Available online at www.sciencedirect.com
SCIENCE DIRECT
Computers and Chemical Engineering 30 (2006) 698–708
Computers & Chemical Engineering
www.elsevier.com/locate/compchemeng

Parameter estimation in continuous-time dynamic models using principal differential analysis

A.A. Poyton [a], M.S. Varziri [a], K.B. McAuley [a,*], P.J. McLellan [a], J.O. Ramsay [b]

[a] *Department of Chemical Engineering, Queen's University, Kingston, Ont., Canada K7L 3N6*
[b] *Department of Psychology, McGill University, Montreal, PQ, Canada H3A 1B1*

Received 29 March 2005; received in revised form 24 November 2005; accepted 29 November 2005
Available online 19 January 2006

Abstract

Principal differential analysis (PDA) is an alternative parameter estimation technique for differential equation models in which basis functions (e.g., B-splines) are fitted to dynamic data. Derivatives of the resulting empirical expressions are used to avoid solving differential equations when estimating parameters. Benefits and shortcomings of PDA were examined using a simple continuous stirred-tank reactor (CSTR) model. Although PDA required considerably less computational effort than traditional nonlinear regression, parameter estimates from PDA were less precise. Sparse and noisy data resulted in poor spline fits and misleading derivative information, leading to poor parameter estimates. These problems are addressed by a new iterative algorithm (iPDA) in which the spline fits are improved using model-based penalties. Parameter estimates from iPDA were unbiased and more precise than those from standard PDA. Issues that need to be resolved before iPDA can be used for more complex models are discussed.

Keywords: Principal differential analysis; Parameter estimation; Dynamic models

1. Introduction

Parameter estimation in dynamic models is important in many fields of science and engineering because many physical, chemical and biological processes are described by systems of ordinary differential equations (ODEs) with unknown parameters. For chemical engineers, the benefits of developing dynamic mechanistic models with accurate parameter estimates have increased in recent years due to the development of process optimization and control technologies that can use fundamental models (Biegler & Grossman, 2004; El-Farra & Christofides, 2003; Nagy & Braatz, 2003).

Parameter estimation is a difficult and important step in the development of models that are consistent with fundamental behavior of the physical process and consistent with available data. Estimates of unknown parameters are obtained using measurements from dynamic experiments, which invariably contain random errors. Numerous system identification techniques are available for estimating parameters in discrete-time dynamic models (Ljung, 1999). The current article is focused on parameter estimation in continuous-time ordinary differential equation (ODE) or differential-algebraic equation (DAE) models that are nonlinear in the parameters.

When estimating model parameters, the objective is to determine appropriate parameter values so that errors between the outputs of the estimated model and the measured data are minimized in some sense, i.e., the predicted response values from the model should match the measurements as closely as possible. Most commonly used parameter estimation techniques are based on least-squares regression, which involves minimization of the sum of squared differences between the measurements and the model predictions. Generalized least-squares regression and multi-response estimation are described in detail by Seber and Wild (1989) and Bates and Watts (1988).

Ogunnaike and Ray (1994) describe the iterative nonlinear least-squares (NLS) procedure that is commonly used to estimate parameters in ODE (or DAE) models that do not have analytical solutions. First the ODEs (or DAEs) are solved numerically, using initial guesses for the parameter values, yielding simulations of the dynamic experiments. Model predictions are

* Corresponding author. Tel.: +1 613 533 2768; fax: +1 613 533 6637.
E-mail address: mcauleyk@chee.queensu.ca (K.B. McAuley).

0098-1354/$ – see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.compchemeng.2005.11.008

compared with measured responses, and an optimization algorithm determines a new set of parameter estimates that should result in better model predictions. Sensitivity equations can be solved along with the model ODEs to obtain the Jacobian of the response variables with respect to the parameters (Leis & Kramer, 1988), so that the optimizer can determine appropriate search directions for improved parameter values. The alternatives are to use either a numerical Jacobian or a direct-search optimizer, requiring additional dynamic simulations with perturbations in each of the parameters. After new parameter values are determined by the optimizer, the ODEs (or DAEs) are solved numerically using the updated parameters and the new predictions are compared with the data. Iteration between parameter updating and computation of numerical solutions continues until convergence criteria for the parameters are met or until no significant improvement in the objective function is obtained. Unfortunately, this method can be time consuming, with most of the computational effort arising from repeated numerical solution of the differential equations. Initial values for all of the states are either assumed to be known, or must be estimated along with the model parameters. Numerous algorithms have been developed within this framework to obtain improved parameter estimates with reduced computational effort (e.g., Bates & Watts, 1985; Biegler, Damiano, & Blau, 1986; Dovì, Arato, & Maga, 1985; Kalogerakis & Luus, 1983; Mansouri & Kernévev, 1998; Stewart, Caracotsios, & Sorensen, 1992).

Varah (1982) used an alternative parameter estimation technique, based on the earlier work of Swartz and Bremermann (1975) and Benson (1979) that does not require repeated numerical solutions of the ODEs. In this methodology, and in a related technique called principal differential analysis (PDA) (Ramsay, 1996), discrete measurements of the output variables, y, are fitted empirically using splines, which are then differentiated with respect to time to obtain estimated time-derivative curves, $\mathrm{d}y/\mathrm{d}t$. This time-derivative information is then substituted into the ODEs, converting the parameter estimation problem from a dynamic optimization problem into a much simpler algebraic optimization problem that can be solved using either linear least-squares (if the ODEs are linear in the parameters) or nonlinear least-squares (if the ODEs are nonlinear in the parameters). PDA techniques (Ramsay, 1996; Varah, 1982) differ from commonly used nonlinear least-squares methods for dynamic models, and from an early spline-based method (Tang, 1971) and its iterative extension by Madar, Abonyi, Roubos, and Szeifert (2003). In PDA, parameter values are selected to minimize squared residuals in the differential form of the model, $(\mathrm{d}y/\mathrm{d}t - \mathrm{d}\hat{y}/\mathrm{d}t)^2$, rather than the traditional integrated form of the model, $(y - \hat{y})^2$. Early on, Swartz and Bremermann (1975) and Varah (1982) identified the main benefits of PDA techniques (they are less computationally expensive than common parameter estimation techniques for dynamic models, and initial conditions for the output variables need not be known) as well as the main problems that limit their use (poor spline fits can result in misleading time-derivative information, which can lead to poor parameter estimates). Computing resources have advanced greatly over the past 30 years, making traditional least-squares parameter estimation techniques easier and less time-consuming for users. It is unlikely that new algorithms for parameter estimation in differential equation models will be adopted on the basis of reduced computation time alone. Additional benefits and applicability to a large class of difficult parameter estimation problems will need to be proven before PDA techniques are adopted by more than a few curious users. The first step toward uncovering potential benefits is to gain an improved understanding of PDA techniques, so that existing problems can be alleviated and so that opportunities can be recognized.

In this article, we examine the influence of various specifications (B-spline knot placement and higher-order derivative penalties) that affect the quality of parameter estimates obtained from PDA, and we propose a new iteratively refined PDA algorithm that ensures good spline fits and parameter estimates. We test this methodology using linearized and nonlinear versions of a simple dynamic continuous stirred-tank reactor (CSTR) model, and we identify issues that need to be resolved before PDA techniques can enjoy widespread use for parameter estimation in more complex models of chemical processes.

2. Principal differential analysis

PDA is a term used by Ramsay (1996) to describe a parameter estimation method wherein coefficients in linear, possibly time-varying, ODEs are fitted empirically from data. Ramsay called his technique principal differential analysis because of analogies to principal component analysis (PCA), in which empirical linear algebraic-equation models are fitted using multivariate data. Ramsay and Silverman (1997, 2005) focused their efforts on problems in which dynamic systems respond to unknown, empirical, time-varying forcing functions. PDA has been used to fit linear differential equation models for a diverse array of applications including handwriting analysis (Ramsay, 2000), analysis of the movement of the lips during speech (Lucero, 2002; Ramsay & Munhall, 1996), economic modeling (Ramsay & Ramsey, 2002), and meteorological modeling (Ramsay & Silverman, 2002, 2005). In this article, we focus on parameter estimation in fundamental ODE models of chemical processes with known forcing inputs (sometimes called exogenous inputs in the system identification literature).

