Bayesian Collocation Tempering and Generalized Profiling for Estimation of Parameters from Differential Equation Models

David Alexander Campbell

Doctor of Philosophy

Department of Mathematics and Statistics

McGill University

Montreal,Quebec

July 2007

A thesis submitted to McGill University in partial fulfilment of the requirements of the degree of Doctor of Philosophy.

DEDICATION

For my bride Tara and our tiny tots Makenzie and Caeden.

ACKNOWLEDGEMENTS

I'd like to thank my supervisor, Jim Ramsay for his insights and guidance and support. His value of family has been instrumental in balancing my own family life with research. None of this would have been possible if he hadn't kept on believing in me even when I had doubts.

I also thank Jiguo Cao and Giles Hooker for their insights and conversations about the problems that I have been working on. I look forward to working with both of them again soon.

I'd like to thank Russ Steele for numerous meetings, insights and Bayesian guidance.

Thanks also to Geneviève Lefebvre who made my French abstract coherent.

The Department of Mathematics and Statistics at McGill University has been full of opportunities and interesting people who have helped to guide my statistical journey over the past four years. They deserve a big thanks for accepting me into the Ph.D. program.

Thanks also to Kim McAuley, Jim McLellan and Saeed Varziri for the nylon data, canoe rides and insightful meetings.

My bride Tara deserves much more than a simple thanks. Five years ago she gently pointed out that maybe statistics would be a good career choice. Since then, her understanding and encouragement keep setting my sights higher, while keeping me firmly planted in reality. Without her tenacious unwillingness to listen when I suggested taking an easier route, this would not have been possible. Makenzie and Caeden played an important role in keeping my priorities in line, and helping me to keep a healthy dose of silliness in my life.

ABSTRACT

The widespread use of ordinary differential equation (ODE) models has long been underrepresented in the statistical literature. The most common methods for estimating parameters from ODE models are nonlinear least squares and an MCMC based method. Both of these methods depend on a likelihood involving the numerical solution to the ODE. The challenge faced by these methods is parameter spaces that are difficult to navigate, exacerbated by the wide variety of behaviours that a single ODE model can produce with respect to small changes in parameter values.

In this work, two competing methods, generalized profile estimation and Bayesian collocation tempering are described. Both of these methods use a basis expansion to approximate the ODE solution in the likelihood, where the shape of the basis expansion, or data smooth, is guided by the ODE model. This approximation to the ODE, smooths out the likelihood surface, reducing restrictions on parameter movement.

Generalized Profile Estimation maximizes the profile likelihood for the ODE parameters while profiling out the basis coefficients of the data smooth. The smoothing parameter determines the balance between fitting the data and the ODE model, and consequently is used to build a parameter cascade, reducing the dimension of the estimation problem. Generalized profile estimation is described with under a constraint to ensure the smooth follows known behaviour such as monotonicity or non-negativity.

Bayesian collocation tempering, uses a sequence of posterior densities with smooth approximations to the ODE solution. The level of the approximation is determined by the value of the smoothing parameter, which also determines the level of smoothness in the likelihood surface. In an algorithm similar to parallel tempering, parallel MCMC chains are run to sample from the sequence of posterior densities, while allowing ODE parameters to swap between chains. This method is introduced and tested against a variety of alternative Bayesian models, in terms of posterior variance and rate of convergence.

The performance of generalized profile estimation and Bayesian collocation tempering are tested and compared using simulated data sets from the FitzHugh-Nagumo ODE system and real data from nylon production dynamics.

ABRÉGÉ

L'utilisation répandue des modèles d'équations différentielles ordinaires (EDO) a depuis longtemps été sous-représentée dans la littérature statistique. Les méthodes les plus communes pour estimer les paramètres des modèles d'EDO sont les moindres carrés non-linéaires et une méthode basée sur les MCMC. Ces méthodes dépendent d'une vraisemblance basée sur la solution numérique de l'EDO. Le défi relevé par ces méthodes est que les espaces de paramètres sont difficiles à naviguer, aggravé par la grande variété de formes fonctionnelles qu'un modèle d'EDO peut produire avec des petits changements de valeurs des paramètres.

Ce travail décrit deux méthodes alternatives, l'estimation généralisée de profil (EGP) et la méthode de lissage bayésienne tempérée (LBT). Ces deux méthodes emploient une expansion de bases pour approximer la solution d'EDO dans la vraisemblance, où la forme de l'expansion est guidée par le modèle d'EDO. Cette approximation de l'EDO lisse la surface de vraisemblance, réduisant ainsi les restrictions de mouvement des paramètres.

L'EGP, maximise le profil de vraisemblance des paramètres d'EDO, tout en profilant les coefficients de l'expansion de bases. Le paramètre de lissage détermine l'équilibre entre l'interpolation des données et l'ajustement au modèle d'EDO. Celui-ci est donc utilisé afin de construire une cascade de paramètres, réduisant ainsi la dimensionnalité du problème d'estimation. L'estimation généralisée de profil est décrite sous des contraintes de lissage connues telles la monotonicité et la non-negativité.

La méthode de LBT utilise une suite de densités postérieures basée sur des approximations lisses à la solution d'EDO. Le niveau de l'approximation est déterminé par la valeur du paramètre de lissage qui contrôle le niveau de rugosité dans la surface de vraisemblance. Dans un algorithme semblable au tempérant parallèle, des chanes MCMC parallèles sont utilisées pour échantillonner de la suite de densités postérieures, tout en permettant aux paramètres d'EDO de permuter entre les chanes. Cette méthode est présentée et examinée contre une variété de modèles bayésiens alternatifs, en terme de variance postérieure et taux de convergence.

La performance de ces méthodes sont examinée et comparée en utilisant des données simulées d'un système d'ODE de FitzHugh-Nagumo et des données réelles de la dynamique en production de nylon.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction to Differential Equation Systems

Ordinary Differential Equation (ODE) models describe rates of change of system components or outputs $Dx(t) = dx(t)/dt$ as a function of the observed component behaviour $x(t)$, giving ODE models of the form $Dx(t) = f\{x(t)\}$. For example, consider a population of rabbits producing on average β offspring per rabbit per unit of time after accounting for gender. The rabbit population beginning at level x_0 rabbits increases to $x(t = 1) = x_0(1 + \beta)$ in a single time unit. With abundant resources and ignoring the occasional predator or other cause for mortality for the moment, gives the rabbit population at any time t,

$$x(t) = x_0(1 + \beta)^t. \tag{1.1}$$

Reparametrization by $\alpha = \log(1 + \beta)$ shows explicitly the unbounded exponential growth of the rabbit population

$$x_t = x_0 \exp(\alpha t). \tag{1.2}$$

The analytical derivative or instantaneous rate of change in population is

$$\frac{dx(t)}{dt} = \alpha x_0 \exp(\alpha t) = \alpha x(t), \tag{1.3}$$

giving a differential equation model for the rate of change in population as a function of the current population size.

Parameter α is often unknown and of interest. The first order linear differential equation model (1.3) has the analytic solution (1.2), and consequently parameter estimation could be performed with log-linear regression model $\log(x_t) = \log(x_0) + \alpha t$. Alternatively, if observations are obtained at evenly spaced time intervals, the model could be re-written as $x(t) = e^{\alpha} x(t - 1)$. With this recursion, α could be obtained from an auto-regressive model using time-series (Brockwell and Davis 1991) or state space (West and Harrison 1997) methods.

Another ODE model arising from physical principles is based on Newton's revelation: $Force = Mass \times Acceleration$. For example, a parachutist experiences the force of gravity by accelerating at constant rate a towards the Earth. The rate of change of the parachutist's position is the velocity and the rate of change in velocity is the acceleration. Consequently, the force of gravity acts on the second derivative of position: $D^2x(t) = a$. This acceleration is eventually balanced by the force of wind resistance, which is itself a function of relative opposing wind speed, and consequently the parachutist's velocity. Also aiding in a safe landing, the wind resistance acting on the parachutist changes drastically once the parachute opens. Labelling constant parameters describing wind resistance $\boldsymbol{\theta}$, and non-constant parameters which change with the status of the parachute $\phi(t)$, produces the model:

$$D^2x(t) = a - f_{wind}\left(Dx(t), \boldsymbol{\theta}, \phi(t)\right). \tag{1.4}$$

Model (1.4) may not have an analytic solution and consequently there may no longer be simple options for estimating parameters. Furthermore, we may only be able to measure position $x(t)$, but be interested in estimating parameters $\boldsymbol{\theta}$ and $\phi(t)$ which act on its second derivative. To simplify somewhat, this second order model with non-constant coefficients can be re-written as a set of linked first order constant coefficients by including a functional input describing the parachute status $u(t)$,

$$\begin{aligned} Dx(t) &= v(t) \\ D^2x(t) = Dv(t) &= a - f_{wind}\left(v(t), u(t), \boldsymbol{\theta}, \phi\right) \end{aligned} \tag{1.5}$$

In most cases, higher order ODE models can be re-expressed as a system of first order ODEs and often non-constant coefficients can be made constant by the inclusion of some additional information about the system inputs. In this particular case, the model simplification has the added advantage that the induced variable v is interpreted as velocity.

Differential equation models arise naturally from principles of conservation of mass, energy, charge or momentum. These systems model the transfer of a conserved currency from one form, or system component, into another. Conservation constraints produce nonlinearities in the

model describing the nature of the interaction between system components. For example, in the Canadian north, the Lynx is a predator of the Hare (a close relative of the rabbit) and their relationship between their population numbers can be described by

$$\begin{aligned} DHare &= \beta_1 \text{Hare} - \beta_2 \text{HareLynx} \\ DLynx &= -\beta_3 \text{Lynx} + \beta_4 \text{HareLynx}. \end{aligned} \tag{1.6}$$

The attraction of the ODE formulation of (1.6) is in the interpretability of the parameters. The net reproduction rate of Hares is β_1, Hare mortality rate per Lynx is β_2. The Lynx's net death rate is β_3, and the Lynx population increase per Hare eaten is β_4. The interaction term HareLynx reflects the need for these two species to connect, in order to influence each others rate of population change. This model suggests a constant amount of energy in the ecological system, and the Lynx and Hare exchange this energy through the process of hunting. Essentially, the Hare's energy is transferred to the Lynx during the hunt, but as the Hare population dwindles, the Lynx will no longer find adequate food and will starve. As the Lynx population in turn crashes, the Hares are able to reproduce under reduced threat of predation, in a sense reclaiming some of the systems energy from the dying Lynx. As the Hare population recovers, they become easier prey which in turn rebuilds the Lynx population. These population cycles are common in ecology and for example are seen in early Canadian Hare and Lynx fur trapper records kept by the Hudson's Bay company (Elton and Nocholson 1914).

While equations (1.6) have no analytic solution, if the system state is known exactly at some point in time, for example if the initial system states $Hare(t = 0)$ and $Lynx(t = 0)$ are the numbers of each animal introduced into a closed system, then a numerical approximation to the solution may be produced using Runge-Kutta or other numerical solvers.

Model building from scientific principles is further described in detail in the next two sections, as an introduction to the two multi-component nonlinear ODE systems that will be used throughout the remainder of this work. The first is based on the conservation of mass producing equations governing the reaction dynamics of nylon production driven at the level of the derivative by external inputs over multiple experimental runs. The second example describes linked

feedback loops inducing another form of nonlinearity into an ODE model through an example from neuro-physiology.

1.1 ODE model for Nylon Production Dynamics

In a heated chemical reactor containing amine (A) and carboxyl (C) chemical groups, A and C react to producing the polymer nylon (L) and water (W) (Zheng, McAuley, Marchildon, and Zhen Yao 2005). At the same time W reacts with L, decomposing it into its constituents A and C. These competing reactions, symbolically summarized by $A + C \rightleftharpoons L + W$, imply that the chemicals change form, but the conservation of mass implies that the total mass in the reactor remains constant. Knowing that the pair (A, C) must react in order to be expended suggests a negative feedback loop reducing the concentrations of A and C at a rate proportional to the interaction of their concentrations. By symmetry of the problem, the self-annihilation of (A, C) will produce the new pair (L, W) giving reaction rates which differ only in sign. Using reaction rate parameters $k_p > 0$ and $K_a > 0$, the dynamics,

$$\begin{aligned} DA = DC &= -k_p(CA - LW/K_a) \\ \text{and } DL = DW &= k_p(CA - LW/K_a), \end{aligned} \tag{1.7}$$

are somewhat reminiscent of the form of the predator prey equations in (1.6). The positive constraint on k_p and K_a imposes the conservation of mass principle used in the formulation of the ODE model.

The time derivative for A is of the form: $DA = -k_1 A + k_2$. As a simplification to explore the meaning behind the equation, consider k_1 and k_2 to be constant. This system has the solution:

$$A = \exp(-k_1 t) + k_2/k_1, \tag{1.8}$$

which decays exponentially towards the asymptote k_2/k_1. Therefore, this system describes a tendency towards equilibrium concentrations of the chemicals[1] . Removing the simplification

[1] By symmetry of the system of equations, any component of C, W or L can be put into the form of (1.8).

and returning to the formulation of (1.7) allows the final equilibrium asymptote to be a function of all outputs. Due to the interaction of all the chemical components, numerical methods are required to map out the trajectories of the components. The equilibrium concentrations are attained when the competing reactions balance, i.e. $CA = LW/K_a$. The rate of the exponential decay is controlled by k_p. Accurate estimates of the reaction rate and asymptote parameters enable chemical engineers to design more efficient nylon production reactors.

Due to the heat of the chemical reactor, when A and C react to produce L and W, W escapes as steam and its concentration therefore dwindles as the reaction proceeds. In an experiment to measure these reaction dynamics (Zheng et al. 2005), input steam was bubbled through molten nylon to maintain an approximately constant concentration of W in the system. The constant W forces A, C and L to move towards equilibrium concentrations coinciding with this input level of W. Within each of the $i = 1, \ldots, 6$ experimental runs, the pressure of input steam was held at a high level until time τ_{i1}, then reduced until time τ_{i2}, when the input pressure returned to its original level for the remainder of the experiment. A high pressure of steam entering the system implies a high concentration of W within the molten mixture, as the entering steam quickly equilibrates with the exiting steam.

The full set of experimental conditions are given in table 1–1. Each experiment was performed at a constant temperature $T_i \in \{536, 544, 554, 557\}$ degrees Kelvin. Using known constants P_c and T_c, the critical temperature and pressure of water, along with the input water pressure P_w, the equilibrium concentration of water in the molten nylon mixture, W_{eq}, is determined through the equations from (Schaffer, McAuley, Cunningham, and Marchildon 2003),

$$W_{eq} = 5.55 \times 10^4 \frac{P_w}{P_w^{sat}} \exp(-9.624 + 3613/T), \tag{1.9}$$

and

$$\ln(P_w^{sat}/P_c) = \frac{T}{T_c}\Big[-7.77244(1-\frac{T}{T_c}) + 1.45684(1-\frac{T}{T_c})^{1.5} - 2.71492(1-\frac{T}{T_c})^3 - 1.41336(1-\frac{T}{T_c})^6\Big]. \tag{1.10}$$

Table 1–1: Nylon Experimental conditions. Temperature T is given in degrees Kelvin. P_w and W_{eq} are given at time zero (equal to the value after time τ_2) and after the first step change at time τ_1 hours. The number of observations are given for A and C. The final column of the table shows the concentration difference between A and C averaged over times when both are observed.

T	$P_w(0)$	$W_{eq}(0)$	τ_1	$P_w(\tau_1)$	$W_{eq}(\tau_1)$	τ_2	# obs A	# obs C	$\overline{C-A}$
536	760	64.3	4.1	181	15.3	8.0	22	22	61.0
544	760	51.3	3.9	58	3.9	7.6	22	22	68.6
554	760	39.0	3.1	205	10.5	6.3	23	23	75.3
557	760	36.0	0.6	152	7.21	3.8	15	12	110.1
557	760	36.0	1.0	152	7.21	5.0	15	13	198.8
557	760	36.0	2.1	152	7.21	5.3	23	12	6.08

Since the input steam forces W towards its equilibrium concentration W_{eq}, the input W_{eq} is including as a forcing function on the right hand side of DW in (1.7) producing the dynamics for this experiment;

$$\begin{aligned} -DL = DA = DC &= -k_p(CA - LW/K_a) \\ \text{and } DW &= k_p(CA - LW/K_a) - 24.3(W - W_{eq}). \end{aligned} \tag{1.11}$$

Using the reference temperature $T_0 = 549.15$, chosen in the middle of the experimentally manipulated temperatures, the reaction rates k_p and K_a are also allowed to change with T and W_{eq} through the relationships with unknown parameters $\boldsymbol{\theta} = [k_{p0}, \gamma, K_{a0}, \Delta H]$,

$$k_p = \frac{k_{p0}}{1000}, \tag{1.12}$$

$$K_a = \left\{1 + W_{eq}\frac{\gamma}{1000}\right\} K_T K_{a0} \ell\left(\frac{\Delta H}{8.314}\right), \tag{1.13}$$

$$\ell(m) = \exp\left(-m10^3\left\{\frac{1}{T} - \frac{1}{T_0}\right\}\right), \tag{1.14}$$

and

$$K_T = 20.97 \exp\left(-9.624 + \frac{3613}{T}\right). \tag{1.15}$$

These equations include scaling factors making all initial parameter estimates used in Zheng et al. (2005) in the range [17.7, 78.1] in absolute value, to ease optimization. Equations (1.12)

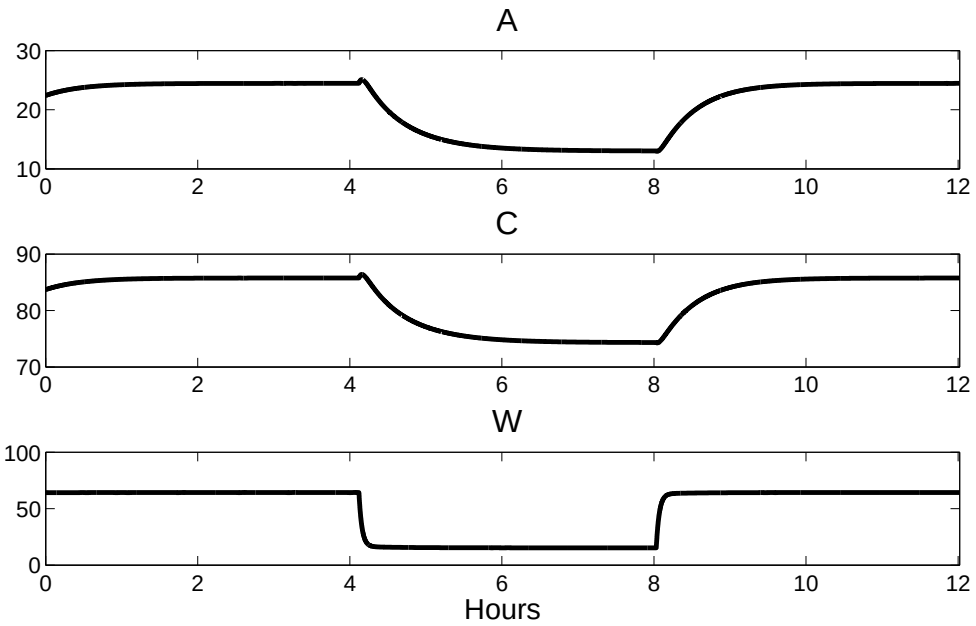

Figure 1–1: A numeric solution to the nylon equations, using values of $T = 554$ Kelvin and $W_{eq} \in \{64.28, 15.31\}$ from one of the experimental runs. The numerical solution is based on the parameters $[k_p, \gamma, K_{a0}, \Delta H] = [20.59, 26.86, 50.22, -36.46]$ and the initial system states $[A(t = 0), C(t = 0), W(t = 0)] = [22.40, 83.70, 64.28]$. Vertical axes are in concentration units and horizontal axes are in hours.

and (1.13) are a simplification of the original 6 parameter model used in Zheng et al. (2005). A discussion of the model reduction process is left for chapter 3.

Figure 1–1 shows the numerical solution to the nylon ODE equations (1.11), following the conditions of one of the experimental runs. Temperature was held constant at $T = 554$ and $W_{eq} \in \{64.28, 15.31\}$ was initially held high until $\tau_1 = 4.12$ then reduced to the lower level and finally returning to its original high level after $\tau_2 = 8.03$. The numerical solution is based on the parameters $[k_p, \gamma, K_{a0}, \Delta H] = [20.59, 26.86, 50.22, -36.46]$ and the initial system states $[A(t = 0), C(t = 0), W(t = 0)] = [22.40, 83.70, 64.28]$. Between times of step changes in input, chemical reactants A, C and W exponentially decay towards an equilibrium level. These equilibrium levels undergo step changes following the step jumps in input W_{eq}.

Figure 1–2 zooms in on the hour of experimental time surrounding the τ_1, the step function drop in W_{eq}. The dependence of K_a on W_{eq} in (1.13) through $\gamma \neq 0$, produces a small amplitude short jump in the levels of A and C immediately after τ_1. For fixed temperature, in (1.13) K_a is a

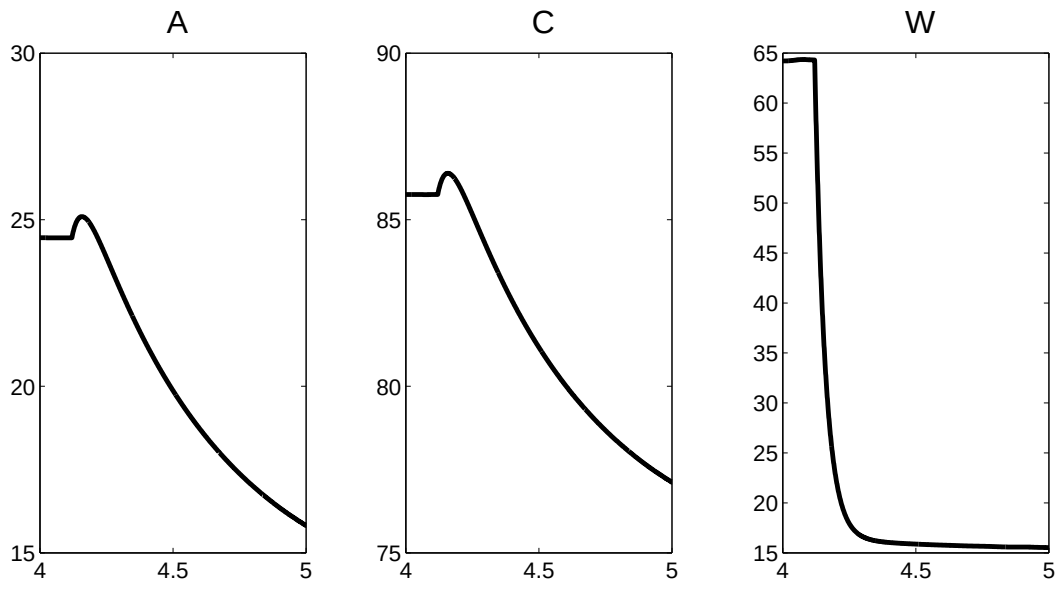

Figure 1–2: An up close view of the hour of experimental time surrounding the step function change in W_{eq} from 64.28 to 15.31. The step change causes a small bump in A and C induced by the dependency in K_{a0} on W_{eq} from the numeric solution in figure 1–1. Vertical axes are in concentration units and horizontal axes are in hours.

linear function of W_{eq} of the form $K_a = \beta_0 + \beta_1 W_{eq}$, $\beta_0, \beta_1 > 0$. Consequently, a step drop in W_{eq} causes a step drop in K_a. The inversion of K_a in (1.11) translates this step drop into the step increase in the rate of production of A and C giving the bump seen in the zoomed figure. Within approximately 15 minutes, this jump in reaction rate is dominated by the strength of the force pulling W towards W_{eq}, allowing excess W to escape as steam. In other words, according to the model, when W_{eq} drops, the system attempts to rid itself of W through all available processes. Most of this excess W escapes as steam but some of it reacts with L to produce this bump in A and C.

Figure 1–3 shows the data for each of the experimental runs. Unfortunately there are no observations close enough to the spike in A and C in order to be able to determine its plausibility. Instead, when $\gamma \neq 0$, the rate of exponential decay and asymptotic equilibrium levels are fine tuned by changes in W_{eq}. The plot shows observed components A and C as well as the input levels of W_{eq}. Vertical lines correspond to times of step changes in input τ_{i1} and τ_{i2}. These observations and the differential equation model produce several potential challenges:

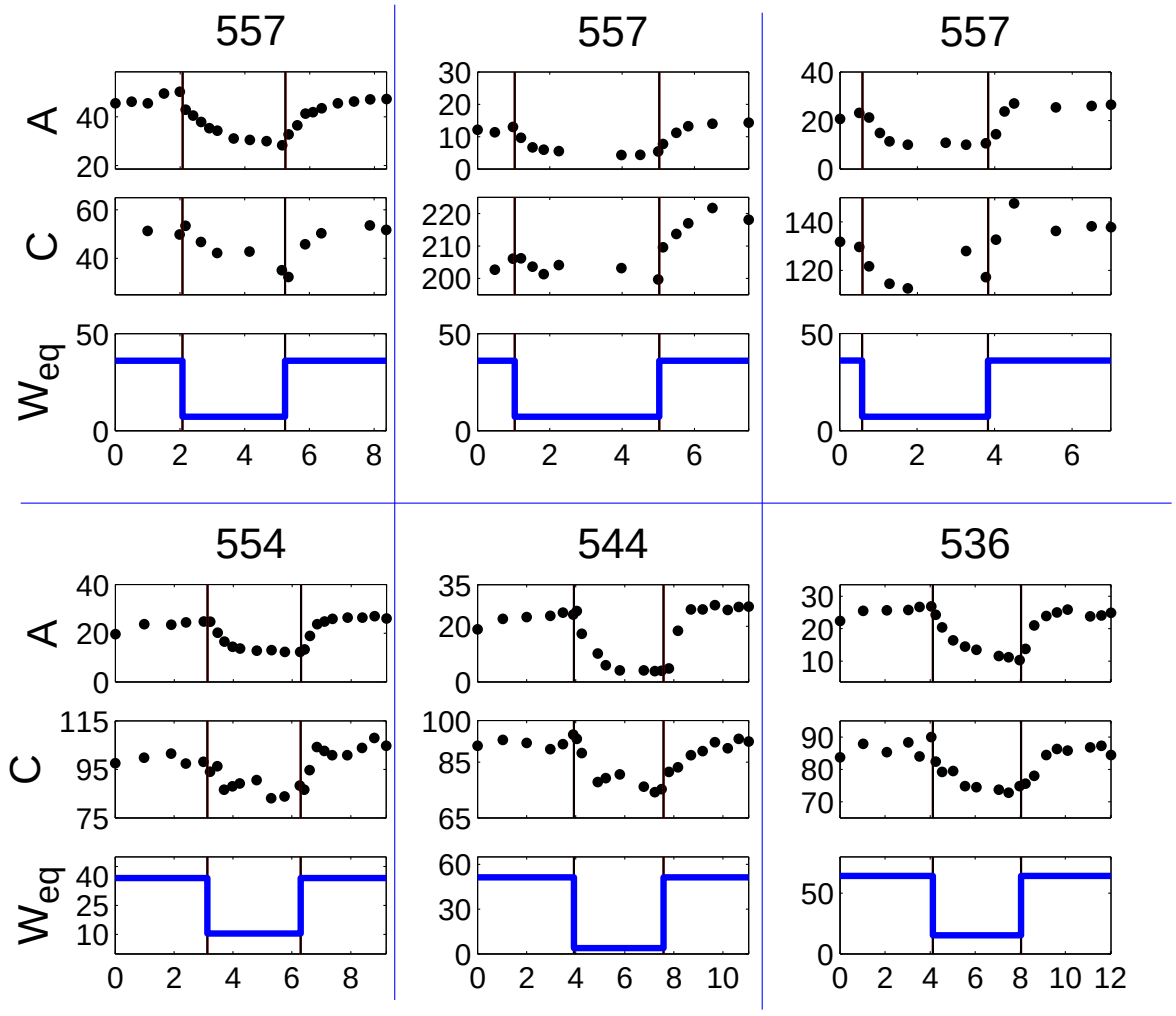

Figure 1–3: The nylon observations for A and C along with the input W_{eq} grouped together within each of the 6 experimental runs. Constant temperatures T of the experimental runs are given above component A in degrees Kelvin. Vertical axes are in concentration units and horizontal axes are in hours. Vertical lines represent times of changes in input W_{eq}.

