Parameter Estimation in Differential Equation Models with Constrained States

David A. Campbell  
Department of Statistics and Actuarial Science, Simon Fraser University  
and  
Giles Hooker  
Department of Statistical Science and  
Department of Biological Statistics and Computational Biology, Cornell University  
and  
Kim B. McAuley  
Department of Chemical Engineering, Queen’s University

Abstract

In this paper, we use an industrial data set with an ordinary differential equation (ODE) model to describe how the methodological work of generalized profile estimation Ramsay et al. (2007) can be adapted to parameter estimation from a challenging noisy data set. The example that we have chosen highlights the estimation process with multiple experimental replications, step function changes in functional inputs, noisy system components observed at different time points and precisions along with unobserved state variables within a nonlinear differential equation model. Additionally, we extend the method of Ramsay et al. (2007) to accommodate constraints such as non-negative values of state variables.

Keywords: Differential Equations, Profiled Likelihood Estimation, Functional Data Analysis, Nylon, Constrained Smoothing

1 Challenges of Parameter Estimation From Differential Equation Models

Ordinary Differential Equations (ODEs) describe the rate of change of system states \( x(t) \) over time \( t \) (or another argument) as a function of \( x(t) \) along with externally controlled inputs and \( p \)
model parameters $\theta$. Using the differential operator $D\mathbf{x} = \frac{d}{dt}\mathbf{x}$, the ODE model:

$$D\mathbf{x}(t) = f\left(\mathbf{x}(t), \theta\right)$$

(1)

is a functional regression model with covariates $\mathbf{x}(t)$. A relatively compact ODE model can be used to describe complex dynamics such as limit cycles, exponential growth or a stiff tendency towards an equilibrium state depending on the values of $\theta$ and the initial system states $\mathbf{x}(0)$. ODE’s are frequently used in pharmacokinetics and chemical reaction models, Gelman et al. (1996), chapter 5 of Bates and Watts (1988), Zheng et al. (2005), Poyton et al. (2006), because of the parsimonious nature of the models and the interpretability of $\theta$.

In general for a nonlinear function $f(\cdot)$ in (1), there is no analytic form for $\mathbf{x}(t)$, but given $\theta$ and $\mathbf{x}(0)$, numerical methods can produce a solution $S(\theta, \mathbf{x}\{0\}) = \mathbf{x}(t)$. However the measurement error model

$$y(t) \sim N\left(S(\theta, \mathbf{x}\{0\}), \sigma^2\right)$$

(2)

will not have a closed form likelihood. Further complicating the model statistically, observations $y(t)$ may only be available from a subset of the states $\mathbf{x}(t)$.

Based on on (2), $\theta$ can be estimated by nonlinear least squares (NLS) through gradient based optimization, Bates and Watts (1988), Seber and Wild (1989). However, difficult topologies including ripples, ridges, flat sections and/or multiple local maxima are characteristic of likelihoods from measurement error models for ODEs, Esposito and Floudas (2000). Knowing $x(0)$ improves reliability of parameter estimates and therefore augmenting the unknown parameter vector from $\theta$ to $\psi = [\theta, \mathbf{x}(0)]$ makes the estimation problems worse, especially when the model trajectory is strongly impacted by changes in initial conditions, Wu et al. (2008).

The alternative to depending on $S(\theta, \mathbf{x}\{0\})$ is to work at the level of the derivative by smoothing the data to produce $\hat{\mathbf{x}}(t) \approx S(\theta, \mathbf{x}\{0\})$ and estimating $\hat{\theta}$ without augmenting the dimension of the parameter space to include $\mathbf{x}(0)$. While some methods use a non-parametric data smooth, Varah (1982), Chen and Wu (2008), others use iterative conditional updating of the smoothing and parameter estimates, Poyton et al. (2006), Ramsay and Silverman (2005). We will focus on the Generalized Smoothing (GS) method of Ramsay et al 2007 where $\hat{\mathbf{x}}(t)$ and $\theta$ are estimated using a multi-level, multi-criterion optimization and parametric, model based smoothing.

GS is a collocation method; it uses a basis expansion with a vector of coefficients $\mathbf{c}$ and basis functions $\phi(t)$ to approximate the system states such that $\hat{\mathbf{x}}(t) = \mathbf{c}'\phi(t) \approx S(\theta, \mathbf{x}\{0\})$. The model
based smooth \( \hat{x}(t) \) is guided by the ODE by penalizing deviation at the level of the derivative. The tradeoff between interpolating \( y(t) \) and perfectly following \( S(\theta, x(0)) \) is controlled by the smoothing parameter \( \lambda \). Consequently the generalized profiling approach requires estimation of \( \hat{c}, \hat{\theta} \) and \( \hat{\lambda} \) but through a multi-criterion three level hierarchical optimization, we define \( c(\theta, \lambda) \) and \( \theta(\lambda) \), such that the process reduces to estimating \( \lambda \). In this paper we extend GS to accommodate constraints on the shape of \( \hat{x}(t) \) and show how GS can be applied to a challenging data set.

Section 2 provides a description of the motivating data set from an industrial chemical engineering application with a nonlinear ODE model. The challenges of the motivating example include multiple experimental replications, step function changes in functional inputs, noisy system components observed at different time points and precisions along with unobserved state variables within a nonlinear differential equation model where system states are constrained to take on non-negative values. Section 3 describes GS along with the extension to constrained state variables. Section 4 describes how GS is adapted to overcome the remaining challenges of the motivating data set. Section 5 provides an analysis of the real data set. Discussion about the method’s performance and further adaptations and generalizations are given in section 6.

2 Nylon system

In a heated reactor, amine (A) and carboxyl (C) groups combine producing polyamide links (L) and water (W) which escapes as steam. Simultaneously, steam bubbled through molten nylon, decomposes L into A and C giving the symbolic competing reactions \( A + C \leftrightarrow L + W \). From any starting values, A,C,L and W will adjust their concentrations until these competing reaction rates balance and equilibrium concentrations are achieved. In the experiment of Zheng et al. (2005) steam is bubbled through molten nylon to maintain a controlled concentration of W in the system. Consequently A, C and L move towards equilibrium concentrations with the imposed level of W. Within each of the \( i = 1, \ldots, 6 \) experimental runs, the amount of input steam followed a step function. Initial high levels of steam were held constant until time \( \tau_{i1} \) then reduced until time \( \tau_{i2} \) and finally returned to the original level for the remainder of the experimental run. Each experimental run was performed at a constant temperature \( T_i \) which, along with the input water pressure, determines the equilibrium concentration of water in the molten nylon mixture, \( W_{eq} \). Experimental conditions are given in table 2 in the appendix. With reaction rates \( k_p \) and \( K_a \), the
reaction dynamics are described with differential equations:

\[ -DL = DA = DC = -k_p(CA - LW/K_a) \] (3)

\[ DW = k_p(CA - LW/K_a) - 24.3(W - W_{eq}) \] (4)

