MBS 7343
13 April 2013
ARTICLE IN PRESS
No. of Pages 10, Model 5G

Mathematical Biosciences xxx (2013) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Mathematical Biosciences

journal homepage: www.elsevier.com/locate/mbs

Maximum profile likelihood estimation of differential equation parameters through model based smoothing state estimates

Q1 D.A. Campbell *, O. Chkrebtii

Department of Statistics and Actuarial Science, Simon Fraser University, Surrey Campus, 13450, 102nd Ave, Surrey BC, Canada V3T 0A3

ARTICLE INFO

Article history:
Received 2 October 2012
Received in revised form 22 March 2013
Accepted 25 March 2013
Available online xxxx

Keywords:
Delay differential equations
Functional data analysis
JAK-STAT
Nonlinear regression
Model based smoothing

ABSTRACT

Statistical inference for biochemical models often faces a variety of characteristic challenges. In this paper we examine state and parameter estimation for the JAK-STAT intracellular signalling mechanism, which exemplifies the implementation intricacies common in many biochemical inference problems. We introduce an extension to the Generalized Smoothing approach for estimating delay differential equation models, addressing selection of complexity parameters, choice of the basis system, and appropriate optimization strategies. Motivated by the JAK-STAT system, we further extend the generalized smoothing approach to consider a nonlinear observation process with additional unknown parameters, and highlight how the approach handles unobserved states and unevenly spaced observations. The methodology developed is generally applicable to problems of estimation for differential equation models with delays, unobserved states, nonlinear observation processes, and partially observed histories.

1. Challenges of parameter estimation from differential equation models

Ordinary Differential Equations (ODEs) relate state functions $\mathbf{x}(t)$ to their rates of change with respect to an index t, such as time, externally controlled forcing functions $\mathbf{u}(t)$, and model parameters $\boldsymbol{\theta} \in \Theta^p$. In statistical terminology, ODEs can be thought of as defining states implicitly by the functional regression model, with differential operator $D = d/dt$:

$$D\mathbf{x}(t) = f(\mathbf{x}(t), \boldsymbol{\theta}, \mathbf{u}(t)), \quad \mathbf{x}(0) = \mathbf{x}_0. \tag{1}$$

The d-dimensional function $\mathbf{x}(t)$ may represent concentrations of d species in a number of system compartments, parameters $\boldsymbol{\theta}$ may represent reaction or metabolism rates, and $\mathbf{u}(t)$ may be an initial input or catalyst. Interest typically lies in estimating unknown parameters $\boldsymbol{\theta}$ required to further our understanding of the biochemical mechanism under study. Recovering state functions $\mathbf{x}(t)$ can help determine when specific concentration levels are attained. In this paper we examine the problem of estimating $\boldsymbol{\theta}$ from experimental data for a biochemical signalling pathway model.

When system (1) can be solved exactly given initial states $\mathbf{x}_0$, state functions $\mathbf{x}(t) = \mathbf{S}(\boldsymbol{\theta}, \mathbf{x}_0, t)$ can be recovered up to the unknown model parameters $\boldsymbol{\theta}$ and used to construct the likelihood $\mathcal{L}(\boldsymbol{\theta}; \mathbf{y}(t))$ of the data. The Gaussian likelihood,

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}; \mathbf{y}(t)) &= \mathrm{N}(\mathbf{S}(\boldsymbol{\theta}, \mathbf{x}_0, t), \boldsymbol{\Sigma}) \\ &= (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left\{\frac{1}{2}(\mathbf{x} - \mathbf{S}(\boldsymbol{\theta}, \mathbf{x}_0, t))'\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mathbf{S}(\boldsymbol{\theta}, \mathbf{x}_0, t))\right\},\end{aligned} \tag{2}$$

will be used for illustration throughout this paper, although the methodology presented is more generally applicable to any likelihood function.

Often f is a nonlinear function of the states for which no closed form solution exists, in which case the likelihood is built around an approximation to $\mathbf{x}(t)$. Classical methods use numerical integration to approximate $\mathbf{S}(\boldsymbol{\theta}, \mathbf{x}_0, t)$, and parameters $\boldsymbol{\theta}$ may be estimated using nonlinear least squares (NLS) regression [1], or Markov Chain Monte Carlo (MCMC) methods [2]. However, exploring this often high-dimensional parameter space can be computationally inefficient, because repeated numerical integration is often disproportionately time-consuming [3]. Furthermore the associated likelihood surface typically exhibits a complex topology including multi-modality, ripples, and other prohibitive features [4]. Estimating $\boldsymbol{\theta}$ is further complicated when state functions are observed through a possibly nonlinear transformation, or when only a subset of states is unobserved.

To alleviate these optimization difficulties, numerous methods have been developed to avoid solving the system numerically by using the state estimator $\hat{\mathbf{x}}(t) \approx \mathbf{S}(\boldsymbol{\theta}, \mathbf{x}_0, t)$ where $D\hat{\mathbf{x}}(t)$ can be readily computed. Smooth and match estimators [5] [6] [7] [8] obtain $\hat{\mathbf{x}}(t)$ by a non-parametric smooth state estimating function in the first stage, which allows conditional estimation of $\boldsymbol{\theta}|\hat{\mathbf{x}}(t)$ in the second stage via:

Q2 * Corresponding author. Tel.: +1 778 782 3730.
E-mail addresses: dac5@sfu.ca, Iamdavecampbell@gmail.com (D.A. Campbell).

0025-5564/$ - see front matter Crown Copyright © 2013 Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.mbs.2013.03.011

$$\hat{\boldsymbol{\theta}}|\hat{\mathbf{x}}(t) = \arg\min_{\boldsymbol{\theta}} \|D\hat{\mathbf{x}}(t) - f(\hat{\mathbf{x}}(t), \boldsymbol{\theta}, \mathbf{u}(t))\|^2.$$

These estimators are flexible and computationally efficient, but problematic for models with derivative discontinuities, unobserved states, sparse data, or rapidly changing dynamics.

Model-based smoothing methods overcome limitations of smooth and match estimators [9] [10]. In particular this paper focuses on the Generalized Smoothing (GS) framework of [11], a model-based smoothing method that estimates $\hat{\mathbf{x}}(t)$ and $\hat{\theta}$. The solution $\mathbf{x}(t) = \mathbf{b}'\boldsymbol{\phi}(t)$ is modelled by a truncated expansion of bases $\boldsymbol{\phi}(t)$ with coefficients $\mathbf{b}$ estimated by optimizing a penalized likelihood criterion that balances state agreement to the data with fidelity to the model. The relative importance of the model fit versus the data fit is regulated by an auxiliary parameter λ. The three different types of parameters $(\lambda, \boldsymbol{\theta}, \mathbf{b})$ are estimated through a parameter hierarchy, where lower hierarchical level parameters are defined as functions of higher levels. Optimization proceeds by holding upper hierarchical level parameters fixed and profiling over lower level parameters where each level has a different optimization criterion.

The model based smoothing of GS has successfully overcome significant optimization challenges faced by numerical solver based methods (see for example the comparison of likelihoods from numerical solvers and model based smoothing state estimates plotted in [11,12]). This paper explores and extends GS to address implementation challenges, with particular emphasis on the additional complexity arising when (1) is a delay differential equation (DDE) system of the form:

$$D\mathbf{x}(t) = f(\mathbf{x}(t), \mathbf{x}(t-\tau), \boldsymbol{\theta}, \mathbf{u}(t)), \quad \text{given } \mathbf{x}(s),\ s \in (-\tau, 0). \tag{3}$$

GS has previously been applied to DDE systems in [13], and we extend their work to the short time domain context of the JAK-STAT model. Among the differences, we develop two new strategies for estimating $\mathbf{x}(s),\ s \in (-\tau, 0)$, introduce an adaptive basis to account for periodic derivative discontinuities in states caused by the delay, and apply a Newey–West variance estimator to account for serial dependencies in the residuals caused by the smoothing process.

Delay dynamics in biological systems typically describe underlying mechanisms that introduce a time lag but cannot be modeled directly. Fixed delay models of this form can describe a large class of often complex behaviours governing gene transcription [14,15], signalling pathways [16], and cell kinetics [17]. Although the system (3) can be studied indirectly through ODE approximations, the long term dynamics can differ substantially from those of the DDE model [18]. The more general distributed delay systems are based on a convolution of the delayed state with a density function. While not explicitly considered in this paper, many of the implementation details can be applied to inference for distributed delay systems.

The GS method is particularly appropriate for inference on DDE models. When delays are introduced as proxies for unmodeled mechanisms, there is some degree of model uncertainty. In addition, DDEs often describe rather complex dynamics, with solutions that can be extremely sensitive to the specification of the functional initial condition. The GS approach effectively uses a model relaxation rather than strictly enforcing a solution to (3), allowing enough flexibility to capture both model dynamics and important features of the data. Importantly, incorporating data into the estimation of the states can result in potential for more reliable parameter estimates by overcoming the very sharp likelihood features arising from solution sensitivity to parameters when model dynamics are strictly enforced.

We organize the paper as follows; Section 2 describes the motivating application, inference for a model of the concentrations of transcription factors in the JAK-STAT intracellular signalling mechanism. Section 3 explains and extend the GS methodology to DDE systems, while Section 4 provides detailed analysis for the JAK-STAT inference problem. Concluding discussion and further extensions are outlined in Section 5.

2. JAK-STAT system

The motivating application is parameter inference for a model of intracellular signal transduction network dynamics. The JAK-STAT system is a signalling pathway in which transcription factors (STATs) undergo biochemical reactions in response to phosphorylation of Janus kinase (JAK) triggered by the binding of Erythropoietin (Epo) hormone to cell surface receptors. A review of this mechanism is available in [19]. Interest lies in inferring rate parameters and the time evolution of four species of STAT-5 transcription factor in two compartments (cytoplasm and nucleus). Current understanding suggests that following gene activation within the nucleus, STAT-5 may revert to the initial state and return to the cytoplasm to be used in the next activation cycle. Whereas factor reactions may be assumed to happen instantaneously, this last stage may introduce a time delay.

The delay model presented in [20] describes the rates of change in concentration of STAT-5 factor in each of four reaction states by a delay differential equation system involving the functional forcing term $u(t)$ representing the concentration of Epo outside of the cell,

$$\begin{aligned}
\dot{x}_1(t) &= -\theta_1 x_1(t)u(t) + 2\theta_4 x_4(t-\tau), \\
\dot{x}_2(t) &= \theta_1 x_1(t)u(t) - \theta_2 x_2^2(t), \\
\dot{x}_3(t) &= -\theta_3 x_3(t) + 0.5\theta_2 x_2^2(t), \\
\dot{x}_4(t) &= \theta_3 x_3(t) - \theta_4 x_4(t-\tau),
\end{aligned} \tag{4}$$

on $t \in \mathbb{R}^+$ with a mixture of known history function $x_2(t<0) = x_3(t<0) = x_4(t<0) = 0$ and constant but unknown $x_1(t<0)$. The unknown reaction rate constants $\theta_1, \theta_2, \theta_3, \theta_4 \in \mathbb{R}^+$ describe the rate at which the STAT factors change states. Following [16], we initially assume $u(t)$ is linearly interpolated between its error-free measurements, or $u(t)$ can be modelled and estimated as though it were a fifth system equation in a manner similar to [21]. While parameter inference in [20] is based on a linear chain ODE approximation of the above model, the present work directly examines the delay model (4).