1. This system describes chemical concentrations which are by definition constrained to take on non-negative values.
2. There are 6 experimental runs and information must be pooled across runs.
3. The equations in (1.11) 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 the first derivative of the underlying process at τ_{i1} and τ_{i2}.
4. Since $DA = DC$ the data in figure 1–3 should only differ by a constant vertical shift. Instead, this figure shows that the variability in the measurement of C is larger than the variability of A. System outputs are measured with different precision.
5. 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 additional state variable must be estimated.

1.2 FitzHugh-Nagumo Non-Linear ODE model

The FitzHugh-Nagumo system of nonlinear differential equations is used in physiology as an approximate model for the voltage $V(t)$ crossing the cell membrane of a giant squid axon. The recovery component $R(t)$ describing outward currents, interacts with V, using the model, defined by parameters $\boldsymbol{\theta} = [\alpha, \beta, \gamma]$,

$$\begin{aligned} DV(t) &= \gamma(V(t) - \tfrac{V(t)^3}{3} + R(t)) \\ \text{and } DR(t) &= -\tfrac{1}{\gamma}(V(t) - \alpha + \beta R(t)). \end{aligned} \tag{1.16}$$

An introduction to the motivation behind these equations and other ODEs for neurophysiology can be found in (Wilson 1999).

The model $DV \propto V - V^3/3$, when V is small, describes a positive feedback loop in which V exhibits exponential unbounded growth. The term $-V^3$ induces a negative feedback loop which takes effect only when the voltage becomes too large allowing voltage to spill across the cell membrane. That is, when $\mid V \mid \geq \sqrt{3}$ the sign changes on DV, returning V back towards zero producing oscillations. The nonlinearity in this equation allows these two feedback loops

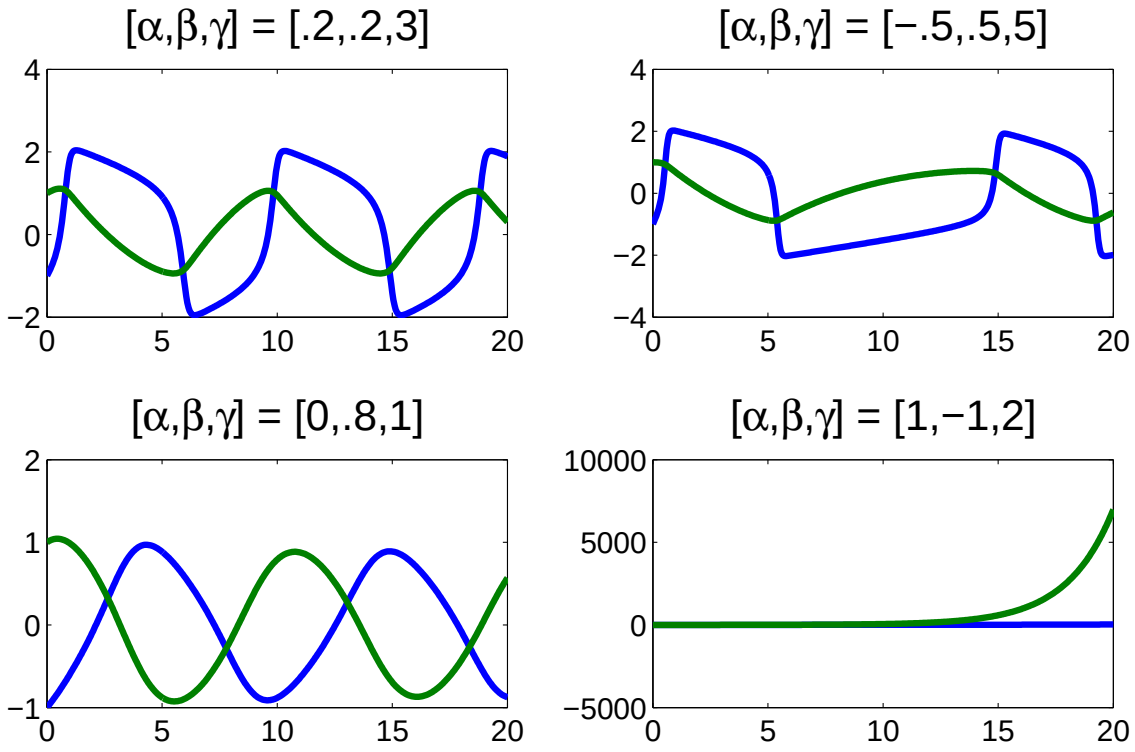

Figure 1–4: Numerical solutions to the FitzHugh-Nagumo equations in (1.16), using initial system states $[V_0, R_0] = [-1, 1]$ and several sets of parameters. Outputs V and R are shown in blue and green respectively as they progress over time.

to compete with the dominant loop being decided by the value of V. Including R as a forcing function in DV adjusts the speed of the transition from increasing to decreasing V, and adjusts the value of V where the sign of DV changes.

The equation $DR \propto -\beta R + \alpha$ has the same form as the nylon equations (1.8), and describes exponential decay towards the asymptote α/β. Including V as a forcing term in DR is equivalent to varying the horizontal asymptote in response to the value of V.

Using the initial system state $\mathbf{X}_0 = [V_0, R_0] = [-1, 1]$, the numerical solutions $S_V(\boldsymbol{\theta}, \mathbf{X}_0, t)$ and $S_R(\boldsymbol{\theta}, \mathbf{X}_0, t)$ to (1.16) are plotted for four sets of parameter values in figure 1–4. The scaled axes represent time and voltage units in the horizontal and vertical dimensions respectively. The shape and period of the cycles vary considerably with the parameter values and the cycles can altogether disappear. The top left panel of figure 1–4 with parameters $[.2, .2, 3]$ highlights the

exponential growth of V, for example occurring over the interval $(6.5, 10)$. Due to the positive value of R, the sign of DV changes at $V(10) = 2$ instead of when $V(t) = \sqrt{3}$, as would be the case without the influence of R. At time $t = 10$ the change in sign of DV produces a sharp bend in the path of V. Compared with this panel, the period roughly doubles by changing to parameters $[-.5, .5, 5]$ (top right panel). Including a negative value for α makes it more difficult for the voltage to build up from negative values, producing the long slow recovery shown in this panel. This value of α also reduces the duration of the positive values of V.

In the bottom left panel, parameters $[0, .8, 1]$ soften the abrupt sharp change of sign in DV to a smaller amplitude, more sinusoidal trajectory. Changing β to a larger negative value with parameters $[1, -1, 2]$ alters the exponential decay of DR into an unregulated positive feedback loop giving unbounded exponential growth for R. The influence of R on DV in (1.16) forces V to also grow exponentially, but because R is the driving force, component R has a head start. Consequently, when both components are plotted on the same vertical scale, as in the bottom right panel, the exponential growth of V is present but not evident. Stable cyclical behaviour occurs when parameters are within the approximate region $-.8 < \alpha, \beta < .8$ and $0 < \gamma < 8$ for a wide range of initial system states.

The variety of behaviour described by a single differential equation model produces flexibility beyond that available from a conventional linear or nonlinear models, and consequently may produce considerable difficulty in parameter optimization.

As a simulation study throughout this work, 50 simulated data sets with observations $V(t)$ and $R(t)$ at times $t \in \{0, .05, \ldots, 20\}$ were produced from the numerical solution to (1.16) using $[V_0, R_0] = [-1, 1]$ and $[\alpha, \beta, \gamma] = [.2, .2, 3]$. Zero mean Gaussian errors were added to observations with $\sigma_V^2 = .5^2$ and $\sigma_R^2 = .4^2$. A representative of the 50 simulated data sets appears in figure 1–5. These same 50 data sets are used to compare a variety of parameter estimation methods in chapters 2, 3 and 4.

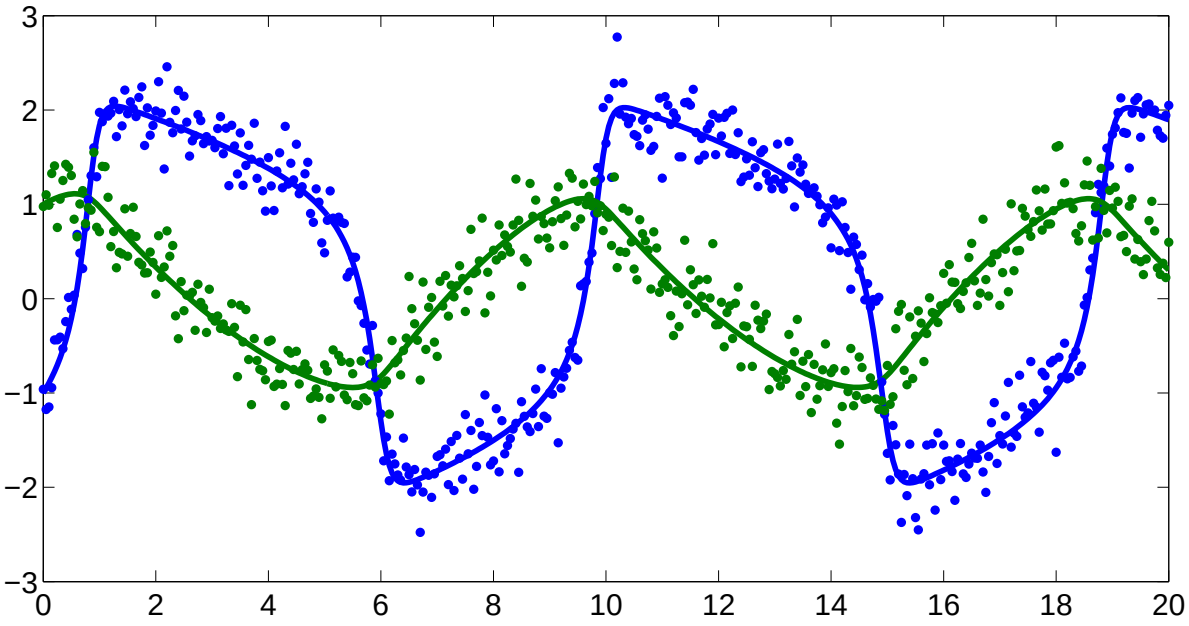

Figure 1–5: The numerical solution to the FitzHugh-Nagumo equations with $[V_0, R_0] = [-1, 1]$ and $[\alpha, \beta, \gamma] = [.2, .2, 3]$ along with simulated observations taken over 20 time units. Outputs V and R are shown in blue and green respectively.

1.3 Overview of this work

An overview of some of the current methodological work is introduced in chapter 2, which demonstrates the performance and limitations of the two most widespread estimation methods for ODE models: nonlinear regression and Markov Chain Monte Carlo. Both of these methods use the numerical solution to the ODE model in the likelihood and subsequent parameter estimation process. This chapter ends by discussing advances to these methods as a mis-en-scene for recent advances.

The next two chapters describe collocation methods, meaning that they use a basis expansion as an approximation to the numerical solution to the ODE model. Centering the likelihood on this basis expansion essentially smooths out the likelihood, simplifying the estimation process. Chapter 3 describes a method based on a generalized version of the maximum profile likelihood estimator allowing for a constrained basis expansion. Chapter 4 develops a Bayesian method using parallel MCMC chains based on collocation approximations to the ODE solution.

The results of these two methods are compared in chapter 5, which highlights the subtle differences in their description of the underlying process. This chapter finishes with discussion about current and future research directions.

CHAPTER 2
Parameter Estimation for ODE Models

The most common methods for estimating the vector of parameters $\boldsymbol{\theta}$ from ODE models of the form

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

with system inputs or forcing functions $\mathbf{u}(t)$, are nonlinear least squares (NLS) and Markov-Chain Monte Carlo (MCMC) methods. These two methods depend on the analytic solution to the ODE, or when none is available, they rely on the numeric solution $S(\boldsymbol{\theta}, \mathbf{X}_0, \mathbf{u}(t), t_i)$ computed from the initial system states $\mathbf{X}_0$ with a Runge-Kutta or other numeric ODE solver. With NLS, $S(\boldsymbol{\theta}, \mathbf{X}_0, \mathbf{u}(t), t_i)$ and gradient with respect to $\boldsymbol{\theta}$ and $\mathbf{X}_0$ are used to guide movement across the parameter space using Gauss-Newton iterative updates as described in section 2.1. With MCMC, the solution is used in a stochastic optimization and density estimation routine described in section 2.2.

The performance of these two methods is demonstrated on the 50 simulated data sets from the FitzHugh-Nagumo equations of section 1.2 and the nylon data from section 1.1 to highlight the strengths and limitations of these methods. Recent extensions to these methods designed to fix some of their well known limitations are described in section 2.3. This chapter motivates the methodological advances described in later chapters.

2.1 The Nonlinear Least Squares Method for Estimating ODE Parameters

Maximum likelihood estimation for ODE models select $\hat{\boldsymbol{\theta}}$ by minimizing the negative log likelihood. When a Gaussian likelihood is used this amounts to minimizing the squared error loss between the (numerical) solution to the ODE and the data:

$$\hat{\boldsymbol{\theta}} = \min_{\boldsymbol{\theta}} SSE = \min_{\boldsymbol{\theta}} \sum_{i}^{N} \left(y(t_i) - S\{\boldsymbol{\theta}, \mathbf{X}_0, \mathbf{u}(t), t_i\}\right)^2.$$

Using this distributional assumption for the measurement error is referred to as Nonlinear Least Squares (NLS), and is described in detail in (Seber and Wild 1989) and (Bates and Watts 1988), but an overview will be given here in order to highlight the nature of the methodological improvements in later chapters.

Since the numerical solution depends on $\mathbf{X}_0 = \mathbf{X}(t = t_0)$, the initial system state of component X, these additional parameters must also be estimated. Using Gauss-Newton iterations, parameter estimates move through the parameter space by incrementally stepping along the steepest decent in the loss function as produced by a Taylor approximation to the loss surface using the gradients:

$$\begin{array}{rcl} \frac{d}{d\boldsymbol{\theta}} SSE & = & -2\sum_i^N \left(y(t_i) - S\{\boldsymbol{\theta}, \mathbf{X}_0, \mathbf{u}(t), t_i\}\right) \left(\frac{d}{d\boldsymbol{\theta}} S\{\boldsymbol{\theta}, \mathbf{X}_0, \mathbf{u}(t), t_i\}\right) \\ \text{and } \frac{d}{d\mathbf{X}_0} SSE & = & -2\sum_i^N \left(y(t_i) - S\{\boldsymbol{\theta}, \mathbf{X}_0, \mathbf{u}(t), t_i\}\right) \left(\frac{d}{d\mathbf{X}_0} S\{\boldsymbol{\theta}, \mathbf{X}_0, \mathbf{u}(t), t_i\}\right). \end{array} \tag{2.1}$$

Numerical estimates of

$$\frac{d}{d\mathbf{X}_0} S(\boldsymbol{\theta}, \mathbf{X}_0, \mathbf{u}(t), t_i), \text{ and } \frac{d}{d\boldsymbol{\theta}} S(\boldsymbol{\theta}, \mathbf{X}_0, \mathbf{u}(t), t_i)$$

may be produced by selecting a small δ and approximating the definition of the integral, for example,

$$\frac{d}{d\boldsymbol{\theta}} S(\boldsymbol{\theta}, \mathbf{X}_0, \mathbf{u}(t), t_i) \approx \frac{S(\boldsymbol{\theta} + \delta, \mathbf{X}_0, \mathbf{u}(t), t_i) - S(\boldsymbol{\theta}, \mathbf{X}_0, \mathbf{u}(t), t_i)}{\delta}.$$

These numerical gradients may perform poorly because of the strong nonlinearities. Instead, the much more stable analytic gradients of $S(\boldsymbol{\theta}, \mathbf{X}_0, \mathbf{u}(t), t_i)$ with respect to parameters $\boldsymbol{\theta}$ and initial conditions $\mathbf{X}_0$ are obtained by numerically solving the sensitivity equations:

$$\begin{array}{rcll} \frac{df(\mathbf{x},\boldsymbol{\theta},\mathbf{u}(t),t)}{d\boldsymbol{\theta}} & = & \frac{\partial f(\mathbf{x},\boldsymbol{\theta},\mathbf{u}(t),t)}{\partial\boldsymbol{\theta}} + \frac{\partial f(\mathbf{x},\boldsymbol{\theta},\mathbf{u}(t),t)}{\partial X} \frac{dX}{d\boldsymbol{\theta}}, & \text{with } \left.\frac{dX(t)}{d\boldsymbol{\theta}}\right|_{t=t_0} = 0 \\ \frac{df(\mathbf{x},\boldsymbol{\theta},\mathbf{u}(t),t)}{d\mathbf{X}_0} & = & \frac{\partial f(\mathbf{x},\boldsymbol{\theta},\mathbf{u}(t),t)}{\partial X} \frac{dX}{d\mathbf{X}_0}, & \text{with } \left.\frac{dX(t)}{d\mathbf{X}_0}\right|_{t=t_0} = 1 \end{array} \tag{2.2}$$

Due to the nonlinear nature of the problem and the diverse behaviours that can be modelled by a single ODE model, the Gauss-Newton increment may overstep the region where the Taylor approximation is reasonable. This may suggest a set of parameters which increase the sum of

squared errors (SSE). When this arises the step size is repeatedly discounted (often halved) and the SSE re-assessed until a step size which decreases the SSE is found. Where the loss surface is highly nonlinear, this can reduce the incremental updating of parameters to a crawl, potentially giving an assessment of convergence based on the minute step size which finally reduces the SSE. Therefore, convergence may actually be indicative of a strong departure from the linear approximation or a local minimum as opposed to a global optimum. Despite warnings about difficulties in navigating the likelihood surface and consequent potentially misleading results (Esposito and Floudas 2000), nonlinear least squares is one of the most used techniques for parameter estimation.

2.1.1 FitzHugh-Nagumo Simulated Data Example Using NLS

In a simulation study, NLS was performed on the 50 FitzHugh-Nagumo data sets, having 401 evenly spaced observations for each of the outputs V and R, as described in section 1.2, following the model with true parameters $\boldsymbol{\theta} = [\alpha, \beta, \gamma] = [.2, .2, 3]$,

$$\begin{aligned} DV(t) &= \gamma(V(t) - \tfrac{V(t)^3}{3} + R(t)) \\ DR(t) &= -\tfrac{1}{\gamma}(V(t) - \alpha + \beta R(t)), \end{aligned} \tag{2.3}$$

and beginning with the initial systems states $\mathbf{X}_0 = [V(t = t_0), R(t = t_0)] = [-1, 1]$. The iterative estimation procedure was initialized using initial system states, and parameters estimated with draws from the prior densities of the Bayesian model described in 2.2. These same initial estimates are used every time these data sets are analyzed in this work.

The NLS algorithm was stopped after 400,000 Gauss-Newton iterations if convergence had not already been attained, as assessed by a relative drop in the SSE of less than 10^{-8} from an additional Gauss-Newton step. In this simulation study, one of the optimization became trapped in a local minima, meeting the convergence criteria while remaining far from the true parameter values. Furthermore, the parameter optimization of another 6 of the 50 simulated data sets did not achieve convergence in the maximum number of allotted iterations. In these 6 cases, the linear approximation to the likelihood surface used by the Gauss-Newton optimization was only

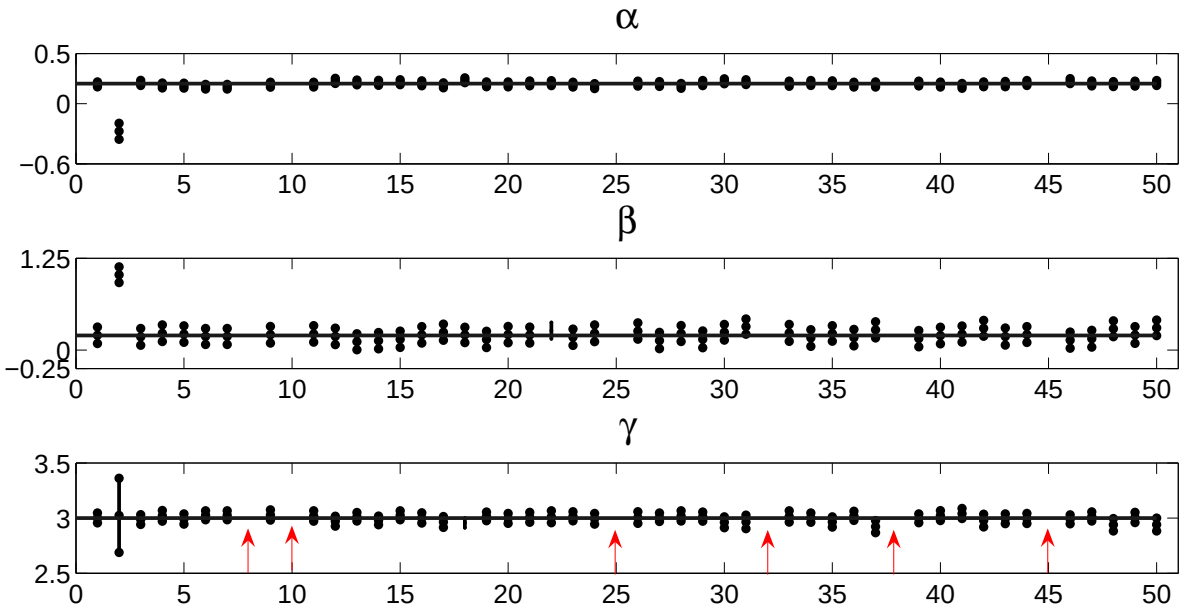

Figure 2–1: 95% confidence intervals for 50 simulated FitzHugh-Nagumo data sets using Nonlinear Least Squares. Vertical axes are the parameter vales while horizontal axes denote the number of the simulated data set. Intervals are shown for optimizations which converged in less than 400,000 Gauss-Newton iterations. Data sets where convergence did not occur are marked with arrows in the plot for γ. The horizontal lines denote the true parameter values.

reasonable over a very small region in the parameter space being explored. Consequently, the parameters moved by only small steps at each iteration, stunting the optimization progress.

Figure 2–1 shows the NLS 95% confidence intervals from the parameter estimates of $\boldsymbol{\theta}$. Using the observed residual variance $\hat{\sigma}^2$, the interval estimates were computed using the standard errors:

$$SE(\hat{\boldsymbol{\theta}}) = \left(\left. \frac{dS(\boldsymbol{\theta}, \hat{\mathbf{X}}_0, \mathbf{u}(t), t_i)}{d\boldsymbol{\theta}} \right|'_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \left. \frac{dS(\boldsymbol{\theta}, \hat{\mathbf{X}}_0, \mathbf{u}(t), t_i)}{d\boldsymbol{\theta}} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \right)^{-1} \hat{\sigma^2}.$$

In the FitzHugh-Nagumo system, the initial system states $\mathbf{X}_0$ determine the phase of the oscillations in $S(\boldsymbol{\theta}^{(true)}, \mathbf{X}_0, \mathbf{u}(t), t_i)$, and when those oscillations begin. Any set of finite values of $\mathbf{X}_0$ will eventually produce limit cycles when $\boldsymbol{\theta} = \boldsymbol{\theta}^{(true)}$ due to the feedback regulation in (2.3). However, if $\mathbf{X}_0$ is far from $\mathbf{X}_0^{(true)}$, trajectory $S(\boldsymbol{\theta}^{(true)}, \mathbf{X}_0, \mathbf{u}(t), t_i)$ will not begin oscillate within the range of observation times. This may produce gradients in (2.1) which make the Gauss-Newton iterations move $\boldsymbol{\theta}$ away from $\boldsymbol{\theta}^{(true)}$ as the optimization navigates the steepest

descent in the loss surface. The nonlinear response to changes initial system state increases the number of required iterations in NLS.

2.1.2 Nonlinear Least Squares and the Nylon Data

Using the nylon data system shown in figure 1–3, NLS was performed to estimate the $\boldsymbol{\theta} = [k_{p0}, \gamma, K_{a0}, \Delta H]$ from the nylon system equations

$$\begin{aligned} -DL = DA = DC &= -\tfrac{k_{p0}}{1000}(CA - LW/K_a), \\ DW &= \tfrac{k_{p0}}{1000}(CA - LW/K_a) - 24.3(W - W_{eq}), \\ K_a &= \left\{1 + W_{eq}\tfrac{\gamma}{1000}\right\} K_T K_{a0} \ell\left(\tfrac{\Delta H}{8.314}\right), \\ \ell(m) &= \exp\left(-m10^3\left\{\tfrac{1}{T} - \tfrac{1}{T_0}\right\}\right), \\ \text{and } K_T &= 20.97\exp\left(-9.624 + \tfrac{3613}{T}\right). \end{aligned} \tag{2.4}$$

The parameter estimates used to initialize the NLS algorithm were $\boldsymbol{\theta} = [k_{p0}, \gamma, K_{a0}, \Delta H] = [17.7, 17.0, 15, -78.1]$, approximations to the parameter estimates originally obtained from linear regression estimation performed on subsets and transformations of the data in (Zheng, McAuley, Marchildon, and Zhen Yao 2005), adjusted for the reparameterization of this model. Initial estimates for $\mathbf{X}_0 = \{\mathbf{X}_{i0} = [A_{i0}, C_{i0}, W_{i0}]; i = 1, \ldots, 6\}$ were the values assumed to be fixed and true in Zheng et al. (2005). Under this assumption, A_{i0} was observed without error, and $C_{i0} = A_{i0} + \hat{\mu}_{C_i - A_i}$, where $\hat{\mu}_{C_i - A_i}$ is the average difference between observations of C_i and A_i. The value of $W_{i0} = (W_{eq})_{i0}$ was chosen based on the assumption that the system was at equilibrium at the first observation time. Weighted NLS was used to account for the relative reduced accuracy in the ability to measure C compared to A. Combining experimental conditions into $\mathbf{u}_i(t)$, with weights w_A and w_C playing the role of relative scaling factors, the likelihood for the model is:

$$[A_i(t), C_i(t)] \,|\, \boldsymbol{\theta}, \mathbf{X}_{i0} \sim N\left(S_i\{\boldsymbol{\theta}, \mathbf{X}_{i0}, \mathbf{u}_i(t), t\}, \sigma^2[1/w_A, 1/w_C]\right).$$

In a Gaussian likelihood, the optimal weights are proportional to the inverse measurement error variances. In Zheng et al. (2005), the weights $w_A = 1./.6^2$ and $w_C = 1./2.4^2$ were obtained from replicate measurements of concentration in an earlier study.

With the nylon data set, output W is unobserved and $\hat{W}_{i0}$ must be estimated in order to produce $S_i\{\hat{\boldsymbol{\theta}}, \mathbf{X}_{i0}, \mathbf{u}_i(t), t\}$. The fit to the data, $S_i\{\hat{\boldsymbol{\theta}}, \mathbf{X}_{i0}, \mathbf{u}_i(t), t\}$, $i = 1, \ldots, 6$ is shown in figure 2–2 using the point estimates given in table 2–1. As shown in the figure, using nonlinear least squares, $\hat{W}_{i0}$ is negative in four of the six experimental runs. Negative values of W eliminate the conservation of mass in (2.4) and have no interpretation.

The estimates of $\hat{W}_{i0}$ are far from equilibrium in all of the experimental runs, causing drastic changes in A, C and W in the first few minutes of the experiment. While these drastic changes are unlikely to have truly occurred, even in the non-negative cases, thanks to the estimates of $\hat{W}_{i0}$, the fit to the data is excellent or near perfect in some of the experimental runs, especially for A. The impact of the estimates of $\hat{W}_{i0}$ and the breach of conservation of mass is quickly overshadowed by the strength of the force pulling W towards W_{eq} in (2.4). Consequently, as shown in figure 2–2, in runs with negative estimates of $\hat{W}_{i0}$, the levels of A and C are declining for only the first few minutes until W is pulled into positive values and the conservation of mass dynamics return. The erratic behaviour described in the early part of figure 2–2 is not a reliable estimate of the fit to the data, because the chemical reactions were allowed to run for some time previous to the first observations in order to bring the components closer to equilibrium. Consequently, the chemical components would have to have been initialized much further from equilibrium for the fit to the data to be accurate even after the reactor was allowed to equilibrate somewhat before the first observation.