Reaction rates \( k_p \) and \( K_a \) are allowed to change with \( T \) and \( W_{eq} \) according to parameter vector \( \theta = [k_p, E, \gamma, \beta, K_{a0}, H] \) and the reference temperature \( T_0 = 549.15 \) Kelvin, through the following relationships:

\[ k_p = \frac{k_p0}{1000} \exp \left( -E \frac{10^3}{8.314} \left\{ \frac{1}{T} - \frac{1}{T_0} \right\} \right) \] (5)

\[ K_a = \left\{ 1 + W_{eq} \frac{\gamma}{1000} \exp \left( -\beta \frac{10^3}{8.314} \left\{ \frac{1}{T} - \frac{1}{T_0} \right\} \right) \right\} K_T[K_{a0}] \exp \left( -H \frac{10^3}{8.314} \left\{ \frac{1}{T} - \frac{1}{T_0} \right\} \right) \] (6)

\[ K_T = 20.97 \exp \left( -9.624 + \frac{3613}{T} \right) \] (7)

These equations include scaling factors making all initial parameter estimates used in Zheng et al. 2005 for \( \theta \) of the same order of magnitude, to ease estimation. Equation (6) is altered from the original paper to ensure consistency of parameter units. Alternatively, 4 or 5 parameters versions of (5)-(7) are also of interest and are described in 5.2.2. Figure 1 shows the data for each of the experimental runs. The plot shows observed components \( A \) and \( C \) and includes vertical lines corresponding to \( \tau_{i1} \) and \( \tau_{i2} \). This data and differential equation model produce several important challenges that are addressed in this paper:

1. Given any three chemical components, the fourth can be determined algebraically using the mass balance of the system, however it was only possible to measure \( A \) and \( C \). Consequently an unobserved state variable must also be estimated.

2. Information must be pooled across the 6 experimental runs. Using a method based on producing \( S_i(\theta, x(t)) \) requires estimating \( 3 \times 6 = 18 \) initial conditions and \( p = 6 \) ODE parameters. Zheng et al. 2005 avoided the increased parameter space by assuming that the observation error model does not apply to initial conditions, that is, they assumed that initial conditions were measured without observation error and that \( W(0) = W_{eq}(0) \). We wish to estimate \( \theta \) without making assumptions about initial conditions.

3. This system describes chemical concentrations which are by definition constrained to take on non-negative values.
Figure 1: The nylon observations along with the fit to the data. Temperatures of the experimental runs are given above component \( A \) in degrees Kelvin. Vertical axes are in concentration units and horizontal axes are in hours.
4. The equations in (3) and (4) suggest that the concentrations of \( A \) and \( C \) are a result of competing exponential growth and decay towards the asymptotic equilibrium level. This level jumps abruptly in response to step changes in \( W_{eq} \) causing discontinuities in (3) and (4) at times \( \tau_{i1} \) and \( \tau_{i2} \). Methods based on non-parameteric smoothing do not easily readily accommodate a mixture of fast and slow changing dynamics.

5. Figure 1 shows that the variability in the measurement of \( C \) is larger than the variability of \( A \), that is \( \sigma^2_A \neq \sigma^2_C \). More generally in ODE models, states \( x(t) \) may measure quantities measured in different units and scales. In addition, the observations for components \( A \) and \( C \) are not evenly spaced, simultaneously measured or observed the same number of times.

6. While the ODE model is based on scientific theory, the model in (3) to (7) is believed to be imperfect.

3 Generalized Profile Estimation

GS estimates \( c, \theta \) and \( \lambda \) through a multi-criterion multi-level hierarchical optimization. A parameter hierarchy is defined by re-writing some parameters as functions of other parameters higher up the hierarchy: \( c(\theta, \lambda) \) and \( \theta(\lambda) \), such that the process reduces the estimation of \( \lambda \) to define all parameters. To describe how this occurs we first define the role of each type of parameter and its place in the hierarchy. The following subsections describe estimation of the parameter hierarchy.

The incidental or local parameters in the sense of Neyman and Scott (1948), are the basis coefficients \( c \) of the data smooth. Representing the lowest level of the hierarchy, \( c \) is defined as a function of parameters \( \theta \) and \( \lambda \). For each \( \lambda \) and \( \theta \), the optimal \( c \) defines a data smooth, balancing the fit between interpolating the data and following the ODE model. The smoothing step described in Section 3.1 allows the data fit to deviate from the ODE model, possibly accounting for some process noise or stochastic variation.

The structural parameters \( \theta \) define the behavior allowed by the ODE model. Changes in these parameters decide between limit cycles, unbounded exponential growth or other behaviors from the ODE model. We are primarily interested in \( \theta \) because of its interpretation and potential use for making decisions and predictions. As described in Section 3.2, for any \( \lambda \), \( \theta(\lambda) \) is estimated
through a profile likelihood maintaining $c(\hat{\theta}\{\lambda\}, \lambda)$ at it’s optimum at every candidate value of $\hat{\theta}$, defining the second level of the hierarchy.

The complexity or smoothing parameter $\lambda$, defines the top level of parameters to estimate. It determines how closely the data follows the ODE model and allows for some model miss-specification. For any $\lambda$ we can find $\theta(\lambda)$ and $c(\theta\{\lambda\}, \lambda)$.

For $n$ data observations, $c$ is a vector of length $m$ and $\theta$ is a vector of length $p$ where typically $m + p > n$. By creating $c(\theta(\lambda), \lambda)$ in the nylon application reduces the dimension of the optimization from 1977 down to 7. By further defining $\theta(\lambda)$ the problem is further reduced to estimating a single parameter $\lambda$ containing all degrees of freedom of the model. Estimation of $\lambda$ is detailed in Section 3.3.

The estimation routine can be thought of as resulting from inner and outer loops to estimate $\hat{c}(\theta, \lambda)$ and $\hat{\theta}(\lambda)$ for fixed $\lambda$. While the estimation process is described in Ramsay et al. 2007, we extend these loops with constraints on the smooth in the remainder of this section. Software to perform the estimation is available.