PDA is part of a broader, relatively new area in statistics called functional data analysis (FDA). The main concept in FDA is to account for the underlying smooth functional behavior of a process response, instead of viewing the output as a collection of discrete points. Ramsay and Silverman (2005) argue that, by using this functional approach, the natural smoothness of the processes from which the data are taken can be exploited, which may allow us to see things that a discrete-data approach would not. In this article we explore the potential benefits and problems associated with using PDA for parameter estimation in a simple continuous stirred-tank reactor (CSTR) model, and we compare the parameter estimation results obtained using PDA to those obtained using traditional nonlinear least-squares estimation. In particular, we investigate the effects of knot placement and higher-order derivative penalties during spline fitting on the quality of the spline fit and the resulting parameter estimates. We also propose a new iteratively refined PDA technique (iPDA).

This technique involves iteration between (1) fitting of splines, which are penalized using the fundamental ODE model (with current estimates of the model parameters) and (2) estimation of the parameters in the fundamental model, using the estimated spline coefficients from step (1). Using the squared residuals from the differential equation as a penalty during spline fitting ensures that the fitted splines are consistent with the behavior of the physical system, and leads to improved parameter estimates.

2.1. PDA terminology and notation

When describing PDA for linear ODE models, Ramsay and Silverman (2005) view the system dynamics as a *linear differential operator* (LDO) acting upon the process variables. For example, consider a simple first-order, single-input single-output (SISO) ODE model:

$$\frac{\mathrm{d}y}{\mathrm{d}t} + w_y y + w_u u = 0 \tag{2.1}$$

If the process gain and time constant are K_p and τ, respectively, then the weighting coefficients w_y and w_u in the output (y) and input (u) variable terms are $w_y = 1/\tau$ and $w_u = -K_\mathrm{p}/\tau$, respectively. In the PDA literature, this system would be described using a linear differential operator L operating on y and u:

$$L[y,u] = D_y^1 y + w_y D_y^0 y + w_u D_u^0 u = 0 \tag{2.2}$$

where the superscripts (1 or 0) on the differentials (D) refer to the first and zeroth derivatives, respectively, with respect to time, and the subscripts (y or u) denote whether the input or output is being differentiated. Eq. (2.2) can be extended to describe higher-order differential equation models (Ramsay & Ramsey, 2002) or multi-input multi-output (MIMO) systems with interacting first-order ODEs (Poyton, 2005). However, for the remainder of this paper we will focus only on systems modeled by a single first-order ODE. We will also forego the use of the LDO notation of the PDA literature, and will use the more familiar notation of Eq. (2.1) instead.

If the model parameters w_y and w_u for the true system were known, and there were no measurement noise or process disturbances, then the left-hand side of Eq. (2.1) would equal zero exactly. The idea of PDA is to approximate the derivative curve using the measured data, and then estimate the parameters so that $(\mathrm{d}y/\mathrm{d}t) + w_y y + w_u u$ is as close as possible to zero. By doing this, we are minimizing some function of the residual curve $e(t)$, where $e(t)$ is defined by

$$\frac{\mathrm{d}y}{\mathrm{d}t} + w_y y + w_u u = e(t) \tag{2.3}$$

In standard PDA, a least-squares approach is used to minimize the function:

$$\mathrm{SSE}_\mathrm{PDA} = \int \left(\frac{\mathrm{d}y}{\mathrm{d}t} + w_y y + w_u u\right)^2 \mathrm{d}t = \int e(t)^2 \,\mathrm{d}t \tag{2.4}$$

where the integration is over the time horizon of the data. For parameter estimation using multiple dynamic experiments or multiple outputs, several integral terms are added together. As in all FDA techniques (Ramsay & Silverman, 2005), the discrete data points for an experimental run $\{y(t_i), u(t_i)\}$ are used to determine approximate continuous-time traces for the underlying process outputs. These empirical curves are expressed in terms of a fixed set of P known basis functions $\psi_P(t)$, $p = 1, \ldots, P$, which are continuous functions of time. There are a number of choices available for the type of basis function to be used: Fourier series, polynomials, wavelets, and splines. B-splines are the most common choice.

2.2. General properties of spline functions

Splines are piecewise polynomial functions that, because of their simplicity and flexibility, are used for a variety of applications such as interpolating between data points and smoothing of noisy data. Spline smoothing involves dividing up the domain of the data into segments separated by knots, with multiple data points between the knots. The function describing the data is approximated as a weighted sum of basis functions, and a least-squares objective function is used to fit this smooth function to the raw data. When using piecewise cubic splines, there are four parameters for each cubic equation ($y_\sim = ax^3 + bx^2 + cx + d$), and three continuity constraints (zeroth, first, and second derivatives match for adjacent splines) at each knot. Throughout this article we use the subscript $\sim$ to indicate the prediction from an empirical spline model, rather than the prediction from a fundamental differential equation model. A restricted least-squares approach that accounts for the continuity constraints can be used to determine the cubic equation coefficients, but this approach can be cumbersome (Seber & Lee, 2003). Sometimes a truncated-power-series method is used instead. Though this approach is simpler, it also has computational problems. Spline predictions in intervals corresponding to large values of the independent variable depend on sums of many polynomial terms, which can lead to an ill-conditioned coefficient-estimation problem (Seber & Lee, 2003).

2.3. B-splines

A computationally better approach is to use B-spline basis functions, consisting of Rth order (or $(R-1)$th degree) piecewise polynomials. Each B-spline is positive within a domain of R intervals (defined by $R+1$ consecutive knots), and zero elsewhere (de Boor, 1978; Seber & Lee, 2003). Since B-spline basis functions are zero everywhere except over a finite interval (referred to as compact support), the weighting-coefficient-estimation problem is well conditioned. B-splines have been used in many engineering applications (e.g., Bahadir, 2003; Kim, 1998; Lainiotis & Deshpande, 1974; Shariff & Moser, 1998; Wang, Keast, & Muir, 2004) including numerical solution of partial differential equations (PDEs) and approximation of probability density functions. B-splines have also been used to aid parameter estimation in fundamental models. For example, Thomaseth, Kautzky-Wilier, Ludvik, Prager, and Pacini (1996) used B-splines to provide an empirical model for a time-varying parameter (or unknown forcing input) in a first-order ODE kinetic model, and then used nonlinear least-squares

to simultaneously estimate the fixed parameters in the fundamental model and the empirical spline coefficients. The more recent PDA applications referenced in this article (Lucero, 2002; Ramsay, 2000; Ramsay & Ramsey, 2002; Varah, 1982) have used B-splines to construct smooth curves through the dynamic response data, whereas the very early work (Benson, 1979; Swartz & Bremermann, 1975) used standard cubic splines.

To illustrate how B-splines can be used to approximate a function, consider the set of data in Fig. 1a. These data are the simulated response of reactant concentration to a step change in reactor temperature set point, for a linearized CSTR model (described in detail in Section 3). Fig. 1 demonstrates how B-spline basis functions are combined to approximate the step-response data as a continuous empirical time trace. The weighting coefficients used for combining the B-spline basis functions in Fig. 1b are shown in Fig. 1c. It is evident, especially during the first 6 min of the simulation, that seven B-spline basis functions (Fig. 1a) are not sufficient to capture this response very well. As we will show in Section 3.1, improved spline fits are obtained when more basis functions are used, and when spline knots are carefully placed.

Unfortunately, poor spline fits can result in misleading information about $\mathrm{d}y/\mathrm{d}t$, causing problems for spline-based parameter estimation methods. Better spline fits can be obtained by careful choice of the location of the knots, or by penalties on high-order derivatives during the spline-fitting procedure (de Boor, 1978; Elfving & Andersson, 1998; Schwetlick & Schütze, 1995, 1997; Varah, 1982).