Inducing the non-negativity constraint $W_{i0} = \exp(\omega_i)$ and subsequent estimation of ω_i forces the initial system states to remain positive and interpretable. Exponentiation of the parameters sharpens the nonlinear loss surface peaks because smaller changes in parameters produce relatively larger changes to the data fit. For the nylon data model, the NLS iterations under the positive parameter constraint failed to converge because the ODE solver could no longer produce an accurate numerical solution to the ODE. This occurred when the Gauss-Newton iterations proposed a particularly poor set of parameter values. Attempts to resolve this problem by re-scaling $\boldsymbol{\theta}$ and $\mathbf{X}_0$ or using an alternative set of initial parameter estimates were not helpful.

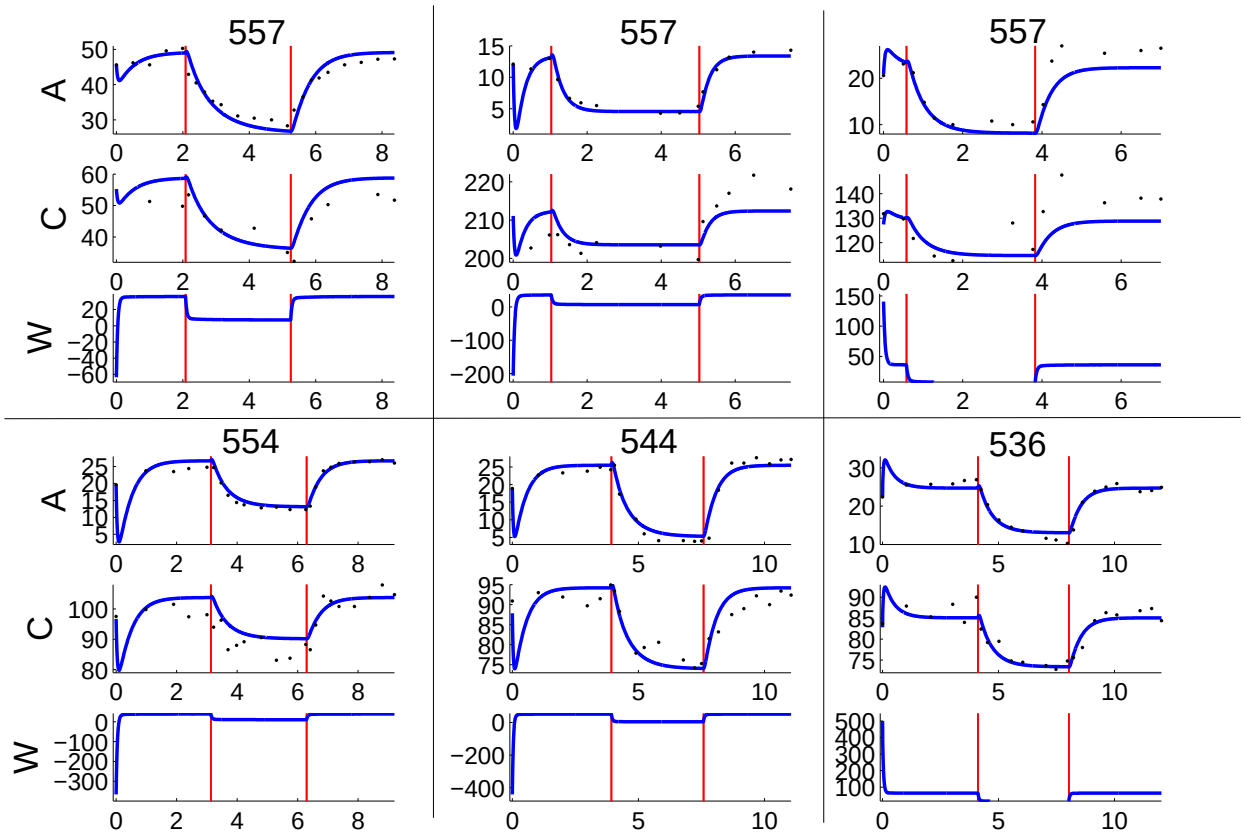

Figure 2–2: The fit to the nylon data using the weighted nonlinear least squares parameter estimates. Observations are shown for A and C. System output W is unobserved but estimated. Note that $\hat{W}_0$ is negative in 4 of the 6 experimental runs. Components are grouped within each experimental run. Constant temperatures are listed in degrees Kelvin component A within each experimental run. Vertical lines show the times of step changes in input W_{eq}. Vertical axes are in concentration units and horizontal axes represent time in hours.

Parameter	k_{p0}	γ	K_{a0}	ΔH
Point	20.47	25.82	50.98	-37.83
SE	1.30	2.90	2.79	4.17

Table 2–1: The top table shows point estimates for the 4 parameter nylon model and the estimated standard errors.

Fixing the initial system states at the initial parameter estimates produces a much improved fit to the data, and parameter estimates nearly identical to the those using NLS in Zheng et al. (2005) subject to the rescaling of the models. However, the assumption of equilibria at the outset of the experiment implies that A and C should not be increasing as suggested by the observations, but rather these observed outputs should be at a steady state.

With the nylon data, the parameter estimation relies on a choice between an uninterpretable fit to the data, strong assumptions about initial system states or a broken methodology. In the FitzHugh-Nagumo simulation study the results were further troubled by estimates that depend on the initial parameter estimates and slow or failed convergence. These problems leave a lot of room for improvement in likelihood based approaches to parameter estimation in differential equation models.

2.2 Bayesian Parameter Density Estimation from ODE models

As with NLS, using ODE parameters $\boldsymbol{\theta}$ and initial conditions $\mathbf{X}_0$, the Bayesian parameter estimation model uses a likelihood centered on the solution to the ODE model, $S(\boldsymbol{\theta}, \mathbf{X}_0, t_i)$. For example, using the FitzHugh-Nagumo equations from (1.16), with $\boldsymbol{\theta} = [\alpha, \beta, \gamma]$ and $\mathbf{X}_0 = [V(t = t_0), R(t = t_0)]$, assuming a Gaussian measurement error structure, the likelihood is

$$P\left([V(t), R(t)] \mid \boldsymbol{\theta}, \mathbf{X}_0, \sigma_V^2, \sigma_R^2\right) = N\left(S\{\boldsymbol{\theta}, \mathbf{X}_0, t\}, [\sigma_V^2, \sigma_R^2]\right). \tag{2.5}$$

When the equations are linear in system components or an analytical solution exists, the problem simplifies to a Bayesian nonlinear regression model as described in (Bates and Watts 1988) or (Seber and Wild 1989). When there is no closed form ODE solution, numerical methods and Monte Carlo sampling must be used to obtain a posterior distribution. Using MCMC methods, this implies that for every set of proposed parameters, the equations must be numerically solved

to compute the likelihood and make a decision about the proposed values (see (Gelman, Bois, and Jiang 1996) and (Gelman, Carlin, Stern, and Rubin 2004) or (Huang and Wu 2006) in the context of mixed effects models).

For the FitzHugh-Nagumo system in (1.16) with observations $V(t)$ and $R(t)$ taken at times $t \in \{t_1, \ldots, t_n\}$ the oscillatory behaviour occurs when parameters $\boldsymbol{\theta} = [\alpha, \beta, \gamma]$ are in the approximate region $.8 < \alpha, \beta < 8$ and $0 < \gamma < 8$ as determined by numerically solving the ODE over a coarse grid of parameter values. This prior knowledge can be expressed by the model:

$$\begin{aligned} \alpha \sim \beta &\sim N\left(0, .4^2\right) \\ \gamma &\sim \chi^2_2 \end{aligned} \tag{2.6}$$

which places approximately 96% of the density of $\boldsymbol{\theta}$ in this region. When $\boldsymbol{\theta}$ is in this approximate prior region, the system will eventually produce oscillations for an extremely wide range of initial system states.

If $\mathbf{X}_0$ is a pair of values taken from the oscillatory behavior of the ODE model, their influence on $S(\boldsymbol{\theta}, \mathbf{X}_0, t)$ is to determine the phase of the system. If $\mathbf{X}_0$ is not a set of values from the oscillations, the ODE solution must first pull V and R towards the stable limit cycles, which may take longer than the window of observed values. Consequently, a reasonable but vague data driven prior for the initial conditions is a Gaussian density centered on the observed initial state, with variance equal to the observed variance of the first 30 observations (from time 0 up to time 1.5) about their mean:

$$\begin{aligned} V_0 &\sim N\left(V[t = o], \mathtt{var}(V\left[t \leq t_{30}\right])\right), \\ \text{and } R_0 &\sim N\left(R[t = o], \mathtt{var}(R\left[t \leq t_{30}\right])\right). \end{aligned} \tag{2.7}$$

This gives considerable uncertainty, due to the rapidly changing values of V and R, however, initial conditions can have a large impact on the shape of the ODE solution and consequently should not be unnecessarily restricted.

The uninformative conjugate prior density on the variance components σ_V^2 and σ_R^2 is the log uniform distribution,

$$P(\sigma_k^2) \propto 1/\sigma_k^2, \quad \text{for} \quad \sigma_k^2 > 0, \quad \text{and} \quad k \in \{V, R\}.$$

Using the FitzHugh-Nagumo example, the standard Metropolis-Hastings MCMC method follows this algorithm Gelman et al. (2004):

1. Initialize the algorithm at draw $i = 0$ with $\boldsymbol{\theta}^{(i)}, \mathbf{X}_0^{(i)}, \sigma_V^{2(i)}$ and $\sigma_R^{2(i)}$. Often these are samples from the prior densities. Use these samples to obtain the numerical solutions $S_{V,R}(\boldsymbol{\theta}^{(i)}, \mathbf{X}_0^{(i)}, t)$ and subsequently the un-normalized posterior.
2. Propose $\mathbf{X}_0^*$ and θ^* from the jumping distribution $J([\boldsymbol{\theta}^*, \mathbf{X}_0^*] \mid \boldsymbol{\theta}^{(i)}, \mathbf{X}_0^{(i)}, \Sigma_J)$ and produce the numerical solution $S(\boldsymbol{\theta}^*, \mathbf{X}_0^*, t)$ and un-normalized posterior.
3. Draw $u \sim U(0, 1)$. If $u < \alpha$, where

$$\alpha = \min\left(1, \frac{P\left(\boldsymbol{\theta}^*, \mathbf{X}_0^* \mid [V\{t\}, R\{t\}]\, \sigma_V^{2(i)}, \sigma_R^{2(i)}\right)}{P\left(\boldsymbol{\theta}^{(i)}, \mathbf{X}_0^{(i)} \mid [V\{t\}, R\{t\}]\,, \sigma_V^{2(i)}, \sigma_R^{2(i)}\right)} \frac{J([\boldsymbol{\theta}^*, \mathbf{X}_0^*] \mid \boldsymbol{\theta}^{(i)}, \mathbf{X}_0^{(i)}, \Sigma_J)}{J([\boldsymbol{\theta}^{(i)}, \mathbf{X}_0^{(i)}] \mid \boldsymbol{\theta}^*, \mathbf{X}_0^*, \Sigma_J)}\right), \quad (2.8)$$

   accept the proposed values by setting $[\boldsymbol{\theta}^{(i+1)}, \mathbf{X}_0^{(i+1)}] = [\boldsymbol{\theta}^*, \mathbf{X}_0^*]$ and otherwise keep the old values for another iteration: $[\boldsymbol{\theta}^{(i+1)}, \mathbf{X}_0^{(i+1)}] = [\boldsymbol{\theta}^{(i)}, \mathbf{X}_0^{(i)}]$. Only the un-normalized posteriors are required to compute α in (2.8).
4. Compute the sum of squared residuals (SSE) from the data to $S_{V,R}(\boldsymbol{\theta}^{(i+1)}, \mathbf{X}_0^{(i+1)}, t)$
5. Use a Gibbs step and draw values for σ_V^2 and σ_R^2 from:

$$P(\sigma_k^{2(i+1)} \mid V, R, \mathbf{X}_0^{(i+1)}, \boldsymbol{\theta}^{(i+1)}) = SSE_k/\chi_{n-1}^2, \quad k = V, R$$

6. Set $i = i + 1$ and return to step 2 until $i = N$ for some large value of N.
7. Discard the first iterations as burn in to correct for over-representation of values close to $\boldsymbol{\theta}^{(0)}, \mathbf{X}_0^{(0)}, \sigma_V^{2(0)}, \sigma_R^{2(0)}$ in the posterior density. Sometimes the remaining posterior draws are further reduced by keeping only every k^{th} draw for some integer k to reduce the autocorrelation between iterations.

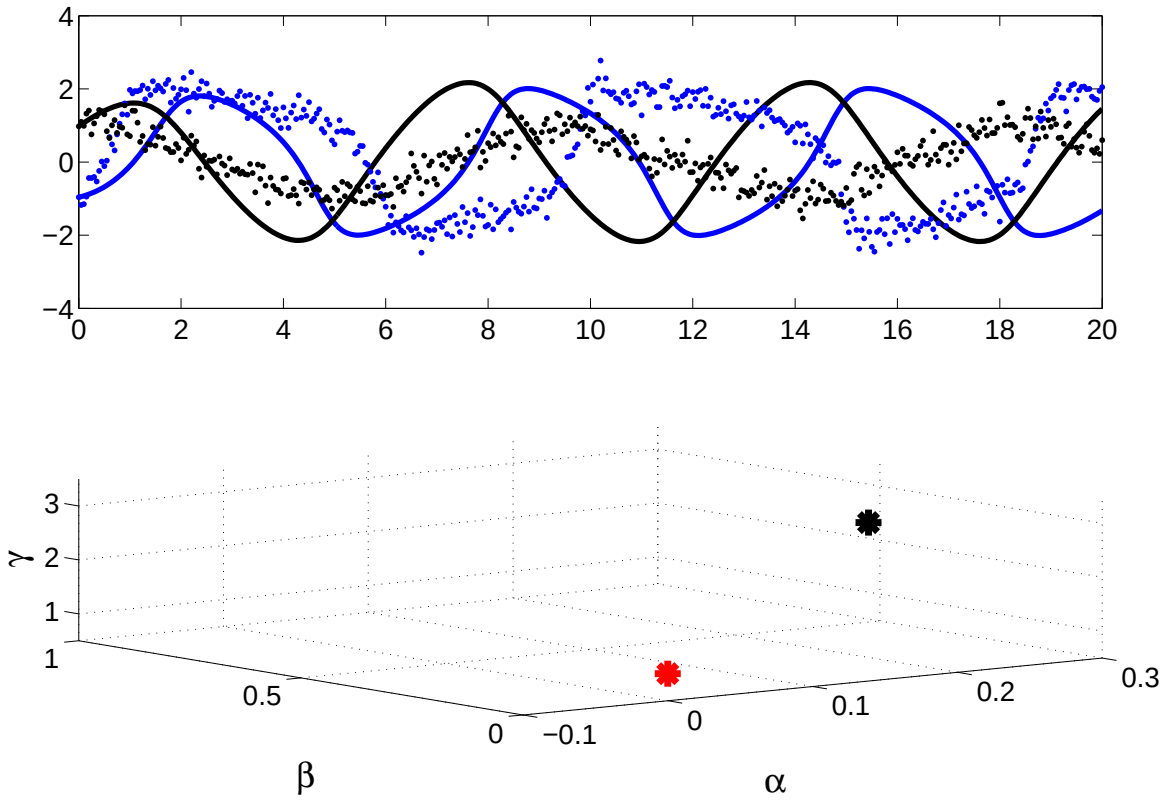

Figure 2–3: The starting point of the MCMC algorithm and the fit to the FitzHugh-Nagumo simulated data. V is in blue and R is in black. The true parameter location is marked by a black star in the bottom plot and the current value is shown with a red star.

For the simulated FitzHugh-Nagumo data sets, the jumping distribution $J([\alpha^*, \beta^*, \gamma^*, V_0^*, R_0^*] \mid [\alpha^{(i)}, \beta^{(i)}, \gamma^{(i)}, V_0^{(i)}, R_0^{(i)},]\Sigma_J)$ was a set of independent normals with

$$\Sigma_j = diag([0.01^2, 0.01^2, 0.05^2, 0.005^2, 0.005^2]),$$

chosen to produce an acceptance rate of 20-25% for the 50 simulated data sets.

To examine how well this method works in practice, figures 2–3 to 2–7 show the path taken by one of the MCMC chains with true parameters $\boldsymbol{\theta}^{(true)} = [\alpha^{(true)}, \beta^{(true)}, \gamma^{(true)}] = [.2, .2, 3]$, from the starting point at $\boldsymbol{\theta}^{(0)} = [0, 0, 1]$ through 80,000 iterations of the chain. The figures show the path through the parameter space as well as the numerical solution to the ODE using the current iteration parameter values as a fit to the data. Figure 2–8 shows the final half of the 200,000 posterior draws and the fit to the data using the posterior mean.

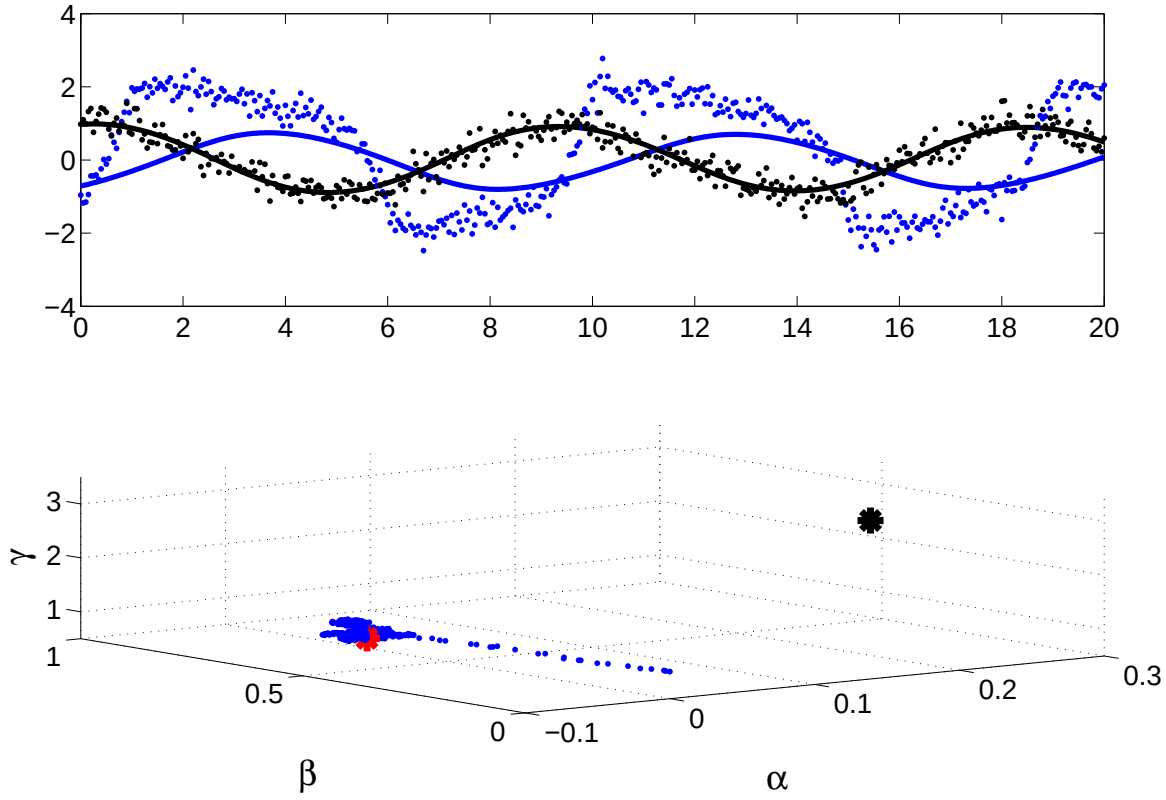

Figure 2–4: The bottom panel shows the first 10,000 posterior MCMC draws from the FitzHugh-Nagumo model. The true parameter location is marked by a black star and the current value is shown with a red star. The top panel shows the fit to the data using the parameters from the $10{,}000^{th}$ draw. V is in blue and R is in black.

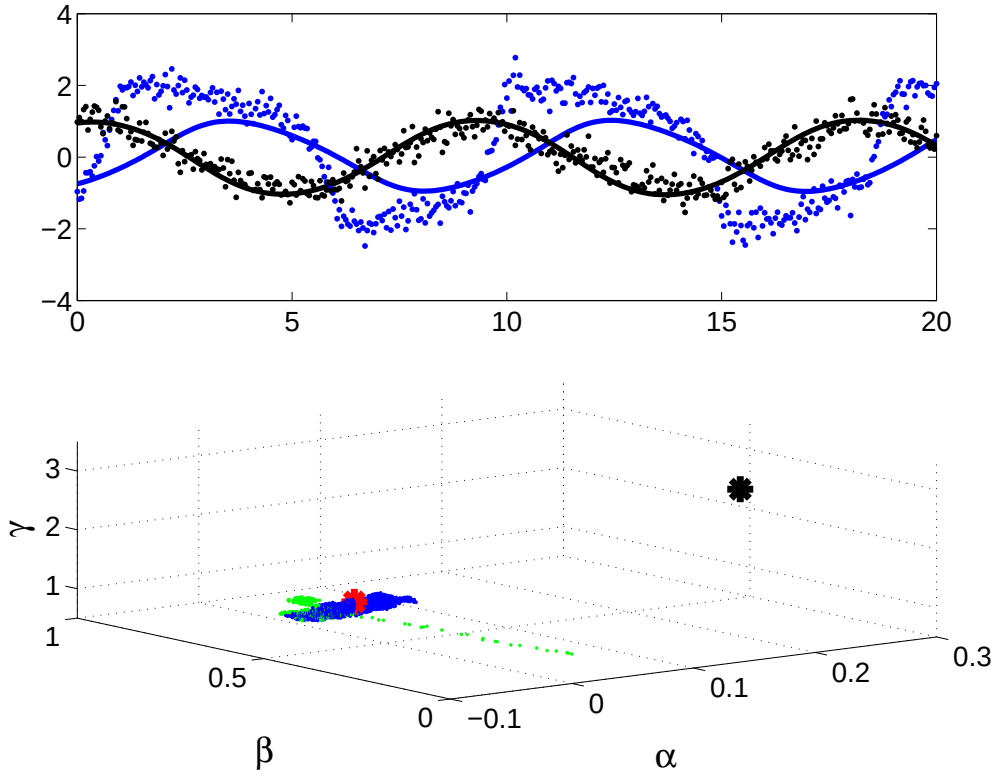

Figure 2–5: The bottom panel shows the first 60,000 posterior MCMC draws from the FitzHugh-Nagumo model. The first 10,000 are in green and the next 50,000 are in blue. The true parameter location is marked by a black star and the current value is shown with a red star. The top panel shows the fit to the data using the parameters from the $60{,}000^{th}$ draw. V is in blue and R is in black.

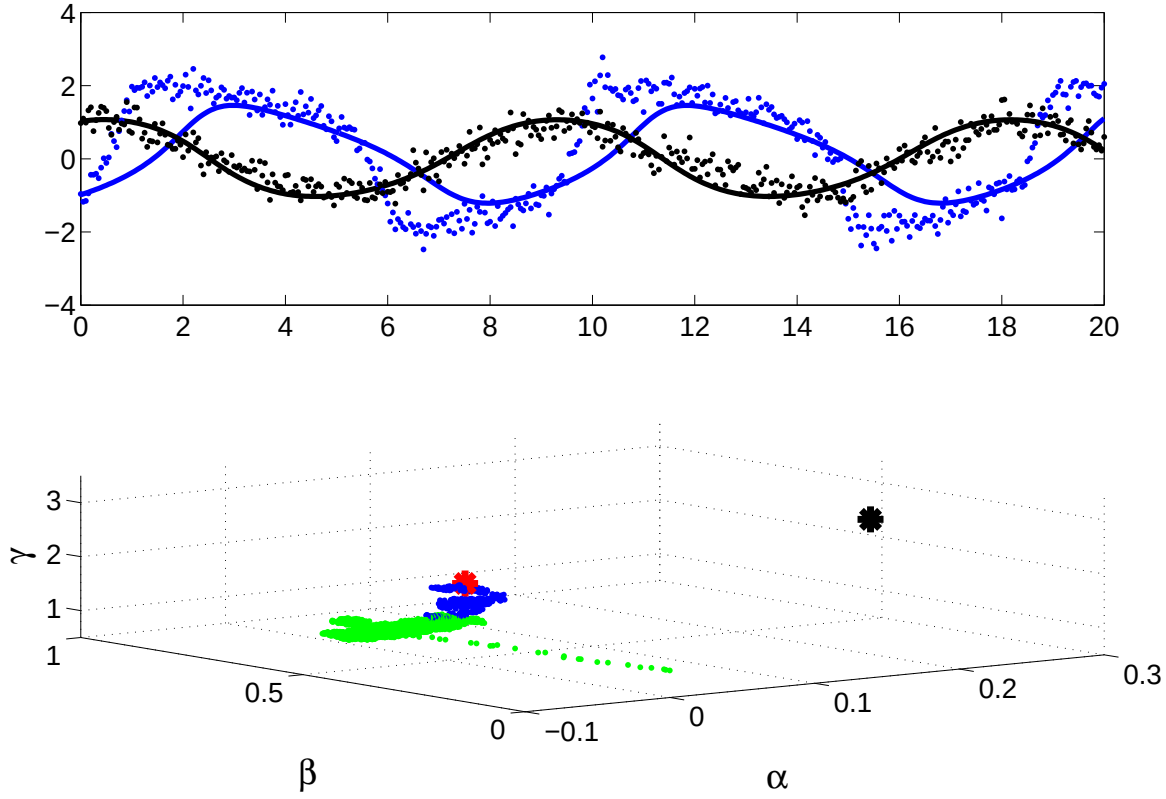

Figure 2–6: The bottom panel shows the first 70,000 posterior MCMC draws from the FitzHugh-Nagumo model. The first 60,000 are in green and the next 10,000 are in blue. The true parameter location is marked by a black star and the current value is shown with a red star. The top panel shows the fit to the data using the parameters from the $70{,}000^{th}$ draw. V is in blue and R is in black.

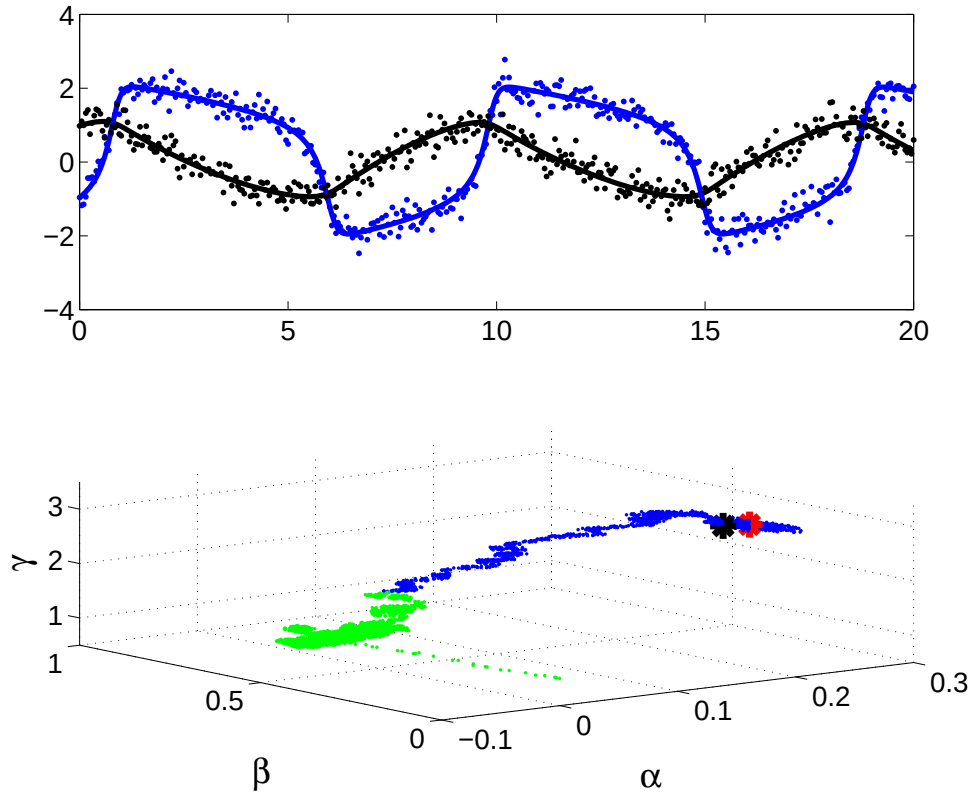

Figure 2–7: the bottom panel shows the first 80,000 posterior MCMC draws from the FitzHugh-Nagumo model. The first 70,000 are in green and the next 10,000 are in blue. The true parameter location is marked by a black star and the current value is shown with a red star. The top panel shows the fit to the data using the parameters from the $80{,}000^{th}$ draw. V is in blue and R is in black.