Direct measurement of the concentrations of the four STAT species is limited by experimental constraints. Observable concentrations are modeled by a partially nonlinear transformation $\mathbf{g}: \mathbb{R}^4 \to \mathbb{R}^4$ of the states,

$$\begin{aligned}
g_1(\mathbf{x}, \boldsymbol{\theta}) &= \theta_5(x_2 + 2x_3), \\
g_2(\mathbf{x}, \boldsymbol{\theta}) &= \theta_6(x_1 + x_2 + 2x_3), \\
g_3(\mathbf{x}, \boldsymbol{\theta}) &= x_1, \\
g_4(\mathbf{x}, \boldsymbol{\theta}) &= x_3/(x_2 + x_3),
\end{aligned} \tag{5}$$

with non-negative scaling factors θ_5 and θ_6. As GS is a maximum profile likelihood estimation method, any appropriate likelihood model can be used but, for expositional simplicity, observations are assumed to have additive Gaussian noise with mean zero and, in this case, known standard deviations based on experimental conditions. The assumption of known standard deviations is not required for GS, it is used here to define likelihood weights that could otherwise be estimated iteratively [22].

$$\mathbf{y_i}(t) = \mathbf{g_i}[\mathbf{x}(t), \boldsymbol{\theta}, \mathbf{u}(t)] + \boldsymbol{\epsilon}(t), \qquad \epsilon_i(t) \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\sigma}_i^2(t)), \quad i = 1, 2, 3, 4. \tag{6}$$

The data, shown in Fig. 1 include 16 observations for each of the first two observation processes, and a single artificial observations for each of the third and fourth processes proposed by [20] in their resolution to the problem of lack of identifiability. The first artificial observation $y_3(0)$ defines measurement units for all four factors,

Please cite this article in press as: D.A. Campbell, O. Chkrebtii, Maximum profile likelihood estimation of differential equation parameters through model based smoothing state estimates, Math. Biosci. (2013), http://dx.doi.org/10.1016/j.mbs.2013.03.011

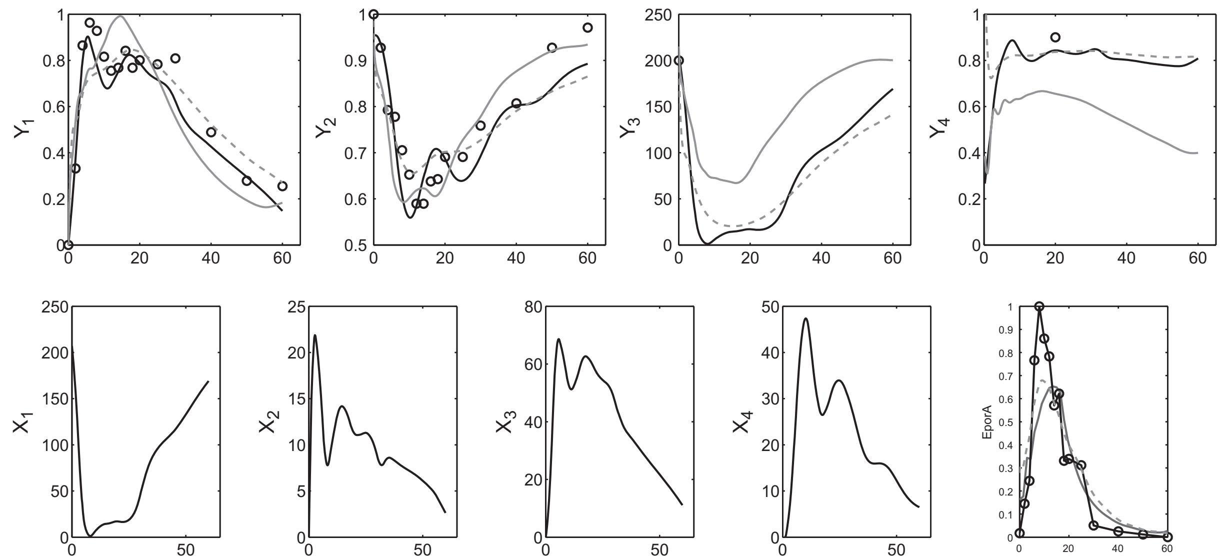

Fig. 1. Top: The JAK-STAT data and estimated observation processes with the model based smooth from approach A (solid black line), approach C (dashed black line), D (solid grey line), and E (dashed grey line). Bottom: The untransformed DDE solution estimates and EpoRa from the same three approaches.

while the artificial observation $y_4(20)$ allows identification between species x_2 and x_3. A thorough identifiability analysis for this system and discussion of the artificial observations is given in [20].

Features of this partially observed system that make inference challenging, are typical in biochemical applications:

1. Discontinuities in the second derivative $d^2\mathbf{x}(t)/dt^2$ of the states occur as a result of a delay in the model, and a piecewise linear forcing function;
2. Known initial states must be enforced when estimating the state functions within the collocation based GS framework;
3. Unknown initial states which define the process history must be estimated together with the other model parameters;
4. Scaling factors θ_5 and θ_6 are not directly involved in the model of the states, but are part of the measurement process which is a partially nonlinear transformation of $\mathbf{x}(t)$;
5. The unknown delay τ plays a role in reducing the model, since un-modelled states likely exist between x_4 and x_1. As this mechanism is not fully understood, an estimated value of the delay may provide some information about the un-modelled reactions.

3. Generalized profile estimation

The GS approach estimates state functions $\mathbf{x}(t)$, parameters $\boldsymbol{\theta}$ and the model relaxation parameter λ by establishing a hierarchy where parameters at each level are expressed in terms of those above them. GS models system states by a basis expansion $\mathbf{x}(t) = \mathbf{b}'(\boldsymbol{\theta},\lambda)\boldsymbol{\phi}(t) \approx \mathbf{S}(\boldsymbol{\theta},\mathbf{x}[0],t)$ weighted by so-called nuisance parameters $\mathbf{b}$. For given basis and order, $\mathbf{b}$ is defined as a function of $\boldsymbol{\theta}$ and λ. For each λ and $\boldsymbol{\theta}$, the optimal $\mathbf{b}$ defines a model-based data smooth, balancing the fit between interpolating the data and solving the ODE model. The smoothing step described in Section 3.1 allows a relaxation from the ODE model to ease optimization and to allow for un-modelled process noise or other model discrepancies.

Structural parameters $\boldsymbol{\theta}$ define the vector field of the ODE model and control features such as limit cycles, exponential decay, or other behaviours. Structural parameters are of primary interest for their interpretability and potential use in making decisions and predictions. Section 3.2 describes the iterative profile likelihood optimization process by which $\boldsymbol{\theta}(\lambda)$ is estimated by profiling over $\mathbf{b}(\hat{\boldsymbol{\theta}}\{\lambda\},\lambda)$. That is, we effectively ensure that for every incremental change in $\boldsymbol{\theta}(\lambda)$, the likelihood is adjusted to its optimum for $\mathbf{b}(\boldsymbol{\theta},\lambda)$.

The complexity or smoothing parameter λ, defines the top level of the hierarchy. The smoothing parameter determines the extent of the model relaxation permitted by the state estimation and consequently allows for some model misspecification. Estimation of λ is detailed in Section 3.3.

Computationally, the estimation routine can be thought of as a multi-level, multi-criterion optimization, that is it proceeds as a series nested optimization loops each with a different optimization criterion to estimate $\hat{\mathbf{b}}(\boldsymbol{\theta},\lambda)$, $\hat{\boldsymbol{\theta}}(\lambda)$, and λ. While the basic estimation process is described in [11], we outline the optimization process with special consideration to the intricacies of the JAK-STAT DDE model. Software to perform the estimation is available on request. Before going into estimation details we present the algorithmic workflow to clarify the estimation stages and nested loops.

Algorithm 1. Nested loops of profile likelihood maximization

```
Initialize: λ, θ, and b
while convergence criteria unmet for λ do
   Update λ by one step
   while convergence criteria unmet for (θ | λ) do
      Update θ by one step
      while convergence criteria unmet for (b | θ, λ) do
         Update b
      end while
   end while
end while
```

3.1. Inner optimization of $\mathbf{b}(\boldsymbol{\theta},\lambda)$; model-based smoothing

We describe a general setting where system states are measured indirectly through the noisy observation process (6) at times $\mathbf{t}_{ik} \in (0,T)$, for states $k = 1,\ldots,K$, and experimental runs $i = 1,\ldots,I$. For fixed $\boldsymbol{\theta}$ and λ, state functions $\hat{\mathbf{x}}_{ik}(t) = \mathbf{b}_{ik}(\boldsymbol{\theta},\lambda)'\boldsymbol{\phi}_{ik}(t)$ are estimated by minimizing the negative log likelihood with re-

Please cite this article in press as: D.A. Campbell, O. Chkrebtii, Maximum profile likelihood estimation of differential equation parameters through model based smoothing state estimates, Math. Biosci. (2013), http://dx.doi.org/10.1016/j.mbs.2013.03.011

spect to $\mathbf{b}(\theta,\lambda)$ subject to the differential equation model based penalty, PEN:

$$\hat{\mathbf{b}}(\boldsymbol{\theta},\lambda) = \arg\min_{\mathbf{b}} J(\mathbf{b}|\mathbf{y},\boldsymbol{\theta}(\lambda),\lambda) = \arg\min_{\mathbf{b}} \sum_{i=1}^{I}\sum_{k=1}^{K}\{-\log[\mathcal{L}(\boldsymbol{\theta}(\lambda);\mathbf{y}_{ik}(t))] + \lambda \mathtt{PEN}_{ik}\}.$$

For expositional simplicity we consider the likelihood from (6) where $-\log[\mathcal{L}(\boldsymbol{\theta}(\lambda);\mathbf{y}_{ik}(t))] \propto w_{ik}SSE_{ik}$ and optionally use weights $w_{ik}(t) = \hat{\sigma}_{ik}^{-2}(t)$ if known or estimated, giving:

$$\hat{\mathbf{b}}(\boldsymbol{\theta}(\lambda),\lambda) = \arg\min_{\mathbf{b}} \sum_{i=1}^{I}\sum_{k=1}^{K}\{w_{ik}SSE_{ik} + \lambda\mathtt{PEN}_{ik}\} = \arg\min_{\mathbf{b}} \sum_{i=1}^{I}\sum_{k=1}^{K} \times \left\{\sum_{t\in\mathbf{t}_{ik}} w_{ik}(t)\left[\mathbf{y}_{ik}(t) - \mathbf{g}_{ik}(\mathbf{b}'_{ik}\boldsymbol{\phi}_{ik}(t),\boldsymbol{\theta}(\lambda),\mathbf{u}(t))\right]^2 + \lambda\mathtt{PEN}_{ik}\right\}. \quad (7)$$

In contrast with non-parametric smoothing, where $\mathtt{PEN}_{ik} = \int (D^2\mathbf{x}_{ik}[t])^2 dt$ penalizes curvature, the GS framework introduces a penalty on the squared functional residual of (1) or (3). Omitting dependence on inputs $\mathbf{u}(t)$, this term,

$$\mathtt{PEN}_{ik} = \int_0^T (D\mathbf{x}_{ik}(s) - f_k(\mathbf{x}_i[s],\boldsymbol{\theta}(\lambda)))^2 ds, \quad (8)$$

penalizes deviations of the states from the model. For fixed $\lambda, \boldsymbol{\theta}$, and $\mathbf{y}_{ik}$ observed at the vector of times $\mathbf{t}_{ik}$, $\hat{\mathbf{b}}(\theta,\lambda)$ is obtained through a nonlinear regression step. It is important to note that the *SSE* term depends on observation function $\mathbf{g}(\cdot)$ while PEN does not, highlighting the fact that $\hat{\mathbf{b}}(\theta,\lambda)$ balances the observation function fit to the data with the state fit to the model. Consequently PEN permits smoothing even for unobserved states, in which case $\hat{\mathbf{x}}(t)$ follows the data only indirectly through $f(\cdot)$. We chose to evaluate the integral in (8) using a Simpson's rule numerical quadrature approximation with 3 evenly spaced points between each basis knot. System states may be estimated with unique bases and irregularly measured variables as long as the quadrature grid is the same for all states within each experimental run.