When using B-splines, interior knots are placed at times τ_1, $\tau_2, \ldots, \tau_\kappa$ (where $\tau_1 < \tau_2 < \cdots < \tau_\kappa$) between the endpoints of the data domain, with exterior knots τ_0 and $\tau_{\kappa+1}$ placed at the ends. The B-splines are generated by starting the first spline $R-1$ artificial intervals to the left of the lower boundary point τ_0; the final spline extends $R-1$ artificial intervals to the right of the upper boundary point, $\tau_{\kappa+1}$. Fig. 1b shows a plot of seven B-spline basis functions of fourth order (cubic B-splines), defined by knots placed at intervals of 10 min, starting at $t = 0$ min. Each of these B-splines is positive over a domain of $R = 4$ intervals, or a time of 40 min, and zero elsewhere. B-spline basis function ψ_1 extends three artificial intervals to the left, with only the fourth interval falling into the domain of the data. ψ_2 extends two artificial intervals to the left and ψ_4 fits entirely within the domain of the data.

For fourth-order B-splines, like those in Fig. 1b, each B-spline is composed of four piecewise cubic polynomials. For example, in Fig. 1b, B-spline ψ_4 is composed of the following cubics:

$$\psi_4(t) = \begin{cases} \dfrac{t^3}{6000} & 0 \le t < 10 \\ -\dfrac{1}{3} + \dfrac{t}{20} + \dfrac{(t-10)^2}{200} - \dfrac{(t-10)^3}{2000} & 10 \le t < 20 \\ \dfrac{2}{3} - \dfrac{(t-20)^2}{100}\dfrac{(t-20)^3}{2000} & 20 \le t < 30 \\ \dfrac{5}{3} - \dfrac{t}{20} + \dfrac{(t-30)^2}{200}\dfrac{(t-30)^3}{6000} & 30 \le t \le 40 \end{cases} \quad (2.5)$$

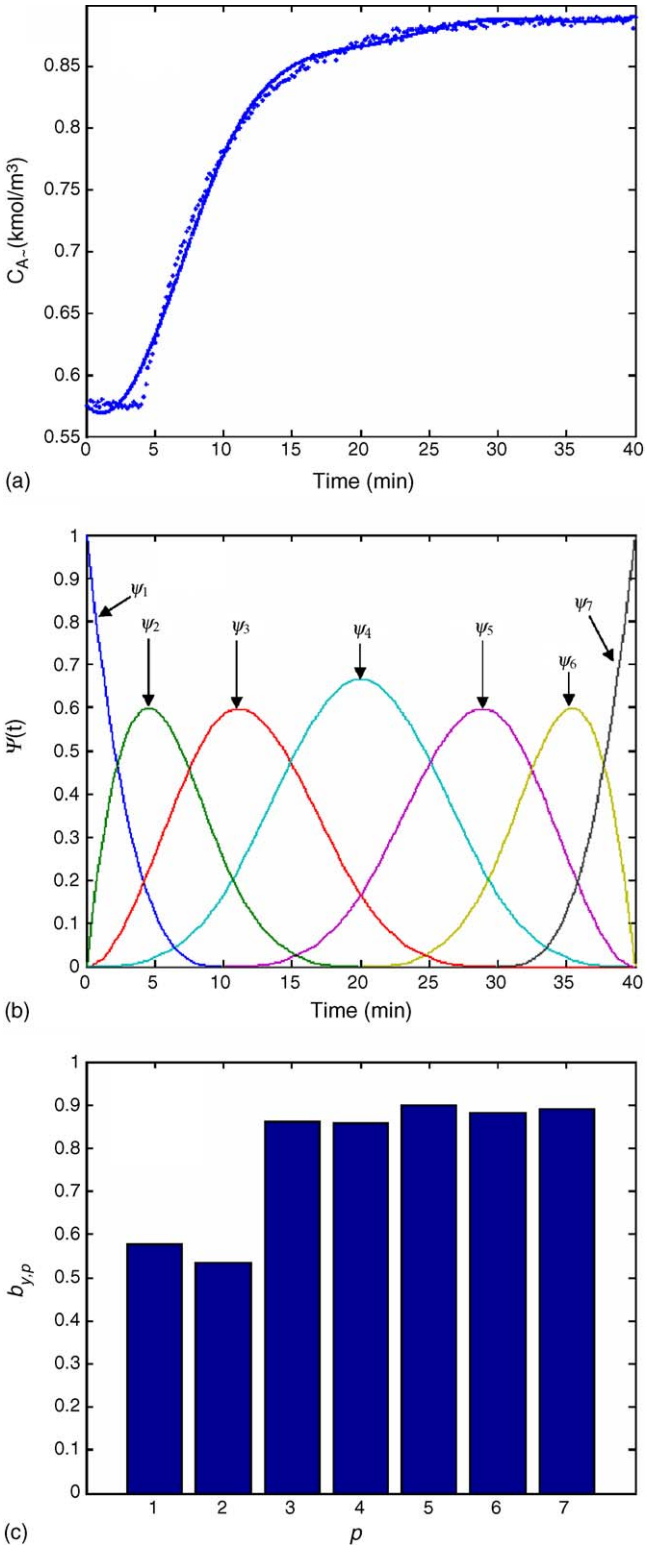

Fig. 1. (a) Spline fit to step-response data using seven B-spline basis functions; (b) seven B-spline basis functions used to construct spline fit; (c) seven estimated weights for B-spline basis functions.

2.4. Determining spline coefficients from discrete data

Observations of functions usually consist of data taken at discrete points. The first step of PDA is to convert these

noisy discrete points into smooth functions of time (Ramsay & Silverman, 2005) by finding appropriate values of the spline weighting coefficients $b_{y,p}$:

$$y_{\sim}(t) = \sum_{p=1}^{P_y} b_{y,p}\psi_{y,p}(t) \tag{2.6}$$

If the input information is also a set of discrete points (rather than a known function) then the input curve can also be fitted using splines:

$$u_{\sim}(t) = \sum_{p=1}^{P_u} b_{u,p}\psi_{u,p}(t) \tag{2.7}$$

In PDA applications, P_y and P_u, the number of B-spline basis functions used to approximate $y(t)$ and $u(t)$, need not be the same. The empirical function $y_{\sim}(t)$ is determined by selecting spline-weighting coefficients to minimize the objective function:

$$\min_{b_y} \sum_i (y(t_i) - y_{\sim}(t_i))^2 \tag{2.8}$$

which is the sum of the squared distances between the data points and the spline functions at the observation times, t_i. Note that there is no requirement for the data to be equally spaced in time; PDA readily accommodates non-uniformly sampled data. The weighting coefficients to be selected are $b_y = [b_{y,1}, b_{y,2}, \ldots, b_{y,P_y}]$.

In converting the raw data to functional form, there is a trade-off between the function being too smooth (so that it does not capture the detailed dynamics of the system) and not being smooth enough (fitting splines to the system noise). In his spline-based ODE parameter estimation approach, Varah (1982) treated this trade-off problem by interactively adjusting the number and position of knots by hand until he was satisfied with the smoothing that was obtained. Alternatively, the extent of smoothing can be controlled by adding a penalty on higher-order derivatives of the splines. One objective of the current work is to investigate the influence of higher-order derivative (HOD) roughness penalties (used by Ramsay, 1996; Ramsay & Ramsey, 2002; Schwetlick & Schütze, 1995, 1997 and others) on the resulting spline fit, and on the subsequent parameter estimates in the fundamental ODE model. In particular, we consider a second-order derivative (curvature) penalty:

$$\min_{b_y} \left[\sum_i (y(t_i) - y_{\sim}(t_i))^2 + \lambda_{\mathrm{HOD}} \int \left(\frac{\mathrm{d}^2 y_{\sim}(t)}{\mathrm{d}t^2} \right)^2 \mathrm{d}t \right] \tag{2.9}$$

Adjusting λ_{HOD} influences the trade-off between the two extremes: a rapidly fluctuating function that passes as closely as possible to the data points, including noise, and an overly smoothed function that cannot capture the detailed dynamic behavior of the system. We also propose a new iteratively refined PDA (iPDA) procedure in which the second-order derivative penalty is replaced by a model-based penalty (Heckman & Ramsay, 2000). For the model in Eq. (2.1), the corresponding iteratively refined PDA objective function for spline fitting is

$$\min_{b_y} \left[\sum_i (y(t_i) - y_{\sim}(t_i))^2 + \lambda_{\mathrm{ODE}} \int \left(\frac{\mathrm{d}y_{\sim}}{\mathrm{d}t} + \hat{w}_y y_{\sim} + \hat{w}_u u_{\sim} \right)^2 \mathrm{d}t \right] \tag{2.10}$$

The penalty term with the weighting coefficient λ_{ODE} uses the residuals of the ODE model (with the initial or current estimates of the model parameters) to prevent over-fitting of the measurement noise and to ensure that the fitted splines are consistent with the fundamental ODE model.