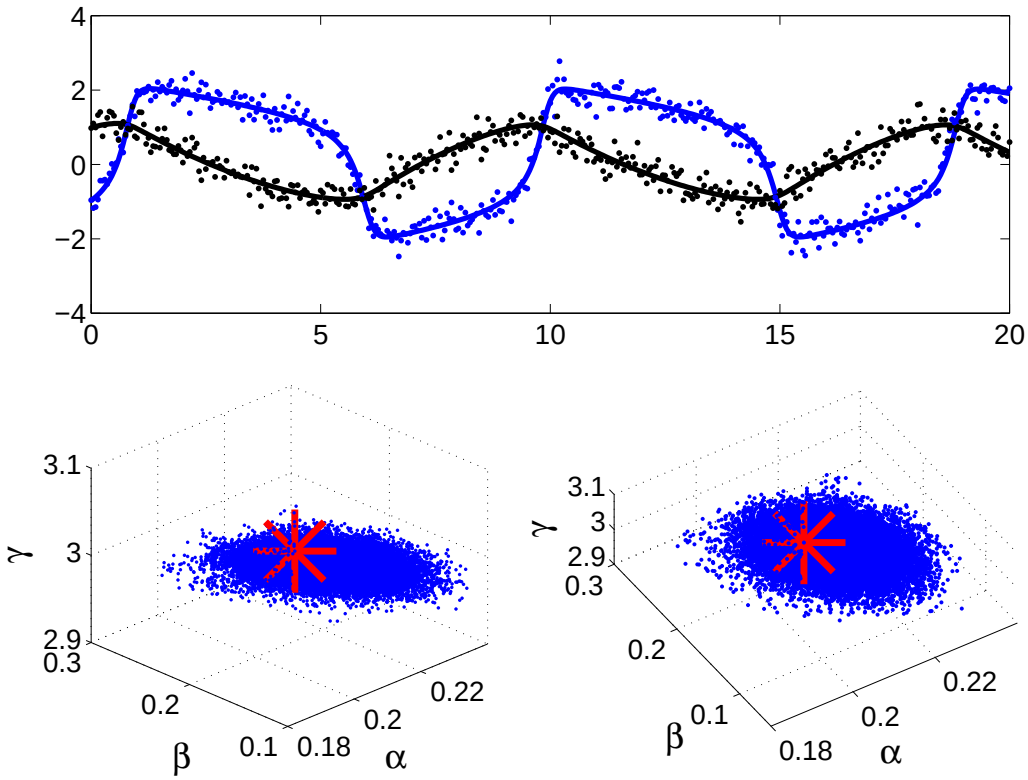

Figure 2–8: The bottom panels offer two perspectives on the final 100,000 posterior draws from the FitzHugh-Nagumo model after discarding 100,000 for burn in. The true value is shown as a red star. The fit to the data using the posterior mean is shown in the top panel where component V is in blue and R is in black.

Figure 2–3 shows that at the starting point, $S(\boldsymbol{\theta}, \mathbf{X}_0, t)$ is periodic with amplitude similar to the data but the data and $S(\boldsymbol{\theta}, \mathbf{X}_0, t)$ are out of phase. Within the first few iterations in figure 2–4, the parameters move towards a location which produces a reasonable fit to component R but a poor fit to V. Although the fit to R is more sinusoidal than the true process, the fit lies approximately in the middle of the observations.

For the next 50,000 draws, shown in figure 2–5, there is almost no change in parameters because the relative gain in fit to V compared to the loss of fit to R makes it an unattractive exchange and consequently occurs with very low probability. As shown in figure 2–6, by the $70{,}000^{th}$ draw, the chain has made it through the worst of this exchange. The shape of the fit to component V has become sharper cornered, closer to the shape of the data, in exchange for moving the sinusoidal fit to R further from the middle of the data. The reduced fit to R is especially notable at the bottom of the valleys of the fit in the time intervals $(3, 6)$ and $(11, 14)$. The next 10,000 draws (figure 2–7) move quickly across the parameter space to the region of the true values. Discarding half the chain as burn in, the final 100,000 draws shown in figure 2–8 produce an excellent fit to the data and describe the relevant portion of the posterior density containing the true parameter value.

The abrupt changes in behaviour of the solution to the ODE with respect to small changes in the parameter values, provide abundant opportunities for the MCMC algorithm to become stuck. With this nonlinear ODE model, there are several of these regions in the parameter space providing a reasonable fit to only a fraction of the data. Improvements in fit from these regions require passing across regions of the parameter space fitting considerably less of the data. While in this case, after 70,000 posterior draws the MCMC was able to eventually move towards the true values, this is not always the case. Figure 2–9 shows the 95% highest posterior density intervals for the 50 simulated FitzHugh-Nagumo data sets, initialized using draws from the prior densities of (2.6) and (2.7). These prior draws are used as initial parameter estimates every time these simulated data sets are analyzed. Intervals shown in figure 2–9 are taken after discarding the first half of the 200,000 MCMC iterations as burn in. In the four cases where initial values

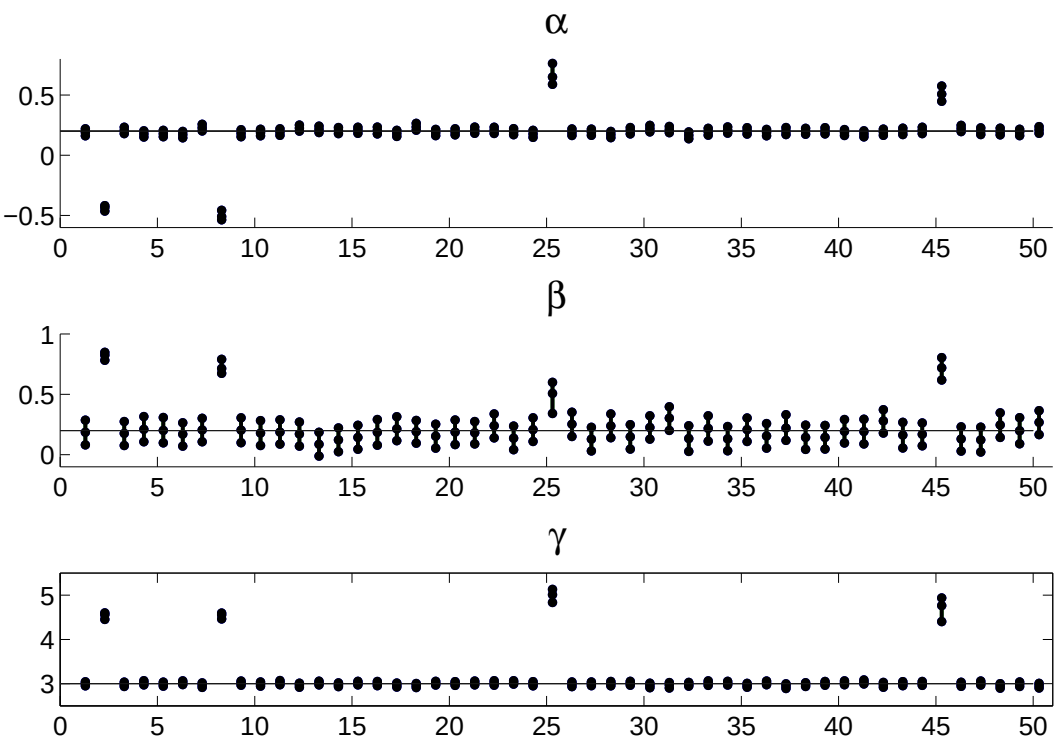

Figure 2–9: 95% Highest posterior density intervals for 50 simulated FitzHugh-Nagumo data sets with true parameters $[\alpha, \beta, \gamma] = [.2, .2, 3]$.

of γ were particularly large or parameters V_0 and R_0 were far from their true values, the chains were not able to move towards the true value within the 200,000 draws. Extending the chains for 1,000,000 draws or restarting the MCMC with the same initial parameter estimates still did not move these chains towards their true values. However, if these chains were initialized with different starting points, reasonable posterior densities and interval estimates are possible.

The slow rate of convergence, the risk of incorrect results and answers that depend on initial estimates leave a lot of room for improvement in Bayesian parameter density estimation for nonlinear ODE models. Furthermore, the Bayesian and NLS methods share many of the same problems due to form of the likelihood, and many opportunities to produce a partial fit to the data.

2.2.1 Prior Specification For Bayesian ODE Models

Uninformative prior densities on the parameters of nonlinear models often produce improper posteriors (Bates and Watts 1988). As an example, consider a simplification of the nylon system of (2.4) with constant parameters and a single model component: $DA = -k_1 A + k_2$. This model describes exponential decay towards the asymptote k_2/k_1. If the only prior information is that both parameters are positive, this information gives the improper priors: $P(k_1) = P(k_2) \propto 1, \;\; k_1 > 0, \; k_2 > 0$. When the data offers little information about the exponential decay rate k_1, but perhaps offers perfect information about the asymptote k_1/k_2, the two parameters do not have a unique posterior mode or proper posterior density.

To avoid uninformative priors producing improper posteriors Bates et al. (1988) suggest placing an uninformative prior on the shape of $S(\boldsymbol{\theta}, \mathbf{X}_0, t)$ and then mapping this density back to the parameter space to obtain prior densities for $\boldsymbol{\theta}$ and $\mathbf{X}_0$. Truncated priors are often used instead, and then the proximity of the posterior mode and density to the truncations is used as a diagnostic tool for determining if the prior or the model require revising.

In the case of the FitzHugh-Nagumo system, the prior knowledge that $S(\boldsymbol{\theta}, \mathbf{X}, t)$ should be oscillating induced priors on $\boldsymbol{\theta}$ and $\mathbf{X}$. However, in some cases prior information exists for the functional form of $S(\boldsymbol{\theta}, \mathbf{X}_0, t)$ which is from a source independent from the prior information on $\boldsymbol{\theta}$. The prior on $S(\boldsymbol{\theta}, \mathbf{X}_0, t)$ induces an additional prior on $\boldsymbol{\theta}$ which may conflict with the prior on $\boldsymbol{\theta}$ directly. A variety of methods to combine these prior information sources are compared in (Poole and Raftery 2000), where this pooling of prior information sources is called Bayesian Melding.

2.2.2 Bayesian Posterior Density Estimation of the Nylon Model Parameters

The conservation of mass principle used to develop the ODE model for nylon dynamics in section 1.1 suggests that components A and C should rise and fall with changes in W_{eq}, producing the constraints $k_{p0} > 0$ and $K_{a0} > 0$ in (2.4). Although the initial parameter estimates used in NLS estimation of section 2.1.2 came from least squares estimates of transformations of subsets of the data and consequently do not really constitute prior knowledge, they were helpful in

determining the scale of the parameters. Consequently, although these parameter estimates are not to be strongly believed, they tell us that the parameters should not be too far from zero. More specifically, in the expression for K_a in (2.4), if $\mid \gamma \mid$ is large then $\left\{1 + W_{eq}\frac{\gamma}{1000}\right\} K_{a0} \approx \left\{W_{eq}\frac{\gamma}{1000}\right\} K_{a0}$ making γ and k_{a0} unidentifiable. This prior information was summarized with the model

$$\begin{aligned} P(k_{p0}) &\sim \Gamma(4,8) \\ P(\gamma) &\sim N(5,15^2) \\ P(K_{a0}) &\sim \Gamma(4,8) \\ \text{and } P(\Delta H) &\sim N(0,50^2), \end{aligned} \tag{2.9}$$

where the gamma density, $\Gamma(A,B)$ is parameterized to have mean AB and variance AB^2.

Priors used on the 18 initial system states Gaussian densities truncated at zero to avoid uninterpretable negative concentration estimates. For A and C, these priors were centered on the observed values with prior variances equal to $(2\tilde{\sigma}_A)^2$ and $(2\tilde{\sigma}_A)^2$, where $\tilde{\sigma}_A = .6$ and $\tilde{\sigma}_C = 2.4$ are the measurement standard deviations determined through replicate measurements in Zheng et al. (2005). These values were also used as weights in the nonlinear least squares routine of section 2.1.2. The quadrupling of the variance reflects potential additional measurement and model mis-specification errors in this study. Since W was unobserved, the prior on its initial system state was a truncated Gaussian centered on the initial value of W_{eq}. Recall that W_{eq} is the expected value of W if the system is at equilibrium. Reflecting considerable uncertainty in the assumption that the systems is beginning at equilibrium, and a desire to avoid undue restrictions of the initial system state of an unobserved component, the prior variance was set at 25^2.

The variance components σ_A^2 and σ_C^2 were assumed to be in the neighbourhood of the values used as weights but independent $\Gamma(3,3)$ densities were used to allow considerable discrepancy . The vagueness of these priors reflects the model uncertainty which is built into this residual error term.

These 24 priors were combined with the information from the 120 observations of A and 104 of C in an MCMC algorithm. Parameter proposals were performed in two vectors, one

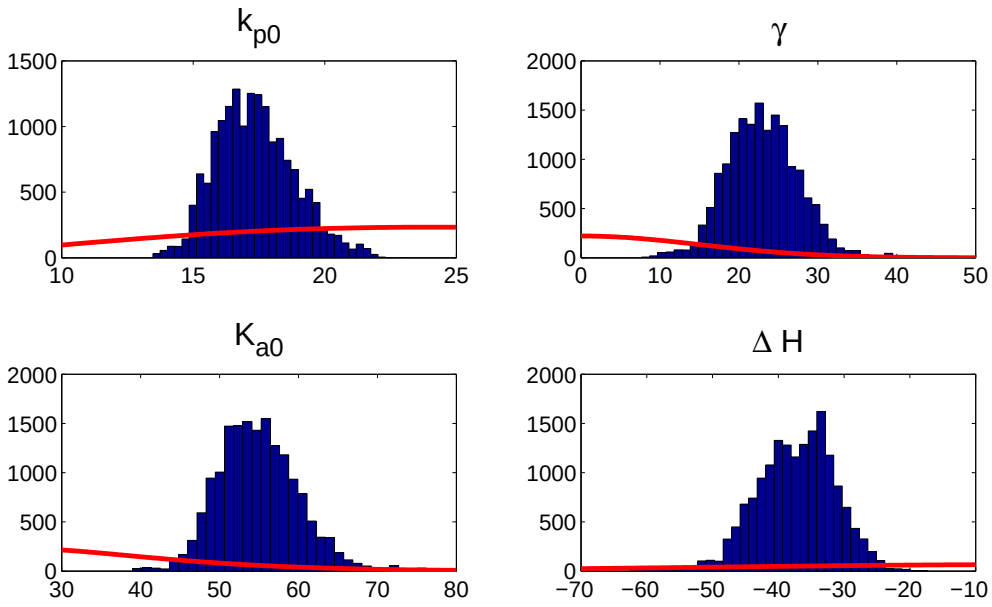

Figure 2–10: Histogram of the final 12,500 draws from the posterior of the 4 nylon ODE parameters (blue). The red lines show the prior densities of the parameters scaled to integrate to 12,500.

for the model parameters $\boldsymbol{\theta} = [k_{p0}, \gamma, K_{a0}, \Delta H]$ and another for the 18 initial system states $\mathbf{X}_0 = \{A_{i0}, C_{i0}, W_{i0}; i = 1, \ldots, 6\}$. As with the FitzHugh-Nagumo example, the posterior draws for σ_A^2 and σ_C^2 were obtained using a Gibbs step. The MCMC algorithm converged quickly. After discarding half of the 25,000 draws for burn in, a histogram density estimate of the final posterior draws for $\boldsymbol{\theta}$ are shown in figure 2–10. A scaled portion of the prior densities are included in this figure for comparison. The posterior draws converged quickly as was confirmed by using 10 independent MCMC runs and computing a Gelman-Rubin diagnostic (Gelman and Rubin 1992) separately on each parameter.

The nylon system results from competing exponential growth and exponential decay describing the production and destruction of L and W. In general, a mixture of exponentials such as this, is an ill-conditioned problem and parameters are difficult to estimate adequately without considerable amounts of data (Gelman et al. 2004). However, in this case the likelihood provides abundant information, fine tuning the posterior considerably.

2.3 Alternative Methods for Parameter Estimation from ODE Models

When problems are well posed, the main problem with parameter estimation from ODE models is the likelihood surface is difficult to navigate. The likelihood surface difficulties are caused by the variety of behaviours which may be produced by a single ODE model, and the dependence on the solution to the ODE which also varies considerably with the initial system states. While this caused problems for NLS and MCMC methods, several attempts have been made to alleviate this root problem from a variety of perspectives. Some of which are outlined in this section.

It was suggested in (Meyer and Christensen 2000) that rather than formulating the ODE models with observation error $\epsilon(t)$ for observations $y(t) = S(\boldsymbol{\theta}, \mathbf{X}_0, t) + \epsilon(t)$, centered on the solution to the dynamic model $S(\boldsymbol{\theta}, \mathbf{X}_0, t)$, the ODE should be reformulated as a Stochastic Differential Equation (SDE) filter. In this model some or all of the observational error is modelled through a random forcing function in the model, so the ODE model defined by $DY = f(Y, \boldsymbol{\theta}, t)$ becomes the SDE model with stochastic error $\zeta(t)$; $DY = f(Y, \boldsymbol{\theta}, t) + \zeta(t)$. The authors show that as a Bayesian formulation, this avoids the need for excessively precise and narrow prior knowledge of the initial system state and therefore avoids being trapped in some of the relatively unimportant regions of the posterior. This is especially shown to be of concern when $f(Y, \boldsymbol{\theta}, \mathbf{u}(t), t)$ describes nonlinear or chaotic behaviour.

State space models and difference equations such as (1.1) can also be used for parameter estimation. From the perspective of sequential data assimilation, the goal is to accurately estimate the (partially) unobserved set of state variables. In (Dowd 2007), an MCMC based method for online estimation of state variables in a nonlinear SDE ecological dynamic model is described. In the example, one of the nonlinear SDE parameters are estimated as time dependent state variables. Estimation from state space models begins by dividing the n data points into the $n-1$ intervals between observations. The SDE is then numerically computed over the intervals and state variables, which may include SDE model parameters $\boldsymbol{\theta}$, are estimated within each interval to produce functional parameter estimates $\boldsymbol{\theta}(t)$. In (Ionides, Bretó, and King 2006),

iterative weighted averaging incrementally restricts the functional form of $\boldsymbol{\theta}(t)$ to obtain the desired estimate of the constant $\boldsymbol{\theta}$.

However not all are in favor of the SDE model formulation. For example, (Judd 2003) points out that it is rare if ever that stochastic errors are representative of the true error structure and "when a stochastic effect appears to be present, such as say thermal noise, it is just complex high dimensional deterministic dynamics and is therefore really model error." Furthermore, Judd (2003) shows that the stochastic formulation performs better than the more realistic error structure model because SDEs have the flexibility to produce pseudo-orbits, approximations to the trajectory of the underlying dynamic process. Rather than using SDEs, the authors suggest an iterative gradient descent method for finding a pseudo-orbit approximation to the solution to the ODE, which is then used in place of the ODE solution in an approach otherwise equivalent to NLS. The pseudo-orbit is initially taken to be a data interpolation and incrementally forced to fit the ODE model. This is similar to a collocation method (a method based on data smoothing or basis expansion) where the smooth is taken as the solution to the ODE model, but profiling out the initial system states. The pseudo orbit algorithm is very similar to the profile estimation algorithm of chapter 3 and (Ramsay, Hooker, Campbell, and Cao 2007) except that Judd (2003) does not use the implicit function theorem to improve convergence, inference and interval estimation.

Avoiding the solution to the ODE model, (Varah 1982) used a collocation method where the data are first smoothed without considering the ODE model. Holding the data smooth fixed, parameter optimization is based on minimizing the discrepancy between the derivative of the smooth and the ODE model. While the smooth removes dependency on the numerical solution to the ODE, the derivative estimate from a general smoothing technique tends to produce a poor estimate of the intricacies available from an ODE model. Alternatively (Arora and Biegler 2004) proposed a collocation method taking into account the ODE model and subsequent optimization of basis coefficients and $\boldsymbol{\theta}$ simultaneously. This large scale high dimensional optimization requires a considerable amount of data in order to produce unique results.

An extension to NLS, similar to an ODE version of the state space model called multiple shooting (Bock 1983) partitions the time domain into M intervals where $M < n$ for n observations. The ODE is numerically solved within each interval using initial system states as additional parameters. Parameters $\boldsymbol{\theta}$ are optimized globally over all intervals subject to the constraint that the numerical solution must be continuous across intervals. Conceptually this is a hybrid between collocation and NLS methods. The similarity to NLS comes from the continuous solution constraint; the need for estimating the initial states in each interval disappears and ultimately the parameter estimation is based on the discrepancy between the data and the numerical solution to the ODE model. As initial system states within each interval are initially allowed to be disjoint, this is also similar to a collocation method where the basis functions are the solutions to the ODE within each of the disjoint time intervals. The local influence of the partitions reduces the threat of poor parameter values or initial system state estimates propagating a poor data fit across the entire time interval.

The notion of collocation and localized parameter estimates has also been explored in a Bayesian context. A Bayesian model using the idea of multiple shooting is described in (Mukhin, Feigin, Loskutov, and Molkov 2006). However instead of constraining the intervals to form a continuous ODE solution, a posterior density is obtained from each of the M intervals, each containing w evenly spaced observations. The results are combined in using the geometric mean of the resulting M posterior densities:

$$\boldsymbol{\theta} \mid \mathbf{y} \sim \left(\prod_{m=1}^{M} P_m(\boldsymbol{\theta} \mid \mathbf{y}(t_{((m-1)*w+1)}, ..., t_{(m*w)})) \right)^{1/(w+1)} . \tag{2.10}$$

This method requires estimation of initial system states for each interval and consequently the number of parameters and number of intervals increases with the amount of data. In the presence of potentially chaotic behaviour from the ODE, this method is shown to perform better than methods using a single interval for the same reasons as the multiple shooting algorithm.

Several recent Bayesian innovations have been developed to ease the movement around the posterior parameter space. While not specifically designed for ODE models, they provide some insightful ways of overcoming the problems of parameter estimation from ODE models.

The Equi-energy sampler (Kou, Zhou, and Wong 2006) allows easier movement across a multi-modal parameter space by allowing jumps to regions with the same posterior height. The method uses M successive approximations to the posterior, where the m^{th} is based on the threshold ϵ_m in

$$P_m(\boldsymbol{\theta} \mid \mathbf{y}) \propto \begin{cases} P(\boldsymbol{\theta} \mid \mathbf{y}) & \text{if } P(\boldsymbol{\theta} \mid \mathbf{y}) > \epsilon_m \\ 1 & \text{otherwise} \end{cases} . \tag{2.11}$$

These approximations fill in the space between modes with a uniform density. The threshold ϵ_m is reduced successively with each approximation until it can be fully removed to sample from the target posterior. To use these approximations, at each iteration with some probability p the usual Metropolis-Hastings step is performed. With probability $1 - p$ an equal energy jump is performed by sampling from the set of points obtained in previous approximations (with larger ϵ_m), that have equal posterior height to the last posterior draw. These jumps move from one region to another identical energy (posterior height) region while passing over the lower energy valleys between them. This method is shown to perform well in multi-modal densities where the posterior heights of the modes are comparable but the near zero energy valleys between modes are prohibitively large. The idea of successive approximations to move around the parameter space more freely is useful with Bayesian ODE models. However, ripples in the posterior surface are much smaller than the mode when they are caused by the model fitting only a small subset of the data. One of the big problems with ODE models is finding the highest posterior mode.

A multigrid MCMC (Liu and Sabatti 1998) also uses a hierarchy of approximations to the posterior surface based on interpolating across a grid of posterior values. Using a finer grid improves the approximation towards the true posterior. The MCMC chains for each approximation are run in parallel such that at each iteration, with probability p the usual MCMC step is taken, but with probability $1-p$ a swap is proposed. If accepted, parameters $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ from chains i and

j swap values. This allows the parameters from coarser approximations to the posterior to travel more easily around the parameter space and then pass along good sets of parameter values into the chain drawing from the true posterior. Well placed grid points reduce the time that coarser chains spend in relatively unimportant parts of the inter-modal valleys. However this method depends on obtaining a reasonable quality, potentially high dimensional grid representation of a rippled surface with an elusive highest posterior mode.

Parallel tempering offers an alternative perspective to Equi-energy jumping and multigrid-MCMC to cross regions of low posterior density. Like Equi-energy sampling, tempering uses a hierarchy of approximations to the posterior of interest, however with tempering each one is smoothed more and more towards a uniform posterior, maintaining smooth transitions between modes and their valleys. Furthermore with parallel tempering the MCMC chain approximations are run in parallel instead of in sequence and parameters are allowed to swap between chains. Parallel tempering is described further and demonstrated in chapter 4.

CHAPTER 3
Profile Estimation with a Constrained Smooth

The generalized profiling method (Ramsay, Hooker, Campbell, and Cao 2007) belongs to the family of collocation methods, using a data smooth in the form of the basis expansion $\mathbf{c}'\boldsymbol{\phi}(t)$ for a vector of coefficients $\mathbf{c}$ and basis functions $\boldsymbol{\phi}(t)$. The basis expansion is used to approximate the solution to the ODE model $S(\boldsymbol{\theta}, \mathbf{X}_0, \mathbf{u}(t), t)$ with model parameters $\boldsymbol{\theta}$,initial system state $\mathbf{X}_0$, and input functions or experimental conditions $\mathbf{u}(t)$. The generalized profiling data smooth is guided by the ODE model by penalizing deviation at the level of the derivative. The tradeoff between interpolating the data and following a solution to the ODE model is controlled by the smoothing parameter λ. The optimization of λ, $\boldsymbol{\theta}$ and $\mathbf{c}$ depends on a hierarchy of parameters layered by their impact on the fit to the data.

The incidental or local parameters in the sense of (Neyman and Scott 1948), are the basis coefficients $\mathbf{c}$. For each λ and $\boldsymbol{\theta}$, the optimal $\mathbf{c}$ defines a data smooth, balancing the fit between the data and the ODE model. Consequently, $\mathbf{c}$ can be written as a function of parameters $\boldsymbol{\theta}$ and λ and will sometimes be written as $\mathbf{c}(\boldsymbol{\theta}, \lambda)$ to emphasize this relationship.

The structural parameters $\boldsymbol{\theta}$ define the behaviour allowed by the ODE model. Changes in these parameters may decide between broad model features such as limit cycles or unbounded exponential growth. Furthermore $\boldsymbol{\theta}$ may determine asymptotic equilibrium levels, rates of decay, or the strength and type of feedback loops. Primary interest is in $\boldsymbol{\theta}$ because of its interpretation and potential use in making decisions. For any λ, the profile likelihood optimization for $\boldsymbol{\theta}(\lambda)$ maintains $\mathbf{c}(\boldsymbol{\theta}, \lambda)$ at its optimum conditional on $\boldsymbol{\theta}$, defining the second level of the hierarchy.

The complexity parameter λ, also known as the smoothing parameter, defines the top level of parameters to optimize. It determines how closely the data follow the ODE model adding flexibility to account for potential model mis-specification. Optimization of λ is performed maintaining $\boldsymbol{\theta}$ and $\mathbf{c}$ at their optimum conditional on λ.

For n data observations, k unknown basis coefficients and p unknown ODE parameters, typically $k+p+1 > n$. By defining $c(\boldsymbol{\theta}, \lambda)$ and $\boldsymbol{\theta}(\lambda)$ as functions of λ, the optimization process is essentially reduced to the single parameter λ. The idea of building a hierarchy of parameters, where lower levels are defined as functions of higher levels is called a parameter cascade (Cao and Ramsay 2007).