3.1.1. Dealing with delay differential equations

When the time domain is large relative to the anticipated value of τ, we propose replacing (7) and (8) with:

$$\hat{\mathbf{b}}(\boldsymbol{\theta},\lambda) = \arg\min_{\mathbf{b}} J(\mathbf{b}|\mathbf{y},\boldsymbol{\theta},\lambda) = \arg\min_{\mathbf{b}} \sum_{i=1}^{I}\sum_{k=1}^{K}\left\{w_{ik}SSE_{ik} + \lambda\mathtt{PEN}_{ik}^{(1)} + \lambda\mathtt{PEN}_{ik}^{(2)}\right\} \quad (9)$$

where

$$\mathtt{PEN}_{ik}^{(1)} = \int_\tau^T (D\mathbf{x}_{ik}(s) - f_k(\mathbf{x}_i[s],\boldsymbol{\theta}(\lambda)))^2 ds \quad (10)$$

$$\mathtt{PEN}_{ik}^{(2)} = \int_0^\tau \left(D^r\mathbf{x}_{ik}(s) - \frac{d^{r-1}}{ds^{r-1}} f_k(\mathbf{x}_i[s],\boldsymbol{\theta}(\lambda))\right)^2 ds \approx \int_0^\tau (D^r\mathbf{x}_{ik}(s))^2 ds \quad (11)$$

The penalty is split into a portion covering the DDE model after the first lag to the end of the time domain in (10), and a second non-parametric penalty on the first τ time in (11) units which borrows from functional data analysis methods and places a penalty on the rth model derivative so as to annihilate or nearly annihilate the lagged variable from the model. Eliminating the lagged variable avoids the infinite recursion associated with model derivatives that depend on even earlier histories that would otherwise also require estimation. It is important to maintain a penalty on the interval $(0,\tau)$ or the optimal delay will become $\hat{\tau} \geqslant T$ resulting in $\mathbf{x}(t)$ interpolating the data and effectively eliminating the flow of data based information to the estimation of $\hat{\theta}$ and λ. All parameter estimation methods for DDE models, including GS, face locally unidentifiable delay parameters when $\tau \geqslant T$. As a precaution a bounded optimization routine should be applied when estimating τ.

When the time domain is short compared to the anticipated value of τ, [23] propose estimating the history by a spline estimator. The states are estimated over interval $(-\tau, 0)$ using a B-spline expansion with the same knot density as the interval $(0,\tau)$ to permit consistent functional flexibility. Corresponding basis coefficients $\mathbf{b}^*$ are appended to $\mathbf{b}$ and estimated in the inner optimization. Unlike in (9)–(11), but following [23], no penalty is placed on the estimated history because data fitting utilizes $\hat{\mathbf{x}}(t\in(-\tau,0))$ indirectly through $f(\cdot)$, so that $\hat{\mathbf{x}}(t\in(-\tau,0))$ is regularized through the model and the data.

3.1.2. Choice of basis

Choice of the basis system $\{\phi_{ik}(t)\}$ for this inner level is an important consideration. In particular, it is important for the basis to span a function space that closely resembles the model dynamics for a wide range of θ. We use B-spline bases throughout this paper because of their flexibility and compact support. In some cases, use of a sufficiently flexible basis capable of matching fast changing features also requires interpretability constraints such as non-negativity or monotonicity, as outlined in [22]. Although JAK-STAT system states must be non-negative in order to maintain interpretability, in this example, we found inclusion of constraints to be unnecessary. Delay differential equation models can exhibit discontinuities in time derivatives higher than one at integer multiples of τ. Consequently, it may be necessary to use an adaptive basis in the inner optimization with multiple spline knots at times $\tau, 2\tau, 3\tau, \ldots$ to allow derivative discontinuities in the model and the basis expansion. The knot placement is then determined in part by $\hat{\tau}$ obtained in the outer optimization loop as part of $\hat{\theta}(\lambda)$, but otherwise $\hat{\mathbf{b}}(\theta,\lambda)$ is estimated as above using (7), consequently we leave adaptive basis details for Section 3.2.2 after discussing estimation of $\hat{\theta}(\lambda)$.

3.2. The outer optimization; estimating $\boldsymbol{\theta}$

The outer optimization provides structural parameter estimates that are used in the inner level. Structural parameters in the JAK-STAT model consist of the reaction rates $\theta_1, \theta_2, \theta_3$, constant time delay τ, initial value $x_1(0)$, and the observation scale factors θ_5, θ_6.

We model the data as observations from the noise process in (6) and choose $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}(\lambda)$ to minimize the negative log profile likelihood:

$$\hat{\theta}(\lambda) = \arg\min_{\theta(\lambda)} H(\boldsymbol{\theta}(\lambda),\mathbf{b}(\boldsymbol{\theta},\lambda)|\mathbf{y},\lambda) = \arg\min_{\theta(\lambda)} \sum_{i=1}^{I}\sum_{k=1}^{K} -\log[\mathcal{L}(\boldsymbol{\theta}(\lambda);\mathbf{y}_{ik}(t))]. \quad (12)$$

Using the Gaussian likelihood from (6) this becomes:

$$H(\boldsymbol{\theta}(\lambda),\mathbf{b}(\boldsymbol{\theta},\lambda)|\mathbf{y},\lambda) = \sum_{i=1}^{I}\sum_{k=1}^{K} w_{ik}SSE_{ik} = \sum_{i=1}^{I}\sum_{k=1}^{K}\sum_{t\in\mathbf{t}_{ik}} w_{ik}(t)[\mathbf{y}_{ik}(t) - \mathbf{g}_{ik}(\mathbf{b}_{ik}\{\boldsymbol{\theta}(\lambda),\lambda\}'\boldsymbol{\phi}_{ik}\{t\},\boldsymbol{\theta}(\lambda))]^2. \quad (13)$$

Variants of this function are widely used in classical parameter estimation methods. For example, replacing $\mathbf{b}_{ik}\{\theta,\lambda\}'\boldsymbol{\phi}_{ik}\{t\}$ in (13) with a numerical ODE solution produces the standard nonlinear regression objective function [1]. A model discrepancy term (like PEN) is not included in (13) because model fitting and deviation thereof is already considered in (7). For any $\lambda, \hat{\theta}(\lambda)$ is a maximum profile likelihood estimate (MPLE) of the vector of structural parameters. Therefore, for every new $\lambda, \hat{\boldsymbol{\theta}}(\lambda)$ is optimized in the outer loop using an iterative algorithm that further updates the nuisance parameters, $\mathbf{b}(\hat{\theta}, \lambda)$, in the inner loop at every step.

For fixed λ, and simplifying notation from $H(\boldsymbol{\theta}(\lambda), \mathbf{b}\{\boldsymbol{\theta}(\lambda), \lambda\}, \mathbf{g}[\mathbf{b}\{\boldsymbol{\theta}(\lambda), \lambda\}'\boldsymbol{\phi}(t), \boldsymbol{\theta}(\lambda)]|\mathbf{y}; \lambda)$ to H; from $\mathbf{g}(\mathbf{b}\{\boldsymbol{\theta}, \lambda\}'\boldsymbol{\phi}, \boldsymbol{\theta}(\lambda))$ to $\mathbf{g}$; from $\mathbf{b}\{\boldsymbol{\theta}(\lambda), \lambda\}$ to $\mathbf{b}$; and from $\boldsymbol{\theta}(\lambda)$ to $\boldsymbol{\theta}$; the total gradient for the profile likelihood is:

$$\frac{dH}{d\theta} = \frac{\partial H}{\partial \theta} + \frac{\partial H}{\partial \mathbf{g}}\frac{\partial \mathbf{g}}{\partial \theta} + \frac{\partial H}{\partial \mathbf{g}}\frac{\partial \mathbf{g}}{\partial \mathbf{b}}\frac{d\mathbf{b}}{d\theta}. \tag{14}$$

Typically, the vector field $f(\cdot)$ is a nonlinear function of $\mathbf{x}(t)$ and/or $\mathbf{x}(t-\tau)$ and no explicit formula for $\hat{\mathbf{b}}(\theta, \lambda)$ is available. Therefore, an expression for $d\mathbf{b}/d\theta$ is obtained below for $\mathbf{b} = \hat{\mathbf{b}}$ by applying the implicit function theorem. Simplifying the notation from $J(\theta, \mathbf{b}, \mathbf{g}|\mathbf{y}; \lambda)$ to J, we assume that H and J are twice continuously differentiable with respect to $\boldsymbol{\theta}, \mathbf{g}$, and $\mathbf{b}$, and that the Hessian matrices,

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

are positive definitive over a nonempty neighbourhood of $\mathbf{y}$ in the data space. The function (7) is optimized at $\mathbf{b} = \hat{\mathbf{b}}$, so that at this point $dJ/d\mathbf{b} = 0$ and

$$\begin{aligned}\frac{d^2 J}{d\mathbf{b}d\theta} &= \frac{d}{d\theta}\left(\frac{dJ}{d\mathbf{b}}\right)\\ &= \frac{\partial^2 J}{\partial \mathbf{g}\partial\theta}\frac{\partial \mathbf{g}}{\partial \mathbf{b}} + \frac{\partial^2 J}{\partial \mathbf{g}^2}\frac{\partial \mathbf{g}}{\partial \theta}\frac{\partial \mathbf{g}}{\partial \mathbf{b}} + \frac{\partial J}{\partial \mathbf{g}}\frac{\partial^2 \mathbf{g}}{\partial \mathbf{b}\partial\theta}\\ &\quad + \left\{\left(\frac{\partial \mathbf{g}}{\partial \mathbf{b}}\right)'\frac{\partial^2 J}{\partial \mathbf{g}^2}\left(\frac{\partial \mathbf{g}}{\partial \mathbf{b}}\right) + \frac{\partial J}{\partial \mathbf{g}}\frac{\partial^2 \mathbf{g}}{\partial \mathbf{b}^2}\right\}\frac{d\mathbf{b}}{d\theta},\end{aligned} \tag{15}$$

and, solving for $d\mathbf{b}/d\theta$ at $\mathbf{b} = \hat{\mathbf{b}}$, we obtain

$$\begin{aligned}\frac{d\mathbf{b}}{d\theta} = &-\left\{\left(\frac{\partial \mathbf{g}}{\partial \mathbf{b}}\right)'\frac{\partial^2 J}{\partial \mathbf{g}^2}\left(\frac{\partial \mathbf{g}}{\partial \mathbf{b}}\right) + \frac{\partial J}{\partial \mathbf{g}}\frac{\partial^2 \mathbf{g}}{\partial \mathbf{b}^2}\right\}^{-1}\\ &\times\left\{\frac{\partial^2 J}{\partial \mathbf{g}\partial\theta}\frac{\partial \mathbf{g}}{\partial \mathbf{b}} + \frac{\partial^2 J}{\partial \mathbf{g}^2}\frac{\partial \mathbf{g}}{\partial \theta}\frac{\partial \mathbf{g}}{\partial \mathbf{b}} + \frac{\partial J}{\partial \mathbf{g}}\frac{\partial^2 \mathbf{g}}{\partial \mathbf{b}\partial\theta}\right\},\end{aligned} \tag{16}$$

which we substitute into (14) to obtain the total gradient for the MPLE:

$$\begin{aligned}\frac{dH}{d\theta} = &\frac{\partial H}{\partial \theta} + \frac{\partial H}{\partial \mathbf{g}}\frac{\partial \mathbf{g}}{\partial \theta} - \frac{\partial H}{\partial \mathbf{g}}\frac{\partial \mathbf{g}}{\partial \mathbf{b}}\left\{\left(\frac{\partial \mathbf{g}}{\partial \mathbf{b}}\right)'\frac{\partial^2 J}{\partial \mathbf{g}^2}\left(\frac{\partial \mathbf{g}}{\partial \mathbf{b}}\right) + \frac{\partial J}{\partial \mathbf{g}}\frac{\partial^2 \mathbf{g}}{\partial \mathbf{b}^2}\right\}^{-1}\\ &\times\left\{\frac{\partial^2 J}{\partial \mathbf{g}\partial\theta}\frac{\partial \mathbf{g}}{\partial \mathbf{b}} + \frac{\partial^2 J}{\partial \mathbf{g}^2}\frac{\partial \mathbf{g}}{\partial \theta}\frac{\partial \mathbf{g}}{\partial \mathbf{b}} + \frac{\partial J}{\partial \mathbf{g}}\frac{\partial^2 \mathbf{g}}{\partial \mathbf{b}\partial\theta}\right\}.\end{aligned} \tag{17}$$

In practice the profile likelihood may be highly nonlinear such that gradient approximations for components of (17) based on finite differences may be numerically unstable if analytic forms are not available. If the assumption of identifiable parameters is not met, (17) may involve inverting rank deficient matrices and practitioners are reminded to consider alternative optimization strategies, several of which are compared in [24].