2.5. ODE parameter estimation

With the data expressed in functional form, the next step of PDA is to estimate the parameters in the ODE model. If the model parameters are assumed to be time-varying (rather than fixed), Ramsay and Silverman's (2005) PDA algorithm can approximate them using B-spline or other basis functions. In this article, however, we consider a simpler problem with fixed parameters and a known input curve $u(t)$. B-spline expressions for y and $\mathrm{d}y/\mathrm{d}t$ can be substituted into the fundamental model, so that the unknown parameters appear in an algebraic expression, resulting in relatively straightforward linear (if the ODE model is linear in the parameters) or nonlinear (if the ODE model is nonlinear in the parameters) least-squares estimation. For the model in Eq. (2.1) the least-squares objective function is

$$\mathrm{SSE}_{\mathrm{PDA}} = \int \left(\frac{\mathrm{d}y_{\sim}}{\mathrm{d}t} + w_y y_{\sim} + w_u u_{\sim} \right)^2 \mathrm{d}t = \int e(t)^2 \, \mathrm{d}t \tag{2.11}$$

evaluated over the time domain of the data. Note that in this parameter estimation step, the spline coefficients are fixed, and optimal values of the model parameters (w_y and w_u) are determined. In the PDA literature, only ODE models that are linear in the parameters have been considered (except for the related work by Varah (1982) which also considers nonlinear models). It is straightforward to extend PDA for parameter estimation in ODE models that are nonlinear in the parameters, using iterative nonlinear least-squares to estimate the model parameters. Note that Varah (1982) used a sum-of-squares objective function $\sum_i e(t_i)^2$ evaluated at times corresponding to the data points, which is different than the integral objective function in Eq. (2.11).

3. PDA example—SISO continuous stirred-tank reactor model

To explore the merits of PDA and iPDA relative to traditional nonlinear least-squares regression, we use a differential equation model (Marlin, 2000) that describes the dynamic response of the concentration of reactant A (C_A) to changes in temperature (T) in a CSTR with constant volume (V):

$$\frac{dC_A}{dt} = \frac{F}{V}(C_{A0} - C_A) - k_{ref} \exp\left(-\frac{E}{R}\left(\frac{1}{T} - \frac{1}{T_{ref}}\right)\right) C_A \tag{3.1}$$

The kinetic parameters to be estimated (k_{ref} and E/R) appear nonlinearly in this ODE model, and the model is also nonlinear in an input/output sense. A centered Arrhenius expression, where $k_{ref} = k_0 \exp(-E/RT_{ref})$ is the value of the kinetic rate constant evaluated at reference temperature T_{ref}, has been used to improve the conditioning of the parameter estimation problem (Watts, 1994). Marlin's (2000) nominal parameter values ($E/R = 8330.1$ K and $k_0 = 1.0 \times 10^{10}$ min^{-1}, corresponding to $k_{ref} = 0.461$ min^{-1} for $T_{ref} = 350$ K) are used as true values in the simulations.

PDA was initially designed for parameter estimation in empirical ODE models that are linear in the parameters and outputs. In this article, PDA is extended to accommodate nonlinear ODE models, but we begin by estimating parameters in a simplified version of the model, which is linearized around steady-state operating point $C_{As} = 0.576$ kmol m^{-3} and $T_s = 332$ K and expressed in deviation variables ($C'_A = C_A - C_{As}$; $T' = T - T_s$):

$$\frac{dC'_A}{dt} + w_C C'_A + w_T T' = 0 \tag{3.2}$$

The constant coefficients w_C and w_T are related to the original model parameters by

$$w_C = \frac{F_s}{V} + k_{ref} \exp\left(-\frac{E}{R}\left(\frac{1}{T_s} - \frac{1}{T_{ref}}\right)\right) \tag{3.3}$$

$$w_T = k_{ref}\frac{E}{R}\frac{C_{As}}{T_s^2} \exp\left(-\frac{E}{R}\left(\frac{1}{T_s} - \frac{1}{T_{ref}}\right)\right) \tag{3.4}$$

PDA parameter estimation in a more complex five-input, two-output, four-parameter nonlinear model has been investigated by Poyton (2005). In our simulated experiments the reactant feed rate is steady at $F_s = 0.05$ m^3 min^{-1}, the inlet reactant concentration C_{A0} is constant at 2.0 kmol m^{-3} and $V = 1.0$ m^3. We assume that a fast temperature controller has been implemented, so that changes in the temperature set point result in nearly instantaneous changes in T (and in T').

3.1. Spline fitting and parameter estimation results

As shown in Fig. 1a, simple step-test experiments were simulated (using the linearized model) in which T was decreased by 10 K at $t = 4$ min. Concentration data were sampled every 10 s. Using PDA to estimate parameters w_C and w_T in the linearized model, the first step is to smooth the noisy output data to obtain $C_{A\sim}(t)$. Knots placed at 4.0 min intervals result in 13 fourth-order B-spline basis functions for which 13 weighting coefficients are determined by minimizing the following objective function:

$$\min_{b_{C_A}} \left[\sum_i (C_A(t_i) - C_{A\sim}(t_i))^2\right] \tag{3.5}$$

The input curve, $T'(t)$, was assumed perfectly known. After fitting the splines, the coefficients w_C and w_T were determined by minimizing:

$$\min_{w_C, w_T} \left[\int \left(\frac{dC'_{A\sim}(t)}{dt} + w_C C'_{A\sim}(t) + w_T T'(t)\right)^2 dt\right] \tag{3.6}$$

Parameters k_{ref} and E/R were calculated from w_C and w_T. The integral in Eq. (3.6) was computed numerically over the 40 min simulation time for the dynamic experiment. A traditional NLS solver (MATLAB routine lsqcurvefit, which uses a Gauss–Newton iteration) was also used to estimate k_{ref} and E/R. Computation times (on a Pentium 4 computer with a 2.0 GHz processor) required to estimate the parameters are shown in Table 1. Initial guesses for the parameters (10% below their true values) were required for NLS parameter estimation, but not for PDA because the ODE model is linear in parameters w_C and w_T. Note that although the parameters appear linearly in the right-hand side of the linearized differential equation, they will appear nonlinearly in the time solution for $C_A(t)$. Traditional NLS estimation was performed in two different ways for the linearized model: (i) using the analytical solution for the ODE (row 1 in Table 1) and (ii) using repeated numerical solution of the ODE (row 2 in Table 1). Similar values for the parameter estimates were obtained using both NLS techniques. The computational effort associated with PDA was slightly higher than that required using traditional NLS and the analytical ODE solution, but was considerably

Table 1
Computation times required for parameter estimation

Method	I (min)	λ_{HOD} (min^3)	λ_{ODE} (min)	Coincident knots	Sampling period (s)	Standard error (kmol m^{-3})	Model type	CPU time (s)
NLSa	–	–	–	–	10	0.002	Linearized	0.08
NLSn	–	–	–	–	10	0.002	Linearized	1.78
PDA	4.0	0	0	No	10	0.002	Linearized	0.14
PDA	1.0	0	0	No	10	0.002	Linearized	0.55
PDA	4.0	0	0	Yes	10	0.002	Linearized	0.14
NLS	–	–	–	–	80	0.016	Nonlinear	1.73
PDA	4.0	0	0	Yes	80	0.016	Nonlinear	0.08
PDA	4.0	1.0	0	Yes	80	0.016	Nonlinear	0.39
PDA	4.0	10.0	0	Yes	80	0.016	Nonlinear	2.16
iPDA	4.0	0	1.0	Yes	80	0.016	Nonlinear	0.48
iPDA	4.0	0	10.0	Yes	80	0.016	Nonlinear	1.78