For fixed λ, the parameter estimation routine can be thought of as resulting from inner and outer loops to estimate $\mathbf{c}(\boldsymbol{\theta}, \lambda)$ and $\boldsymbol{\theta}(\lambda)$. While the estimation process is described in detail in Ramsay et al. (2007), these loops are described with constraints on the data smooth in sections 3.1 and 3.2. Section 3.3 also offers a suggestion for the optimization of λ. Section 3.4 describes how this method deals with the challenges of the nylon system. Section 3.5 describes the results of the profile estimation routine on the nylon system as well as a simulation study based on the nylon system. Section 3.6 contains the estimation details and results from the 50 simulated FitzHugh-Nagumo data sets.

3.1 The Inner Optimization; ODE Model-Based Data Smoothing

Using the continuously differentiable one to one function $g\{\cdot\}$, the data smooth $\mathbf{x}_{ki}(t) = g_k\left\{\mathbf{c}'_{ki}\boldsymbol{\phi}_{ki}(t)\right\}$ for the $i = 1, \ldots, I$ experimental runs and $k = 1, \ldots, K$ system components, approximates the solution to the ODE. Function $g\{\cdot\}$ constrains the smooth to follow known behaviour. Some examples from (Ramsay and Silverman 2005) include the following:

$$\begin{aligned} g\{a\} &= \exp\{a\} && \text{for a positive smooth,} \\ g\{a\} &= \exp\{a\}/[1+\exp\{a\}] && \text{for a bounded smooth,} \\ g\{a\} &= \textstyle\int_0^t \exp\{a\}ds && \text{for a monotone smooth,} \\ \text{and } g\{a\} &= a && \text{for an unconstrained smooth.} \end{aligned}$$

Model based smoothing, also known as *L-spline* smoothing (Heckman and Ramsay 2000), uses a penalty term (`PEN`) to enforce fidelity of the smooth to the ODE model rather than simply interpolating the data. For fixed λ, $\boldsymbol{\theta}$ and data $\mathbf{y}_{ki}$ observed at the vector of times $\mathbf{t}_{ki}$, the coefficients $\mathbf{c}(\boldsymbol{\theta}, \lambda)$ defining the data smooth for the ith experimental run, possibly using weights

w_{ki}, are the minimizers of

$$J_i(\mathbf{c} \mid \mathbf{y}, \boldsymbol{\theta}, \lambda) = \sum_{k=1}^{K} \left\{ \sum_{t \in \mathbf{t}_{ki}} w_{ki} \left[\mathbf{y}_{ki}(t) - g\left\{ \mathbf{c}'_{ki} \boldsymbol{\phi}_{ki}(t) \right\} \right]^2 + \lambda \texttt{PEN}_{ki} \right\}. \tag{3.1}$$

This is a nonlinear least squares estimation problem corresponding to minimizing the negative log likelihood penalized through PEN to fit an ODE model of the form $D\mathbf{x} = f(\boldsymbol{\theta}, \mathbf{x}, \mathbf{u}(t), t)$, depending on inputs, experimental conditions or forcing functions $\mathbf{u}(t)$. The penalty term PEN penalizes discrepancy between the smooth and the ODE model at the level of the derivative through:

$$\texttt{PEN}_{ki} = \int_{\mathbf{T}_i} \left(D\mathbf{x}_{ki}(s) - f_k(\mathbf{x}_i, \mathbf{u}_i, s, \hat{\boldsymbol{\theta}}) \right)^2 ds. \tag{3.2}$$

The integral in (3.2) is taken over the interval $\mathbf{T}_i = [\min_k(\mathbf{t}_{ki}), \max_k(\mathbf{t}_{ki})]$, the maximum range of observation times over all K observed variables in the i^{th} run. Section 3.4.2 describes how to approximate this integral. The choice of weights is discussed in section 3.4.3.

3.2 The Outer Optimization; Estimating ODE parameters

For fixed λ, the estimator $\hat{\boldsymbol{\theta}}(\lambda)$ is the maximum of the profile likelihood maintaining $\mathbf{c}(\hat{\boldsymbol{\theta}}, \lambda)$ at its optimum value. When a Gaussian error structure is a reasonable assumption, the profile likelihood is defined by

$$\mathbf{y}_{ik} \mid \boldsymbol{\theta}(\lambda) \sim N\left(g\{\mathbf{c}(\boldsymbol{\theta}, \lambda)\boldsymbol{\phi}\}, \sigma^2_{ki} \right).$$

Minimizing the negative log likelihood is equivalent to minimizing squared error loss applied to the discrepancy between the data and the smooth pooling information across the I experimental runs and K system components:

$$H\left(\boldsymbol{\theta}(\lambda), \hat{\mathbf{c}}(\boldsymbol{\theta}, \lambda) \mid \mathbf{y}\right) = \sum_{i}^{I} \sum_{k=1}^{K} \sum_{t \in \mathbf{t}_{ki}} w_{ki} \left[\mathbf{y}_{ki}(t) - g_k \left\{ \hat{\mathbf{c}}_{ki}\left(\boldsymbol{\theta}, \lambda\right)' \boldsymbol{\phi}_{ki}(t) \right\} \right]^2. \tag{3.3}$$

There is no need to include another penalty term enforcing fidelity to the ODE model since $\mathbf{c}(\boldsymbol{\theta}, \lambda)$ is already encouraged to fit the model in the inner optimization. Although the loss function in equation (3.3) is based on a Gaussian model, the method is not restricted by the distributional assumptions.

To obtain the maximum profile likelihood estimate, $\mathbf{c}(\boldsymbol{\theta}, \lambda)$ is maintained at its optimum. Consequently, during the estimation of $\hat{\boldsymbol{\theta}}(\lambda)$, $\mathbf{c}(\hat{\boldsymbol{\theta}}, \lambda)$ must also be updated at every iteration conditional on the latest value of $\hat{\boldsymbol{\theta}}(\lambda)$. Simplifying notation from $H(\boldsymbol{\theta}, \mathbf{c}(\boldsymbol{\theta}, \lambda) \mid \mathbf{y})$ to H, assuming that λ is fixed and simplifying $g(\mathbf{c}'\boldsymbol{\phi})$ to g, this gives the total gradient for the outer optimization:

$$\frac{dH}{d\theta} = \frac{\partial H}{\partial \theta} + \frac{\partial H}{\partial g}\frac{dg}{d\mathbf{c}}\frac{d\mathbf{c}}{d\boldsymbol{\theta}}. \tag{3.4}$$

Using this gradient, optimization of H with respect to $\boldsymbol{\theta}$ can then be performed using Gauss-Newton iterations.

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

$$\frac{\partial^2 H}{\partial \boldsymbol{\theta}^2}, \frac{\partial^2 H}{\partial \mathbf{c}^2} \text{ and } \frac{\partial^2 J}{\partial \boldsymbol{\theta}^2}, \frac{\partial^2 J}{\partial \mathbf{c}^2}$$

are positive definitive over a nonempty neighbourhood of $\mathbf{y}$ in the data space. Also, recall at the optimal value for $\hat{\mathbf{c}}$ from (3.1) we have $dJ/d\mathbf{c} = 0$, consequently at $\hat{\mathbf{c}}$,

$$\begin{aligned}
\left.\frac{d^2 J}{d\hat{\mathbf{c}} d\boldsymbol{\theta}}\right|_{\mathbf{c}=\hat{\mathbf{c}}} &= \left.\frac{d}{d\boldsymbol{\theta}}\left(\frac{dJ}{d\mathbf{c}}\right)\right|_{\mathbf{c}=\hat{\mathbf{c}}} \\
&= \left.\frac{d}{d\boldsymbol{\theta}}\left(\frac{\partial J}{\partial g}\frac{dg}{d\mathbf{c}}\right)\right|_{\mathbf{c}=\hat{\mathbf{c}}} \\
&= \left.\left[\frac{\partial^2 J}{\partial g \partial \boldsymbol{\theta}}\frac{dg}{d\mathbf{c}} + \left\{\left(\frac{dg}{d\mathbf{c}}\right)'\frac{\partial^2 J}{\partial g^2}\frac{dg}{d\mathbf{c}} + \frac{\partial J}{\partial g}\frac{d^2 g}{d\mathbf{c}^2}\right\}\frac{d\mathbf{c}}{d\boldsymbol{\theta}}\right]\right|_{\mathbf{c}=\hat{\mathbf{c}}} \\
&= 0.
\end{aligned} \tag{3.5}$$

Solving for $\frac{d\mathbf{c}}{d\boldsymbol{\theta}}$ produces

$$\left.\frac{d\mathbf{c}}{d\boldsymbol{\theta}}\right|_{\mathbf{c}=\hat{\mathbf{c}}} = -\left\{\left(\frac{dg}{d\mathbf{c}}\right)'\frac{\partial^2 J}{\partial g^2}\frac{dg}{d\mathbf{c}} + \frac{\partial J}{\partial g}\frac{d^2 g}{d\mathbf{c}^2}\right\}^{-1}\left\{\frac{\partial^2 J}{\partial g \partial \boldsymbol{\theta}}\frac{dg}{d\mathbf{c}}\right\}\Bigg|_{\mathbf{c}=\hat{\mathbf{c}}}, \tag{3.6}$$

which is substituted into (3.4) to obtain the total gradient

$$\left.\frac{dH}{d\theta}\right|_{\mathbf{c}=\hat{\mathbf{c}}} = \left[\frac{\partial H}{\partial \theta} - \frac{\partial H}{\partial g}\frac{dg}{d\mathbf{c}}\left\{\left(\frac{dg}{d\mathbf{c}}\right)'\frac{\partial^2 J}{\partial g^2}\frac{dg}{d\mathbf{c}} + \frac{\partial J}{\partial g}\frac{d^2 g}{d\mathbf{c}^2}\right\}^{-1}\left\{\frac{\partial^2 J}{\partial g \partial \boldsymbol{\theta}}\frac{dg}{d\mathbf{c}}\right\}\right]\Bigg|_{\mathbf{c}=\hat{\mathbf{c}}} \tag{3.7}$$

for the Gauss-Newton iterative estimation.

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

Interval estimates obtained using the delta method,

$$var(\boldsymbol{\theta}) \approx \frac{d\boldsymbol{\theta}}{d\mathbf{y}} var(\mathbf{y}) \frac{d\boldsymbol{\theta}}{d\mathbf{y}} \tag{3.8}$$

require the implicit function theorem once again to define $d\boldsymbol{\theta}/d\mathbf{y}$. Using the fact that at the maximum profile likelihood estimate $\hat{\boldsymbol{\theta}}$ from (3.3) $dH/d\boldsymbol{\theta} = 0$, and solving for $d\boldsymbol{\theta}/d\mathbf{y}$ in

$$\frac{d}{d\mathbf{y}} \left(\frac{dH}{d\boldsymbol{\theta}} \right) \Big|_{\hat{\boldsymbol{\theta}}} = \left[\frac{d^2 H}{d\mathbf{y} d\boldsymbol{\theta}} + \frac{d^2 H}{d\boldsymbol{\theta}^2} \frac{d\boldsymbol{\theta}}{d\mathbf{y}} \right] \Big|_{\hat{\boldsymbol{\theta}}} = 0. \tag{3.9}$$

using

$$\frac{d^2 H}{d\boldsymbol{\theta}^2} = \left[\frac{\partial^2 H}{\partial \boldsymbol{\theta}^2} + 2 \frac{\partial^2 H}{\partial \boldsymbol{\theta} \partial g} \frac{dg}{d\mathbf{c}} \frac{d\mathbf{c}}{d\boldsymbol{\theta}} + \left(\frac{dg}{d\mathbf{c}} \frac{d\mathbf{c}}{d\boldsymbol{\theta}} \right)' \frac{\partial^2 H}{\partial g^2} \left(\frac{dg}{d\mathbf{c}} \frac{d\mathbf{c}}{d\boldsymbol{\theta}} \right) + \left(\frac{\partial \mathbf{c}}{\partial \boldsymbol{\theta}} \right)' \frac{\partial H}{\partial g} \frac{d^2 g}{d\mathbf{c}^2} \left(\frac{\partial \mathbf{c}}{\partial \boldsymbol{\theta}} \right) + \frac{\partial H}{\partial g} \frac{dg}{d\mathbf{c}} \frac{d^2 \mathbf{c}}{d\boldsymbol{\theta}^2} \right] \Big|_{\mathbf{c}=\hat{\mathbf{c}}, \boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}, \tag{3.10}$$

and

$$\frac{d^2 H}{d\boldsymbol{\theta} d\mathbf{y}} = \left[\frac{d^2 H}{d\boldsymbol{\theta} d\mathbf{y}} + \frac{d^2 H}{d\boldsymbol{\theta} dg} \frac{dg}{d\mathbf{c}} \frac{d\mathbf{c}}{d\mathbf{y}} + \frac{d^2 H}{dg d\mathbf{y}} \frac{dg}{d\mathbf{c}} \frac{d\mathbf{c}}{d\boldsymbol{\theta}} + \left(\frac{dg}{d\mathbf{c}} \frac{d\mathbf{c}}{d\mathbf{y}} \right)' \frac{d^2 H}{dg^2} \frac{dg}{d\mathbf{c}} \frac{d\mathbf{c}}{d\boldsymbol{\theta}} + \left(\frac{d\mathbf{c}}{d\mathbf{y}} \right)' \frac{dH}{dg} \frac{d^2 g}{d\mathbf{c}^2} \frac{d\mathbf{c}}{d\boldsymbol{\theta}} + \frac{dH}{dg} \frac{dg}{d\mathbf{c}} \frac{d^2 \mathbf{c}}{d\boldsymbol{\theta} d\mathbf{y}} \right] \Big|_{\mathbf{c}=\hat{\mathbf{c}}, \boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}. \tag{3.11}$$

These last two equations involve the terms $d^2\mathbf{c}/d\boldsymbol{\theta}^2$, $d^2\mathbf{c}/d\boldsymbol{\theta} d\mathbf{y}$ and $d\mathbf{c}/d\mathbf{y}$, all of which are obtained from further use of the implicit function theorem. These terms are given in the appendix.

Equation (3.6) for point estimates and equations (3.10) and (3.11) used in obtaining confidence intervals simplify when there are no constraints on the smooth. In that case $g(\mathbf{c}'\boldsymbol{\phi}) = \mathbf{c}'\boldsymbol{\phi}$ which gives $dg/d\mathbf{c} = \boldsymbol{\phi}$ and $d^m g/d\mathbf{c}^m = 0$ for all $m > 1$ while a positively constrained smooth has $g(\mathbf{c}'\boldsymbol{\phi}) = \exp(\mathbf{c}'\boldsymbol{\phi})$ and $d^m g(\mathbf{c}'\boldsymbol{\phi})/d\mathbf{c}^m = \boldsymbol{\phi}^m g(\mathbf{c}'\boldsymbol{\phi})$ for all natural numbers m.

3.3 Choosing the Smoothing Parameter λ

When λ is small, the optimal smooth from (3.1) ignores the model based penalty term `PEN`, and interpolates the data by minimizing $\sum_t w \{\mathbf{y}(t) - \mathbf{x}(t)\}^2$. As λ increases, emphasis shifts

from fitting the data to fitting the ODE model, by instead minimizing

$$\int_T \left\{ D\mathbf{x}(s) - f\left(\mathbf{x}, \mathbf{u}(s), s, \hat{\boldsymbol{\theta}}\right)\right\}^2 ds.$$

Consequently, the data smooth evolves from interpolation towards the solution to the ODE. This shift away from the data causes an increase in the squared error discrepancy between the data and the smooth (SSE). This behaviour is shown in figure 3–1 for the 50 simulated FitzHugh-Nagumo data sets and figure 3–2 for the nylon system. The SSE plateaus as λ increases when the basis sufficiently captures the features of the ODE system. For the FitzHugh-Nagumo simulated data this occurs after about $\lambda = 10^3$ and for the nylon data the region of stability in the SSE occurs for $\lambda \in (10^2, 10^4)$.

When λ is pushed too large, bias in $\hat{\boldsymbol{\theta}}$ is induced by the inability of the basis to accommodate the ODE features to the accuracy imposed by λ. This pulls the data smooth and hence the parameters towards an alternative solution where the discrepancy between the smooth and ODE can be further reduced at the detriment of the SSE (Cao 2006). This is only beginning to occur at the extremely high values of λ shown in figure 3–1 for the FitzHugh-Nagumo simulated data, although its degree is very small at this vertical scale. In figure 3–2 the deterioration of the SSE for the nylon system is much more pronounced and occurs earlier (after $\lambda = 10^4$) because the basis used for the nylon example is coarser than the basis used for the FitzHugh-Nagumo model. Details of the profile estimation process for the FitzHugh-Nagumo and nylon systems are given in sections 3.5 and section 3.6 respectively.

The corresponding decline in `PEN`, due to increasing λ can be seen in figures 3–3 and 3–4 for the FitzHugh-Nagumo data sets and the 4 parameter nylon model respectively. The behaviour of the $\log_{10}(\texttt{PEN})$ roughly corresponds to exponential decay with increasing $\log_{10}(\lambda)$, showing that it becomes exponentially more difficult to decrease the discrepancy between the smooth and the ODE model. When the data is interpolated by the smooth it is easy to obtain large gains from `PEN` by increasing λ but these gains decrease quickly once a reasonable approximation to the ODE has been attained.

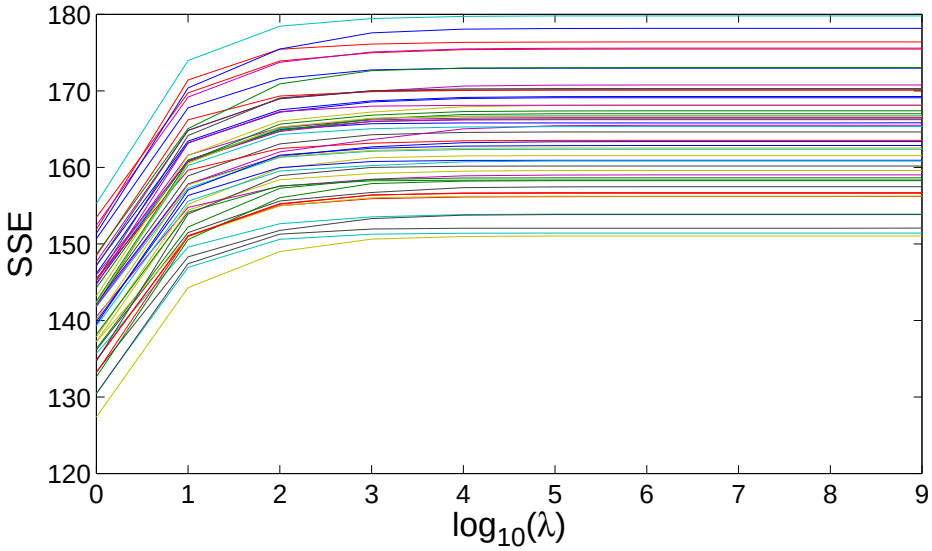

Figure 3–1: The change in SSE with $\log_{10}(\lambda)$ for the 50 simulated FitzHugh-Nagumo data sets.

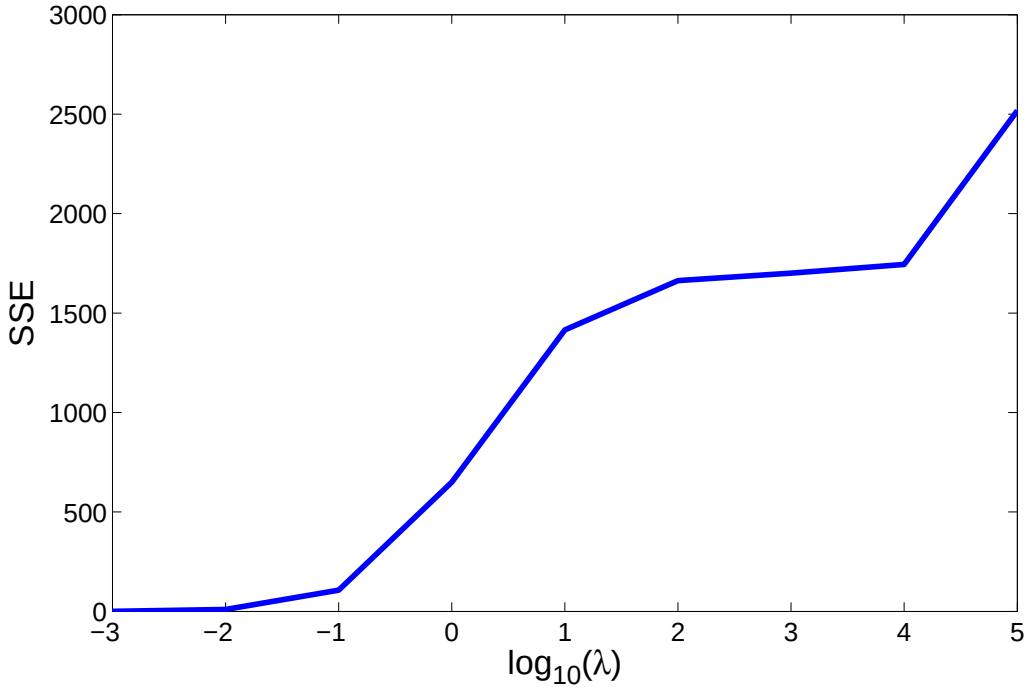

Figure 3–2: The change in SSE with $\log_{10}(\lambda)$ for the 4 parameter nylon model.

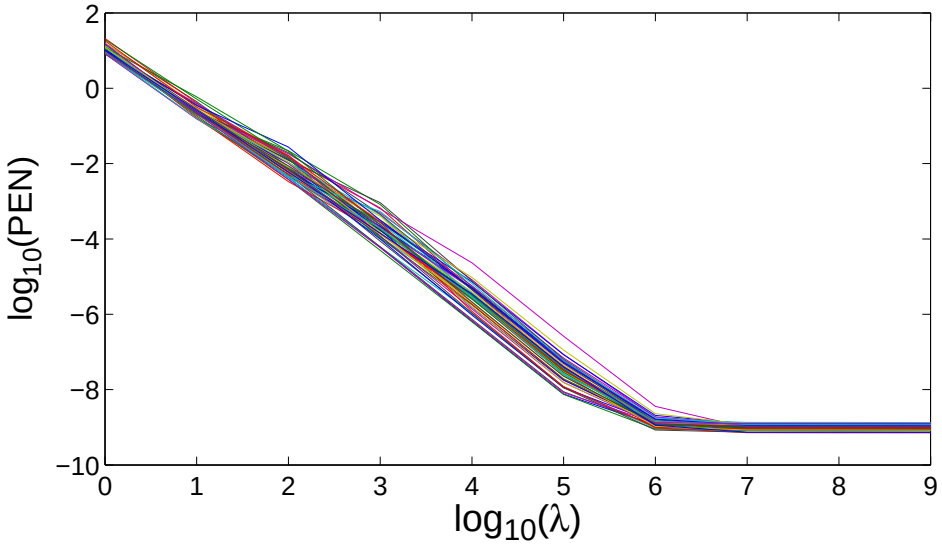

Figure 3–3: The change in $\log_{10}(\texttt{PEN})$ with $\log_{10}(\lambda)$ for the 50 simulated FitzHugh-Nagumo data sets.

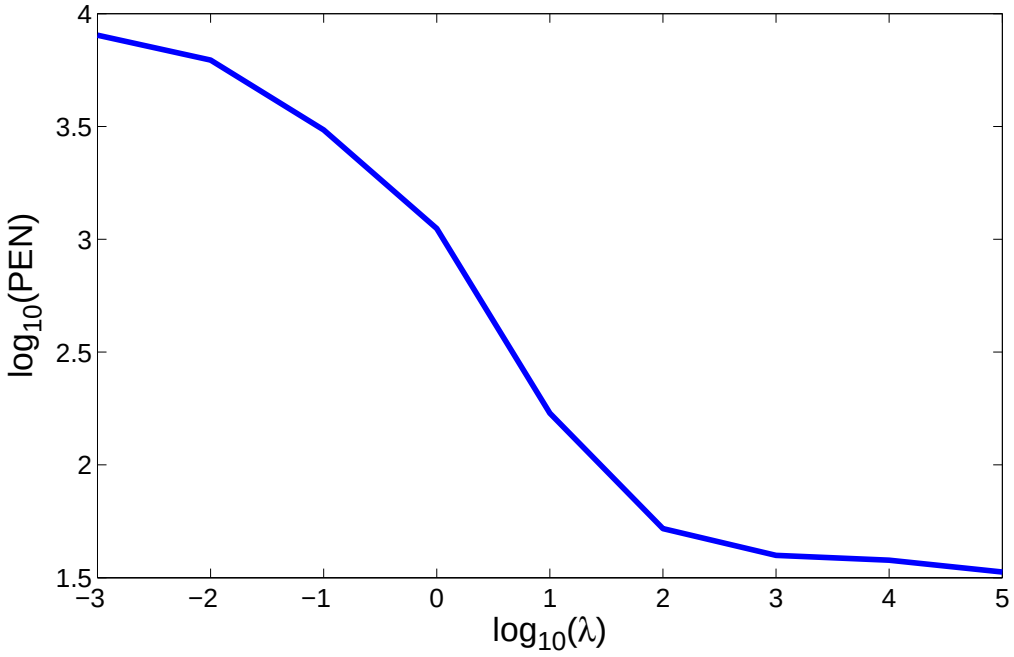

Figure 3–4: The change in $\log_{10}(\texttt{PEN})$ with $\log_{10}(\lambda)$ for the 4 parameter nylon model.

Since profile estimation attempts to fit an optimal balance between the data and the model under the accuracy constraints of the basis, $SSE + \texttt{PEN}$ could be thought of as a composite measure of fit. Choosing an optimal $\log(\lambda)$ then amounts to increasing λ until it minimizes

$$K = \frac{dSSE}{d\log(\lambda)} + \frac{dPEN}{d\log(\lambda)}. \tag{3.12}$$

In other words, increase λ to find the best overall fit to the data but stop before the basis approximation breaks down causing an increase in SSE which can not be offset by the improvement to PEN.

Using (3.12), for the 50 FitzHugh-Nagumo simulated data sets, figure 3–5 shows that $\hat{\lambda} = 10^8$ is optimal. In the 4 parameter nylon example, the optimum occurs at $\hat{\lambda} = 10^3$ as seen in figure 3–6. In the nylon example when $\lambda < 10^1$, the value of K is increasing due to the large jumps in SSE as the smooth moves from interpolating a sparse data set towards fitting a model. This is further enhanced by the model mis-specification inherent in any real data system and compounded by fitting multiple experimental runs. Figures 3–5 and 3–6 were made by taking the difference of $(SSE_i + \texttt{PEN}_i) - (SSE_{i+1} + \texttt{PEN}_{i+1})$ and plotting this quantity at $\log_{10}(\lambda_{i+1})$ as a crude estimate of the effect of having increased from λ_i to λ_{i+1}.

3.4 Overcoming Challenges From The Nylon Data Set

The nylon real data system has many challenges. This section describes how profile estimation overcomes them while highlighting some of the features and details behind this estimation process.

3.4.1 Multiple Experimental Runs

The data smoothing step of (3.1) is performed separately on each experimental run. However, within each run, information is pooled across the $k = 1, \ldots, K$ components to compute $f_{ki}(\mathbf{x}_i, \mathbf{u}_i(t), t, \boldsymbol{\theta})$. Estimation of $\boldsymbol{\theta}(\lambda)$ is performed by pooling data $\mathbf{y}_i(t)$ and smooth fits $\hat{\mathbf{y}}_{ki}(t) = g_k\{\mathbf{c}'_{ki}\boldsymbol{\phi}_{ki}(t)\}$ from all I experimental runs and K observed system components.