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

In general, under the mild conditions specified in [25], the GS maximum profile likelihood estimator is asymptotically normal with covariance matrix equal to that of the maximum likelihood estimator, meaning that one could use the observed inverse Fisher information:

$$\mathrm{var}\left(\hat{\theta}(\lambda)\right) = I(\theta)^{-1} = \left(\frac{d^2}{d\theta^2}H\right)^{-1} \tag{18}$$

as an asymptotic covariance estimate. The total derivative in (18) can be determined using the implicit function theorem through derivatives specified in the appendix. This is the strategy we used for obtaining interval estimates.

Alternatively one may consider interval estimates based on a first order Taylor expansion of the likelihood. Linearization strategies for interval estimates can over or underestimate the actual sampling variance since the quality of the approximation will depend on the remainder of the Taylor expansion. The choice of likelihood and the nonlinearity of the model determine the utility of the linearization method, so we present it here and caution the reader to consider them as a crude approximation and instead advocate the Fisher Information Intervals above. The advantage to the linearization is the ability to incorporate the data covariance structure as outlined below.

In the Gaussian likelihood example from (6), under the assumption of independent errors and minor non-linearities in the model around the parameter estimates, with covariance matrix $\mathrm{Cov}(\mathbf{y}) = \mathrm{diag}(\boldsymbol{\sigma}^2)$, sampling variances for $\hat{\boldsymbol{\theta}}$ can be obtained using a linearization approximation to the variance estimator:

$$\mathrm{Var}(\hat{\theta}) \approx \frac{d\hat{\theta}'}{d\mathbf{y}}\mathrm{Cov}(\mathbf{y})\frac{d\hat{\theta}}{d\mathbf{y}}, \tag{19}$$

An expression for $d\boldsymbol{\theta}/d\mathbf{y}$ is derived below using the fact that $dH/d\boldsymbol{\theta} = 0$ at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$.

$$\frac{d}{d\mathbf{y}}\left(\frac{dH}{d\theta}\right)\bigg|_{\theta=\hat{\theta}} = \left[\frac{d^2 H}{d\mathbf{y}d\theta} + \frac{d^2 H}{d\theta^2}\frac{d\theta}{d\mathbf{y}}\right]\bigg|_{\theta=\hat{\theta}} = 0, \tag{20}$$

and, using the implicit function theorem at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$,

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

where $d^2H/d\theta^2$ and $d^2H/d\theta d\mathbf{y}$ are provided in the appendix.

More generally, the process of smoothing often introduces serial dependence between the residuals at consecutive time points. One way to account for this is by using the asymptotic Newey–West covariance sandwich estimator introduced [26] [27]:

$$Cov(\theta) \approx V_0^{-1}\left(V_0 + \sum_{\ell=1}^{l}\left(1 - \frac{\ell}{l+1}\right)(V_k + V_k')\right)V_0^{-1}, \tag{21}$$

where,

$$V_k = \frac{\hat{\boldsymbol{\sigma}}^2}{n}\sum_{t=k+1}^{n}\left[\phi(t-k)'\frac{d\mathbf{b}'}{d\theta}\frac{\partial g'}{\partial \mathbf{b}} + \frac{\partial g(t-k)'}{\partial\theta}\right]\left[\frac{\partial g(t)}{\partial\theta} + \frac{\partial g}{\partial \mathbf{b}}\frac{d\mathbf{b}}{d\theta}\phi(t)\right]. \tag{22}$$

The resulting covariance estimate is equivalent to (19) when the errors are independent, but is a more conservative estimate when serial correlations exist. In practice l can be chosen to permits lags spanning the full dataset, however in practice that is not necessary, and [27] suggest setting $l = n/5$ as a rule of thumb.

A tempting alternative is to consider parametric bootstrap standard errors based on repeatedly simulating data from the likelihood centred around $\mathbf{g}\left(\mathbf{S}\left[\hat{\theta}, \hat{\mathbf{x}}(0), t\right], \hat{\theta}\right)$ and re-estimating $\hat{\boldsymbol{\theta}}$. However, the generalized smoothing approach is designed to provide flexibility by relaxing the DDE model under the assumption that the model is useful but imperfect. Consequently, parametric bootstrap based on the numeric DDE solution can generate data from the incorrect model producing invalid standard error estimates when model discrepancies are present.

The asymptotic confidence intervals for $\boldsymbol{\theta}$ are based on the likelihood curvature. However, [20] show that without the addition of the artificial observations in the JAK-STAT system, some

Please cite this article in press as: D.A. Campbell, O. Chkrebtii, Maximum profile likelihood estimation of differential equation parameters through model based smoothing state estimates, Math. Biosci. (2013), http://dx.doi.org/10.1016/j.mbs.2013.03.011

parameters are not identifiable. The identifiability diagnostics of [20] and other likelihood contour intervals [1] will in some cases be more appropriate than the asymptotic intervals based on the delta method. These interval finding routines can be performed using the likelihood based on the smooth state estimate in (12). Lack of identifiability is a complex issue and a thorough treatment extends beyond the scope of this paper, we therefore refer the reader to [28] for an overview of the reasons behind and possible solutions for lack of identifiability.

3.2.2. Adaptive basis

To permit additional flexibility in $\mathbf{x}(t)$, the state estimator should have discontinuous higher derivatives to match those imposed in the model by the delay. In the present application, the history is considered to be constant giving rise to a discontinuity in $\dot{\mathbf{x}}(t=0)$. This propagates through the system leading to discontinuities in $\ddot{\mathbf{x}}(t=\tau), \ddot{\mathbf{x}}(t=2\tau), \ldots$, which can be accommodated by using a B-spline basis of order n_{order} and placing $n_{\text{order}}-2$ spline knots at $\tau, 2\tau, \ldots$

As τ is estimated iteratively, the basis must be modified at each algorithm iteration. It is important to note that changes in the number of bases can potentially confound our estimate of $\boldsymbol{\theta}(\lambda)$, as they allow $\mathbf{x}(t)$ more or less flexibility to accommodate the dynamics described by a particular $\boldsymbol{\theta}(\lambda)$. As a result, $\hat{\boldsymbol{\theta}}(\lambda)$ could be pushed towards values that are biased but ensure a restricted basis that will reduce (8). To avoid attributing this flexibility to the choice of parameters, we suggest using a number of basis functions large enough that addition or subtraction of $n_{\text{order}}-2$ bases has negligible impact on the flexibility of $\mathbf{x}(t)$ to match model features between lag intervals. We also note that even when using an adaptive basis, fixed quadrature points must be maintained for the numerical approximation of (8) to ensure that the quality of the approximation remains comparable between iterations. Consequently a dense quadrature grid will be required to regularize the adaptive basis expansion.

3.3. Estimating the complexity parameter

A variety of approaches have been proposed for choosing the complexity parameter λ, and here we offer an overview of the main ideas. We divide the discussion into the case where $\boldsymbol{\sigma}^2$ is unknown and the case when it is known.

3.3.1. Unknown $\boldsymbol{\sigma}^2$

Standard cross-validation is designed for non-parametric smoothing and tends to produce good estimates of $\mathbf{x}(t)$, but suggests less than optimal values for estimating $\boldsymbol{\theta}$. A few methods have been proposed for estimating λ such as, minimizing by choice of λ, the squared deviation between the estimated states $\mathbf{x}(t)$ and the numerical ODE solution $\mathbf{S}(\hat{\boldsymbol{\theta}}, \hat{\mathbf{x}}(0), t)$ [29], minimizing forward prediction error [27], or choosing the best model fit available to the basis expansion [22]. The first case is useful if the model and the data are thought to be in agreement. In all cases it is important to consider the role of λ:

- When $\lambda = 0, \mathbf{x}(t)$ interpolates the data;
- When $\lambda \to \infty$, $\mathbf{x}(t) \to \mathbf{S}(t)$.

In general, λ controls the flow of information from $\mathbf{y}$ to $\boldsymbol{\theta}$. When $\lambda = 0$, $\mathbf{x}(t)$ does not depend on $\boldsymbol{\theta}$, and results in a likelihood that is completely flat (uninformed) in $\boldsymbol{\theta}$. As $\lambda \to \infty$, GS becomes a variant of nonlinear least squares regression based on $\mathbf{S}(t)$. The role of λ is complicated by the fact that the function space of $\mathbf{x}(t)$ is limited by the choice of basis, and consequently $\mathbf{x}(t) \to \mathbf{S}(t)$ only occurs in the intersection of the function spaces of model solutions and basis expansions. Generally this intersection occurs over a limited set of $\boldsymbol{\theta}(\lambda)$, and it is unwise to consider the limiting case as results will differ substantially from those based directly on $\mathbf{S}(\boldsymbol{\theta}, \mathbf{x}[0], t)$.

The shift in focus of $\mathbf{x}(t)$ and changes in the likelihood of $\boldsymbol{\theta}$ as a result of increasing λ suggest using the annealing-type strategy of beginning with a small λ and gradually increasing it until some optimality criterion is met.

The cross-validation based forward prediction error method of [27] minimizes:

$$\sum_{m=0}^{M}\sum_{t=t_m}^{t_m+h}\Big(\mathbf{y}(t) - S(\hat{\boldsymbol{\theta}}(\lambda), \hat{\mathbf{x}}(t))\Big)^2,$$

the squared difference between the observations and the model solution, initialized at the smooth estimate, over an interval of size h. Effectively this criterion assumes that the model is correct over a small interval but permits model deviations to propagate as the time domain increases. A special case of forward prediction error is when $m=0$ and $h=T$, which assumes that the DDE model is an accurate long term data representation [29]. The computational complexity of the forward prediction error method may be substantial, and consequently we outline an alternative algorithm from [22]:

1. Begin with small λ
2. Compute *SSE* and `PEN`.
3. Increase λ by an order of magnitude, as a result *SSE* increases and *PEN* decreases.
4. Increase λ until `PEN` and *SSE* stabilize, and choose the largest λ before `PEN` decreases sharply at the expense of large increase in *SSE*.