3.1 The Inner Optimization; Model-Based Smoothing

Using a piecewise differentiable one to one function $g(\cdot)$ to constrain the smooth to follow known behaviour, the basis expansion $x_{ik}(t) = g_{ik}(c'_{ik}\phi_{ik}(t)) \approx S(\theta, x\{0\})$ for the $i = 1, \ldots, I$ experimental runs and $k = 1, \ldots, K$ system components smooths the data and approximates the ODE trajectory. Some examples of $g(\cdot)$ given in Ramsay and Silverman (2005) include the following:

- $g(a) = \exp(a)$ positive smooth
- $g(a) = \exp(a)/(1 + \exp(a))$ bounded smooth
- $g(a) = \int_0^t \exp(a)ds$ monotone smooth
- $g(a) = a$ unconstrained smooth

$L$-spline smoothing in Ramsay et al. 2005 estimates $\hat{c}$ from a (weighted) penalized likelihood regularizing the data fit with the ODE model. Using weights $w_{ik}$ and the measurement error model in (2), the estimates $\hat{c}$ minimize the weighted penalized negative log likelihood:

$$\hat{c} = \min_c J_i(c \mid y, \theta, \lambda) = \min_c \sum_{k=1}^K w_{ik} \{SSE_{ik} + \lambda \text{PEN}_{ik}\}$$

$$= \min_c \sum_{k=1}^K w_{ik} \{\sum_{t \in t_{ik}} [y_{ik}(t) - g(c'_{ik}\phi_{ik}(t))]^2 + \lambda \text{PEN}_{ik}\}$$ (8)
In non-parameteric smoothing literature, often $\text{PEN} = \int D^2 x(t) dt$ implying a model with little or no curvature. In GS, $\text{PEN}$ penalizes the residual of (1) and here takes the form:

$$\text{PEN}_{ik} = \int_{T_i} (Dx_{ik}(s) - f_k(x_i, \theta))^2 ds.$$ (9)

For fixed $\lambda$ and $\theta$ and data $y_{ik}$ observed at the vector of times $t_{ik}$, $\hat{c}(\theta, \lambda)$ is obtained through nonlinear regression. The integral in equation (9) is over the interval $T_i = [\min_k(t_{ik}), \max_k(t_{ik})]$, the maximum range of observation times over all $K$ observed variables in the $i^{th}$ run. Section 4.3 describes how to approximate this integral.

### 3.2 The Outer Optimization; Estimating ODE parameters

The maximum profile likelihood estimate (MPLE) is used for $\hat{\theta}$, where, for example using the error model (2), the MPLE is equivalent to minimizing the negative log profile likelihood:

$$H(\theta, \hat{c}(\theta, \lambda) \mid y) = \sum_{i} \sum_{k=1}^{I} \sum_{j=1}^{K} w_{ik} SSE_{ik} = \sum_{i} \sum_{k=1}^{I} \sum_{j=1}^{K} \sum_{t \in t_{ik}} w_{ik} [y_{ik}(t) - g_k(\hat{c}_{ik}(\theta, \lambda)' \phi_{ik}(t))]^2.$$ (10)

A penalized likelihood is not used at this level of the parameter hierarchy since fidelity to the ODE model is already enforced in (8).

For fixed $\lambda$, finding $\hat{\theta}$ is a nonlinear regression problem applied to the profile likelihood while ensuring $c(\hat{\theta}, \lambda)$ is at its optimum. Consequently, as $\hat{\theta}$ is updated through Gauss-Newton iterations in the outer loop, $c(\hat{\theta})$ must be updated for each increment through an inner optimization step. Simplifying notation from $H(\theta, c(\theta, \lambda) \mid y)$ to $H$, assuming that $\lambda$ is fixed and simplifying $g(c' \phi)$ to $g$, gives the total gradient for the likelihood

$$\frac{dH}{d\theta} = \frac{\partial H}{\partial \theta} + \frac{\partial H}{\partial g} \frac{dc}{dc} \frac{dc}{d\theta}.$$ (11)

When $f(x(t), \theta)$ is a nonlinear function of $x(t)$ there is no explicit function for $\hat{c}(\theta, \lambda)$ and consequently $dc/d\theta$ must be obtained using the implicit function theorem. Assume that $H$ and $J$ are twice continuously differentiable with respect to $\theta$ and $c$ and that the Hessian matrices

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

are positive definitive over a nonempty neighborhood of $y$ in the data space. Using $dJ/dc = 0$ at $c = \hat{c}$ from (8);
\[
\begin{align*}
\frac{d^2 J}{d \mathbf{c} d \theta} & = \frac{d}{d \theta} \left( \frac{d J}{d \mathbf{c}} \right) \\
& = \frac{d}{d \theta} \left( \frac{d J}{d g} \frac{d g}{d \mathbf{c}} \right) \\
& = \left[ \frac{\partial^2 J}{\partial \theta \partial \mathbf{c}} \frac{d g}{d \mathbf{c}} + \left( \frac{d g}{d \mathbf{c}} \right) \frac{\partial^2 J}{\partial g^2 \partial \mathbf{c}} + \frac{\partial^2 J}{\partial g \partial \mathbf{c}^2} \right] \frac{d \mathbf{c}}{d \theta}. 
\end{align*}
\]

(12)

At \( \mathbf{c} = \hat{\mathbf{c}} \), solving for \( \frac{d \mathbf{c}}{d \theta} \) we obtain

\[
\frac{d \mathbf{c}}{d \theta} = - \left\{ \left( \frac{d g}{d \mathbf{c}} \right) \frac{\partial^2 J}{\partial g^2 \partial \mathbf{c}} + \frac{\partial J}{\partial g} \frac{d^2 g}{d \mathbf{c}^2} \right\}^{-1} \left\{ \frac{\partial^2 J}{\partial g \partial \mathbf{c}} \frac{d g}{d \mathbf{c}} \right\}
\]

(13)

which we substitute into (11) to obtain the total gradient for the MPLE:

\[
\frac{d H}{d \theta} = \frac{\partial H}{\partial \theta} - \frac{\partial H \frac{d g}{d \mathbf{c}}}{\partial g} \left\{ \left( \frac{d g}{d \mathbf{c}} \right) \frac{\partial^2 J}{\partial g^2 \partial \mathbf{c}} + \frac{\partial J}{\partial g} \frac{d^2 g}{d \mathbf{c}^2} \right\}^{-1} \left\{ \frac{\partial^2 J}{\partial g \partial \mathbf{c}} \frac{d g}{d \mathbf{c}} \right\}.
\]

(14)

3.2.1 Interval Estimates for \( \hat{\theta}(\lambda) \)

Interval estimates obtained using the delta method approximation,

\[
\text{var}(\theta) \approx \left. \frac{d \theta}{d y} \right| \text{var}(y) \left. \frac{d \theta}{d y} \right|
\]

(15)

require the implicit function theorem once again to define \( d \theta/d y \). Using the fact that at \( \theta = \hat{\theta} \), \( d H/d \theta = 0 \) from (10), \( d \theta/d y \) can be found using

\[
\left. \frac{d}{d y} \left( \frac{d H}{d \theta} \right) \right|_{\theta = \hat{\theta}} = \left. \left[ \frac{d^2 H}{d y d \theta} + \frac{d^2 H}{d \theta^2} \frac{d \theta}{d y} \right] \right|_{\theta = \hat{\theta}} = 0,
\]