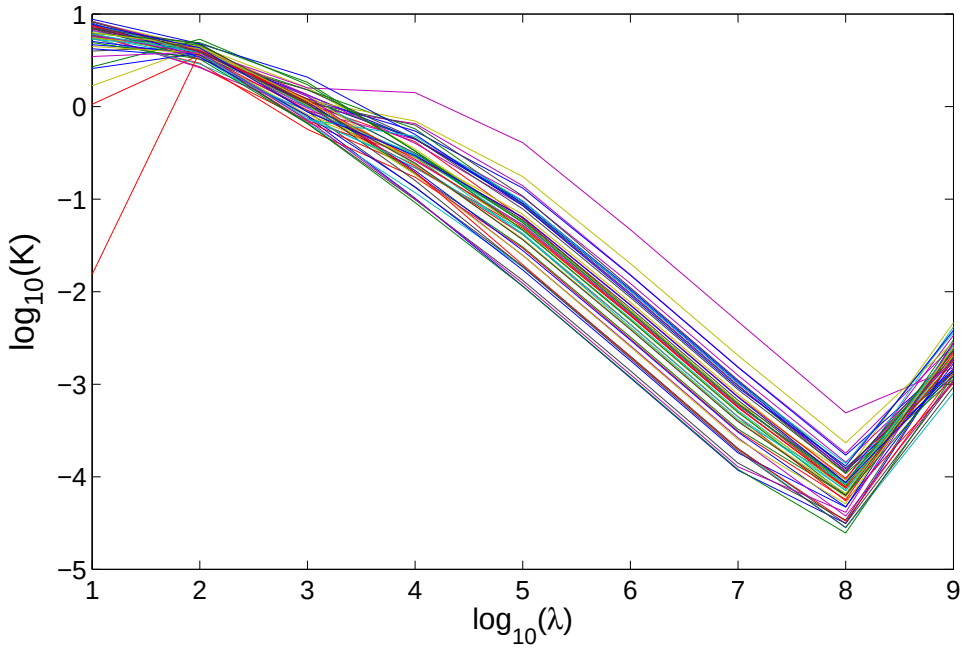

Figure 3–5: The change in composite fitting criteria with changes in $\log \lambda$ for the 50 FitzHugh-Nagumo simulated data sets.

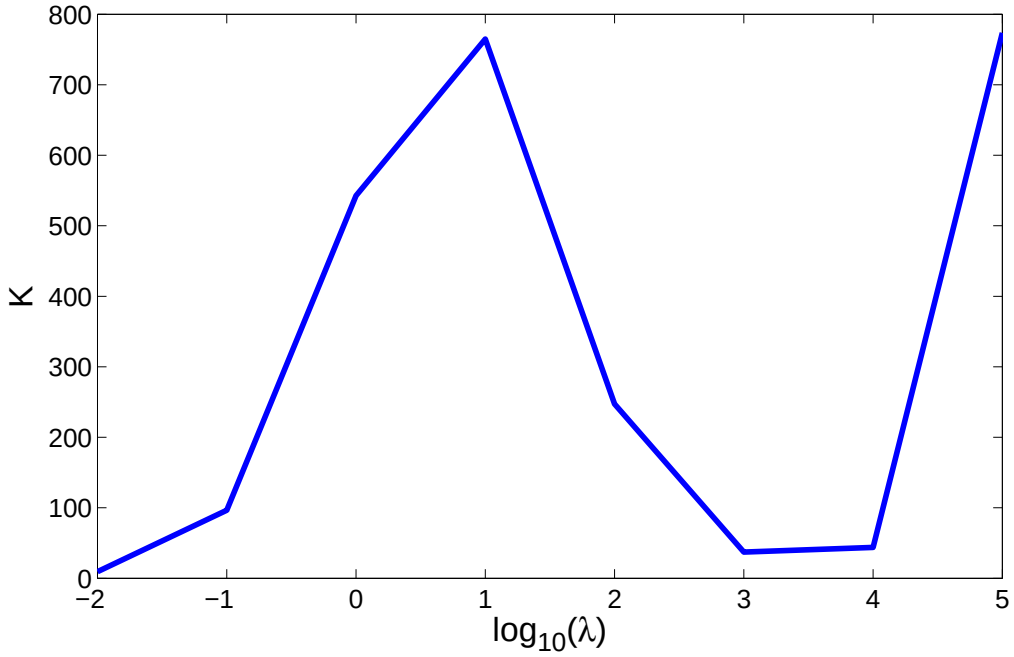

Figure 3–6: The change in composite fitting criteria with changes in $\log \lambda$ for the four parameter nylon data set.

3.4.2 Step Function System Inputs

The integral in (3.2) 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 a run so that information from each component is available to compute $f(\mathbf{x}_i, \mathbf{u}_i(t), t, \boldsymbol{\theta})$ over the same quadrature grid. When using b-spline bases for all components, Ramsay et al. (2007) suggests creating the quadrature grid by dividing $\mathbf{T}_i$ into a set of small intervals whose boundaries are the unique knot locations compiled over the bases of all K components within an experimental run. Denoting the location of the ℓ^{th} such point by ξ_ℓ, 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.

At points of discontinuity in the first derivative $\boldsymbol{\tau}_i$, such as when the system inputs undergo step changes, the derivatives in $\texttt{PEN}_{ki}$ are undefined. The integration avoids these points by removing a small δ sized neighbourhood denoted $\boldsymbol{\tau}_i^\delta$ around them giving the PEN term

$$\begin{aligned}
\texttt{PEN}_{ki} &= \textstyle\int_{\mathbf{T}_i \cap \bar{\boldsymbol{\tau}}_i^\delta} \left(D\mathbf{x}_{ki}(s) - f_k(\mathbf{x}_i, \mathbf{u}_i, s \mid \boldsymbol{\theta})\right)^2 ds \\
&= \textstyle\int_{T_i^{\min}}^{\tau_{i,1}-\delta/2} \left(D\mathbf{x}_{ki}(s) - f_k(\mathbf{x}_i, \mathbf{u}_i, s \mid \boldsymbol{\theta})\right)^2 ds \\
&+ \textstyle\int_{\tau_{i,1}+\delta/2}^{\tau_{i,2}-\delta/2} \left(D\mathbf{x}_{ki}(s) - f_k(\mathbf{x}_i, \mathbf{u}_i, s \mid \boldsymbol{\theta})\right)^2 ds \\
&+ \textstyle\int_{\tau_{i,2}-\delta/2}^{T_i^{\max}} \left(D\mathbf{x}_{ki}(s) - f_k(\mathbf{x}_i, \mathbf{u}_i, s \mid \boldsymbol{\theta})\right)^2 ds.
\end{aligned} \tag{3.13}$$

This integral is approximated by shifting the quadrature interval boundaries at times $\boldsymbol{\tau}_i$ to the points defining the boundaries of $\boldsymbol{\tau}_i^\delta$ and omitting quadrature weights across $\boldsymbol{\tau}_i^\delta$.

If $\boldsymbol{\tau}_i^\delta$ is not omitted from the integral in (3.1), minimizing the discrepancy between the smooth and the ODE, as λ increases, may amount to minimizing the discrepancy between the right and left hand derivatives at $\boldsymbol{\tau}$. This implies that the criteria flattens out the resulting smooth which in turn pulls $\boldsymbol{\theta}$ in the outer optimization away from its true value. If the neighbourhood is too large, the basis will be able to push an extremely poor fit into the neighbourhood which effectively allows a discontinuous smooth across $\boldsymbol{\tau}_i^\delta$. For this reason a small neighbourhood,

such as $10^{-6} \times \min_\ell(\xi_{\ell+1} - \xi_\ell)$ while ensuring that no observations fall within this neighbourhood, appears adequate.

If an observation occurs at the $\boldsymbol{\tau}_i$, a small shift in the location of the $\boldsymbol{\tau}_i$ may be required to avoid this problem from occurring, although care must be taken to check the impact on the final results of moving $\boldsymbol{\tau}_i$ forward or back. Alternatively the left or right hand derivatives could be used at times $\boldsymbol{\tau}_i$

3.4.3 Outputs Measured With Different Precision

In ODE systems, often components are measured in different units, scales and precisions consequently it is important for the optimization process that weights w_{ki} in (3.3) and (3.1) be chosen to bring the residual loss of all system components to approximately the same scale. In some cases this may include using a vector of weights $\mathbf{w}_{ki}$ which allow observations to be weighted to accommodate autocorrelations in the data. For a Gaussian likelihood, the ideal choice of weights is $w_{ki} = \sigma_{ki}^2$, the inverse of the measurement error variance. 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 $n = 0$ initialize $w_{ki}^{(n)} = 1$ or use another value consistent with prior information.
2. Perform the profile estimation to obtain $\hat{\mathbf{c}}, \hat{\boldsymbol{\theta}}$ using weights $w_{ki}^{(n)}$.
3. Using the fitted data values $\hat{\mathbf{y}} = g\{\mathbf{c}(\boldsymbol{\theta}, \lambda)'\boldsymbol{\phi}\}$ obtain the vectors of residuals $\mathbf{r}_{ki} = \hat{\mathbf{y}}_{ki} - \mathbf{y}_{ki}$ and estimate new weights $\hat{w}_{ki}^{(n+1)} = \left(\frac{n_{ki}}{\mathbf{r}_{ki}'\mathbf{r}_{ki}}\right)$, the inverse of the residual variance estimate.
4. If $\mid \hat{w}_{ki}^{(n+1)} - \hat{w}_{ki}^{(n)} \mid > \epsilon$ for some convergence tolerance $\epsilon > 0$, return to step 2.

This is process is described further in the context of the nylon system in section 3.5.4.

3.4.4 Unobserved Outputs

When system components $\mathbf{x}_{ij}, \ldots, \mathbf{x}_{ki}$ are unobserved, their smooth estimate from (3.1) is the solution to the differential equation but data regularized through $f(\mathbf{x}, \mathbf{u}(t), t, \boldsymbol{\theta})$ in the `PEN` term by the smooth of the observed components. This is equivalent to setting $\lambda \to \infty$ for the unobserved components within their experimental runs or estimating the initial conditions to produce the best numerical solution to the ODE constrained to fit the observed components.

3.5 Nylon Results and model selection

The positively constrained data smooth is derived from a fifth order B-spline basis with knots at each observation of component A was used to fit the nylon data of section 1.1. This basis was expanded with additional knots at a rate of 5 per hour of the experimental duration removing any non unique knots resulting from these two knot placement steps. Furthermore, additional knots at the times of step function changes in inputs were included to allow for the discontinuity in the first derivative of the smooth arising in the model when W_{eq} undergoes a step change. Whenever C was observed there is also an observation of A, although the opposite is not true in all experimental runs. Consequently, this same set of knots was used for C and unobserved W. This strategy produced between 56 and 88 unique interior knots per component. The inverse of the measurement error, as determined from previous experiments and used in (Zheng, McAuley, Marchildon, and Zhen Yao 2005), were used as weights $w_A = 1/\sigma_A^2 = 1/0.6^2$ and $w_C = 1/\sigma_C^2 = 2.4^2$ in (3.1).

Profile estimation was initialized with parameter values used by Zheng et al. (2005) or approximations to those parameters based on the discrepancy between parameterizations. Initially $\lambda = 10^{-3}$ was used in the inner-optimization criteria of (3.1) until the profile estimation converged as determined by a relative drop in SSE from one additional Gauss-Newton step of less than 10^{-8}. The profiling process was then repeated with $\lambda_{new} = \lambda_{old} * 10$ and initialized with $\boldsymbol{\theta}_{new}^{(0)} = \boldsymbol{\theta}_{old}^{(final)}$, the final parameter values obtained with λ_{old}. This process was repeated, increasing λ by integer increments on the $\log_{10}$ scale until the optimal value of λ was found.

As the nylon example uses real data there is likely to be some model mis-specification error and consequently the model is scrutinized in the following subsections describing the results.

3.5.1 Profile Estimation and the Six Parameter Nylon Model

Initially a 6 parameter model with $\boldsymbol{\theta} = [k_{p0}, E, \gamma, \beta, K_{a0}, \Delta H]$ was examined for the nylon data. This model,

$$\begin{aligned} -DL = DA = DC &= -k_p(CA - LW/K_a), \\ DW &= k_p(CA - LW/K_a) - 24.3(W - W_{eq}), \\ k_p &= \tfrac{k_{p0}}{1000}\ell\left(\tfrac{E}{8.314}\right), \\ K_a &= \left\{1 + W_{eq}\tfrac{\gamma}{1000}\ell(\beta)\right\} K_T K_{a0}\ell\left(\tfrac{\Delta H}{8.314}\right), \\ \ell(m) &= \exp\left(-m10^3\left\{\tfrac{1}{T} - \tfrac{1}{T_0}\right\}\right) \\ \text{and } K_T &= 20.97\exp\left(-9.624 + \tfrac{3613}{T}\right), \end{aligned} \tag{3.14}$$

is slightly altered from the model in Zheng et al. (2005) in order to re-scale the initial parameter estimates to the same order of magnitude. Furthermore the equation for K_a in (3.14) was slightly adjusted to fix an inconsistency in the units of the equations.

Using the profile estimation strategy described in section 3.5 found $\hat{\lambda} = 10^3$. Parameter estimates were nearly identical to the point estimates to corresponding parameters via NLS in Zheng et al. (2005). Their initial system states were assumed to be known without error but related to observed quantities in the data and the assumption that the system was at steady state. However, generalized profile parameter estimates were highly dependent on the initial parameter estimates. Additionally, by changing the convergence tolerance values of the outer optimization in (3.3), the parameter estimates destabilized and produced even more erratic results with changes in initial parameter estimates. Over numerous estimation attempts, the final parameter estimates contained one pair of parameters with extremely strong correlations ($r^2 > .99999$). While the actual pair of parameters changed with choice of initial parameter estimates, one of the parameters was always β and the other was either ΔH or γ. In equation (3.14) parameters γ and ΔH are essentially main effect terms for W_{eq} and T respectively, while β acts as an interaction between T and W_{eq}. In this model the interaction parameter is excessively correlated with one of the main effect terms to the point where β can not be properly identified. This may be due to insufficient data immediately after the step function input change in W_{eq} to properly

identify the characteristic spike in A and C. This spike, described in section 1.1, is caused by including a dependency on W_{eq} in K_a. The β parameter effectively allows the spike to vary in size with changes in temperature, however without observations to measure the spike, this amounts to having multiple parameters to identify the temperature dependent rate of decay in A and C. For this reason it makes sense to retry the model with β set to 0 giving the reduced model in section 3.5.2

3.5.2 Profile Estimation and the Five Parameter Nylon Model

The six parameter model (3.14) is reduced by replacing the expression for K_a with

$$K_a = \left\{1 + W_{eq}\frac{\gamma}{1000}\right\} K_T K_{a0} \ell \left(\frac{\Delta H}{8.314}\right). \tag{3.15}$$

Using the same estimation process as described in section 3.5, the 5 parameter model gave stable estimates with respect to initial parameter estimates from Zheng et al. (2005) or random draws from a uniform $[0, 100]$. Furthermore, removing the high correlation with β sped up the optimization by a factor of 20.

In a relationship nearly identical to the one in figure 3–6, the optimal smoothing parameter was chosen to be $\hat{\lambda} = 10^3$. Figure 3–7 shows how the parameter point and interval estimates change with $\log_{10}(\lambda)$. All of the point and interval estimates are most stable with respect to $\log_{10}(\lambda)$ between values of 1 and 4 with the smallest change overall occurring between values of 2 and 3. The jump in $\hat{\boldsymbol{\theta}}$ at $\lambda > 10^3$ is due to the bias induced by the smooth attempting to reduce its inability to match the ODE model to the accuracy defined by λ at the detriment to the fit to the data. Consequently, the behaviour shown in figure 3–7 is consistent with the behaviour of K in figure 3–6 which decreases and stabilizes over the same interval but begins to increase after $\lambda = 10^3$. The actual best value of λ is a function of the number and resolution of data points, basis functions and the actual differential equations in the model.

The interval estimate for E in figure 3–7 overlaps zero suggesting that a simpler model would be just as effective at fitting the data. Setting E to zero removes the temperature dependence from k_p. This model and the final results are described next.

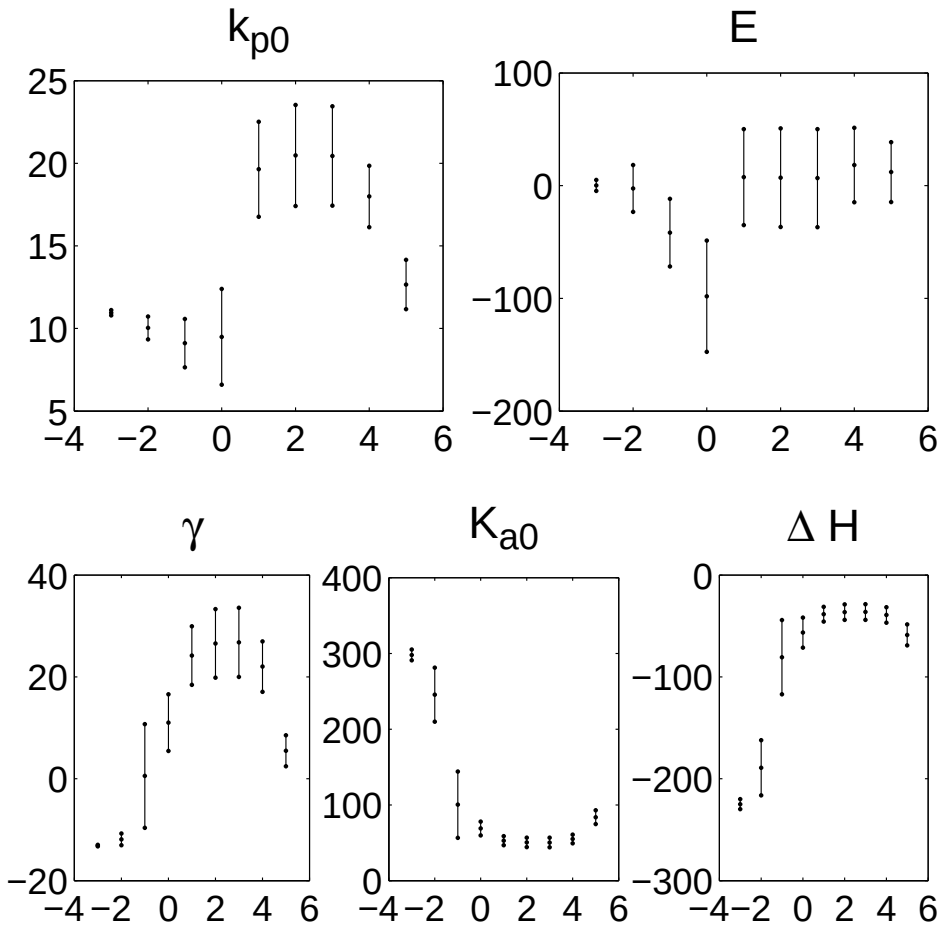

Figure 3–7: 95% confidence intervals for the parameters of the 5 parameter model as a function of the smoothing parameter λ. Horizontal axis is in units of $\log_{10}(\lambda)$ and the vertical axis units are specific to the parameter. Parameters in the top row are used to estimate k_p and in the bottom row are used to estimate K_a in equations (3.14) revised with (3.15).

3.5.3 Profile Estimation and the Four Parameter Nylon Model

The 4 parameter nylon model,

$$\begin{aligned} -DL = DA = DC &= -\tfrac{k_{p0}}{1000}(CA - LW/K_a,) \\ DW &= \tfrac{k_{p0}}{1000}(CA - LW/K_a) - 24.3(W - W_{eq}), \\ K_a &= \left\{1 + W_{eq}\tfrac{\gamma}{1000}\right\} K_T K_{a0} \ell\left(\tfrac{\Delta H}{8.314}\right), \\ \ell(m) &= \exp\left(-m10^3\left\{\tfrac{1}{T} - \tfrac{1}{T_0}\right\}\right), \\ \text{and} \quad K_T &= 20.97\exp\left(-9.624 + \tfrac{3613}{T}\right), \end{aligned} \tag{3.16}$$

is reduced to have a single temperature dependent parameter, compared to the three Arrhenius type of temperature relationships for parameters β, E and ΔH in (3.14).

The profile estimation routine of section 3.5 was performed with the initial parameter estimates of Zheng et al. (2005) but the process was robust to this choice over a wide range of values. Some particularly bad choices of initial estimates were also attempted, such as $k_{p0}, K_{a0} < 0$ or $\mid \gamma \mid$ large. Allowing negative values in the rate parameters k_{p0} and K_{a0} produces positive feedback loops suggesting chemical components are self replicating without bound. These estimation attempts required an additional reduction in the initial choice of λ however if a sufficiently small value was chosen, the data prevailed in the initial smooth and those parameter values quickly moved the model back to the negative feedback system. Through incremental increases in λ, these attempts still converged to the same final result. However, the problem caused by initializing γ far from zero was not so easily fixed. As $\mid \gamma \mid$ increases, $\left\{1 + W_{eq}\frac{\gamma}{1000}\right\} K_{a0} \to \left\{W_{eq}\frac{\gamma}{1000}\right\} K_{a0}$. The lack of identifiability between γ and K_{a0} was not resolved by a reduction in initial λ. Once parameters become unidentifiable, the damage appears to be irreparable.

Figure 3–6 shows the behaviour of the composite fitting criteria K from (3.12) with respect to λ. This relationship did not change substantially in the model reduction process. The behaviour of K is echoed in figure 3–8, which shows the point and 95% interval estimates as a function of $\log_{10}(\lambda)$. The optimal estimates barely change across the window $\lambda \in (10, 10^4)$ suggesting little need to further refine the estimate for $\hat{\lambda} = 10^3$. The final parameter point estimates and the standard errors from $\lambda = 10^3$ are given in table 3–1 along with the estimates using NLS without

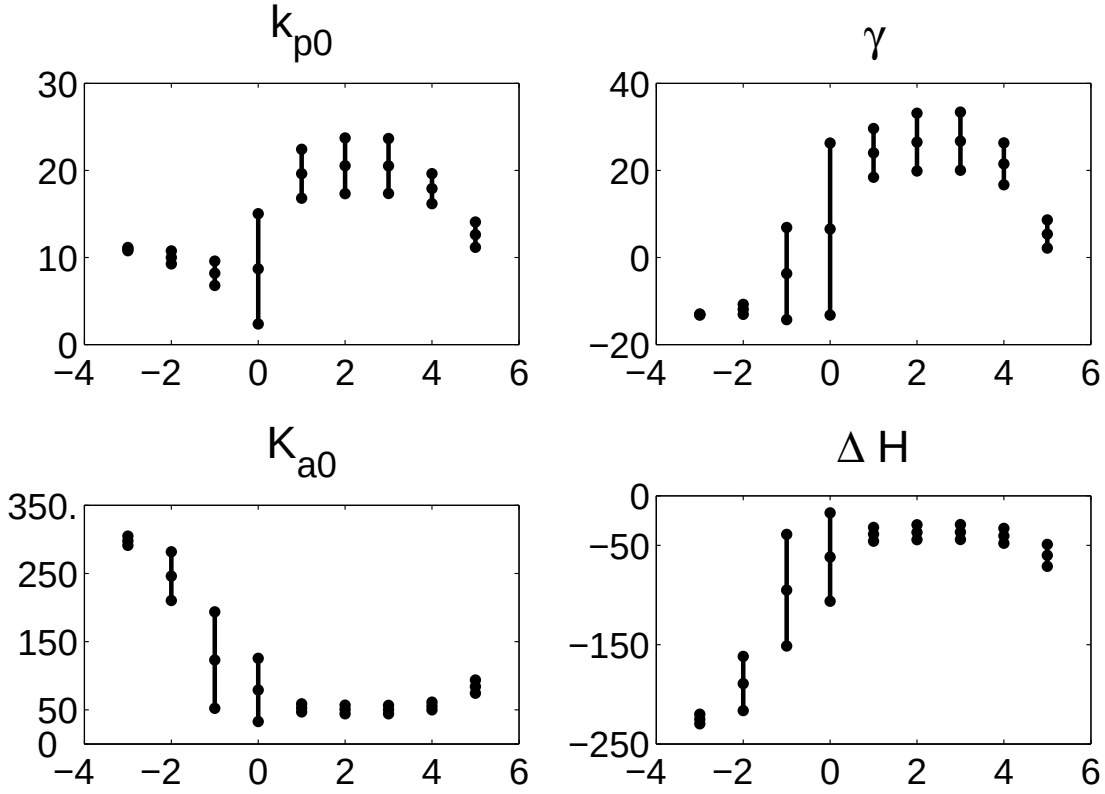

Figure 3–8: 95% confidence intervals for the parameters of the 4 parameter model as a function of the smoothing parameter λ. Horizontal axis is in units of $\log_{10}(\lambda)$ and the vertical axis units are specific to the parameter.

parameter	$\hat{\boldsymbol{\theta}}$	SE($\hat{\boldsymbol{\theta}}$)
k_{p0}	20.59 (20.47)	1.66 (1.30)
γ	26.86 (25.82)	3.48 (2.90)
K_{a0}	50.22 (50.98)	3.24 (2.79)
ΔH	-36.46 (-37.83)	3.86 (4.17)

Table 3–1: 95% confidence intervals and point estimates for the generalized profile estimated parameters using $\lambda = 10^3$. The values obtained using NLS without constraint on the initial system states are shown in brackets.

constraint on the initial system states. Recall from section 2.1, using NLS either produced negative estimates for the $\hat{W}_0$ in 4 out of the 6 experimental runs or under the non-negative constraint, the optimization broke down. The influence of the weights is further explored in section 3.5.4.

The parameter correlation matrix for the profile estimation is

$$\begin{array}{cccccc} & k_{p0} & \gamma & K_{a0} & \Delta H \\ \gamma & 0.441 & 1 & & \\ K_{a0} & -0.407 & -0.981 & 1 & \\ \Delta H & 0.219 & 0.446 & -0.358 & 1. \end{array} \tag{3.17}$$

Correlations between K_{a0} and γ remain strongly negative because γ is moderately large and approaching the brink of unidentifiability. The overall data fit is shown in figure 3–9. This figure shows that the residuals within a system component of an experimental run tend to be autocorrelated. This suggests a deterministic block effect on the experimental runs or some other mild form of model mis-specification but the overall fit to the data features is quite good.

The data fit shown in figure 3–9 is the solution to the differential equation using the initial system state estimated by the data smooth and the final parameter estimates from the four parameter model. At the scale of this figure, there is no clear difference between the smooth at $\lambda = 10^3$ and the solution to the ODE. Figure 3–10 highlights the difference between the smooth fit to the data using $\lambda \in \{10^2, 10^3, 10^4\}$ and the numerical solution to the ODE from the initial system state estimates from the data smooth at time zero. The integral of the square of these plotted curves is equal to $\texttt{PEN}_{ki}$ used in (3.1). Since there are no observations for W, the quality

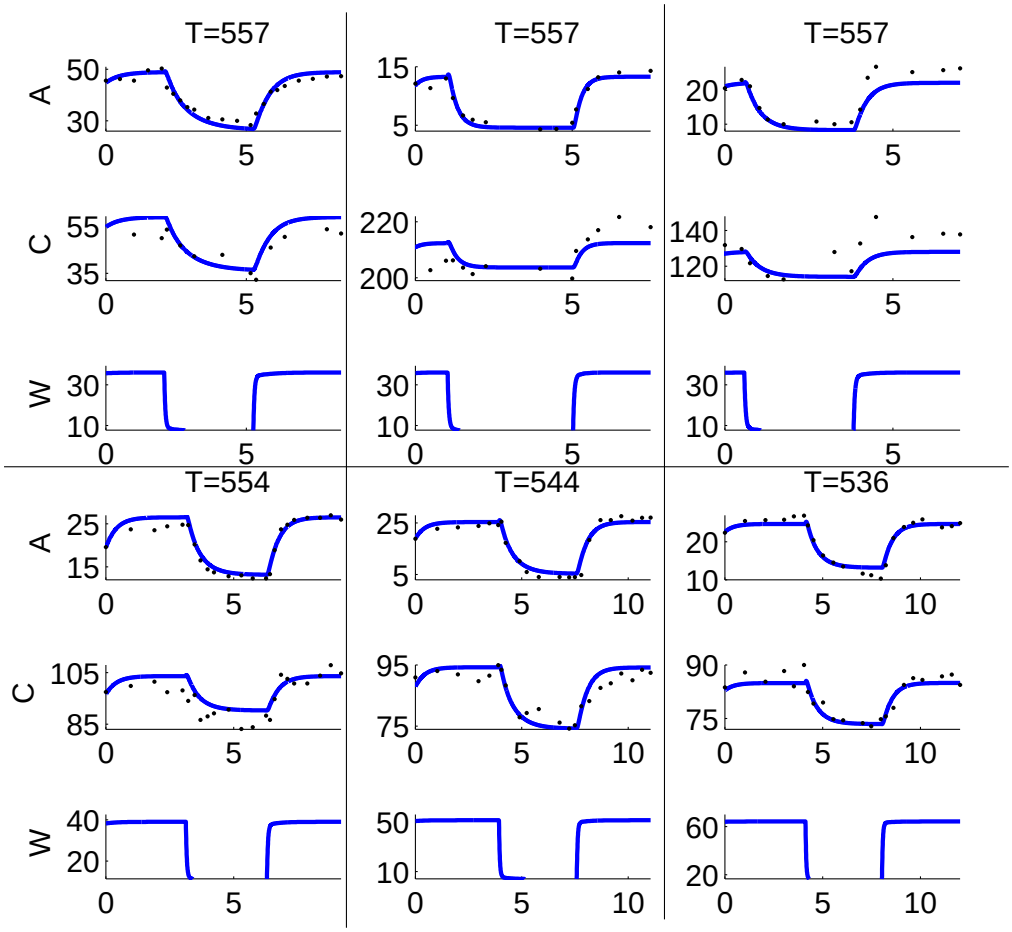

Figure 3–9: The Data fit from the four parameter nylon model using generalized profile estimation.

of the fit at all three λ values is essentially constant. The discrepancies between the smooth and the ODE solutions are notably spiked. This feature represents the inability of the basis to accommodate the short term spike suggested by the ODE model following the step input change in W_{eq}, as discussed in section 1.1. The omitted δ-neighbourhoods around the times of step input change, shown as red lines on the figure, are much too small compared to the basis resolution for the lack of fit between the two models to be pushed within these ignored regions of the `PEN` integral. However with too large a δ-neighbourhood this may have become a problem.