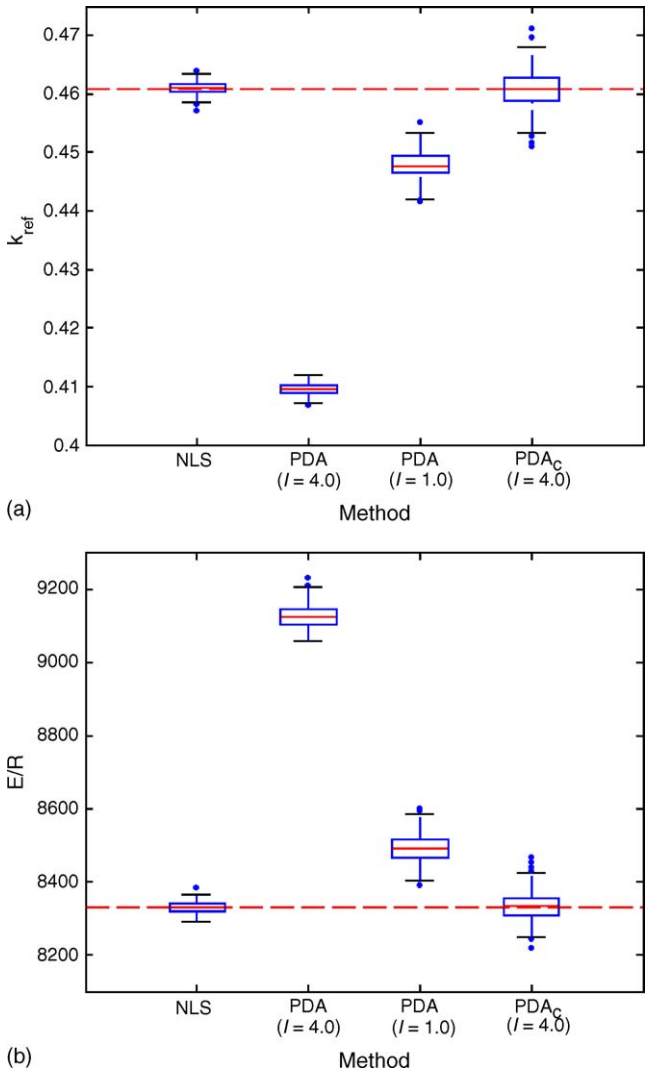

Fig. 2. Precision and bias of parameter estimates from traditional NLS and PDA estimation using simulated data from the linearized CSTR model with a sampling rate of 10 s and noise standard deviation of 0.002 kmol m^{-3} (a) k_{ref} estimates and (b) E/R estimates. PDA results are shown for uniform knot spacings of 4.0 and 1.0 min. PDA$_c$ refers to PDA parameter estimates obtained using three coincident knots at the time of the step change in the reactor temperature.

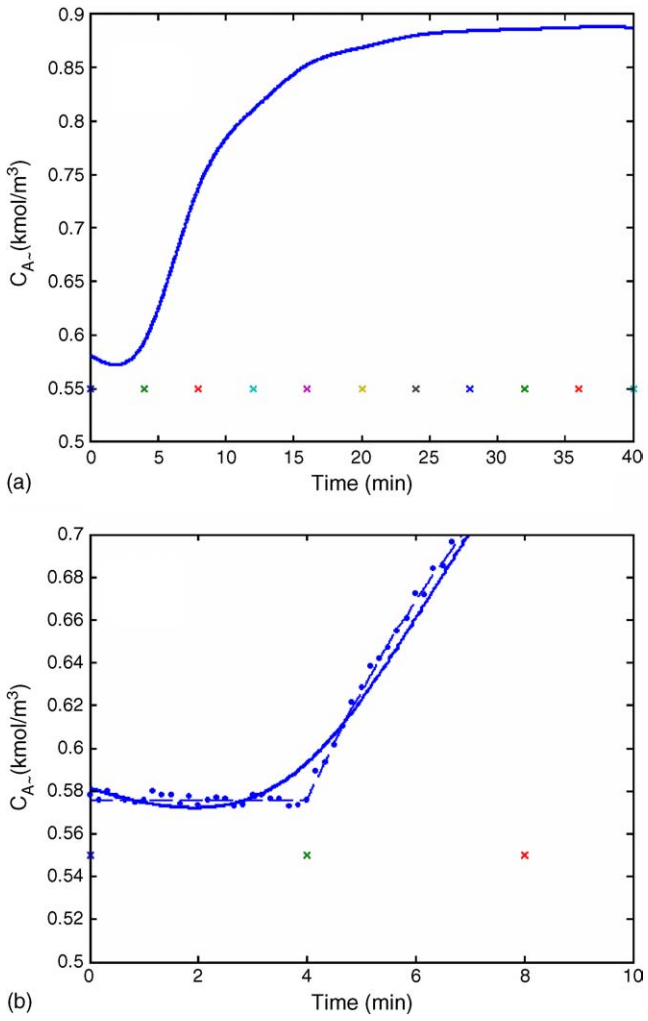

Fig. 3. B-spline approximation to output response from linearized CSTR model for a step change in temperature (with $I = 4.0$ min, $\lambda_{HOD} = 0$, $\lambda_{ODE} = 0$). Concentration was sampled every 10 s with a standard deviation of 0.002 kmol m^{-3} (●, simulated data; – – –, true response; —, fitted spline).

less than when numerical ODE solutions were used for NLS.

Box plots in Fig. 2 summarize the distribution of the estimates for parameters k_{ref} and E/R obtained using 500 simulated experiments, each with different random noise sequences. Results are shown for traditional NLS estimation and PDA with two different knot spacings ($I = 4.0$ min and $I = 1.0$ min) during the spline-fitting step. Bias in parameter estimates obtained from PDA improved considerably when closer knot placement was used but, unfortunately, this simplest form of the PDA algorithm, with uniformly spaced knots, did not do a very good job of estimating the parameters compared to traditional NLS.

The B-spline approximation that was obtained by minimizing Eq. (3.5) is shown in Fig. 3 for a uniform knot spacing of 4.0 min (the '×' symbols indicate the knot positions). The spline fit is especially poor near the sharp change in slope immediately following the step change. Decreasing I, the time interval between knots from 4.0 to 1.0 min, increased the number of basis functions used to express $C_{A\sim}$, improving the spline fit and the resulting parameter estimates (see Fig. 2). However, the B-spline fit to the data (not shown) was still poor near the sharp change in C_A.

Spline fitting and parameter estimation were repeated using two additional knots placed at $t = 4.0$ min. Using three coincident knots (τ_{c1}, τ_{c2}, τ_{c3}) where the output changes abruptly creates two artificial splines with no length, so that first and second derivatives immediately to the left of τ_{c1} do not need to match the corresponding derivatives immediately to the right of τ_{c3}. The parameter estimates (right-most box plots in Fig. 2) obtained from this improved spline fit (Fig. 4) were considerably better than those obtained without the coincident knots. These results are consistent with those of Varah (1982) who advocated using coincident knots to improve spline fits when response variables change abruptly.