(16)

where

\[
\frac{d^2 H}{d \theta^2} = \frac{\partial^2 H}{\partial \theta^2} + 2 \frac{\partial^2 H \frac{d g}{d \mathbf{c}}}{\partial \theta \partial g} \frac{d \mathbf{c}}{d \theta} + \left( \frac{d g}{d \mathbf{c}} \right) \frac{\partial^2 H}{\partial g^2 \partial \mathbf{c}} \frac{d \mathbf{c}}{d \theta} + \left( \frac{d g}{d \mathbf{c}} \right) \frac{\partial^2 H}{\partial g \partial \mathbf{c}^2} \frac{d \mathbf{c}}{d \theta} + \frac{\partial H \frac{d^2 g}{d \mathbf{c}^2}}{\partial g} \frac{d \mathbf{c}}{d \theta} \frac{d^2 \mathbf{c}}{d \theta^2},
\]

(17)

and

\[
\frac{d^2 H}{d \theta d y} = \frac{d^2 H}{d \theta d y} + \frac{d^2 H \frac{d g}{d \mathbf{c}}}{d \theta d y} \frac{d \mathbf{c}}{d \theta} + \frac{d^2 H \frac{d g}{d \mathbf{c}}}{d \theta d y} \frac{d \mathbf{c}}{d \theta} + \left( \frac{d g}{d \mathbf{c}} \right) \frac{\partial^2 H}{\partial g^2 \partial \mathbf{c}} \frac{d \mathbf{c}}{d \theta} \frac{d \mathbf{c}}{d \theta} + \frac{\partial H \frac{d^2 g}{d \mathbf{c}^2}}{d \theta} \frac{d \mathbf{c}}{d \theta} \frac{d^2 \mathbf{c}}{d \theta^2} + \frac{\partial H \frac{d^2 g}{d \mathbf{c}^2}}{d \theta} \frac{d \mathbf{c}}{d \theta} \frac{d^2 \mathbf{c}}{d \theta^2}.
\]

(18)

These last two equations involve the terms \( d^2 \mathbf{c}/d \theta^2 \), \( d^2 \mathbf{c}/d \theta d y \) and \( d \mathbf{c}/d y \), all of which arise from further calls to the implicit function theorem. These terms are given in the appendix.
Equation (13) for point estimates and equations (17) and (18) used in obtaining confidence intervals (as described in section 3.2.1) simplify when there are no constraints on the smooth. In that case \( g(c'\phi) = c'\phi \) giving \( dg/dc = \phi \) and \( d^ng/dc^m = 0 \) for all \( m > 1 \) while a positively constrained smooth has \( g(c'\phi) = \exp(c'\phi) \) and \( d^mg(c'\phi)/dc^m = \phi^mg(c'\phi) \).

3.3 Estimating the Complexity Parameter

Smoothing parameter \( \lambda \) controls the flow of information between \( y(t) \) and \( \theta(\lambda) \) such that small \( \lambda \) allows limited influence from \( \theta(\lambda) \) on the shape of \( x(t) \). When \( \lambda = 0 \), the differential equation model is ignored and (8) is minimized non-uniquely along the manifold of function space where \( x(t) \) interpolates the data in any way permitted by the basis functions. Small values of \( \lambda \) limit the impact of \( \theta(\lambda) \) on \( x(t) \) producing a wide basin of attraction for \( \hat{\theta}(\lambda) \) allowing it to traverse difficult likelihood surfaces and avoid local likelihood maxima. As \( \lambda \) increases, the SSE term in (8) increases, as shown in the left plot of figure 2 using 100 simulated data sets based on the 4 parameter version of the nylon model (3) and (4) that is described in section 5.2.2. Simulated observations were obtained by adding observational noise \( N(0,\sigma^2_k) \), where \( [\sigma^2_A,\sigma^2_C] = [.6^2,2.4^2] \) to the ODE solution with \( \theta, x(0) \) equal to their estimated values from the real data. The right column of figure 2 shows how \( \text{PEN} \) decreases with increasing \( \lambda \), reducing the discrepancy between \( Dg[c(\theta\{\lambda\},\lambda)'\phi(t)] \) and \( f(x(t),\theta) \). Increasing \( \lambda \) increases the impact of changes in \( \theta \) on \( x(t) \) and re-introduces the sharper profile likelihood features. Consequently \( \lambda \) has a similar role to a temperature parameter in annealing as both are used to ease movement around the parameter space and both are used to refine point and interval estimates.

When \( \lambda = \infty \), \( \text{PEN} = 0 \) by forcing the residual from (1) to 0. Since \( x(t) \) is then a solution to the ODE, at \( \lambda = \infty \) GS is equivalent to NLS while profiling over \( x(0) \). However, setting \( \lambda = \infty \) requires a basis capable of perfectly matching \( S(\theta,x\{0\}) \) over a wide range of \( \theta \).

For most practical choices of \( \phi(t) \), as \( \lambda \to \infty \), the range of values of \( \theta \) for which \( \text{PEN} \to 0 \) is limited. For instance when \( \lambda \) becomes ‘too large’ for \( \phi(t) \), \( x(t) \) is forced to find reductions in \( \text{PEN} \) at the resolution demanded by \( \lambda \) by moving \( x(t) \) towards a trajectory which is closer to the function space of ODE solutions. This results in \( x(t) \) moving away from the dynamics in the data, but reduces \( \text{PEN} \) by moving \( \theta(\lambda) \) towards a value where the function spaces of the basis and ODE are closer to intersecting.
Figure 2: Weighted SSE (left) and PEN (right) as a function of $\log_{10}(\lambda)$ for 100 simulated nylon data sets.

Figure 3: The 4 parameter estimates from the 4 parameter nylon model from 100 simulated data sets at different values of $\log_{10}(\lambda)$. The horizontal line is the true parameter value.
As an extreme example, consider the situation where the chosen basis is the set of piecewise linear functions with discontinuities at fixed locations and the ODE model describes a sinusoid with frequency $\theta$. As $\lambda$ increases, to reduce PEN, $\theta$ must move towards a frequency of 0, to eliminate the periodicity so that the ODE dynamics can match the basis. This will occur as $\lambda \to \infty$ regardless of the dynamics expressed by the data. However increasing the number of basis functions will postpone the problem towards a larger $\lambda$ if the basis and the model do not span intersecting function spaces. The parameter estimates from the 100 simulated data sets as a function of $\lambda$ and the onset of the basis induced bias are shown in figure 3. Figure 2 also shows a corresponding sharp increase in SSE and an abrupt decrease in PEN at $\lambda > 10^4$, which is too large for the basis.

Although Ramsay et al. (2007) admit choosing an optimal $\lambda$ remains an open problem, figures 2 and 3 show that $\lambda \in (10^2, 10^4)$ produces $\theta(\hat{\lambda})$ is close to $\theta_{true}$ and gives stable values of SSE and PEN. Within this range of $\lambda$, $x(t)$ is not visually different from $S(\theta, x\{0\})$. Consequently, we suggest the following practical estimation strategy:

1. Start with a small $\lambda_j$ to obtain a low bias but potentially high variance initial estimate $\hat{\theta}(\lambda_j)$.
2. Increase $\lambda_{j+1} = \lambda_j \times 10$ to remove smoothness in the likelihood and obtain the refined estimate $\hat{\theta}(\lambda_{j+1})$.
3. Repeat step 2 until parameter estimates converge or de-stabilize at $\lambda_J$. Keep the estimates from $\theta(\lambda_{J-1})$ since larger $\lambda$ enforces the ODE model as much as possible for the given basis.