In practice the above algorithm will select the largest λ that the basis can accommodate without introducing bias due to differences between the function spaces of $\phi(t)$ and $\mathbf{S}(t)$. When the model is close to that describing the true dynamics, the above algorithm will terminate when $\mathbf{x}(t)$ is near $\mathbf{S}(t)$. When the model describes dynamics that are not reflected in the data, then a much smaller λ will be produced allowing extraction of some useful information about $\boldsymbol{\theta}(\lambda)$ from the data through a model relaxation.

3.3.2. Known $\boldsymbol{\sigma}^2$

Discrepancies between the data and model can arise for many reasons, most of which could be approximated by a form of process error or functional disturbance, where λ can then be interpreted as the ratio of process noise to measurement noise in the case that (1) is forced by a stochastic Wiener process. When extensive experiments have been performed and $\boldsymbol{\sigma}^2$ is fairly well known, $\hat{\lambda}$ can be selected so as to match the sample variance $\hat{\boldsymbol{\sigma}}^2(\lambda)$, computed after fitting the model under fixed λ, with the known $\boldsymbol{\sigma}^2$ as outlined in [30]:

$$\hat{\lambda} = \arg\min_{\lambda}\left(\frac{\hat{\sigma}^2(\lambda)}{\sigma^2} - 1\right)^2.$$

This results in selection of $\hat{\lambda}$ that produces the best fit to the data while ensuring the model discrepancy matches prior knowledge of the observation error.

Due to the hierarchical structure of $\hat{\theta}(\lambda)$ and the difficult topological features of differential equation model likelihoods, it is still strongly recommended that λ be initialized at a small value, and then incrementally increased to refine $\hat{\theta}(\lambda)$.

4. Estimation details and results for the JAK-STAT data

We use the JAK-STAT data from [16], including the artificial observations and parameter reduction of [20]. Although [20] use

Please cite this article in press as: D.A. Campbell, O. Chkrebtii, Maximum profile likelihood estimation of differential equation parameters through model based smoothing state estimates, Math. Biosci. (2013), http://dx.doi.org/10.1016/j.mbs.2013.03.011

the simplification $\tau = 10/\theta_4$ as part of a linear chain ODE approximation with 10 chain steps, the present work directly examines the delay model (4) and uses the simplification $\tau = 1/\theta_4$ to assist with parameter estimability.

We consider two types of estimation based on considering $u(t)$ known and linearly interpolated, and $u(t)$ unknown and estimated.

4.1. Basis details

Three smoothing approaches were applied to the JAK-STAT data set to examine the impact of changes in the basis. Because the JAK-STAT system states have discontinuities in the second derivatives at times $\tau, 2\tau, \ldots$, approaches *A* and *B* are constructed with a much rougher higher derivative structure than is needed to satisfy this requirement. These approaches use fixed cubic B-spline bases with 19 and 39 evenly spaced interior knots respectively. In contrast, approach *C* uses the smoother quintic spline basis functions with 59 interior knots, but with the additional interior knots placed at $\tau, 2\tau, \ldots$ to provide a basis with second derivative discontinuities matching those of the JAK-STAT model.

All of the approaches use more knots than observations; states y_1 and y_2 have 16 observations each, while y_3 and y_4 have only 1 observation each. In general, more knots are required than observations so that the basis can approximate the model dynamics over a wide range of $\boldsymbol{\theta}$ values. A denser basis should be used if the model dynamics are expected to exhibit particularly sharp features.

Another reason for using a B-spline basis system is that only the first basis function, $\phi_0(0) = 1$, is nonzero at $t = 0$. Consequently with JAK-STAT, known initial conditions $x_2(0) = x_3(0) = x_4(0) = 0$ are enforced by maintaining the first coefficients in $\mathbf{b}$ fixed at the known values while optimizing (7).

The remaining initial state $x_1(0)$ is unknown and must be estimated. GS typically estimates initial states in the same way as states at any point on the domain. In contrast to the nonlinear regression methodology where numerical solver error propagates over the time domain, GS spreads the approximation uncertainty more or less uniformly. As a result states are estimated at every time point, including $t = 0$, in the same way by $\mathbf{b}'\boldsymbol{\phi}(t)$.

In delay systems where the function $\mathbf{x}(t < 0)$ is considered constant, the initial state defines the system history. In such cases one could consider $\mathbf{x}(0)$ as a structural parameter, including it in the vector $\boldsymbol{\theta}$ and estimating remaining basis coefficients as in the case where initial states are known. JAK-STAT only requires the history for x_4, which is known, so the initial state for $x_1(0)$ can be estimated with the other $\mathbf{b}$. However, to demonstrate ability to be consistent with other literature studies of the JAK-STAT system, we estimated $x_1(0)$ along with $\boldsymbol{\theta}$.

Approaches A, B, and C, were performed using a linear interpolator for $u(t)$ as per [16]. However if the forcing functions $\mathbf{u}(t)$ are not known perfectly they should be estimated and treated as the additional state. Following [21], we also use attempt using a cubic spline smoother defining the additional state $\mathbf{x}_5(t) = \mathbf{u}(t)$ where:

$$\mathbf{y}_5(t) = \mathbf{g}_5[\mathbf{x}(t), \boldsymbol{\theta}] = \mathbf{x}_5(t).$$

Approaches D and E have the same knot basis as A but treat the estimation of $\mathbf{x}_5(t)$ in two different ways. Approach D follows [21] in modelling the additional state as

$$\frac{d^2}{dt^2}\mathbf{x}_5(t) = 0.$$

The resulting component of PEN in (8) is equivalent to standard spline smoothing methods with a curvature based penalty, except that $\hat{\mathbf{x}}_5(t)$ is also guided by the model to help fit the other states as well. As $\lambda \to \infty$, PEN ensures that the estimated Epo concentration will therefore be linear, however the data clearly show an increase and subsequent decrease. We therefore also consider approach E, a model that will permit an additional change in curvature through the higher order penalty:

$$\frac{d^4}{dt^4}\mathbf{x}_5(t) = 0 \tag{23}$$

4.2. Optimization details

The inner optimization was performed using gradient based optimization. Although it is technically possible to perform the outer optimization through gradient based methods using the total gradient (17), we selected a gradient-free algorithm in our analysis. Ordinary differential equation models without delays, and subject to a linear observation function $\mathbf{g}(\cdot)$ require simpler derivatives which can usually be determined analytically, in contrast to the JAK-STAT system where numeric derivatives were highly unstable. For this reason, the outer optimization was implemented using the genetic optimization algorithm, *ga*, in Matlab (version 2012a, The MathWorks Inc., Natick, MA, 2012). As is common in general optimization, parameters were rescaled by multiplicative factors so that their anticipated optimal values were in the interval $(0, 10)$. The genetic algorithm is a parallel-coded, bounded optimization routine with lower bounds for all parameters set to zero to be consistent with interpretability. Upper bounds were set to 50 after rescaling.

Decreasing λ results in a smoother likelihood surface [11], behaviour which is exploited in the annealing type of approach to optimizing λ. Consequently GS is inherently robust to parameter values used to initialize the algorithm. The random starting values of the genetic algorithm enable efficient exploration of the parameter space, further adding to the robustness to initialization. The genetic algorithm was initially performed with 100 random starting points, holding $\lambda = 1$ fixed. Optimization continued in sequence by setting $\lambda_{\text{new}} = 10 \times \lambda_{\text{old}}$ and setting $\hat{\theta}(\lambda_{\text{old}})$ from all previous λ_{old} values as deterministic points among the otherwise random starting values for the genetic algorithm. Because measurement error standard deviations for the JAK-STAT data are provided in [16], we estimate $\hat{\lambda}$ by considering $\sigma_{ik}^2(t)$ to be known, as outlined in Section 3.3.2, using the weights $w_{ik}(t) = 1/\sigma_{ik}^2(t)$ in (7) and (13). In general weights and error term variances can be estimated using iteratively re-weighting [22].

4.3. Results

Resulting point and standard error estimates, constructed using (21), are given in Table 1. The Newey–West estimator was constructed with $l = 10$. Parameter estimates were fairly robust to $\hat{\lambda}$, giving stable $\hat{\theta}(\lambda)$ over a wide range of values. Fig. 1 shows $\hat{\mathbf{x}}(t)$ and $\mathbf{S}(\hat{\theta}, \hat{\mathbf{x}}(0), t)$ obtained using approach A, although $\hat{\mathbf{x}}(t)$ is visually nearly indistinguishable under approaches A and B. While impacting $\hat{\lambda}$, an increase in the basis density had little noticeable effect on the model fit. Denser bases permit more flexibility and require more quadrature weights in (8) to control the fit to the model. For this reason $\hat{\lambda}$ does not have a straightforward interpretation that is comparable between models.

Approach C uses a higher order basis, more basis coefficients, and an adaptive basis to accommodate discontinuous second derivatives at $\tau, 2\tau, \ldots$ Since $\mathbf{x}(t)$ is an approximation to the DDE solution, the added model matching ability of the adaptive basis approach does not seem to have a strong impact on the results and consequently the fit to approaches A and C are nearly indistinguishable in Fig. 1. In all cases the richness of the basis was suffi-

Please cite this article in press as: D.A. Campbell, O. Chkrebtii, Maximum profile likelihood estimation of differential equation parameters through model based smoothing state estimates, Math. Biosci. (2013), http://dx.doi.org/10.1016/j.mbs.2013.03.011

Table 1
Point estimates and standard errors (in brackets) for $\boldsymbol{\theta}$ from the different approaches. Approaches A, B, and C use a linearly interpolated $u(t)$, where A has a low knot density basis, B has a high knot density basis, and C uses an adaptive knot density basis. Approaches D and E use a low knot density basis but estimate $u(t)$ with a second and fourth derivative penalty, respectively.

Approach	θ_1	θ_2	θ_3	$x_1(0)$	θ_5	θ_6	τ	λ
A	2.43 (0.43)	0.11 (0.04)	0.128 (0.039)	207 (25)	0.0059 (0.031)	0.0046 (0.008)	3.91 (0.60)	1000
B	2.64 (0.37)	0.34 (0.13)	0.150 (0.064)	166 (22)	0.0071 (0.0084)	0.0051 (0.008)	4.63 (0.64)	1000
C	2.38 (0.41)	0.28 (0.09)	0.228 (0.051)	215 (18)	0.0062 (0.009)	0.0045 (0.008)	4.50 (0.61)	10000
D	1.02 (0.35)	0.13 (0.10)	0.456 (0.15)	215 (14)	0.0134 (0.018)	0.0043 (0.010)	2.89 (0.16)	1000
E	1.82 (0.07)	0.77 (0.16)	0.174 (0.009)	207 (7.4)	0.0068 (0.024)	0.0048 (0.011)	3.18 (0.25)	1000

cient to approximate the DDE trajectory and estimate parameters. However, approach C shows considerably more flexibility in matching the DDE solution as the function space spanned by $\{\phi_{ik}\}$ is closer to that spanned by $\mathbf{S}(t)$. As a result, approach C was able to follow the DDE model to a higher degree of fidelity, which led to a much wider range of λ values over which $\hat{\boldsymbol{\theta}}(\lambda)$ was nearly constant.

Approaches A, D, and E had the fewest basis functions and therefore the computation of $\mathbf{b}(\boldsymbol{\theta},\lambda)$ was faster than under approaches B and C. In other models, richness of the basis can have a substantial impact on the results, for example compare the nylon example in [11,22].

The minimal differences in the estimates between approaches A, B, and C shown in Table 1 are most likely due to the near unidentifiability of the system under study. The benefit of GS is that it relaxes the model dynamics towards the data while extracting the available information about the structural parameters. Additionally, robustness of the method to parameter initializations and complex likelihood topologies makes the GS a useful tool for estimating complex models. As was observed by [21], large differences in parameter and state estimates resulted from differences in how $u(t)$ was treated. Treating $u(t)$ as a state and estimating it along with the others means that a single λ is selected to capture all of the features in all states. This is in contrast to [21], who fitted a non-parametric smooth to $u(t)$ and then used a numerical solver for the remaining states.

