge University Press.

Li, L., Brown, M. B., Lee, K. H. and Gupta, S. (2002) Estimation and inference for a spline-enhanced population pharmacokinetic model. *Biometrics*, **58**, 601–611.

Li, L., Lin, X., Brown, M., Gupta, S. and Lee, K. H. (2004) A population pharmacokinetic model with time-dependent covariates measured with errors. *Biometrics*, **60**, 451–460.

Li, Z., Osborne, M. and Prvan, T. (2005) Parameter estimation in ordinary differential equations. *IMA J. Numer. Anal.*, **25**, 264–285.

Liu, Y. H. and Wang, X. J. (2001) Spike-frequency adaptation of a generalized leaky integrate-and-fire model neuron. *J. Computnl Neursci.*, **10**, 25–45.

London, W. and Yorke, J. A. (1973) Recurrent outbreaks of measles, chickenpox and mumps: i, seasonal variation in contact rates. *Am. J. Epidem.*, **98**, 453–468.

McSharry, P. E. and Smith, L. A. (2004) Consistent Nonlinear Dynamics: identifying model inadequacy. *Physica* D, **192**, 1–22.

Mendes, P., Moles, C. G. and Banga, J. R. (2003) Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res.*, **13**, 2467–2474.

Mitchell, T., Morris, M. and Ylvisaker, D. (1994) Asymptotically optimum experimental designs for prediction of deterministic functions given derivative information. *J. Statist. Planng. Inf.*, **41**, 377–389.

Molenaar, P. C. M. and Newell, K. M. (2003) Direct fit of a theoretical model of phase transition in oscillatory finger motions. *Br. J. Math. Statist. Psychol.*, **56**, 199–214.

Mood, A. M. (1940) The distribution theory of runs. *Ann. Math. Statist.*, **11**, 367–392.

Morris, M. D., Mitchell, T. J. and Ylvisaker, D. (1993) Bayesian design and analysis of computer experiments—use of derivatives in surface prediction. *Technometrics*, **35**, 243–255.

Nocedal, J. and Wright, S. (2006) *Numerical Optimization*, 2nd edn. New York: Springer.

O'Hagan, A. (1992) Some Bayesian numerical analysis. In *Bayesian Statistics 4*, pp. 345–363. New York: Oxford University Press.

Pillai, G., Mentre, F. and Steimer, J. (2005) Non-linear mixed effects modeling—from methodology and software development driving implementation in drug development science. *J. Pharmkin. Pharmdyn.*, **32**, 161–183.

Prinz, A., Bucher, D. and Marder, E. (2004) Similar network activity from disparate circuit parameters. *Nat. Neursci.*, **7**, 1345–1352.

Ramsay, J. O. (1998) Estimating smooth monotone functions. *J. R. Statist. Soc.* B, **60**, 365–375.

R Core Development Team (2006) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Robert, C. P. and Titterington, D. M. (1998) Reparameterization strategies for hidden Markov models and Bayesian approaches to maximum likelihood estimation. *Statist. Comput.*, **8**, 145–158.

Ruppert, D., Wand, M. P. and Carroll, R. J. (2005) *Semiparametric Regression*. Cambridge: Cambridge University Press.

Särkkä, S. (2006) On sequential Monte Carlo sampling of discretely observed stochastic differential equations. In *Proc. Nonlinear Statistical Signal Processing Wrkshp, Cambridge, Sept.*

Schaffer, W. M. (1985) Can nonlinear dynamics elucidate mechanisms in ecology and epidemiology? *IMA J. Math. Appl. Med. Biol.*, **2**, 221–252.

Schwartz, I. B. and Smith, H. L. (1983) Infinite subharmonic bifurcation in an seir model. *J. Math. Biol.*, **18**, 233–253.

Singer, H. (1993) Continuous-time dynamical systems with sampled data, errors of measurement and unobserved components. *J. Time Ser. Anal.*, **14**, 527–545.

Smith, L. A. (2000) Disentangling uncertainty and error: on the predictability of nonlinear systems. In *Nonlinear Dynamics and Statistics* (ed. A. I. Mees), pp. 31–64. Boston: Birkhäuser.

Solak, E., Murray-Smith, R., Leithead, W. and Leith, D. (2003) Derivative observations in gaussian process models of dynamic systems. In *Advances in Neural Information Processing Systems*, vol. 16. Cambridge: MIT Press.

Stengel, R. F. (1994) *Optimal Control and Estimation*. London: Dover Publications.

Tanartkit, P. and Biegler, L. T. (1995) Stable decomposition for dynamic optimization. *Industrl Engng Chem. Res.*, **34**, 1253.

Tanartkit, P. and Biegler, L. T. (1996) Reformulating ill-conditioned DAE optimization problems. *Industrl Engng Chem. Res.*, **35**, 1853.

Tarantola, A. (2005) *Inverse Problem Theory*. Philadelphia: Society for Industrial and Applied Mathematics.

Thompson, K. R., Dowd, M., Lu, Y. and Smith, B. (2000) Oceanographic data assimilation and regression analysis. *Environmetrics*, **11**, 183–196.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B, **58**, 267–288.

Tien, J. H. (2007) Optimization for bursting neural models. *PhD Thesis*. Cornell University, Ithaca.

Tien, J. H. and Guckenheimer, J. (2007) Parameter estimation for bursting neural models. Submitted to *J. Computnl Neursci.*

Tjoa, I.-B. and Biegler, L. (1991) Simultaneous solution and optimization strategies for parameter estimation of differential-algebraic equation systems. *Industrl Engng Chem. Res.*, **30**, 376–385.

Turchin, P. (2003) *Complex Population Dynamics: a Theoretical/Empirical Synthesis*. Princeton: Princeton University Press.

de Valpine, P. (2004) Monte Carlo state space likelihoods by weighted posterior kernel density estimation. *J. Am. Statist. Ass.*, **99**, 523–536.

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

Wahba, G. (1978) Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. R. Statist. Soc.* B, **40**, 364–372.

Wahba, G. (1990) *Spline Models for Observational Data*. Philadelphia: Society for Industrial and Applied Mathematics.

Wahba, G. and Wang, Y. (1990) When is the optimal regularization parameter insensitive to the choice of the loss function? *Communs Statist. Theory Meth.*, **19**, 1685–1700.

Wallinga, J. and Teunis, P. (2004) Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *Am. J. Epidem.*, **160**, 509–516.

Wood, S. N. (2006) *Generalized Additive Models: an Introduction with R*. Boca Raton: Chapman and Hall–CRC.

Wu, H., Zhu, H., Miao, H. and Perelson, A. S. (2007) Parameter identifiability and estimation of hiv/aids dynamics models. To be published.

Zenker, S., Rubin, J. and Clermont, G. (2006) Towards a model-based medicine: integration of probabilistic inference with mechanistic knowledge. *J. Crit. Care*, **21**, 350.

Zimmer, C. (2002) Life after chaos. *Science*, **284**, 83–86.