The blue line in figure 3–10 representing the discrepancy for $\lambda = 10^2$ strays the furthest from 0 and deviates for the longest duration. This is especially notable for C where the smooth compensates for the smaller relative weights compared to A. For example in the middle experiment in the top half of the figure, the blue line for C strays above zero as the smooth at this λ attempts to accommodate the larger than expected series of observations. In general, the compensation for the preferential fit to A is seen by the larger deviation of the blue line from zero for C compared to A. When $\lambda = 10^3$, shown in green, stronger emphasis is placed on the fit to the ODE leaving only the spikes and their recovery as the notable discrepancy. Increasing to $\lambda = 10^4$ as shown in black produces very little improvement in this `PEN` term. However to produce this improvement the smooth moves detrimentally away from the data outweighing the improved `PEN` with a jump in SSE. To put these relative discrepancies into perspective the vertical scale of this figure is on the order of 1% of that of the data fit in figure 3–9.

3.5.4 Iteratively Re-weighted Profile Estimation for the Nylon System

Iterative re-weighting of section 3.4.3 was also attempted to accommodate potential external factors influencing the measurement dispersion. The following 3 schemes were applied to estimate the weights and compare their impact:

- Method 1) The assumed fixed weights of Zheng et al. (2005), $w_a = 1/\sigma_A^2 = 1/.6^2$ and $w_c = 1/\sigma_C^2 = 1/2.4^2$ were used in the profile estimation process as performed in section 3.5.3.

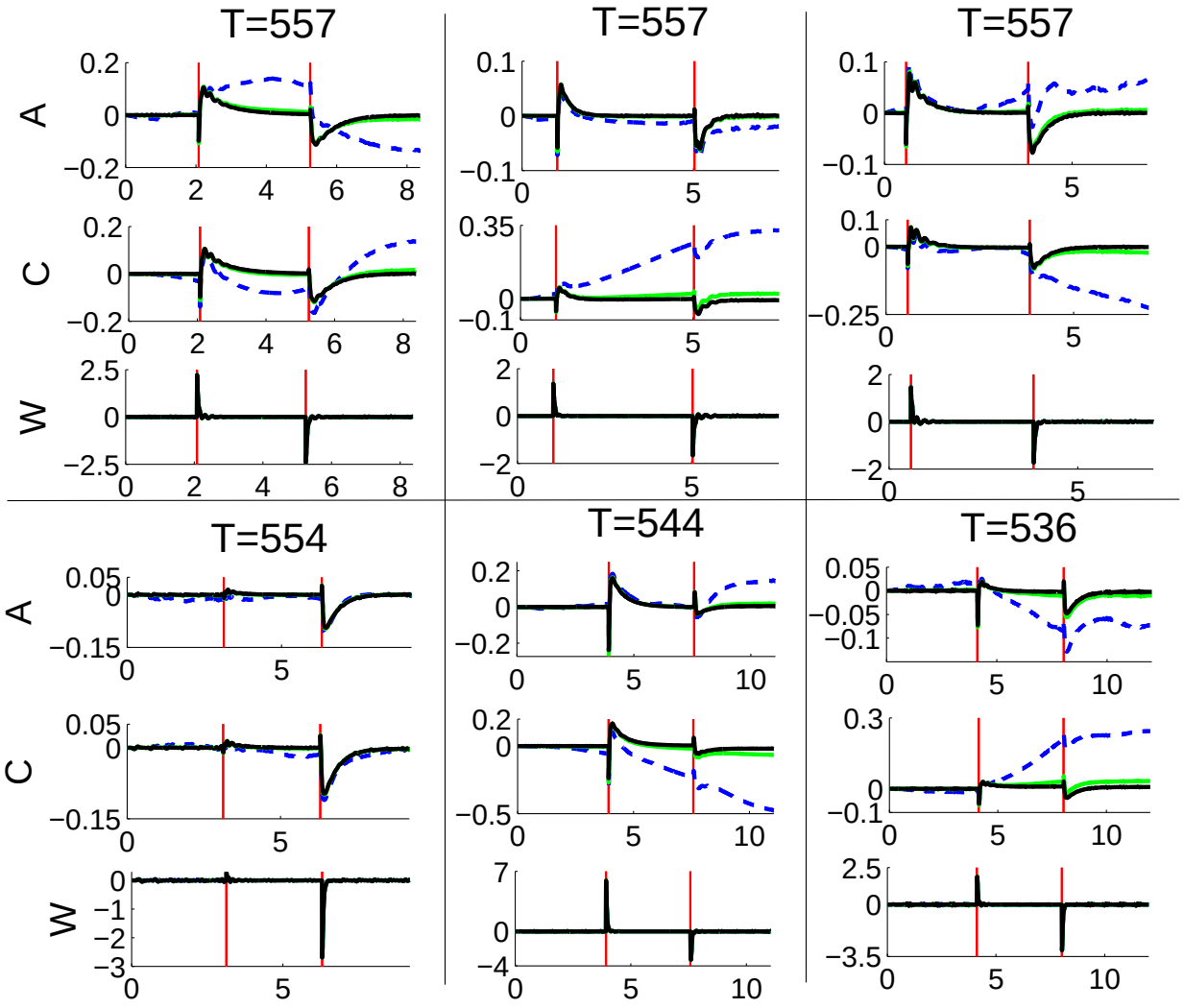

Figure 3–10: The discrepancy between the data smooth and the solution to the ODE system using the final estimates of the 4 parameter model and initial system states equal to the values of the data smooth at time 0. The blue lines are the results for $\lambda = 10^2$, green is for $\lambda = 10^3$ and black represents $\lambda = 10^4$. Red lines denote the times of the changes in input P_w.

- Method 2) Maintaining the weight structure of model 1, the relative accuracy of measurements A and C are estimated. Iterative re-weighting was performed to estimate weights $\hat{w}_A$ and $\hat{w}_C$ pooling all information across the $i = 1, \ldots, 6$ experimental runs. Estimates were $\hat{w}_k = 1/\hat{\sigma}_k^2$, $k \in \{A, C\}$, where

$$\hat{\sigma}_k^2 = \left[\sum_{i=1}^{6} \sum_{t \in t_{ki}} \left(y_{ki}(t) - \mathbf{c}'_{ki} \boldsymbol{\phi}(t) \right)^2 \right] / \sum_i n_{ki}.$$

- Method 3) Iterative re-weighting was performed to estimate the 12 weights $\hat{w}_{ki} = 1/\hat{\sigma}_{ki}^2$, $k \in \{A, C\}$ unique to each system component- experimental run combination:

$$\hat{\sigma}_{ki}^2 = \left[\sum_{t \in t_{ki}} \left(y_{ki} - \mathbf{c}'_{ki} \boldsymbol{\phi} \right)^2 \right] / n_{ki}.$$

Iterative re-weighing was performed by incorporating a weight estimation step into the estimation process of 3.5 as described in section 3.4.3, except that λ was adjusted along with the weights. Initially $\lambda = 10^{-3}$ was chosen and the profiling process was initialized with the assumed weights suggested in Zheng et al. (2005). After $\hat{\boldsymbol{\theta}}(\lambda)$ converged, weights were updated using the estimate of method 2 or method 3 above. Using these new weights, the smoothing parameter was updated by $\lambda_{new} = \lambda_{old} * 10$ and the profile method performed again. The weights were updated and λ increased incrementally until $\lambda = 10^3$. Further iterations to refine weights were performed maintaining λ at this maximum level, known to perform well with this basis and data set, until the weights converged as assessed by $\max_{ki} \mid (\hat{w}_{ki}^{(new)} - \hat{w}_{ki}^{(old)}) \mid < 10^{-4}$.

This estimation process is outlined in table 3–2 for method 2. The weights converged after 10 iterations to $\hat{\sigma}_A = 2.1056$ and $\hat{\sigma}_C = 4.2637$. The estimated ratio of weights was $\hat{w}_C/\hat{w}_A = \hat{\sigma}_A^2/\hat{\sigma}_C^2 = 2.1056^2/4.2637^2 = 0.2438$ compared to the ratio in method 1 of $w_C/w_A = 0.6^2/2.4^2 = 0.0625$. This suggests that in method 2 the observations of C were deemed relatively more precise compared to A, than suggested by method 1. Recall that the weights in method 1 were determined through repeated measurements in additional experiments which were therefore not subject to model mis-specification. Consequently, the relative size of weights in method 2 may

Iteration	1	2	3	4	5	6	7	8	9	10
$\hat{\sigma}_A$	0.6	1.046	1.870	2.052	2.094	2.103	2.105	2.105	2.106	2.106
$\hat{\sigma}_C$	2.4	3.970	4.810	4.360	4.294	4.271	4.266	4.264	4.264	4.264
λ	1	10	100	1000	1000	1000	1000	1000	1000	1000

Table 3–2: Iteratively updating weights for the four parameter nylon model.

be more an indication of the relative ability of the model to accommodate the features in A and C.

The fit to the data using method 1 and method 2 appears in figure 3–11. This figure shows the numerical solution to (3.16) using the estimated initial system states from the smooth fit at the optimal $\lambda = 10^3$. As expected by the relative weights, the main impact is a shift in the relative importance of fitting C. In the experiments plotted in the top right and bottom left corners of the figure, the impact of fitting C more closely is evident by the shifted fits. The final parameter estimates from method 2 were slightly influenced by changes in weights as reported in table 3–3. The discrepancy in parameters is also shown in figure 3–11. For example in the bottom left panel of the figure, the estimated initial system state is nearly identical using both weighting schemes, however the altered reaction rate parameters from method 2 allow an improved fit to C.

Method 3 allows for the possibility that there is additional variability in the measurement of a component from a particular experimental run. Essentially, this model of iterative re-weighting determines the largest subset of the functional data observations which could best be fit by the model. Allowing unique weights to each experimental run-system component combination highlights functional observations which might be considered influential outliers, and therefore are less likely to be adequately fit by the model. These functional outliers may be due to model mis-specification or measurement errors. If both A and C in a single experimental run are heavily down-weighed it may be indicative of model mis-specification errors, possibly fixed by including a block effect due to the day that the experimental run was performed. If a single component within a run is strongly down-weighed it may suggest correlated measurement errors induced by the process of taking measurements.

Parameter	lower	point	upper
Method 1 (Fixed weights)			
k_p	23.849	20.587	17.325
γ	33.679	26.859	20.039
K_{a0}	56.568	50.222	43.876
ΔH	-28.887	-36.462	-44.036
Method 2 (2 weights)			
k_p	23.400	20.432	17.463
γ	31.238	25.216	19.193
K_{a0}	58.531	52.483	46.434
ΔH	-26.885	-34.238	-41.591
Method 3 (12 weights)			
k_p	20.012	18.216	16.413
γ	22.891	19.443	15.995
K_{a0}	64.618	60.10	55.590
ΔH	-26.378	-31.593	-36.808

Table 3–3: 95% Confidence intervals for the nylon data using iteratively re-weighted profile estimation and the weights suggested from additional experiments.

The final parameter values for method 3 are shown in table 3–3. As expected these parameter estimates are further from the method 1 values than the method 2 estimates. Figure 3–11 compares the fit to the data from the ODE solution and the parameter estimates from all three methods. The estimated standard deviations (recall that $\hat{\sigma}_{ki} = 1/\sqrt{\hat{w}_{ki}}$) for method 3 are listed on the figure. Component W is unobserved and therefore only influenced by the weights through their impact on the parameter estimates. However, W is much more strongly governed by the input W_{eq} in (3.16). Consequently the fit to W is omitted from this plot but the sake of reference is shown for method 1 on figure 3–9.

The main impact of method 3 compared to the other weighting schemes is that method 3 nearly ignores the top right experimental run in figure 3–11 by heavy down-weighing. This experimental run has notably large observations for A at times $t \in \{4.25, 4.5\}$ and for C at times $t \in \{3.25, 4, 4.5\}$ which show strong deviation from the model and therefore promote larger weights for this experimental run.

While in general it is to be expected that the experimental runs where the data line up smoothly should have the lowest $\hat{\sigma}_{ki}$, this is only true if this alignment of observations follows

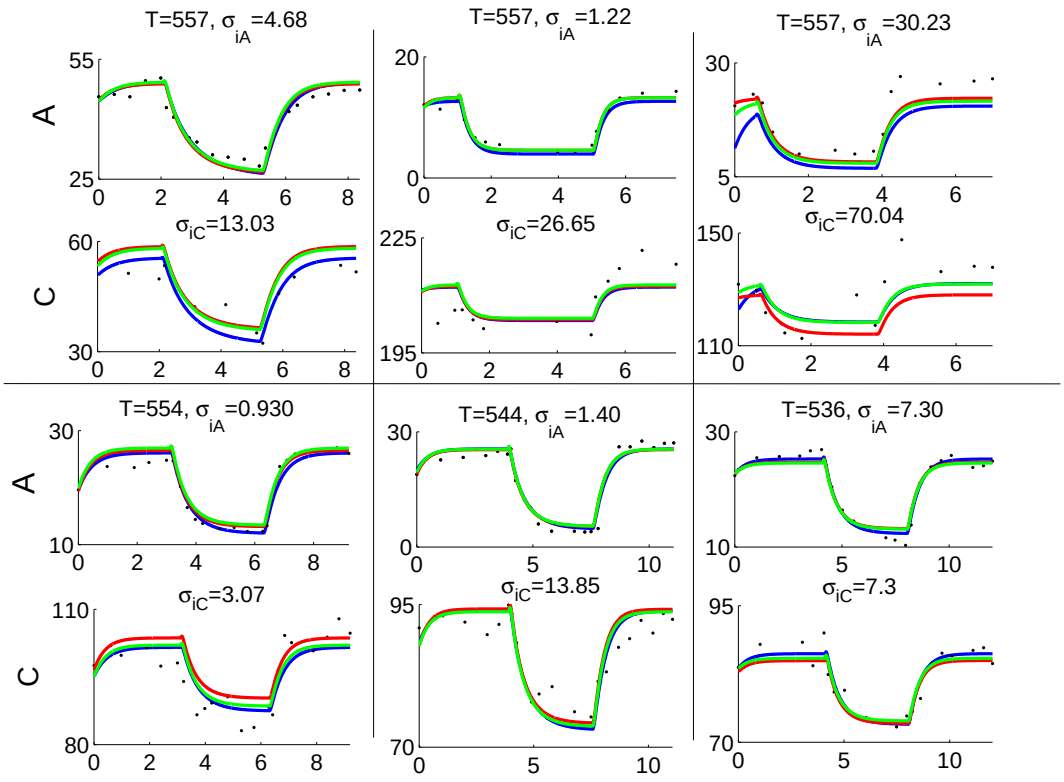

Figure 3–11: The ODE solution fit to the data using the 4 parameter nylon model with 12 iteratively estimated weights (in blue) with 2 iteratively selected weights (in green) and the fit using the assumed weights $w_a = 1/.6^2$ and $w_c = 1/2.4^2$ (in red). The 12 estimated standard deviations are shown on the figure, where $\hat{\sigma}_{ki} = 1/\sqrt{w_{ki}}$. The temperature T is given in degrees Kelvin for the run.

the prescribed model. The observed values of A in the top left panel of figure 3–11 tightly follow a smooth line but this data is given a moderate weight because the parameter estimates causing the model to flow through these points harms the overall fit to the other experiments.

To determine the quality of these estimates and to decide which sets of parameters in table 3–3 should be finally believed and reported, a simulation study is explored in the next section.

3.5.5 Nylon Iterative Re-Weighted Simulation Results

The impact of miss-specifying the weights in the profile estimation process of section 3.5 is explored through a simulation study using the nylon system. One hundred simulated data sets were produced with additive errors to observations from the ODE solutions S_A and S_C as shown in (3.18) under the experimental conditions in the $i = 1, \ldots, 6$ nylon experimental runs.

Temperature:	557	557	557
$\tilde{\sigma}_{1A}$	2.154	0.804	2.162
$\tilde{\sigma}_{1C}$	3.728	5.095	7.517
Temperature:	554	544	536
$\tilde{\sigma}_{1A}$	1.026	2.009	1.295
$\tilde{\sigma}_{1C}$	3.679	2.717	1.874

Table 3–4: The true standard deviations of the noise in the simulated nylon data sets.

The true parameters in this study were $\boldsymbol{\theta}_{true} = [k_{p0}, \gamma, K_{a0}, \Delta H] = [20.59, 26.86, 50.22, -36.46]$, the final parameter estimates from section 3.5.3 using the method 1 weights $w_A = 1/.6^2$ and $w_C = 1/2.4^2$. Numerical solutions $S_{Ai}(\boldsymbol{\theta}, A_{i0}, \mathbf{u}(t), t)$ and $S_{Ci}(\boldsymbol{\theta}, C_{i0}, \mathbf{u}(t), t)$ were obtained using $A_{i0} = \hat{A}_i(t = 0)$ and $C_{i0} = \hat{C}_i(t = 0)$ estimated from the data smooth of the method 1 profile estimation results under the optimal $\lambda = 10^3$. These values were taken to be true for the simulation study producing the observations,

$$\begin{aligned} A_i(t) &= S_{Ai}(\boldsymbol{\theta}, A_{i0}, t) + \epsilon_{Ai}(t), \\ \text{and } C_i(t) &= S_{Ci}(\boldsymbol{\theta}, C_{i0}, t) + \epsilon_{Ci}(t). \end{aligned} \tag{3.18}$$

The variance of the additive random Gaussian noise, $\epsilon_{Ai} \sim N(0, \tilde{\sigma}^2_{Ai})$ and $\epsilon_{Ci} \sim N(0, \tilde{\sigma}^2_{Ci})$, is specific to each component within each experimental run such that method 3 in section 3.5.4 could potentially accurately estimate the true measurement variance structure. The variance of the simulated data $\tilde{\sigma}^2_{ki}$, was based on the centered, estimated residual variance from residuals r_{ki} left over from the profile estimation with method 1 in section 3.5.3,

$$\begin{aligned} \tilde{\sigma}^2_{ki} &= \textstyle\sum_{t=t_1}^{T_{ki}} \left(r_{ki}(t) - \mu_{ki}\right)^2 / n_{ki}, \\ \mu_{ki} &= \textstyle\sum_{t=t_1}^{T_{ki}} \left(r_{ki}(t)\right) / n_{ki}. \end{aligned} \tag{3.19}$$

The simulated data standard deviations used are listed in table 3–4 given in the order that the experimental runs are shown in figure 3–11.

Generalized profile estimation was performed on these 100 simulated data sets under the three different weighting methods used in section 3.5.4 to evaluate the impact of potentially miss-specifying weights. The 95% confidence intervals resulting from the three methods appear in figure 3–12 for all three methods. Method 1, using two assumed and fixed weights, generally

		Method 1			Method 2			Method 3		
Parameter	True value	Bias	var	MSE	bias	var	MSE	bias	var	MSE
k_{p0}	20.47	0.365	2.37	2.50	0.036	1.86	1.86	-0.363	1.03	1.16
γ	26.86	0.241	8.06	8.12	-0.369	6.41	6.54	-1.038	3.95	5.03
K_{a0}	50.22	-0.077	6.80	6.81	0.447	5.59	5.79	0.999	3.81	4.80
ΔH	-36.46	0.186	11.58	11.61	-0.120	10.18	10.19	-0.489	6.29	6.52

Table 3–5: The average parameter bias and observed variance in the point estimates from the 100 simulated nylon data sets.

produced the widest confidence intervals. Method 2, which estimated the two weights, produced noticeably narrower intervals and finally method 3, which estimates one weight for each experimental run-component combination, produced the narrowest intervals. The average parameter estimate bias, variance and mean squared error (MSE) of the observed estimates from the 100 simulated runs are shown in table 3–5. In moving from fixed weights to estimating two weights to 12 estimated weights, the variability in the estimated values of the parameters declines suggesting that the iteratively re-weighting improves stability in the parameter estimates. This is consistent with re-weighting in general which is used to improve robustness of the estimator. While there appears to be an increase in magnitude of the bias in moving from method 1 to method 3, an in depth examination of the reason is left for future work. However the bias appears to be offset by the decrease in MSE.

In method 1 the assumed standard deviation of the data is $\sigma_A = .6$ and $\sigma_C = 2.4$. In Method 2 the inverse square root of the two average weights (an estimate of the standard deviations) were $\hat{\sigma}_A = 1.5831$ and $\hat{\sigma}_C = 3.9188$. The method 3 standard deviation estimated are shown as histograms in figure 3–13, listed in the same order as the experiments are shown in table 3–4 and figure 3–11. The true values are shown in red on the figure and are placed nicely near the middle of the observed density suggesting that iterative re-weighting produced good estimates of the true weights. For reference the observed average is shown in green on the figure.

Due to the reduction in the width of the parameter confidence intervals, higher consistency of the results, reduced MSE and excellent weight estimates when examining the simulated nylon data sets, this suggests that the best set of final parameter estimates to consider for the real

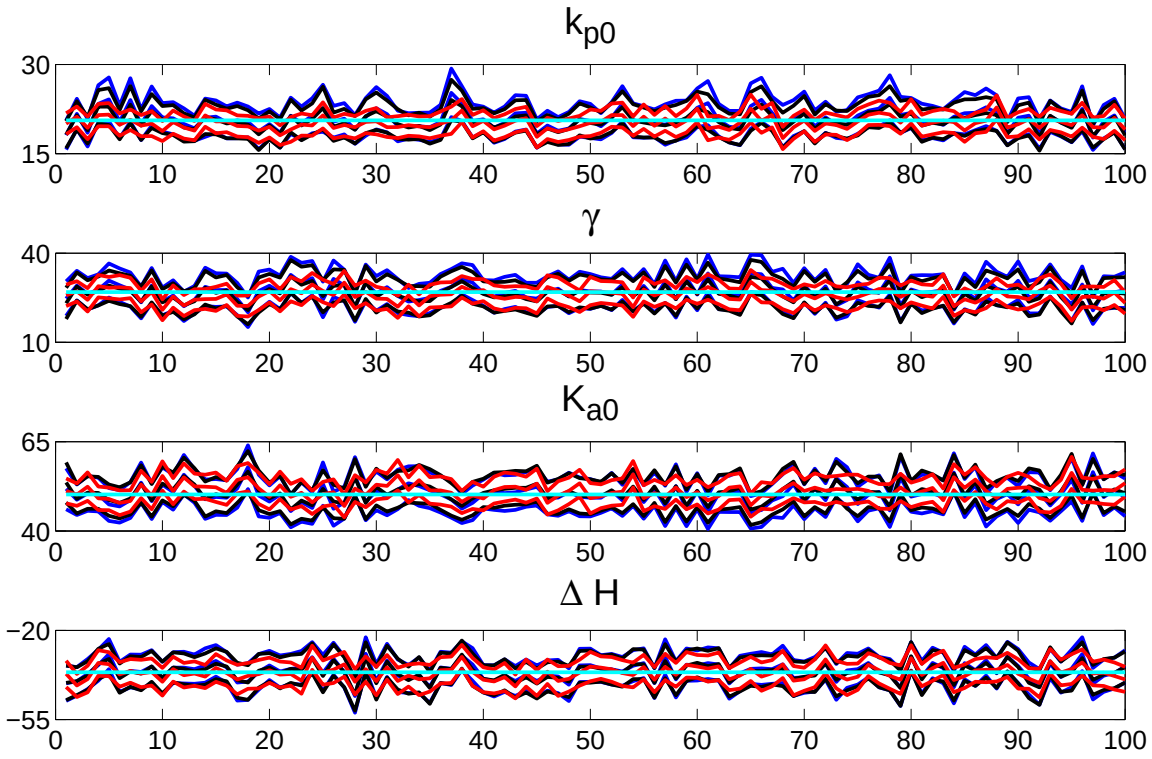

Figure 3–12: The 95% confidence intervals for the four parameters from the 100 simulated nylon data sets using method 1 weights (blue), method 2 weights (black) and method 3 weights (red). The true parameter value is shown in cyan.

nylon data comes from method 3. Furthermore, final weight estimates from method 3 in the real data set are useful indicators of potential functional observations or entire experimental runs worth re-examining due to their small weights.

The parameter estimates from methods 1,2 and 3 are summarized in table 3–5.

3.6 Profile Estimation of the Simulated FitzHugh-Nagumo Data Sets

A cubic b-spline basis with one knot at each of the 399 interior observation times was used for both V and R. Profiling was initialized with $\lambda = 1$ and parameters $\boldsymbol{\theta}^{(0)}$ were the same initial parameter estimates used in previous methods. These originated from draws from the prior densities of the Bayesian estimation process of section 2.2. The profile estimation procedure was run until convergence as assessed by a relative drop in the SSE of less than 10^{-8} from an additional Gauss-Newton iteration. Then the smoothing parameter was updated by $\lambda_{new} = 10 \times \lambda_{old}$ and

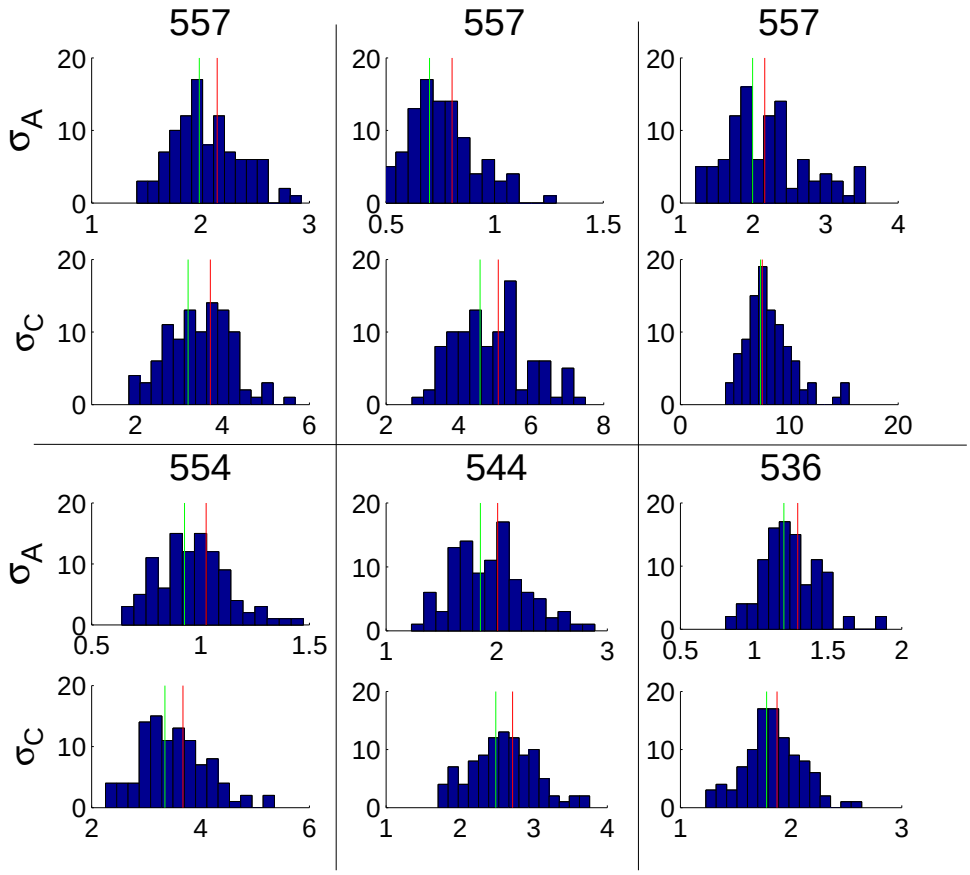

Figure 3–13: Histograms of the standard deviation estimates for the method 3 iterative reweighting. True values are shown as red lines and the mean of the 100 simulated runs is shown in green.