5. Discussion

We outline the GS methodology, originally designed for estimating parameters from ordinary differential equation models, and show extensions to account for real data challenges, including partially known initial conditions, unobserved system states, nonlinear observation functions, and systems where delays are modelled directly rather than being implemented through modelling additional states.

As with any method it is important to consider the strengths and limitations of the methodology.

In the JAK-STAT example the history is assumed constant, with $x_2(t)$, $x_3(t)$ and $x_4(t)$ known to be zero over the interval $t \in (-\tau, 0)$, and $x_1(t)$ constant but unknown. However, in many cases the function $\mathbf{x}(t \in (-\tau, 0))$ must be estimated. Estimation of the history is an important problem as DDE systems may exhibit chaotic behaviour with small changes in τ and/or $\mathbf{x}(t \in (-\tau, 0))$. Chaotic behaviour translates into complex likelihood topologies and the model relaxation towards the data enables the GS approach to smooth out the likelihood, simplifying optimization.

The GS method uses a model based smooth state estimate, which relaxes the solution to the differential equation towards the data. This feature permits parameter estimation when the model is somewhat mis-specified, and can ease optimization by avoiding some topological pitfalls associated with using a numerical differential equation solver.

The inability to observe all modelled states is overcome in the GS approach by feeding information from observed states through the model to guide the fit to unobserved ones. However if all states are sparsely observed, there is not enough information in the data to guide the model relaxation permitted by the smooth. Severe sparsity requires stronger assumptions on the underlying process and consequently we recommend numerical solver based methods instead. As a rule of thumb, exploratory plots of the data should reveal a clear signal, at least in the observed modelled states. Although the JAK-STAT model has the same number of modelled states as observation processes, this need not be the case. While smoothing based methods overcome topological constraints in the likelihood surface, GS has the further advantage of pooling information from the model and the observed states to estimate unobserved states. GS requires optimization of basis coefficients and therefore will generally be slower than producing a numerical solution.

Different values of λ may be useful for estimating different states or may enhance or hinder different model features. Consequently an area of future research is to consider a scale space approach to state estimation. This approach has been very successful in estimating local features from data [31] and may be an efficient way to consider the roles of different types of modelled states, for example consider the distinction between $\mathbf{x}_1(t), \ldots, \mathbf{x}_4(t)$ with a well devised DDE versus the convenient penalty placed on $\mathbf{x}_5(t)$.

Acknowledgements

The authors thank Shota Gugushvili and two anonymous reviewers for helpful comments and discussion in the preparation of this manuscript.

Appendix A. Additional implicitly defined derivatives

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

A.1. $d^2H/d\theta^2$

$$\begin{aligned}
\frac{d^2H}{d\theta^2} = {} & \frac{\partial^2 H}{\partial\theta^2} + \frac{\partial^2 H}{\partial\theta\partial\mathbf{g}}\frac{\partial\mathbf{g}}{\partial\theta} + \frac{\partial^2 H}{\partial\theta\partial\mathbf{g}}\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\frac{d\mathbf{b}}{d\theta} + \frac{\partial^2 H}{\partial\mathbf{g}\partial\theta}\frac{\partial\mathbf{g}}{\partial\theta} + \left(\frac{\partial\mathbf{g}}{\partial\theta}\right)'\frac{\partial^2 H}{\partial^2\mathbf{g}}\left(\frac{\partial\mathbf{g}}{\partial\theta}\right) \\
& + \frac{\partial^2 H}{\partial^2\mathbf{g}}\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\frac{d\mathbf{b}}{d\theta}\frac{\partial\mathbf{g}}{\partial\theta} + \frac{\partial H}{\partial\mathbf{g}}\frac{\partial^2\mathbf{g}}{\partial\theta^2} + \frac{\partial H}{\partial\mathbf{g}}\frac{\partial^2\mathbf{g}}{\partial\theta\partial\mathbf{b}}\frac{d\mathbf{b}}{d\theta} + \frac{\partial^2 H}{\partial\mathbf{g}\partial\theta}\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\frac{d\mathbf{b}}{d\theta} \\
& + \frac{\partial^2 H}{\partial\mathbf{g}^2}\frac{\partial\mathbf{g}}{\partial\theta}\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\frac{d\mathbf{b}}{d\theta} + \left(\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\frac{d\mathbf{b}}{d\theta}\right)'\frac{\partial^2 H}{\partial\mathbf{g}^2}\left(\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\frac{d\mathbf{b}}{d\theta}\right) + \frac{\partial H}{\partial\mathbf{g}}\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}\partial\theta}\frac{d\mathbf{b}}{d\theta} \\
& + \frac{\partial H}{\partial\mathbf{g}}\left(\frac{d\mathbf{b}}{d\theta}\right)'\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}^2}\left(\frac{d\mathbf{b}}{d\theta}\right) + \frac{\partial H}{\partial\mathbf{g}}\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\frac{d^2\mathbf{b}}{d\theta^2},
\end{aligned} \tag{A.1}$$

A.2. $d^2H/d\theta d\mathbf{y}$

$$\begin{aligned}\frac{d^2H}{d\theta d\mathbf{y}} &= \frac{\partial^2 H}{\partial\theta\partial\mathbf{y}} + \frac{\partial^2 H}{\partial\theta\partial\mathbf{g}}\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\frac{\partial\mathbf{b}}{\partial\mathbf{y}} + \frac{\partial^2 H}{\partial\mathbf{g}\partial\mathbf{y}}\frac{\partial\mathbf{g}}{\partial\theta} + \frac{\partial^2 H}{\partial\mathbf{g}^2}\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\frac{\partial\mathbf{b}}{\partial\mathbf{y}}\frac{\partial\mathbf{g}}{\partial\theta}\\
&\quad + \frac{\partial H}{\partial\mathbf{g}}\frac{\partial^2\mathbf{g}}{\partial\theta\partial\mathbf{b}}\frac{\partial\mathbf{b}}{\partial\mathbf{y}} + \frac{\partial^2 H}{\partial\mathbf{g}\partial\mathbf{y}}\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\frac{d\mathbf{b}}{d\theta} + \left(\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\frac{\partial\mathbf{b}}{\partial\mathbf{y}}\right)'\frac{\partial^2 H}{\partial\mathbf{g}^2}\left(\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\frac{d\mathbf{b}}{d\theta}\right)\\
&\quad + \left(\frac{\partial\mathbf{b}}{\partial\mathbf{y}}\right)'\frac{\partial H}{\partial\mathbf{g}}\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}^2}\frac{\partial\mathbf{b}}{\partial\theta} + \frac{\partial\mathbf{g}}{\partial\mathbf{b}}\frac{d^2\mathbf{b}}{d\theta d\mathbf{y}}.\end{aligned} \tag{A.2}$$

Expressions (A.1) and (A.2) require $d^2\mathbf{b}/d\theta^2$, $d^2\mathbf{b}/d\theta d\mathbf{y}$ and $d\mathbf{b}/d\mathbf{y}$, which are derived next.

A.3. $d^2\mathbf{b}/d\theta d\theta_k$

Differentiating $\mathbf{b}$ with respect to θ and then θ_k is equivalent to differentiating (15) with respect to θ_k. The resulting expression (A.3) can then be solved for $d^2\mathbf{b}/d\theta d\theta_k$ in (A.4), using the implicit function theorem and the fact that $\partial J/\partial\mathbf{b} = 0$ when $\mathbf{b} = \hat{\mathbf{b}}$.

$$\begin{aligned}\frac{\partial}{\partial\theta_k}\left(\frac{d^2J}{d\mathbf{b}d\theta}\right) &= \frac{\partial}{\partial\theta_k}\left(\frac{\partial^2 J}{\partial\mathbf{g}\partial\theta}\frac{\partial\mathbf{g}}{\partial\mathbf{b}} + \frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial\mathbf{g}}{\partial\theta}\frac{\partial\mathbf{g}}{\partial\mathbf{b}} + \frac{\partial J}{\partial\mathbf{g}}\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}\partial\theta}\right.\\
&\qquad\left. + \left\{\left(\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\right)'\frac{\partial^2 J}{\partial\mathbf{g}^2}\left(\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\right) + \frac{\partial J}{\partial\mathbf{g}}\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}^2}\right\}\frac{d\mathbf{b}}{d\theta}\right)\\
&= \left(\frac{\partial^3 J}{\partial\mathbf{g}\partial\theta\partial\theta_k} + \frac{\partial^3 J}{\partial\mathbf{g}\partial\theta\partial\mathbf{b}}\frac{\partial\mathbf{b}}{\partial\theta_k} + \frac{\partial^3 J}{\partial\mathbf{g}\partial\theta\partial\mathbf{g}}\frac{\partial\mathbf{g}}{\partial\theta_k}\right.\\
&\qquad\left. + \frac{\partial^3 J}{\partial\mathbf{g}\partial\theta\partial\mathbf{g}}\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\frac{\partial\mathbf{b}}{\partial\theta_k}\right)\frac{\partial\mathbf{g}}{\partial\mathbf{b}} + \frac{\partial^2 J}{\partial\mathbf{g}\partial\theta}\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}\partial\theta_k} + \frac{\partial^2 J}{\partial\mathbf{g}\partial\theta}\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}^2}\frac{\partial\mathbf{b}}{\partial\theta_k}\\
&\quad + \left(\frac{\partial^3 J}{\partial\mathbf{g}^2\partial\theta_k} + \frac{\partial^3 J}{\partial\mathbf{g}^2\partial\mathbf{b}}\frac{\partial\mathbf{b}}{\partial\theta_k} + \frac{\partial^3 J}{\partial\mathbf{g}^3}\frac{\partial\mathbf{g}}{\partial\theta_k} + \frac{\partial^3 J}{\partial\mathbf{g}^3}\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\frac{\partial\mathbf{b}}{\partial\theta_k}\right)\\
&\quad \times\frac{\partial\mathbf{g}}{\partial\theta}\frac{\partial\mathbf{g}}{\partial\mathbf{b}} + \frac{\partial^2 J}{\partial^2\mathbf{g}}\frac{\partial^2\mathbf{g}}{\partial\theta\partial\theta_k}\frac{\partial\mathbf{g}}{\partial\mathbf{b}} + \frac{\partial^2 J}{\partial^2\mathbf{g}}\frac{\partial^2\mathbf{g}}{\partial\theta\partial\mathbf{b}}\frac{\partial\mathbf{b}}{\partial\theta_k}\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\\
&\quad + \frac{\partial\mathbf{g}}{\partial\theta}\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}\partial\theta_k} + \frac{\partial\mathbf{g}}{\partial\theta}\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}^2}\frac{\partial\mathbf{b}}{\partial\theta_k}\\
&\quad + \left(\frac{\partial^2 J}{\partial\mathbf{g}\partial\theta_k} + \frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial\mathbf{g}}{\partial\theta_k} + \frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\frac{\partial\mathbf{b}}{\partial\theta_k}\right)\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}\partial\theta}\\
&\quad + \frac{\partial J}{\partial\mathbf{g}}\frac{\partial^3\mathbf{g}}{\partial\mathbf{b}\partial\theta\partial\theta_k} + \frac{\partial J}{\partial\mathbf{g}}\frac{\partial^3\mathbf{g}}{\partial\mathbf{b}\partial\theta\partial\mathbf{b}}\frac{\partial\mathbf{b}}{\partial\theta_k}\\
&\quad + \left(\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}\partial\theta_k} + \frac{\partial^2\mathbf{g}}{\partial\mathbf{b}^2}\frac{\partial\mathbf{b}}{\partial\theta_k}\right)'\frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial\mathbf{g}}{\partial\mathbf{b}} + \left(\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\right)'\\
&\quad \times\left(\frac{\partial^3 J}{\partial\mathbf{g}^2\partial\theta_k} + \frac{\partial^3 J}{\partial\mathbf{g}^2\partial\mathbf{b}}\frac{\partial\mathbf{b}}{\partial\theta_k} + \frac{\partial^3 J}{\partial\mathbf{g}^3}\frac{\partial\mathbf{g}}{\partial\theta_k} + \frac{\partial^3 J}{\partial\mathbf{g}^3}\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\frac{\partial\mathbf{b}}{\partial\theta_k}\right)\left(\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\right)\\
&\quad + \left(\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\right)'\left(\frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}\partial\theta_k} + \frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}^2}\frac{\partial\mathbf{b}}{\partial\theta_k}\right)\\
&\quad + \left(\frac{\partial^2 J}{\partial\mathbf{g}\partial\theta_k} + \frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial\mathbf{g}}{\partial\theta_k} + \frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\frac{\partial\mathbf{b}}{\partial\theta_k}\right)\left(\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}^2}\right)\\
&\quad + \left(\frac{\partial J}{\partial\mathbf{g}}\right)\left(\frac{\partial^3\mathbf{g}}{\partial\mathbf{b}^2\partial\theta_k} + \frac{\partial^3\mathbf{g}}{\partial\mathbf{b}^3}\frac{\partial\mathbf{b}}{\partial\theta_k}\right)\frac{\partial\mathbf{b}}{\partial\theta}\\
&\quad + \left\{\left(\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\right)'\frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial\mathbf{g}}{\partial\mathbf{b}} + \frac{\partial J}{\partial\mathbf{g}}\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}^2}\right\}\frac{d^2\mathbf{b}}{d\theta d\theta_k}\end{aligned} \tag{A.3}$$