## 4 Overcoming Challenges of a Real Data System

Now that we have addressed constrained estimation, this section addresses the additional challenges of the motivating data set outlined in section 2.

### 4.1 Unobserved Outputs

The model (1) depends on observed and unobserved system states. The unobserved states in GS are estimated by $\hat{x}(t)$ while following the ODE model through PEN. Estimated unobserved states are data regularized by their impact on the fit to $y(t)$ through the observed states in (8).
Therefore, the shape of $\hat{x}(t)$ for the unobserved states is the function that helps to produce the best fit to the observed system states while regularized to follow the model dynamics.

### 4.2 Multiple Experimental Runs

The data smoothing step of (8) is performed separately on each experimental run. The information for $\theta(\lambda)$ is combined by multiplying together the likelihoods from the independent experiments. Consequently, the outer optimization is performed by pooling the $y_{ik}$ and $x_{ik}(t)$ over $i$ and $k$ treating each experimental run as an additional functional observation used to estimate $\theta$.

### 4.3 Computing PEN and Discontinuous Derivatives

The integral in (9) is evaluated using a numerical quadrature approximation which allows each component to have a unique differentiable basis and does not require variables to have been measured at the same times. It does however require that the quadrature points be the same for all components within the $k^{th}$ experimental run so that information from each component is available to compute $\text{PEN}_{ik}$ over the $k^{th}$ quadrature grid. Ramsay et al. 2007 describe using quadrature approximation by dividing $T_i$ into a set of small intervals whose boundaries are the unique knot locations compiled over the bases of all $k$ components. Denoting the location of the $\ell^{th}$ such knot location by $\xi_\ell$, intervals are then split into four equal-sized subintervals, and Simpson’s rule weights $[1, 4, 2, 4, 1](\xi_{\ell+1} - \xi_\ell)/5$ are used to approximate the integral over each interval. Other quadrature rules may be appropriate in some situations.

At points of discontinuity in the first derivative $\tau_i$, the integral in (9) is undefined. At $\tau_i$, the left or right hand derivatives could be defined at $\tau_i$. Alternatively, the integration can be set up to avoid $\tau_i$ by integrating over the intersection of $T_i$ and the compliment of a small $\delta$ sized neighborhood, $\tau_i^\delta$:

$$\text{PEN}_{ik} = \int_{T_i \cap \tau_i^\delta} (Dx_{ik}(s) - f_k(x_i, u_i, s | \theta))^2 \, ds$$

$$= \int_{\tau_{i_{\min}} - \delta/2}^{\tau_{i_{\min}}} (Dx_{ik}(s) - f_k(x_i, u_i, s | \theta))^2 \, ds$$

$$+ \int_{\tau_{i_{\max}} + \delta/2}^{\tau_{i_{\max}}} (Dx_{ik}(s) - f_k(x_i, u_i, s | \theta))^2 \, ds$$

(19)

This integral is approximated by shifting the quadrature interval boundaries at times $\tau_i$ to the points defining the boundaries of $\tau_i^\delta$ and omitting quadrature weights across $\tau_i^\delta$. To avoid enabling
the basis to push an extremely poor fit into \( \tau_i^\delta \), effectively allowing a discontinuous smooth across \( \tau_i^\delta \), we recommend using a small neighborhood such as \( 10^{-6} \times \min_t(\xi_{t+1} - \xi_t) \).

### 4.4 Outputs Measured With Different Precision

In ODE systems, often components are measured in different units, scales and precisions consequently it is important for parameter estimation that weights \( w_{ij} \) bring \( SSE_k \) to approximately the same scale. In some cases this may include using a vectors for \( w_{ij} \) to accommodate autocorrelations in the data. For a Gaussian likelihood, the optimal choice is \( w_{ki} = 1/\sigma^2_{ki} \), the inverse of the measurement error variance, Sen and Srivastava (1990). As with any regression or smoothing problem, iterative re-weighting can be applied when the relative importance of the weights is unknown. This can be performed using the following steps.

1. At iteration \( m = 0 \) initialize \( w_{ki}^{(m)} = 1 \) or use another value consistent with prior information.

2. Perform the profile estimation to obtain \( \hat{c}, \hat{\theta}, \hat{\lambda} \) using weights \( w_{ki}^{(m)} \).

3. Obtain the vectors of residuals \( r_{ki} = \hat{y}_{ki} - y_{ki} \) and estimate new weights \( \hat{w}_{ki}^{(m+1)} = \left( \frac{n_{ki}}{F_{ki}^{-1}g_{ki}} \right) \), the inverse of the residual variance estimate for the \( n_{ki} \) observations. Other robust weight estimators may also be appropriate.

4. If \( | \hat{w}_{ki}^{(m+1)} - \hat{w}_{ki}^{(m)} | > \epsilon \) for some convergence tolerance \( \epsilon > 0 \), return to step 2.

### 4.5 Model Imperfection

When the model is not correct, \( \lambda \to \infty \) may not be reasonable. In which case a smaller \( \lambda \) balances the fit to the model dynamics with the fit to the data. We used a simulation study to explore model uncertainty. We altered model (3) and (4) such that when \( W_{eq} \) decreases after \( \tau_{i1} \), it moves to a level \( \tilde{W}_{eq} = W_{eq} \times \xi_i \), where \( \xi_i \sim N(1, 2^2) \), but we continue to assume it is actually at level \( W_{eq} \). At time \( \tau_{i2} \), the level is returned to it’s original known value. This models the situation that could arise if the gauge used to calibrate \( W_{eq} \) was malfunctioning. Allowing \( \tilde{W}_{eq} \) to differ from \( W_{eq} \) by 20% or 40% results in the model over or under-estimating the values of \( A \) or \( C \) during the intervals \( (\tau_{i1}, \tau_{i2}) \).
Figure 4: The 4 parameter estimates from the mis-specified 4 parameter nylon model from 100 simulated data sets at different values of $\log_{10}(\lambda)$. The horizontal line is the true parameter value.

GS was performed using the 4 parameter nylon model of Section 5. Simulated data have Gaussian measurement error noise $[\sigma_A^2, \sigma_C^2] = [2.26^2, 4^2]$. The resulting parameter estimates are in figure 4. Parameter estimates remain close to the true values, despite the miss-specification in the value of $W_{eq}$. The parameters in this model are involved in determining the rate at which the system approaches equilibrium after $\tau$ and in determining the equilibrium concentration levels. Although $\tilde{W}_{eq}$ alters some of this information, the equilibrium levels before $\tau_{i1}$ and after $\tau_{i2}$ provide accurate information but with wider spread than in figure 3. Naturally, the quality of the estimation is dependent on the type of model miss-specification and more extreme cases will require model refinements as described in Hooker (2009).
5 Estimation Details and Results for the Nylon data.