In the results presented so far, the output measurements had only a small amount of error and sampling was frequent, so good

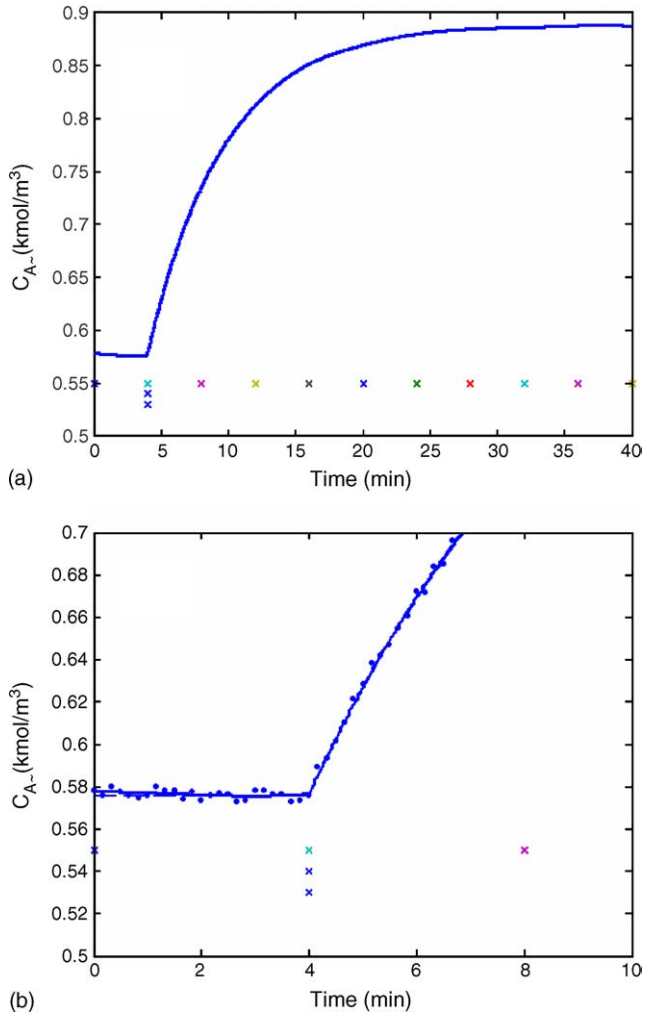

Fig. 4. B-spline approximation to output response from linearized CSTR model for a step change in temperature (with $I = 4.0$ min, $\lambda_{HOD} = 0$, $\lambda_{ODE} = 0$) and three coincident knots at the time of the step change. Concentration was sampled every 10 s with a standard deviation of 0.002 kmol m^{-3}.

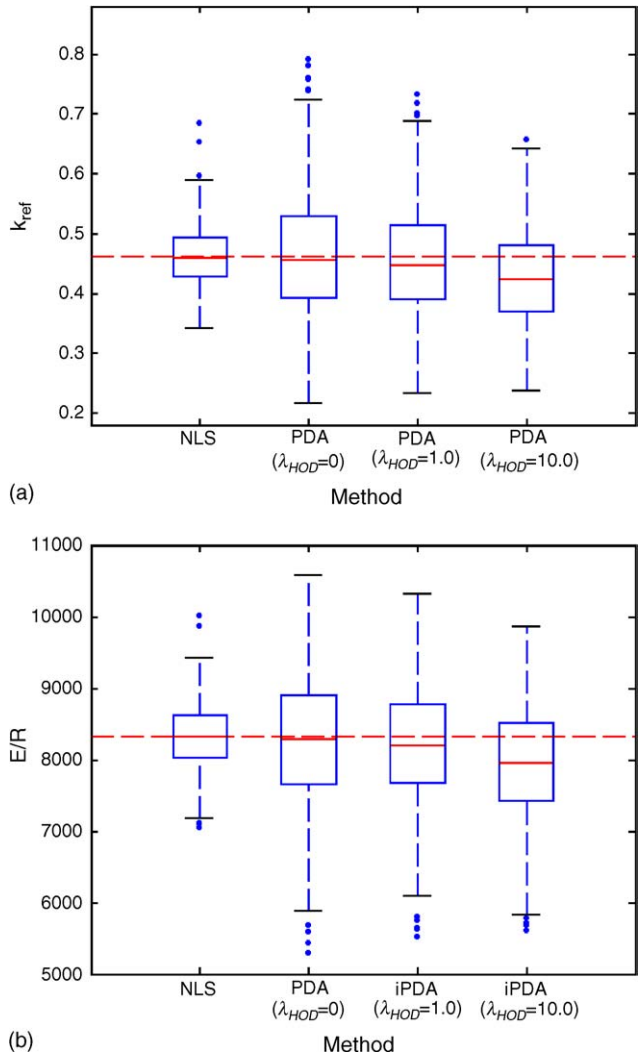

Fig. 5. Effect of second-order derivative penalties on estimates of (a) k_{ref} and (b) E/R obtained using the nonlinear CSTR model. Concentration was measured every 80 s with a standard deviation of 0.016 kmol m^{-3} ($I = 4.0$ min, three coincident knots at $t = 4.0$ min for PDA).

parameter estimates were obtained using regular PDA (with coincident knots and $I = 4.0$ min), but the parameter estimates were not as precise as those obtained using traditional NLS. Fig. 5 shows box plots for parameter estimates obtained using step-response data with the noise standard deviation increased to 0.016 kmol m^{-3} (from 0.002 in Fig. 2), and less frequent sampling (80 s rather than 10 s). As expected, noisier data lead to higher variances in the parameter estimates from both traditional NLS and PDA. Parameter estimates in Fig. 5 were obtained using data generated from the nonlinear CSTR model (Eq. (3.1)), rather than the linearized model. Because the right-hand side in this model is nonlinear in the parameters, initial guesses for parameter values were required for both NLS and PDA. Numerical Jacobians were used to obtain traditional NLS parameter estimates, but PDA estimates were obtained using an analytical Jacobian. One of the benefits of using PDA techniques is that analytical Jacobians can easily be obtained by differentiating the right-hand side of the ODE(s) with respect to the parameters. The ease with which analytical Jacobians are obtained may make PDA especially useful for parameter estimation in larger-scale ODE and DAE models.

The two right-most box plots in Fig. 5a and b show the effect of using a second-order-derivative penalty ($\lambda_{HOD} = 1$ and $\lambda_{HOD} = 10$, respectively) in an attempt to reduce the amount of rippling when splines are fitted to the noisier data. Using a larger value of λ_{HOD} during spline fitting decreased the variance of the resulting model parameter estimates, but increased the bias. Unfortunately, the penalty term forced the second derivative of the fitted splines to be smaller than the true second derivative of the output in time intervals where the curvature is large. Computing times for PDA with the nonlinear model and higher-order derivative penalty terms are shown in Table 1. PDA generally requires only a fraction of the time required by traditional NLS to estimate the parameters. When second-order derivative penalties ($\lambda_{HOD} \neq 0$) are included, the computation time increases a little due to the additional effort required in fitting the splines.

4. Iteratively refined principal differential analysis

In the previous section, a simple two-step PDA procedure was used to estimate the parameters. B-splines were fitted to the data, and then model parameters were estimated using the resulting splines and their derivatives. Unfortunately, poor spline fits from the first step sometimes led to poor model parameter estimates, especially when output data were noisy and sparse. Here we describe a new *iteratively refined* PDA technique (iPDA), which iterates between the smoothing and estimation steps as follows, so that good spline fits, and hence good parameter estimates can be obtained:

1. Estimate the model parameters using the fitted splines and their derivatives as in standard PDA.
2. Obtain an improved spline fit using a model-based roughness penalty (using values of the parameter estimates from step one and an objective function like Eq. (2.10)) to ensure that the fitted splines are smooth and physically reasonable.
3. Iterate between steps one and two until parameter estimates converge.

Box plots in Fig. 6 show the effect that using iPDA with an ODE-based penalty has on the PDA estimates. Improvements in bias and precision of the estimates are obtained as λ_{ODE} increases. Computing times become longer when larger penalties are used (Table 1) because more iterations between spline-fitting and parameter estimation are required for the parameter estimates to converge.

In Fig. 7, we can see the effect of the ODE-based penalty on the spline fit for a particular set of nonlinear step-response data. The regular PDA spline fit in Fig. 7a contains many fluctuations. As λ_{ODE} increases (Fig. 7b and c), these fluctuations are smoothed out, and the spline fit approximates the true output more accurately.