Next, solve for $d^2\hat{\mathbf{b}}/d\theta d\theta_k$ using the fact that the above expression is zero at $\mathbf{b} = \hat{\mathbf{b}}$. The implicit function theorem allows us to take the inverse of the last term of (A.3).

$$\begin{aligned}\frac{d^2\hat{\mathbf{b}}}{d\theta d\theta_k} &= -\left[\left(\frac{\partial\mathbf{g}}{\partial\hat{\mathbf{b}}}\right)'\frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial\mathbf{g}}{\partial\hat{\mathbf{b}}} + \frac{\partial J}{\partial\mathbf{g}}\frac{\partial^2\mathbf{g}}{\partial\hat{\mathbf{b}}^2}\right]^{-1}\\
&\quad \times\left[\left(\frac{\partial^3 J}{\partial\mathbf{g}\partial\theta\partial\theta_k} + \frac{\partial^3 J}{\partial\mathbf{g}\partial\theta\partial\hat{\mathbf{b}}}\frac{\partial\hat{\mathbf{b}}}{\partial\theta_k} + \frac{\partial^3 J}{\partial\mathbf{g}\partial\theta\partial\mathbf{g}}\frac{\partial\mathbf{g}}{\partial\theta_k} + \frac{\partial^3 J}{\partial\mathbf{g}\partial\theta\partial\mathbf{g}}\frac{\partial\mathbf{g}}{\partial\hat{\mathbf{b}}}\frac{\partial\hat{\mathbf{b}}}{\partial\theta_k}\right)\right.\\
&\quad \times\frac{\partial\mathbf{g}}{\partial\hat{\mathbf{b}}} + \frac{\partial^2 J}{\partial\mathbf{g}\partial\theta}\frac{\partial^2\mathbf{g}}{\partial\hat{\mathbf{b}}\partial\theta_k} + \frac{\partial^2 J}{\partial\mathbf{g}\partial\theta}\frac{\partial^2\mathbf{g}}{\partial\hat{\mathbf{b}}^2}\frac{\partial\hat{\mathbf{b}}}{\partial\theta_k}\\
&\quad + \left(\frac{\partial^3 J}{\partial\mathbf{g}^2\partial\theta_k} + \frac{\partial^3 J}{\partial\mathbf{g}^2\partial\hat{\mathbf{b}}}\frac{\partial\hat{\mathbf{b}}}{\partial\theta_k} + \frac{\partial^3 J}{\partial\mathbf{g}^3}\frac{\partial\mathbf{g}}{\partial\theta_k} + \frac{\partial^3 J}{\partial\mathbf{g}^3}\frac{\partial\mathbf{g}}{\partial\hat{\mathbf{b}}}\frac{\partial\hat{\mathbf{b}}}{\partial\theta_k}\right)\frac{\partial\mathbf{g}}{\partial\theta}\frac{\partial\mathbf{g}}{\partial\hat{\mathbf{b}}}\\
&\quad + \frac{\partial^2 J}{\partial^2\mathbf{g}}\frac{\partial^2\mathbf{g}}{\partial\theta\partial\theta_k}\frac{\partial\mathbf{g}}{\partial\hat{\mathbf{b}}} + \frac{\partial^2 J}{\partial^2\mathbf{g}}\frac{\partial^2\mathbf{g}}{\partial\theta\partial\hat{\mathbf{b}}}\frac{\partial\hat{\mathbf{b}}}{\partial\theta_k}\frac{\partial\mathbf{g}}{\partial\hat{\mathbf{b}}} + \frac{\partial\mathbf{g}}{\partial\theta}\frac{\partial^2\mathbf{g}}{\partial\hat{\mathbf{b}}\partial\theta_k}\\
&\quad + \frac{\partial\mathbf{g}}{\partial\theta}\frac{\partial^2\mathbf{g}}{\partial\hat{\mathbf{b}}^2}\frac{\partial\hat{\mathbf{b}}}{\partial\theta_k} + \left(\frac{\partial^2 J}{\partial\mathbf{g}\partial\theta_k} + \frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial\mathbf{g}}{\partial\theta_k} + \frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial\mathbf{g}}{\partial\hat{\mathbf{b}}}\frac{\partial\hat{\mathbf{b}}}{\partial\theta_k}\right)\frac{\partial^2\mathbf{g}}{\partial\hat{\mathbf{b}}\partial\theta}\\
&\quad + \frac{\partial J}{\partial\mathbf{g}}\frac{\partial^3\mathbf{g}}{\partial\hat{\mathbf{b}}\partial\theta\partial\theta_k} + \frac{\partial J}{\partial\mathbf{g}}\frac{\partial^3\mathbf{g}}{\partial\hat{\mathbf{b}}\partial\theta\partial\hat{\mathbf{b}}}\frac{\partial\hat{\mathbf{b}}}{\partial\theta_k} + \left(\frac{\partial^2\mathbf{g}}{\partial\hat{\mathbf{b}}\partial\theta_k} + \frac{\partial^2\mathbf{g}}{\partial\hat{\mathbf{b}}^2}\frac{\partial\hat{\mathbf{b}}}{\partial\theta_k}\right)'\frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial\mathbf{g}}{\partial\hat{\mathbf{b}}}\\
&\quad + \left(\frac{\partial\mathbf{g}}{\partial\hat{\mathbf{b}}}\right)'\left(\frac{\partial^3 J}{\partial\mathbf{g}^2\partial\theta_k} + \frac{\partial^3 J}{\partial\mathbf{g}^2\partial\hat{\mathbf{b}}}\frac{\partial\hat{\mathbf{b}}}{\partial\theta_k} + \frac{\partial^3 J}{\partial\mathbf{g}^3}\frac{\partial\mathbf{g}}{\partial\theta_k} + \frac{\partial^3 J}{\partial\mathbf{g}^3}\frac{\partial\mathbf{g}}{\partial\hat{\mathbf{b}}}\frac{\partial\hat{\mathbf{b}}}{\partial\theta_k}\right)\left(\frac{\partial\mathbf{g}}{\partial\hat{\mathbf{b}}}\right)\\
&\quad + \left(\frac{\partial\mathbf{g}}{\partial\hat{\mathbf{b}}}\right)'\left(\frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial^2\mathbf{g}}{\partial\hat{\mathbf{b}}\partial\theta_k} + \frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial^2\mathbf{g}}{\partial\hat{\mathbf{b}}^2}\frac{\partial\hat{\mathbf{b}}}{\partial\theta_k}\right)\\
&\quad + \left(\frac{\partial^2 J}{\partial\mathbf{g}\partial\theta_k} + \frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial\mathbf{g}}{\partial\theta_k} + \frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial\mathbf{g}}{\partial\hat{\mathbf{b}}}\frac{\partial\hat{\mathbf{b}}}{\partial\theta_k}\right)\left(\frac{\partial^2\mathbf{g}}{\partial\hat{\mathbf{b}}^2}\right)\\
&\quad \left. + \left(\frac{\partial J}{\partial\mathbf{g}}\right)\left(\frac{\partial^3\mathbf{g}}{\partial\hat{\mathbf{b}}^2\partial\theta_k} + \frac{\partial^3\mathbf{g}}{\partial\hat{\mathbf{b}}^3}\frac{\partial\hat{\mathbf{b}}}{\partial\theta_k}\right)\frac{\partial\hat{\mathbf{b}}}{\partial\theta}\right].\end{aligned} \tag{A.4}$$

A.4. $\partial\hat{\mathbf{b}}/\partial\mathbf{y}$

Expression (A.2) requires an expression for $d\hat{\mathbf{b}}/d\mathbf{y}$. Differentiating $dJ/d\mathbf{b}$ with respect to $\mathbf{y}$ produces (A.5). We again use the fact that $\partial J/\partial\mathbf{b} = 0$ at $\mathbf{b} = \hat{\mathbf{b}}$ and solve for $\partial\hat{\mathbf{b}}/\partial\mathbf{y}$ in (A.6), using the implicit function theorem.

$$\frac{d}{d\mathbf{y}}\left(\frac{dJ}{d\mathbf{b}}\right) = \frac{\partial^2 J}{\partial\mathbf{g}\partial\mathbf{y}}\frac{\partial\mathbf{g}}{\partial\mathbf{b}} + \left(\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\right)'\frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\frac{\partial\mathbf{b}}{\partial\mathbf{y}} + \frac{\partial J}{\partial\mathbf{g}}\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}^2}\frac{\partial\mathbf{b}}{\partial\mathbf{y}}, \tag{A.5}$$

$$\frac{\partial\hat{\mathbf{b}}}{\partial\mathbf{y}} = \left\{\left(\frac{\partial\mathbf{g}}{\partial\hat{\mathbf{b}}}\right)'\frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial\mathbf{g}}{\partial\hat{\mathbf{b}}} + \frac{\partial J}{\partial\mathbf{g}}\frac{\partial^2\mathbf{g}}{\partial\hat{\mathbf{b}}^2}\right\}^{-1}\left\{\frac{\partial^2 J}{\partial\mathbf{g}\partial\mathbf{y}}\frac{\partial\mathbf{g}}{\partial\hat{\mathbf{b}}}\right\}. \tag{A.6}$$

A.5. $\partial^2\hat{\mathbf{b}}/\partial\mathbf{y}\partial\theta$

We obtain this derivative by differentiating (A.5) with respect to θ_k to produce Eq. (A.7). Solving for $\partial^2\hat{\mathbf{b}}/\partial\mathbf{y}\partial\theta$ gives us the expression (A.8).