5.1 Details of the Basis and Weights

Data smoothing was performed using a fifth order B-spline basis with knots linearly spaced at a rate of 10 per experiment hour and one at each observation of component $A$. Additional knots at the times of changes in input $W_{eq}$ were used to allow a discontinuity in the first derivative of the smooth at $\tau$. This same basis was also used for $C$ and unobserved $W$. This strategy produced between 84 and 142 unique interior knots per component with the actual numbers given in the last column of table 2 in the appendix. Although there are at most 23 observations for any component, the density of this basis avoids producing a badly behaved smooth in the times of fast dynamics immediately after $\tau_i$ and $\tau_{i+2}$. A denser basis should be used if the ODE solution is expected to exhibit finer resolution features or faster dynamics. Since all of the components are concentrations, negative estimates of concentration have no interpretation and we use the positively constrained data smooth $x(t) = g(c'\phi(t)) = \exp(c'\phi(t))$.

In the nylon example, all components are concentrations with the same scale and units however $A$ was measured with more precision than $C$. In what we will refer to as method 1 below, we use the weights determined by additional experiments and used by Zheng et al. (2005) in their weighted NLS. In accordance with their weights, for all $i = 1, \ldots, 6$ experimental runs we use $w_{iA} = 1/\sigma_A^2 = 1/6^2$ and $w_{iC} = 1/\sigma_C^2 = 1/2.4^2$ held constant across all runs. In method 2 we challenge this assumption and estimate $\sigma_A^2$ and $\sigma_C^2$ through iterative re-weighting.

5.2 Parameter Estimation Details

5.2.1 Estimation Procedure

The parameter estimation process for the nylon model was initialized with the parameters all set to the value 10. This choice is discussed in section 6.1. GS was initially performed holding $\lambda^{(1)} = 10$ fixed until $\hat{\theta}(\lambda^{(1)})$ converged. The iterative refinement of $\hat{\theta}(\lambda)$ was used following the strategy of section 3.3.
<table>
<thead>
<tr>
<th>Parameter</th>
<th>lower</th>
<th>point</th>
<th>upper</th>
</tr>
</thead>
<tbody>
<tr>
<td>Method 1 (Fixed weights), $\sigma_A = .6$ &amp; $\sigma_C = 2.4$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$k_p$</td>
<td>14.63 (20.54)</td>
<td>20.50 (20.67)</td>
<td>26.37 (20.80)</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>14.44 (21.86)</td>
<td>26.71 (26.91)</td>
<td>38.99 (31.97)</td>
</tr>
<tr>
<td>$K_{a0}$</td>
<td>39.04 (45.58)</td>
<td>50.35 (50.23)</td>
<td>61.66 (54.87)</td>
</tr>
<tr>
<td>$H$</td>
<td>-49.91 (-43.48)</td>
<td>-36.56 (-37.06)</td>
<td>-23.20 (-30.64)</td>
</tr>
<tr>
<td>Method 2 (2 weights) $\sigma_A = 2.13(2.26)$ &amp; $\sigma_C = 4.37(4.00)$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$k_p$</td>
<td>15.95 (15.64)</td>
<td>18.72 (17.72)</td>
<td>21.48 (20.16)</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>14.92 (17.09)</td>
<td>22.10 (22.66)</td>
<td>29.29 (28.67)</td>
</tr>
<tr>
<td>$K_{a0}$</td>
<td>47.46 (49.51)</td>
<td>55.68 (55.65)</td>
<td>63.90 (62.49)</td>
</tr>
<tr>
<td>$H$</td>
<td>-47.08 (-44.28)</td>
<td>-36.74 (-35.18)</td>
<td>-26.40 (-26.57)</td>
</tr>
</tbody>
</table>

Table 1: 95% Confidence intervals for the nylon data using Method 1 with values from NLS with $X(0)$ fixed in brackets and Method 2 values with MCMC results in brackets.

5.2.2 Results

The nylon model in (3) to (7) differs from that of Zheng et al. 2005 to correct an inconsistency in the balance of units. However, point estimates in our model were plagued by extremely high correlation ($r^2 > .99999$) between $\beta$ and $\gamma$ removing the ability to statistically uniquely identify these parameters. Consequently $\beta$ was eliminated from the model so that instead of (6) we use:

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

(20)

A similar simplification was also recently developed by Varziri et al. (2008) to address the high correlations of $\beta$ and $\gamma$ in a stochastic differential equation version of this system. In this reduced model using the procedure of section 5.2.1, the final 95% confidence interval estimates for $E$ overlap zero suggesting that a simpler model would be effective at fitting the data. Setting $E$ to zero in (5) is equivalent to removing the effect of temperature on $k_p$ in (3) and (4), replacing (5) with:

$$k_p = \frac{k_{p0}}{1000}.$$

(21)

Final point and interval estimates for the 4 parameter model are given under method 1 in
table 1 along with the estimates using NLS assuming that the initial system states are known and equal to those estimated by the smoothing step of GS. Note that the point estimates for NLS and GS are very similar but the assumed knowledge about the initial states reduces the overall uncertainty in $\hat{\theta}(\lambda)$, Wu et al. (2008). Consequently NLS intervals are considerably smaller than those of GS. Estimating the initial system states in NLS adds 3 new parameters to the model for each experimental run and produces some negative estimates for $W(0)$. Furthermore with $x(0)$ unknown, NLS was more sensitive to choice of parameters used to initialize the algorithm.

The fit to the data is shown in figure 1. Due to the weighted least squares criteria the fit to the ODE for $A$ is slightly better than the fit to $C$, the latter is presumed to have a larger measurement variance. The data fit shown in figure 1 is the solution to the differential equation using $\hat{x}(0)$ from GS. The difference between the smooth at $\lambda = 10^3$ and the solution to the differential equation based on the smooth estimates $\hat{x}(0)$ is less than 1% of the variability of the data fit at its worst point.

5.3 Estimating Unknown Weights

In this section we use method 2 to challenge the assumption of method 1, that $\sigma_A^2$ and $\sigma_C^2$ as determined from additional experiments are indicative of the relative precisions of the additive error component. Here we use iterative re-weighting as described in section 4.4 assuming that $\sigma_A^2$ and $\sigma_C^2$ are unknown but constant across experimental runs.