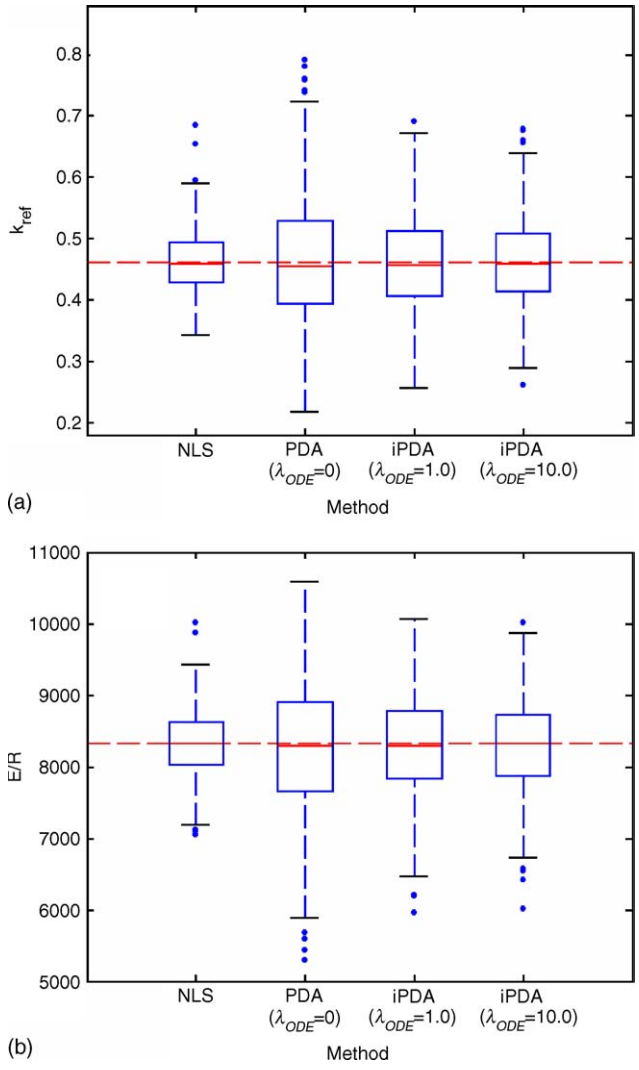

Fig. 6. Effect of iterative PDA penalty weight on estimates of (a) k_{ref} and (b) E/R obtained using the nonlinear CSTR model. Concentration was measured every 80 s with a standard deviation of 0.016 kmol m^{-3} (I = 4.0 min, three coincident knots at t = 4.0 min for PDA).

5. Conclusions and future work

Principal differential analysis can be used for estimating parameters in continuous dynamic models that describe chemical processes. Standard PDA consists of two steps: (i) fitting B-splines (or other basis functions) to dynamic data, and (ii) using the resulting empirical spline curves and their derivatives to convert the differential equations to algebraic expressions that are used for parameter estimation. Several benefits arise when PDA is used. Since the resulting parameter estimation problem is algebraic, repeated numerical simulation of differential equations is not required, and it is easy to determine analytical Jacobians for PDA parameter estimation by simply differentiating the right-hand side of the model ODE(s). As a result, PDA requires considerably less computational effort than traditional NLS estimation. As well, PDA does not require initial values for the dynamic output variables to be either known or estimated as additional parameters. Many PDA parameter estimation problems may be better behaved than their NLS counterparts; parameters (like kinetic rate constants) that appear linearly in differential equations behave nonlinearly in the integrated form of the model, which is used in traditional NLS estimation.

Nevertheless, some difficulties can arise when PDA is used. Poor spline fits (especially when data are sparse or noisy) give misleading derivative information, which results in inaccurate parameter estimates. PDA parameter estimates for a nonlinear CSTR model were less precise than those obtained using traditional NLS estimation, even when coincident spline knots were placed at points corresponding to abrupt changes in the output response. Penalties on second-order derivatives, which have been used to prevent over-fitting of noise, resulted in biased parameter estimates because the fitted splines were not consistent with the underlying behavior of the true process. As such, we do not recommend that higher-order-derivative penalties be used for PDA parameter estimation in fundamental models.

Instead, we recommend a new iteratively refined PDA algorithm (iPDA) which ensures that spline fits are consistent with the fundamental process behavior. The iPDA algorithm

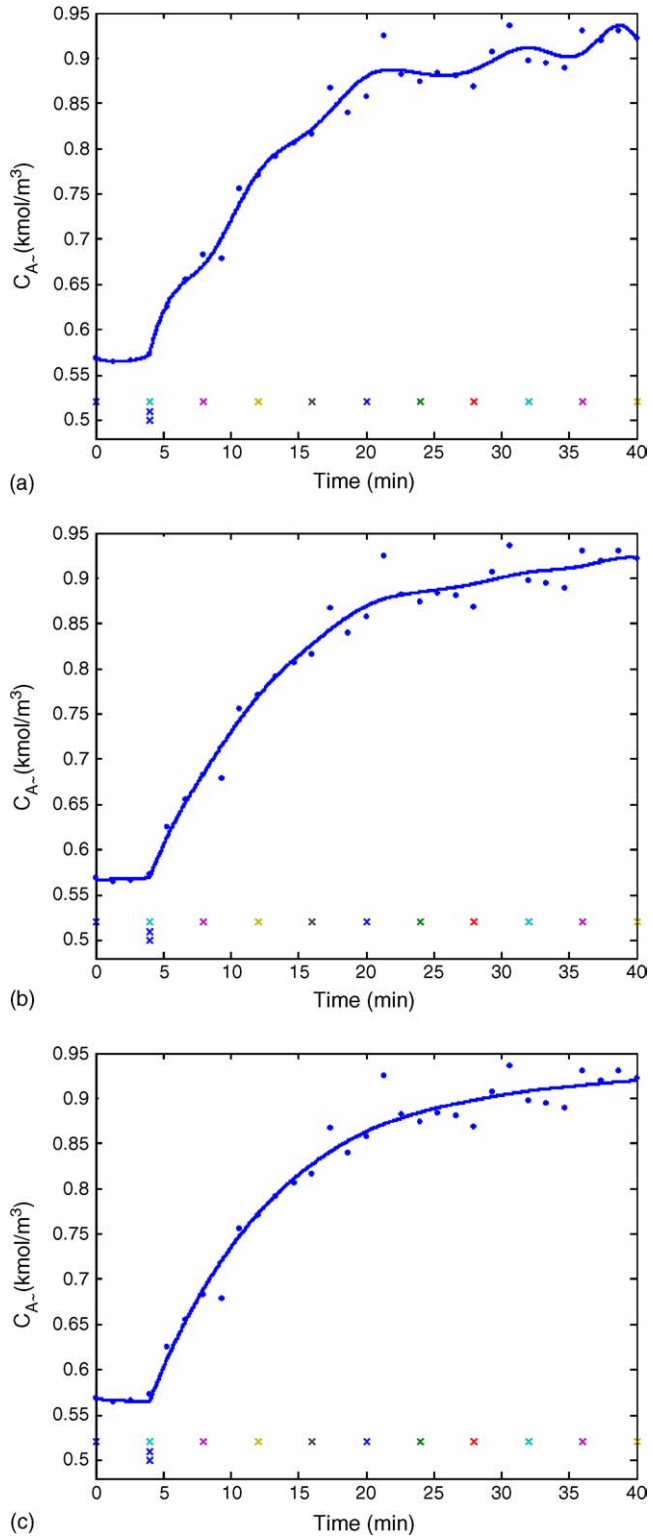

Fig. 7. Spline approximations using (a) regular PDA, (b) iPDA with $\lambda_{ODE} = 1.0$, and (c) iPDA with $\lambda_{ODE} = 10.0$. Concentration was measured every 80 s with a standard deviation of 0.016 kmol m^{-3} ($I = 4.0$ min, three coincident knots at $t = 4.0$ min).

improves the initial spline fit using a model-based penalty term. When iPDA was applied to the CSTR model using sparse and noisy data, the resulting parameter estimates were unbiased and were much more precise than those obtained using standard PDA. As the size of the model-based penalty coefficient increased, precision of the parameter estimates approached the precision obtained from standard NLS estimation; computation times increased as more iPDA iterations were required for the parameters to converge.