$$\begin{aligned}
\frac{\partial}{\partial\theta_k}\left(\frac{\partial^2 J}{\partial\mathbf{b}\partial\mathbf{y}}\right) &= \left(\frac{\partial^3 J}{\partial\mathbf{g}\partial\mathbf{y}\partial\theta_k}+\frac{\partial^3 J}{\partial\mathbf{g}\partial\mathbf{y}\partial\mathbf{b}}\frac{\partial\mathbf{b}}{\partial\theta_k}+\frac{\partial^3 J}{\partial\mathbf{g}\partial\mathbf{y}\partial\mathbf{g}}\frac{\partial\mathbf{g}}{\partial\theta_k}\right.\\
&\quad\left.+\frac{\partial^3 J}{\partial\mathbf{g}\partial\mathbf{y}\partial\mathbf{g}}\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\frac{\partial\mathbf{b}}{\partial\theta_k}\right)\frac{\partial\mathbf{g}}{\partial\mathbf{b}}+\frac{\partial^2 J}{\partial\mathbf{g}\partial\mathbf{y}}\left(\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}\partial\theta_k}+\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}^2}\frac{\partial\mathbf{b}}{\partial\theta_k}\right)\\
&\quad+\left\{\left(\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}\partial\theta_k}+\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}^2}\frac{\partial\mathbf{b}}{\partial\theta_k}\right)'\frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial\mathbf{g}}{\partial\mathbf{b}}+\left(\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\right)'\right.\\
&\quad\times\left(\left[\frac{\partial^3 J}{\partial\mathbf{g}^2\partial\theta_k}+\frac{\partial^3 J}{\partial\mathbf{g}^2\partial\mathbf{b}}\frac{\partial\mathbf{b}}{\partial\theta_k}+\frac{\partial^3 J}{\partial\mathbf{g}^3}\frac{\partial\mathbf{g}}{\partial\theta_k}+\frac{\partial^3 J}{\partial\mathbf{g}^3}\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\frac{\partial\mathbf{b}}{\partial\theta_k}\right]\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\right.\\
&\quad\left.+\frac{\partial^2 J}{\partial\mathbf{g}^2}\left[\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}\partial\theta_k}+\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}^2}\frac{\partial\mathbf{b}}{\partial\theta_k}\right]\right)\\
&\quad+\left(\frac{\partial^2 J}{\partial\mathbf{g}\partial\theta_k}+\frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial\mathbf{g}}{\partial\theta_k}+\frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\frac{\partial\mathbf{b}}{\partial\theta_k}\right)\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}^2}\\
&\quad\left.+\frac{\partial J}{\partial\mathbf{g}}\left(\frac{\partial^3\mathbf{g}}{\partial\mathbf{b}^2\partial\theta_k}+\frac{\partial^3\mathbf{g}}{\partial\mathbf{b}^3}\frac{\partial\mathbf{b}}{\partial\theta_k}\right)\right\}\frac{d\mathbf{b}}{d\mathbf{y}}\\
&\quad+\left\{\left(\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\right)'\frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial\mathbf{g}}{\partial\mathbf{b}}+\frac{\partial J}{\partial\mathbf{g}}\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}^2}\right\}\frac{\partial^2\mathbf{b}}{\partial\mathbf{y}\partial\theta_k},
\end{aligned}\tag{A.7}$$

$$\begin{aligned}
\frac{\partial^2\hat{\mathbf{b}}}{\partial\mathbf{y}\partial\boldsymbol{\theta}} &= -\left\{\left(\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\right)'\frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial\mathbf{g}}{\partial\mathbf{b}}+\frac{\partial J}{\partial\mathbf{g}}\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}^2}\right\}^{-1}\\
&\quad\times\left\{\left(\frac{\partial^3 J}{\partial\mathbf{g}\partial\mathbf{y}\partial\theta_k}+\frac{\partial^3 J}{\partial\mathbf{g}\partial\mathbf{y}\partial\mathbf{b}}\frac{\partial\mathbf{b}}{\partial\theta_k}+\frac{\partial^3 J}{\partial\mathbf{g}\partial\mathbf{y}\partial\mathbf{g}}\frac{\partial\mathbf{g}}{\partial\theta_k}+\frac{\partial^3 J}{\partial\mathbf{g}\partial\mathbf{y}\partial\mathbf{g}}\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\frac{\partial\mathbf{b}}{\partial\theta_k}\right)\right.\\
&\quad\times\frac{\partial\mathbf{g}}{\partial\mathbf{b}}+\frac{\partial^2 J}{\partial\mathbf{g}\partial\mathbf{y}}\left(\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}\partial\theta_k}+\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}^2}\frac{\partial\mathbf{b}}{\partial\theta_k}\right)\\
&\quad+\left\{\left(\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}\partial\theta_k}+\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}^2}\frac{\partial\mathbf{b}}{\partial\theta_k}\right)'\frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\right.\\
&\quad+\left(\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\right)'\left(\left[\frac{\partial^3 J}{\partial\mathbf{g}^2\partial\theta_k}+\frac{\partial^3 J}{\partial\mathbf{g}^2\partial\mathbf{b}}\frac{\partial\mathbf{b}}{\partial\theta_k}+\frac{\partial^3 J}{\partial\mathbf{g}^3}\frac{\partial\mathbf{g}}{\partial\theta_k}+\frac{\partial^3 J}{\partial\mathbf{g}^3}\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\frac{\partial\mathbf{b}}{\partial\theta_k}\right]\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\right.\\
&\quad\left.+\frac{\partial^2 J}{\partial\mathbf{g}^2}\left[\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}\partial\theta_k}+\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}^2}\frac{\partial\mathbf{b}}{\partial\theta_k}\right]\right)\\
&\quad+\left(\frac{\partial^2 J}{\partial\mathbf{g}\partial\theta_k}+\frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial\mathbf{g}}{\partial\theta_k}+\frac{\partial^2 J}{\partial\mathbf{g}^2}\frac{\partial\mathbf{g}}{\partial\mathbf{b}}\frac{\partial\mathbf{b}}{\partial\theta_k}\right)\frac{\partial^2\mathbf{g}}{\partial\mathbf{b}^2}\\
&\quad\left.\left.+\frac{\partial J}{\partial\mathbf{g}}\left(\frac{\partial^3\mathbf{g}}{\partial\mathbf{b}^2\partial\theta_k}+\frac{\partial^3\mathbf{g}}{\partial\mathbf{b}^3}\frac{\partial\mathbf{b}}{\partial\theta_k}\right)\right\}\frac{d\mathbf{b}}{d\mathbf{y}}\right\}.
\end{aligned}\tag{A.8}$$

References

[1] D. Bates, D. Watts, Nonlinear Regression Analysis and Its Applications, Wiley, New York, 1988.

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

[3] E.O. Voit, J. Almeida, Decoupling dynamical systems for pathway identification from metabolic profiles, Bioinformatics 20 (2004) 1670.

[4] W.R. Esposito, C. Floudas, Deterministic global optimization in nonlinear optimal control problems, Journal of Global Optimization 17 (2000) 97.

[5] J. Varah, A spline least squares method for numerical parameter estimation in differential equations, SIAM Journal on Scientific and Statistical Computing 3 (1982) 28.

[6] N. Brunel, Parameter estimation of odes via nonparametric estimators, Electronic Journal of Statistics 2 (2008) 1242.

[7] H. Liang, H. Wu, Parameter estimation for differential equation models using a framework of measurement error in regression models, Journal of the American Statistical Association 103 (2008) 1570.

[8] B. Calderhead, M. Girolami, N. Lawrence, Accelerating bayesian inference over nonlinear differential equations with gaussian processes, Advances in Neural Information Processing Systems 21 (2009). Q3

[9] A. Poyton, M. Varziri, K. McAuley, P. McLellan, J.O. Ramsay, Parameter estimation in continuous-time dynamic models using principal differential analysis, Computers and Chemical Engineering 30 (2006) 698.

[10] J.O. Ramsay, B. Silverman, Functional Data Analysis, second edition., Springer, New York, 2005.

[11] J.O. Ramsay, G. Hooker, D. Campbell, J. Cao, Parameter estimation for differential equations: A generalized smoothing approach (with discussion), Journal of the Royal Statistical Society, Series B 69 (2007) 1.

[12] D. Campbell, R.J. Steele, Smooth functional tempering for nonlinear differential equation models, Statistics and Computing (2011). Q4

[13] L. Wang, J. Cao, Estimating parameters in delay differential equation models, Journal of Agricultural, Biological, and Environmental Statistics (2012) 1.

[14] J. Lewis, Autoinhibition with transcriptional delay: A simple mechanism for the zebrafish somitogenesis oscillator, Current Biology 13 (2003) 1398.

[15] S. Bernard, B. Ajavec, L. Pujo-Menjouet, M.C. Mackey, H. Herzel, Modelling transcriptional feedback loops: the role of Gro/TLE1 in Hes1 oscillations, Philosophical Transactions: Mathematical, Physical and Engineering Sciences 364 (2006) 1155–1170.

[16] I. Swameye, T. Muller, J. Timmer, O. Sandra, U. Klingmuller, Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling, Proceedings of the National Academy of Sciences 100 (2003).

[17] S. Busenberg, B. Tang, Mathematical models of the early embryonic cell cycle: the role of mpf activation and cyclin degradation, Journal of Mathematical Biology 32 (1994) 573.

[18] G.A. Bocharov, F.A. Rihan, Numerical modelling in biosciences using delay differential equations, Journal of Computational and Applied Mathematics 125 (2000) 183.

[19] S. Pellegrini, I. Dusanter-Fourt, The structure, regulation and function of the janus kinases (jaks) and the signal transducers and activators of transcription (stats), European Journal of Biochemistry 248 (1997) 615.

[20] A. Raue, C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmuller, J. Timmer, Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood, Bioinformatics 25 (2009).

[21] M. Schelker, A. Raue, J. Timmer, C. Kreutz, Comprehensive estimation of input signals and dynamics in biochemical reaction networks, Bioinformatics 28 (2012) i529.

[22] D.A. Campbell, G. Hooker, K.B. McAuley, Parameter estimation in differential equation models with constrained states, Journal of Chemometrics (2012).

[23] W. Horbelt, J. Timmer, H.U. Voss, Parameter estimation in nonlinear delayed feedback systems from noisy data, Physics Letters A 299 (2002) 513.

[24] C.G. Moles, J.R. Banga, K. Keller, Solving nonconvex climate control problems: pitfalls and algorithm performances, Applied Soft Computing 5 (2004) 35.

[25] X. Qi, H. Zhao, Asymptotic efficiency and finite-sample properties of the generalized profiling estimation of parameters in ordinary differential equations, Annals of Statistics 38 (2010) 435.

[26] W.K. Newey, K.D. West, A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix, Econometrica: Journal of the Econometric Society (1987) 703–708. Q5

[27] G. Hooker, S.P. Ellner, L.D.V.e.V. Roditi, D.J.D. Earn, Parameterizing state-space models for infectious disease dynamics by generalized profiling: measles in Ontario, Journal of the Royal Society Interface 8 (2011) 961–974.

[28] H. Miao, X. Xia, A.S. Perelson, H. Wu, On identifiability of ode models and applications in viral dynamics, SIAM Review 53 (2011) 3.

[29] J. Cao, L. Wang, J. Xu, Robust estimation for ordinary differential equation models, Biometrics 67 (2011) 1305.

[30] M. Varziri, K. McAuley, P. McLellan, Approximate maximum likelihood parameter estimation for nonlinear dynamic models; application to a laboratory-scale nylon reactor model, Industrial and Engineering Chemistry Research 47 (2008) 7274.

[31] P. Chaudhuri, J.S. Marron, Scale space view of curve estimation, The Annals of Statistics 28 (2000) 408.