For comparison with model 2, the table includes Bayesian point and interval estimates using a likelihood centered on the solution to the ODE model, $S_i(\theta, x_i(0), t)$ (see Gelman, Bois and Jiang 1996 or Huang and Wu 2006):

$$P(y \mid \theta, x_i(0), \sigma_A^2, \sigma_C^2) \sim TRN \left( S_i(\theta, x_i(0)), \begin{bmatrix} \sigma_A^2 & 0 \\ 0 & \sigma_C^2 \end{bmatrix} \right),$$

where TRN is the normal distribution truncated to prevent negative chemical concentrations. The
Bayesian model uses the variance structure of method 2, and the following priors:

\[
P(k_p | 0) = P(K_a | 0) = \Gamma(4, 8),
\]
\[
P(\gamma) = N(0, 15^2),
\]
\[
P(H) = N(0, 50^2),
\]
\[
P(\sigma_A^2) = P(\sigma_C^2) = \Gamma(3, 3),
\]
\[
P(A_i(0)) = N(A_{observed}(0), (2 * \tilde{\sigma}_A)^2), \quad \tilde{\sigma}_A = .6,
\]
\[
P(C_i(0)) = N(C_{observed}(0), (2 * \tilde{\sigma}_C)^2), \quad \tilde{\sigma}_C = 2.4,
\]
\[
P(W_i(0)) = N(W_{eq,i}(0), 25^2).
\]

These priors represent the expected order of magnitude of the point estimates along with considerable uncertainty. While Bayesian estimates depend on the prior specification, the priors in (22) and the Bayesian results are included for illustrative purposes and comparison with the results of GS. Large discrepancies between MCMC and GS results would suggest cause for concern. MCMC was performed using the above model with 85,000 iterations, where the first 25,000 were discarded as burn in.

Method 2 estimated standard deviations (\(\hat{\sigma}_k = 1/\sqrt{w_k}\)) using iteratively re-weighted GS and the corresponding marginal posterior means of the Bayesian model are included in table 1. Point and interval estimates for GS and MCMC are close with differences in interval estimates due to philosophy, choice of prior and uncertainty in \(\sigma^2\) which is not accounted for in the intervals of GS.

6 Discussion

6.1 Sensitivity to Initial Parameter Estimates

Parameter estimation was attempted with the poor choice of initial parameter estimates \(k_{p0} < 0\) and/or \(K_{a0} < 0\). Negative values of these parameters alter (3) and (4) from describing exponential decay to unbounded exponential growth, ignoring the mass balance of the system. When \(\theta\) prevents the dynamics of (1) from matching the dynamics of the data, a small value of \(\lambda\) reduces the influence of \(f(\theta, x)\) and improves the basin of attraction for \(\theta\) by shifting \(x(t)\) towards the data. Decreasing the starting value of \(\lambda\) provides considerable robustness to poor initial parameter estimates by allowing \(x(t)\) to approach a data interpolant.
Another type of poor initial parameter estimates also arises when $\gamma$ is large and consequently in (20) the term
\[
(1 + \frac{\gamma}{1000} W_{eq} K_{a0}) \approx \frac{\gamma}{1000} W_{eq} K_{a0}
\]
eliminating the ability to uniquely determine $\gamma$ and $K_{a0}$. This feature of the parametrization may not be overcome by altering $\lambda$ and highlights the potential problem of the parameterization of ODE models.

For the nylon example, a poor choice of initial parameter estimates did not affect the converged values if GS succeeded but impacted whether or not the parameters converged to finite values.

### 6.2 Choice of Basis and Model

The ODE in (3) and (4) can be reformulated by noting that $A$ and $C$ have identical derivatives and consequently the difference in concentration $C_i(t) - A_i(t) = \omega_i$ is constant within each experimental run. This suggests that the ODE system could be rewritten as a differential algebraic equation (DAE):

\[
-DL = DA = -k_p(CA - LW/K_a) \\
C = \omega + A \\
DW = k_p(CA - LW/K_a) - 24.3(W - W_{eq})
\] (23)

The DAE model requires the augmenting $\theta$ by including $\omega = [\omega_1, \ldots, \omega_I]$. Using GS, in the ODE version of the nylon system, $\omega$ is estimated implicitly by the smooth in the same way that the smooth estimates $x(0)$. However these values are profiled out in the multi-level optimization. As $\omega$ and $x(0)$ are not of primary interest we prefer to keep this system as an ODE model.

An alternative basis system which exploits the DAE structure of the system uses the smooth for $A$ as one of two basis functions for $C$. The second basis function is equivalent to estimating the constant $\omega_i$. At large values of $\lambda$, this accelerates the computation by reducing the number of basis coefficients. However, at small values of $\lambda$ this basis hinders the flexibility of the smooth to simultaneously interpolate the data for $A$ and $C$ and therefore the estimation process may lose robustness to poor initial parameter estimates. While there is little lost by using too many basis functions, too few will not permit the model dynamics to be adequately represented by the basis expansion.
6.3 Conclusion

While our example data set models the chemical reactions of components of nylon production, these equations could just as easily be describing the dynamics of supply and demand by manipulating price, the number of viruses in an infected person while manipulating their medications or a variety of other situations. Consequently, this test problem was chosen to highlighting the advantages and intricacies of the generalized smoothing approach to parameter estimation from ODE models.

Although, (1) uses a measurement error model, using a finite $\lambda$ in estimating $\hat{x}(t)$ allows some deviation from the deterministic system that may prove useful when applied to stochastic differential equation models. Furthermore, although the paper assumes a Gaussian error structure, GS can be altered to suit another error distribution by altering the log likelihood optimization steps in (8) and (10).

The generalized profiling method is computationally efficient and robust to initial parameter estimates. Furthermore, the estimation procedure allows the incorporation of additional information through constraints on the data smooth. The parameter estimates are comparable in accuracy to those obtained through NLS and MCMC but computationally faster and more stable.

Acknowledgements

The authors would like to thank NSERC, MITACS and FFF grant NYC-150446 for funding to support this work.

Appendix

A Nylon Experiment Details

Table 2 shows the experimental conditions manipulated during each run. Temperatures are given in degrees Kelvin. Values of the input pressure $P_w$ and calculated values of $W_{eq}$ using equations in Zheng et al. 2005 are given at the initial value and after they drop. The times $\tau$ of the change in $P_w$ to the mid or final level are given in hours after the initial observation. The table also shows the number of observations within each run for components $A$ and $C$. The last column is the difference between $C$ and $A$ averaged over all observations times where both were measured within each run.
Table 2: T is given in degrees Kelvin. $P_w$ and $W_{eq}$ are given at time zero (equal to the value after time $\tau_2$) and after the first step change at time $\tau_1$. Times of the step changes are given in hours after the first observation time. The number of observations are given for $A$ and $C$ along with the number of basis functions used to smooth each of $A$, $C$ and $W$. The final column of the table shows the concentration difference between $A$ and $C$ averaged over times when both are observed.

## B Additional Implicitly defined derivatives

In this section we give the remaining implicitly defined derivatives required to obtain the confidence interval estimates of section 3.2.1. All of these derivatives simplify considerably when the smooth is unconstrained.