To date, PDA (and iPDA) has only been used for parameter estimation in simple dynamic models, for which traditional NLS estimation works well. PDA may have considerable benefits for parameter estimation in larger-scale models, either on its own, or as a computationally attractive means of obtaining good initial parameter estimates to be used in traditional NLS estimation. However, before PDA can be used for complex problems, several issues need to be resolved. The most serious drawback is that, in their current forms, PDA and iPDA are restricted to dynamic models in which all of the states are measured (all of the derivatives that appear in the model need to be converted to algebraic expressions). Other issues that need attention are related to algorithm tuning. Better knowledge is required so that PDA users can select enough B-spline knots (but not too many) so that good spline fits can be obtained efficiently, and so that additional knots can be added automatically in intervals where they are needed to obtain good spline fits. Users also need to know how to select appropriate weighting parameters for the model-based penalties used in iPDA, so that a desirable balance between accuracy of parameter estimates and computational effort is achieved. Furthermore, it will be helpful to understand the implications of the peculiar error structure in PDA (and iPDA) parameter estimation. Residuals for the parameter estimation step of PDA (see Eq. (2.11)) are in the differentiated rather than the integrated form of the model, whereas residuals for the spline-fitting step are in the regular output variables. PDA and iPDA may be well suited for dynamic parameter estimation problems in which different types of error arise from a variety of sources. Uncorrelated errors associated with measurement noise are consistent with the residuals from the spline-fitting step; whereas, correlated errors due to random disturbances that pass through the dynamic process are consistent with the residuals that are minimized during the parameter estimation step (residuals in the differentiated form of the model).

It will be important to generate information about the uncertainty of the parameter estimates that are obtained using iPDA. In the limiting case where the modeler assumes (as when using NLS) that the random errors are independent (no correlated errors due to disturbances), it is appropriate to use a very large value of the weighting coefficient, λ_{ODE}, to obtain parameter estimates that are unbiased and spline fits that closely approximate the ODE solution. In this situation, confidence intervals for model parameters obtained using iPDA become identical to those obtained using NLS (Varziri, 2006). In more complex situations, when uncorrelated measurement errors and correlated process disturbances are both present, iPDA users should select a smaller value of λ_{ODE} that takes into account the relative variances of the two types of error. In our current research we are developing methods for appropriate selection of λ_{ODE} and expressions to describe parameter uncertainty for multi-response ODE parameter estimation problems with both types of random error.

Acknowledgments

Financial support from the Natural Sciences and Engineering Research Council of Canada (NSERC), Mathematics of Information Technology and Complex Systems (MITACS) council, and the School of Graduate Studies, Queen's University, is gratefully acknowledged.

References

Bahadir, A. R. (2003). Application of cubic B-spline finite element technique to the thermistor problem. *Applied Mathematics and Computation*, *149*, 379–387.

Bates, D. M., & Watts, D. G. (1985). Multiresponse estimation with special application to linear systems of differential equations. *Technometrics*, *27*, 329–339.

Bates, D. M., & Watts, D. G. (1988). *Nonlinear regression analysis and its applications*. New York: John Wiley and Sons, Inc.

Benson, M. (1979). Parameter fitting in dynamic models. *Ecological Modelling*, *6*, 97–115.

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

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

de Boor, C. (1978). *A practical guide to splines*. New York: Springer.

Dovì, V. G., Arato, E., & Maga, L. (1985). A more general formulation of separable least squares. *Mathematical Modelling*, *18*, 20–23.

El-Farra, N. H., & Christofides, P. D. (2003). Bounded robust control of constrained multivariable nonlinear processes. *Chemical Engineering Science*, *58*, 3025–3047.

Elfving, T., & Andersson, L. E. (1998). An algorithm for computing constrained smoothing spline functions. *Numerische Mathematik*, *52*, 583–595.

Heckman, N. E., & Ramsay, J. O. (2000). Penalized regression with model-based penalties. *Canadian Journal of Statistics*, *28*, 241–258.

Kalogerakis, N., & Luus, R. (1983). Improvement of Gauss–Newton method for parameter estimation through the use of information index. *Industrial and Engineering Chemistry Fundamentals*, *22*, 436–445.

Kim, E. (1998). A mixed Galerkin method for computing the flow between eccentric rotating cylinders. *International Journal for Numerical Methods in Fluids*, *26*, 877–885.

Lainiotis, D., & Deshpande, J. G. (1974). Parameter estimation using splines. *Information Sciences*, *7*, 291–315.

Leis, J. R., & Kramer, M. A. (1988). ALGORITHM 658: ODESSA—An ordinary differential equation solver with explicit simultaneous sensitivity analysis. *ACM Transactions on Mathematical Software*, *14*, 61–67.

Ljung, L. (1999). *System identification—Theory for the user* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Lucero, J. C. (2002). Identifying a differential equation for lip motion. *Medical Engineering and Physics*, *24*, 521–528.

Madar, J., Abonyi, J., Roubos, H., & Szeifert, F. (2003). Incorporating prior knowledge in a cubic spline approximation—Application to the identification of reaction kinetic models. *Industrial and Engineering Chemistry Research*, *42*, 4043–4049.

Mansouri, N., & Kernévev, J. P. (1998). Estimation of parameters in nonlinear problems. *Numerical Algorithms*, *17*, 333–343.

Marlin, T. E. (2000). *Process control: Designing processes and control systems for dynamic performance* (2nd ed.). McGraw-Hill.

Nagy, Z. K., & Braatz, R. D. (2003). Robust nonlinear model predictive control of batch processes. *AIChE Journal*, *49*, 1776–1786.

Ogunnaike, B. A., & Ray, W. H. (1994). *Process dynamics, modeling, and control*. New York, NY: Oxford University Press, Inc [Chapter 12].

Poyton, A. (2005). *Application of principal differential analysis to parameter estimation in fundamental dynamic models*. M.Sc. Thesis, Queen's University, Kingston, Canada.

Ramsay, J. O. (1996). Principal differential analysis: Data reduction by differential operators. *Journal of the Royal Statistical Society, Series B*, *58*, 495–508.

Ramsay, J. O. (2000). Functional components of variation in handwriting. *Journal of the American Statistical Association*, *95*, 9–16.

Ramsay, J. O., & Munhall, K. G. (1996). Functional data analyses of lip motion. *Journal of the Acoustical Society of America*, *99*, 3718–3727.

Ramsay, J. O., & Ramsey, J. B. (2002). Functional data analysis of the dynamics of the monthly index of nondurable goods production. *Journal of Econometrics*, *107*, 327–344.

Ramsay, J. O., & Silverman, B. W. (1997). *Functional data analysis*. New York: Springer.

Ramsay, J. O., & Silverman, B. W. (2002). *Applied functional data analysis: Methods and case studies*. New York: Springer-Verlag.

Ramsay, J. O., & Silverman, B. W. (2005). *Functional data analysis* (2nd ed.). New York: Springer.

Schwetlick, H., & Schiitze, T. (1995). Least squares approximation by splines with free knots. *BIT*, *35*, 361–384.

Schwetlick, H., & Schiitze, T. (1997). Constrained approximation by splines with free knots. *BIT*, *37*, 105–137.

Seber, G. A. F., & Wild, C. J. (1989). *Nonlinear regression*. John Wiley and Sons, Inc.

Seber, G. A. F., & Lee, A. J. (2003). *Linear regression analysis*. New Jersey: Wiley.

Shariff, K., & Moser, R. D. (1998). Two-dimensional mesh embedding for B-spline methods. *Journal of Computational Physics*, *145*, 471–488.

Stewart, W. E., Caracotsios, M., & Sorensen, J. P. (1992). Parameter estimation from multiresponse data. *AIChE Journal*, *38*, 641–655.

Swartz, J., & Bremermann, H. (1975). Discussion of parameter estimation in biological modeling: Algorithms for estimation and evaluation of the estimates. *Journal of Mathematical Biology*, *1*, 241–275.

Tang, Y. P. (1971). On the estimation of rate constants for complex kinetic models. *Industrial and Engineering Chemistry Fundamentals*, *10*, 321–322.

Thomaseth, K., Kautzky-Wilier, A., Ludvik, B., Prager, R., & Pacini, G. (1996). Integrated mathematical model to assess β-cell activity during the oral glucose test. *American Journal of Physiology*, *270*, E522–E531.

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

Varziri, M. S. (2006). *Parameter estimation in continuous dynamic models using principal differential analysis techniques*. Ph.D. Thesis, Queen's University, in preparation.

Wang, R., Keast, P., & Muir, P. (2004). A high-order global spatially adaptive collocation method for 1-D parabolic PDEs. *Applied Numerical Mathematics*, *50*, 239–260.

Watts, D. G. (1994). Estimating parameters in nonlinear rate equations. *Canadian Journal of Chemical Engineering*, *72*(4), 701–710.