### B.1 $\partial^2 c/\partial \theta \partial \theta_k$

The implicit function theorem is required to define $\partial^2 c/\partial \theta \partial \theta_k$ in (17). The term $\partial^2 c/\partial \theta \partial \theta_k$ comes from the fact that $\partial J/\partial c = 0$ at the optimal choice of $c = \hat{c}$. Then differentiating twice with respect to $\theta$ and $\theta_k$, equivalent to differentiating (12) with respect to $\theta_k$ produces (24) which is then rearranged to give the derivative in (25).
produces (26) which is then rearranged to give the derivative in (27).

\[
\frac{\partial}{\partial \theta_k} \left( \frac{\partial^2 J}{\partial c \partial \theta} \right) = \frac{\partial}{\partial \theta_k} \left( \frac{\partial^2 J}{\partial c \partial \theta} \right) + \left( \frac{dg}{dc} \right) \frac{\partial}{\partial c} \left( \frac{\partial^2 J}{\partial c \partial \theta} \right) + \left( \frac{\partial f}{\partial g} \right) \frac{\partial}{\partial g} \left( \frac{\partial^2 J}{\partial c \partial \theta} \right)
\]

(24)

\[
\frac{\partial^2 c}{\partial \theta \partial \theta_k} = - \left[ \left( \frac{dg}{dc} \right) \frac{\partial^2 J}{\partial c^2} + \frac{\partial f}{\partial g} \left( \frac{\partial^2 J}{\partial c \partial \theta} \right) \right]^{-1} \left[ \frac{\partial^3 J}{\partial g \partial c \partial \theta_k} \frac{dg}{dc} + \left( \frac{dg}{dc} \right) \frac{\partial^3 J}{\partial g \partial c \partial \theta_k} \right] + \left( \frac{dg}{dc} \right) \frac{\partial^3 J}{\partial g \partial c \partial \theta_k} \frac{dg}{dc} + \left( \frac{dg}{dc} \right) \frac{\partial^3 J}{\partial g \partial c \partial \theta_k} \frac{dg}{dc}
\]

(25)

### B.2 \( \partial \hat{c} / \partial y \)

The implicit function theorem is required to define \( \partial \hat{c} / \partial y \) in (18). This derivative again uses the fact that \( \partial J / \partial c = 0 \) at the optimal choice of \( c = \hat{c} \). Then differentiating twice with respect to \( y \) produces (26) which is then rearranged to give the derivative in (27).

\[
\frac{\partial}{\partial y} \left( \frac{\partial J}{\partial c} \right) = \frac{\partial^2 J}{\partial y \partial c} + \left( \frac{dg}{dc} \right) \frac{\partial^2 J}{\partial y \partial c} + \left( \frac{\partial f}{\partial g} \right) \frac{\partial^2 J}{\partial y \partial c} + \frac{\partial f}{\partial g} \frac{\partial^2 J}{\partial y \partial c}
\]

(26)

\[
\frac{\partial \hat{c}}{\partial y} = \left( \frac{dg}{dc} \right) \frac{\partial^2 J}{\partial y \partial c} + \left( \frac{\partial f}{\partial g} \right) \frac{\partial^2 J}{\partial y \partial c} + \left( \frac{\partial f}{\partial g} \right) \frac{\partial^2 J}{\partial y \partial c}
\]

(27)

### B.3 \( \partial^2 \hat{c} / \partial y \partial \theta \)

We obtain this derivative by differentiating (26) with respect to \( \theta_k \) to produce equation (28). Solving for \( \partial^2 \hat{c} / \partial y \partial \theta \) gives us the results in (29).
\[
\frac{\partial}{\partial \theta_k} \left( \frac{\partial^2 \ell}{\partial c \partial y} \right) = \frac{\partial}{\partial \theta_k} \left( \frac{\partial^2 \ell}{\partial y \partial c} \right) + \left( \frac{\partial y}{\partial c} \right)' \frac{\partial^2 \ell}{\partial y^2} \frac{\partial y}{\partial c} + \frac{\partial^2 \ell}{\partial y \partial c} \left( \frac{\partial y}{\partial c} \right)' \frac{\partial c}{\partial y} \left( \frac{\partial c}{\partial y} \right) \frac{\partial^2 \ell}{\partial c \partial y} + \frac{\partial^2 \ell}{\partial c \partial y} \frac{\partial^2 \ell}{\partial c \partial y} + \left( \frac{\partial c}{\partial y} \right)' \frac{\partial^2 \ell}{\partial c^2} \frac{\partial c}{\partial y} + \frac{\partial^2 \ell}{\partial c \partial y} + \frac{\partial^2 \ell}{\partial c \partial y} + \frac{\partial^2 \ell}{\partial c \partial y} \left( \frac{\partial c}{\partial y} \right)' \frac{\partial c}{\partial y} \left( \frac{\partial c}{\partial y} \right) \frac{\partial^2 \ell}{\partial c \partial y} + \frac{\partial^2 \ell}{\partial c \partial y} \frac{\partial^2 \ell}{\partial c \partial y} + \left( \frac{\partial c}{\partial y} \right)' \frac{\partial^2 \ell}{\partial c^2} \frac{\partial c}{\partial y} + \frac{\partial^2 \ell}{\partial c \partial y} + \frac{\partial^2 \ell}{\partial c \partial y} + \frac{\partial^2 \ell}{\partial c \partial y} \left( \frac{\partial c}{\partial y} \right)' \frac{\partial c}{\partial y} \left( \frac{\partial c}{\partial y} \right) \frac{\partial^2 \ell}{\partial c \partial y} \right) \quad (28)
\]

\[
\frac{\partial^2 \ell}{\partial y \theta} = - \left\{ \left( \frac{\partial y}{\partial c} \right)' \frac{\partial^2 \ell}{\partial y \partial c} + \frac{\partial^2 \ell}{\partial y \partial c} \left( \frac{\partial y}{\partial c} \right)' \frac{\partial^2 \ell}{\partial y \partial c} \right\}^{-1} \left\{ \frac{\partial^3 \ell}{\partial y^2 \partial c} \frac{\partial y}{\partial c} + \frac{\partial^3 \ell}{\partial y^2 \partial c} \frac{\partial y}{\partial c} + \frac{\partial^3 \ell}{\partial y^2 \partial c} \frac{\partial y}{\partial c} + \frac{\partial^3 \ell}{\partial y^2 \partial c} \frac{\partial y}{\partial c} \right\} + \left\{ \frac{\partial^2 \ell}{\partial y \partial c} \left( \frac{\partial y}{\partial c} \right)' \frac{\partial^2 \ell}{\partial y \partial c} \right\} + \left\{ \frac{\partial^2 \ell}{\partial y \partial c} \left( \frac{\partial y}{\partial c} \right)' \frac{\partial^2 \ell}{\partial y \partial c} \right\} + \left\{ \frac{\partial^2 \ell}{\partial y \partial c} \left( \frac{\partial y}{\partial c} \right)' \frac{\partial^2 \ell}{\partial y \partial c} \right\} + \left\{ \frac{\partial^2 \ell}{\partial y \partial c} \left( \frac{\partial y}{\partial c} \right)' \frac{\partial^2 \ell}{\partial y \partial c} \right\} \quad (29)
\]

References