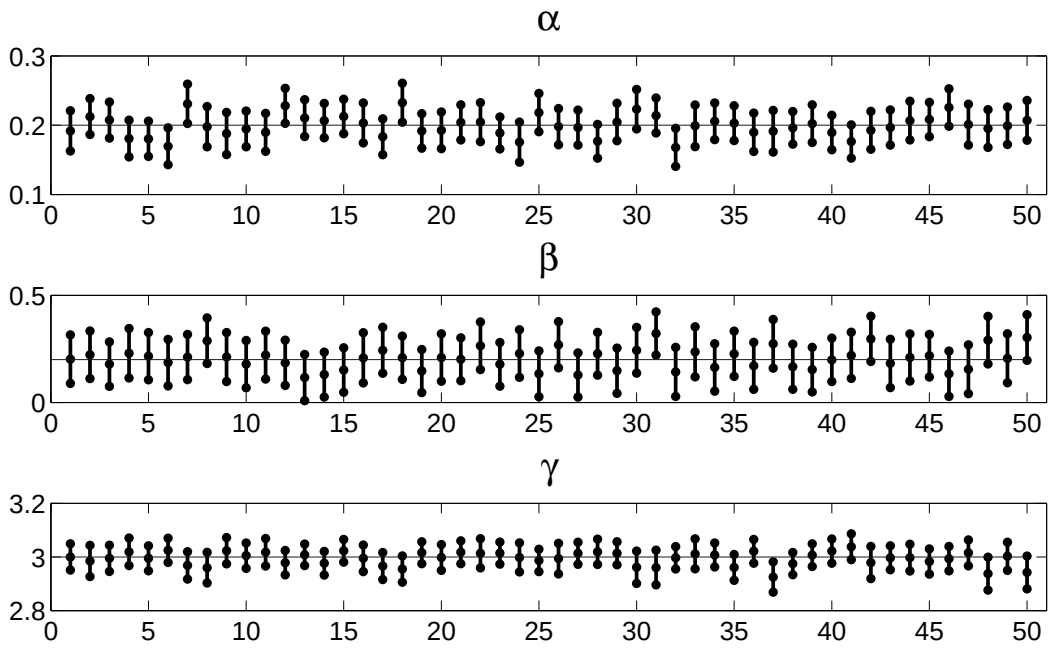

Figure 3–14: 95% Confidence intervals for the profile estimation of the 50 simulated FitzHugh-Nagumo simulated data sets. Horizontal lines mark the true values.

the profile estimation was rerun using the initial parameter estimates $\boldsymbol{\theta}_{new}^{(0)} = \hat{\boldsymbol{\theta}}_{old}^{(final)}$. This was continued until $\lambda = 10^9$ at which point it was deemed that $\hat{\lambda} = 10^8$ as seen in figure 3–5. However, as reflected by the stability of the SSE and `PEN` in figures 3–1 and 3–3 respectively, parameter estimates essentially do not change for any $\lambda \in (10^4, 10^8)$ due to the high resolution of the basis knots and lack of model mis-specification error.

Parameter estimates and intervals for the FitzHugh-Nagumo system are shown in figure 3–14. Note that unlike in the previous methods, all of the point estimates converged to the neighbourhood of the true parameter values. Furthermore, marginal interval estimates are nearly identical to those of the MCMC method of 2.2 when those estimates converged to the correct location.

3.6.1 Iteratively Re-Weighted Profile Estimation for the FitzHugh-Nagumo System

In producing the 50 simulated FitzHugh-Nagumo data sets, Gaussian noise with variance $.5^2$ and $.4^2$ was added for components V and R respectively. In this section the iterative re-weighting is applied to the FitzHugh-Nagumo data simulated data sets.

	equal weights			re-weighted		
parameter	bias	var	MSE	bias	var	MSE
α	$-.00080$	.00019	.00041	$-.00091$	.00020	.00043
β	.00474	.00304	.00549	.00483	.00299	.00546
γ	$-.00581$	0.00064	.00130	$-.00642$	.00119	.00213

Table 3–6: The observed average of the bias, variance and mean square error (MSE) of the observed parameter estimates for the 50 simulated FitzHugh-Nagumo data sets.

Using the same b-spline basis from section 3.6, the re-weighted profile estimation process was initialized with initial weights $w_V^{(1)} = w_R^{(1)} = 1$. The weighted profile estimation routine was performed at fixed $\lambda = 10^4$ and the same initial parameter estimates consistently used with these data sets. After parameters $\hat{\boldsymbol{\theta}}^{(1)}$ converged, the weights were updated with the inverse of the observed residual variance as described in section 3.4.3. Using the new weights, weighted profile estimation was again performed initialized with $\boldsymbol{\theta}_{initial}^{(new)} = \boldsymbol{\theta}_{final}^{(old)}$. Weights were continually updated until

$$max_k |w_k^{(i)} - w_k^{(i-1)}| < 10^{-4}.$$

Figure 3–15 shows a histogram of the inverse weight estimates, the estimated data variances, for the 50 simulated data sets. The red line shows the true data variance which lies towards the middle of the observed densities. The average of the parameter estimate bias, variance and mean square error (MSE) are given in table 3–6. The average bias and variance of the parameter estimates is small, suggesting that both methods perform well at estimating parameters from the high resolution of the observations. Furthermore the MSE estimates are nearly identical between the two methods except for parameter γ whose MSE increased with iterative re-weighting.

Parameter γ is important in determining the rate of exponential growth in V and, through iterative re-weighting component V is down weighted by a larger residual variance estimate. While γ appears in DR in equation (1.16), its impact on R is diluted and confounded by parameter β. Consequently, the information used to identify γ is reduced through unequal weighting. While iterative re-weighting provides additional information about the relative precisions of measurements from V and R, it occurs at the expense of precision in the ability to estimate parameters tied closely to the down-weighed component.

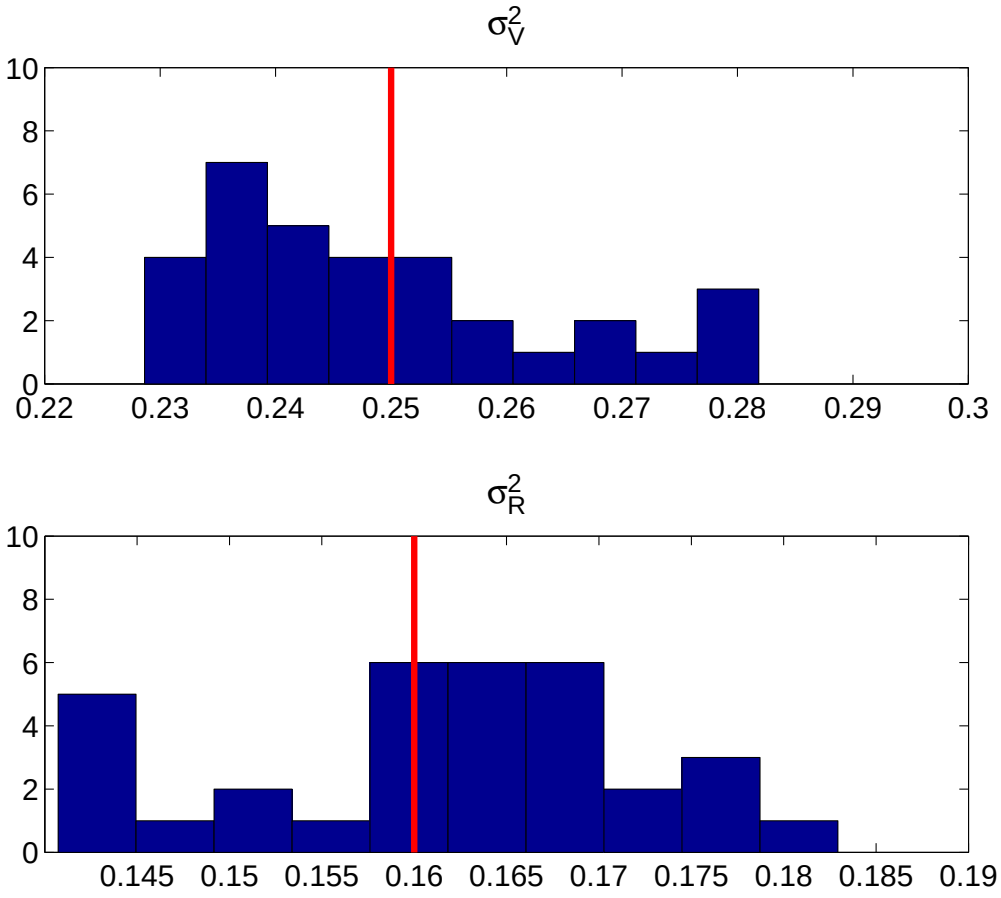

Figure 3–15: The estimated data variance for V and R, where weights $w_k = \hat{\sigma}_k^2$ for $k \in \{V, R\}$. The red lines denote the true values.

3.6.2 Conclusions

This chapter presents a collocation based profile likelihood approach to point and interval estimates for parameters from difficult to navigate likelihood surfaces. The use of a collocation routine improves the convergence properties of the algorithm and improves robustness to the initial parameter guesses. Interval estimates are obtained using a delta method approximation and assumes that the differential equation model is parameterized to produce a single dominant mode. In some cases the delta method approximation may produce poor interval estimates. Furthermore, the differential equation model may produce behaviour close to the data trajectory with a variety of parameter values, leading to a likelihood with multiple important modes. When the likelihood based inference and delta method intervals alone do not provide a reasonable description of the parameters, and when frequentist philosophical logic is considered inappropriate, a Bayesian method for parameter estimation would be worthwhile. Improvements to the method of section 2.2 are described in the next chapter, taking into account the insights and improvements available from generalized profile estimation.

CHAPTER 4
Bayesian Collocation Methods for Differential Equation Models

In section 2.3 many improvements were proposed to work around the potential pitfalls of parameter estimation in ODE models with nonlinear least squares (NLS) and MCMC. The most promising of these methods use collocation, pseudo-orbits or other local approximations to the ODE solution to ease movement around the parameter space, and reduce the dependency on the initial system states. Furthermore, using these approximations has the potential to relax the unforgiving parameter space pitfalls caused by drastic behavioural changes in the ODE model from small changes in parameter values.

While chapter 3 describes a method based on maximizing the profile likelihood to overcome these challenges, this chapter focuses on Bayesian methods. Section 4.1 describes a first attempt at producing a collocation based Bayesian model and outlines its successes and shortcomings through a linear ODE example. Section 4.2 extends this method to nonlinear ODE models. Section 4.3 describes the well established Bayesian parallel tempering, originally developed for sampling from multi-modal distributions, and shows its benefit in the context of single modal ODE models. The advantages of these methods are combined to produce Bayesian Collocation Tempering, described in section 4.4. The remainder of the chapter focuses on the performance of this method using the simulated FitzHugh-Nagumo data sets and the nylon real data example.

4.1 Bayesian Collocation ODE models

This section develops an alternative Bayesian model (Campbell and Cao 2006) for estimating posterior densities of parameters $\boldsymbol{\theta}$ from the ODE model $D\mathbf{y} = f(\mathbf{y}, \boldsymbol{\theta}, \mathbf{u}(t), t)$ with time varying system outputs $\mathbf{y}(t)$ and input functions $\mathbf{u}(t)$. Rather than depending on the ODE solution, $S(\boldsymbol{\theta}, y_0, \mathbf{u}(t), t)$, this method constructs a hierarchical collocation model using the smooth $X(t) = g\{\mathbf{c}'\boldsymbol{\phi}(t)\} \approx S(\boldsymbol{\theta}, Y_0, \mathbf{u}(t), t)$ with coefficients $\mathbf{c}$, basis functions $\boldsymbol{\phi}(t)$ and constraint function $g\{\cdot\}$.

The resulting model is similar to the smoothing model used for Generalized Additive Models (Hastie and Tibshirani 2000) and measurement error problems (Berry, Carroll, and Ruppert 2002), however these methods use a penalty on the first or second derivative as opposed to the more informative ODE model in Campbell et al. (2006). Here, the hierarchical Bayesian model uses the ODE parameters $\boldsymbol{\theta}$ as hyper parameters describing the shape of the smooth approximation of the solution to the ODE. For example, the model

$$\begin{aligned} \mathbf{y}(t) \mid \mathbf{c}, \boldsymbol{\theta}, \sigma^2 &\sim N(\mathbf{c}'\boldsymbol{\phi}(t), \sigma^2), \\ \mathbf{c} \mid \gamma, \boldsymbol{\theta} &\sim \gamma^{-m/2} \exp\left(-\gamma \mathtt{PEN}(\boldsymbol{\theta})\right), \\ \gamma &\sim G(A_\gamma, B_\gamma), \\ \sigma^2 &\sim IG(A_\sigma, B_\sigma), \\ \mathtt{PEN}(\boldsymbol{\theta}) &= \textstyle\int_T \left\{DX - f(X, \boldsymbol{\theta}, u(s), s)\right\}^2 ds, \end{aligned} \tag{4.1}$$

produces a posterior density for the m basis functions $\mathbf{c}$, the smoothing parameter $\lambda = \gamma/\sigma^2$ and hyper parameters of interest $\boldsymbol{\theta}$.

The density $G(A, B)$ is the gamma distribution parameterized to have mean AB and IG is the inverse gamma. Parameters $A_\gamma, B_\gamma, A_\sigma$ and B_σ are known in advance based on prior information. The conditional prior density on $\mathbf{c}$ is increasing as the smooth moves closer to the solution to the ODE, penalized at the level of the derivative, similar to the PEN term in profile estimation. The priors for γ and σ^2 in model (4.1) are the conjugate priors. Conjugate priors for $\boldsymbol{\theta}$ depend on the form of PEN, however in most cases this will not have a closed form solution and therefore a conjugate prior will not be available.

If the ODE model is linear in system outputs, that is, for linear differential operator $L(\cdot)$ the model is $f(\mathbf{c}'\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{u}(t), t) = L(\mathbf{c}'\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{u}(t), t) = \mathbf{c}'L(\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{u}(t), t)$ then PEN can be rewritten:

$$\mathtt{PEN} = \int \left(\mathbf{c}'(D\boldsymbol{\phi}) - \mathbf{c}'\left[L(\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{u}(t), t)\right]\right)^2 dt = \mathbf{c}'\left[\int \left((D\boldsymbol{\phi}) - \left[L(\boldsymbol{\phi}, \boldsymbol{\theta}, \mathbf{u}(t), t)\right]\right)^2 dt\right]\mathbf{c} = \mathbf{c}'R\mathbf{c},$$

for matrix R, a function of $\boldsymbol{\theta}$ and $\mathbf{u}(t)$. Then the posterior for $\mathbf{c}$ has the closed form conditional posterior:

$$\begin{aligned} P(\mathbf{c} \mid \mathbf{y}, \boldsymbol{\theta}, \sigma^2, \gamma) &\propto P(\mathbf{y}(t) \mid \mathbf{c}, \boldsymbol{\theta}, \sigma^2) P(\mathbf{c} \mid \gamma, \boldsymbol{\theta}) \\ &\propto \exp\left\{ -\tfrac{(\mathbf{y}-\mathbf{c}'\boldsymbol{\phi})'(\mathbf{y}-\mathbf{c}'\boldsymbol{\phi})}{2\sigma^2} - \tfrac{\gamma}{2}\mathtt{PEN} \right\} \\ &= \exp\left\{ -\tfrac{(\mathbf{y}-\mathbf{c}'\boldsymbol{\phi})'(\mathbf{y}-\mathbf{c}'\boldsymbol{\phi})}{2\sigma^2} - \tfrac{\gamma}{2}\mathbf{c}'R\mathbf{c} \right\} \\ &\sim N\left(M\boldsymbol{\phi}'\mathbf{y}, M\boldsymbol{\phi}'\sigma^2\right) \\ M &= [\boldsymbol{\phi}'\boldsymbol{\phi} + \gamma R\sigma^2]^{-1}. \end{aligned} \tag{4.2}$$

This is the familiar conditional posterior used in many general Bayesian smoothing applications. The posterior mean is the usual least squares penalized likelihood smooth estimator.

Samples from the conditional posterior for $\mathbf{c}$ can then be obtained using a Gibbs sampler. Similarly a Gibbs sampler may be used to obtain a sample from the conditional posteriors of γ and σ^2 using:

$$\begin{gathered} P(\sigma^2 \mid \mathbf{y}, \boldsymbol{\theta}, \mathbf{c}, \gamma) = IG\left(A_\sigma + n/2, [(\mathbf{y}-\mathbf{c}'\boldsymbol{\phi})'(\mathbf{y}-\mathbf{c}'\boldsymbol{\phi})/2 + 1/B_\sigma]^{-1}\right) \\ \text{and} \quad P(\gamma \mid \mathbf{y}, \boldsymbol{\theta}, \mathbf{c}, \sigma^2) = G\left(A_\gamma + m/2, [\mathtt{PEN}/2 + 1/B_\gamma]^{-1}\right) \end{gathered} \tag{4.3}$$

This is a hierarchical model so the posterior for $\boldsymbol{\theta}$ is only indirectly affected by the data through $\mathbf{c}$ via the smooth in PEN:

$$\begin{aligned} P(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{c}, \sigma^2\gamma) &\propto P(\mathbf{y} \mid \mathbf{c}, \boldsymbol{\theta}, \sigma^2) P(\mathbf{c} \mid \gamma, \boldsymbol{\theta}) P(\boldsymbol{\theta}) \\ &\propto \exp\left\{ -\tfrac{(\mathbf{y}-\mathbf{c}'\boldsymbol{\phi})'(\mathbf{y}-\mathbf{c}'\boldsymbol{\phi})}{2\sigma^2} - \tfrac{\gamma\mathtt{PEN}}{2} \right\} P(\boldsymbol{\theta}) \\ &\propto \exp\left\{ -\tfrac{\gamma\mathtt{PEN}}{2} \right\} P(\boldsymbol{\theta}) \end{aligned} \tag{4.4}$$

The strength of the information flow from $\mathbf{y}$ to $\mathbf{c}$ and onto $\boldsymbol{\theta}$ is regulated by the value γ.

4.1.1 Exponential Growth Example

In Campbell et al (2006), this method was tested using the simple linear ODE model for unbounded exponential growth: $DX(t) = \theta X(t) = f(X, \theta, t)$. Although this model has the solution $X(t) = exp(\theta t)$, the goal of this exercise was to avoid solving the ODE model. Using $\theta = 1.2$, 50 evenly spaced observations were simulated over the interval $[1, 3]$ with added $N(0, .75^2)$

Gaussian noise. The Bayesian model used was

$$\begin{aligned}
\mathbf{y}(t) \mid \mathbf{c}, \theta, \sigma^2 &\sim N(\mathbf{c}'\boldsymbol{\phi}(t), \sigma^2), \\
\mathbf{c} \mid \gamma, \theta &\sim \gamma^{-m/2} \exp\left(-\gamma \mathtt{PEN}(\theta)\right), \ m = 11, \\
\gamma &\sim G(A_\gamma, B_\gamma), \ A_\gamma = B_\gamma = 2, \\
\sigma^2 &\sim IG(A_\sigma, B_\sigma), \ A_\sigma = B_\sigma = 2, \\
\mathtt{PEN}(\theta) &= \textstyle\int_T \left\{DX - f(X, \theta, s)\right\}^2 ds \\
\theta &\sim N(10, 25^2).
\end{aligned} \tag{4.5}$$

The linear differential equation $DX = \boldsymbol{\theta} X$, with the unconstrained basis expansion $\mathbf{x}(t) = \mathbf{c}'\boldsymbol{\phi}(t)$ can be written as $DX = L(X) = \mathbf{c}'L(\boldsymbol{\phi}) = \mathbf{c}'\boldsymbol{\theta}\boldsymbol{\phi}$. Consequently the penalty term can be expressed simply with the matrix R:

$$\mathtt{PEN} = \mathbf{c}' \left\{ \int \left[D\boldsymbol{\phi} - \theta\boldsymbol{\phi}\right]^2 dt \right\} \mathbf{c} = \mathbf{c}' R \mathbf{c}. \tag{4.6}$$

In this example the posterior for θ can be therefore be written as

$$\begin{aligned}
P(\theta \mid \mathbf{y}, \mathbf{c}, \sigma^2, \gamma) &\propto P(\mathbf{y} \mid \mathbf{c}, \theta, \sigma^2) P(\mathbf{c} \mid \gamma, \theta) P(\theta) \\
&\propto \exp\left\{ -\gamma \textstyle\int_{t=1}^{3} \left[\mathbf{c}' D\boldsymbol{\phi} - \theta \mathbf{c}'\boldsymbol{\phi}\right]^2 dt - \frac{(\theta - 10)^2}{2(25^2)} \right\} \\
&\propto \exp\left\{ -\theta^2 \left[\gamma \mathbf{c}' \textstyle\int_{t=1}^{3} \boldsymbol{\phi}'\boldsymbol{\phi} dt \mathbf{c} + \frac{1}{2(25^2)} \right] + \theta \left[\gamma \mathbf{c}' \textstyle\int_{t=1}^{3} \left(\boldsymbol{\phi}' D\boldsymbol{\phi} + D\boldsymbol{\phi}'\boldsymbol{\phi}\right) dt \mathbf{c}' + \frac{10}{25^2} \right] \right\}.
\end{aligned} \tag{4.7}$$

This is in the form of an exponentiated quadratic function of θ and therefore $P(\theta \mid \mathbf{y}, \mathbf{c}, \sigma^2, \gamma)$ is a normal distribution with mean:

$$\left[\gamma \mathbf{c}' \int_{t=1}^{3} \left(\boldsymbol{\phi}' D\boldsymbol{\phi} + D\boldsymbol{\phi}'\boldsymbol{\phi}\right) dt \mathbf{c} + \frac{10}{25^2} \right] \left[2\gamma \mathbf{c}' \int_{t=1}^{3} \boldsymbol{\phi}'\boldsymbol{\phi} dt \mathbf{c} + \frac{1}{25^2} \right]^{-1}$$

and variance

$$\left[2\gamma \mathbf{c}' \int_{t=1}^{3} \boldsymbol{\phi}'\boldsymbol{\phi} dt \mathbf{c} + \frac{1}{25^2} \right]^{-1}$$

and a Gibbs sampler can be used to obtain posterior draws.

The parameters were initialized with draws from their priors. One hundred thousand posterior draws were performed from the Gibbs samplers. The first 50,000 posterior draws are shown

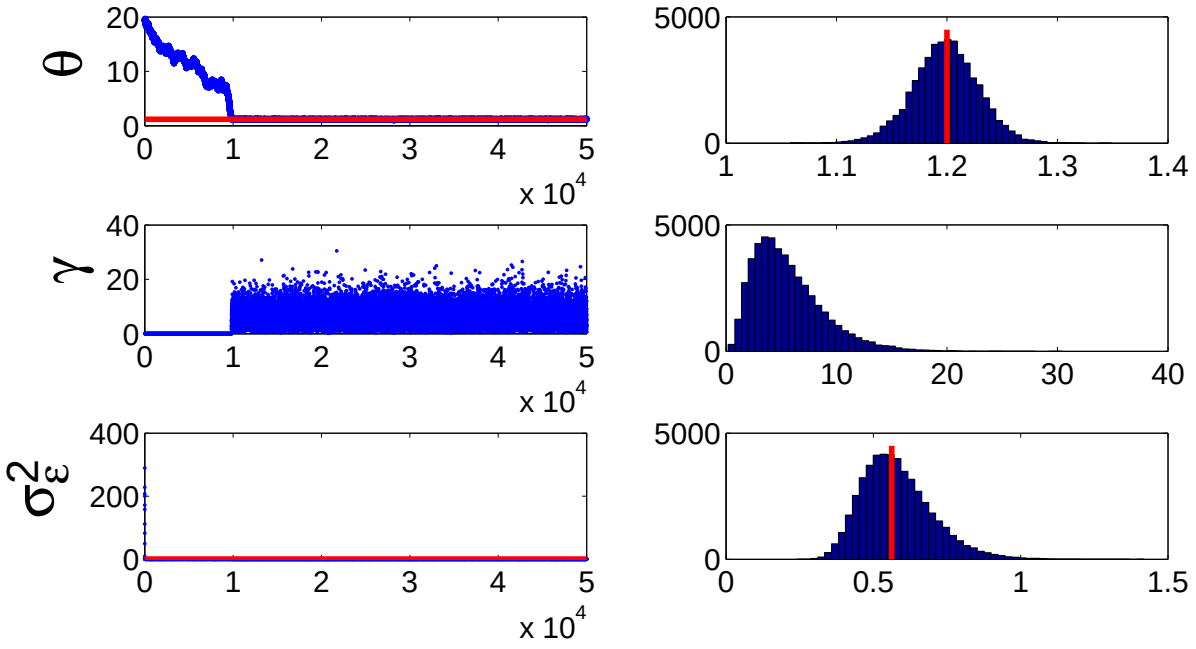

Figure 4–1: The first 50,000 posterior draws for the unbounded exponential growth example of section 4.1.1 and histograms of the second 50,000. Parameter $\boldsymbol{\theta}$ is the ODE parameter, γ is the smoothing parameter and σ_ϵ^2 is the measurement error variance. Red lines indicate the true values. Note that the true value of γ is a function of the type and number of basis functions, the time scale of observations, measurement and model error as well as the quadrature rule. In this case the true marginal posterior mean for the hyper-parameter γ would be 26 if the basis could perfectly accommodate the features of the ODE model, but the imperfections of the basis cause the posterior for λ to be shifted towards zero.

for σ^2, γ and θ in the left hand column of figure 4–1. The right hand column shows a histogram of the posterior draws from the second 50,000 draws. Initially, the posterior draws have a small γ due to a large value of PEN, which in turn is caused by a poor value of θ. When θ moves close to its true value, in this case near draw number 10,000, the smoothing parameter γ jumps suggesting that the smooth and the ODE model agree and PEN substantially decreases.

In this simple example, θ did travel to the neighbourhood of its true value. However, the impact of the data on $\boldsymbol{\theta}$ is softened due to γ and the data smoothing process. Consequently, the posterior for $\boldsymbol{\theta}$ has a much larger variance than would be found by a Bayesian model depending on the numerical solutions to the ODE from section 2.2. While this method may occasionally work nicely with simple systems, in non-linear ODE models, $\texttt{PEN} \neq \mathbf{c}'R\mathbf{c}$ and consequently there may not be a closed form posterior or conjugate prior for $\mathbf{c}$. Using a Metropolis Hastings MCMC sampler for $\mathbf{c}$ includes fine tuning the algorithm for hundreds or thousands of basis functions, in order to be able to adequately model the shape of the ODE model. Since basis coefficients are local parameters, this problem increases in complexity with increasing numbers of observations.

In a more general model, Metropolis Hastings MCMC instead of a Gibbs sampler will be also required to draw from the posterior of $\boldsymbol{\theta}$, because a closed form solution for $P(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{c}, \sigma^2, \gamma)$ will not be available. The MCMC posterior sampling could then be thought of itself as being governed by a dynamic system, because PEN controls a feedback loop influencing γ and θ in the posterior draws. If PEN is large, the posterior mean and variance of γ are reduced due to the inverse of PEN in (4.3). This in turn reduces the information flow back to $\boldsymbol{\theta}$ in (4.4), returning $P(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{c}, \sigma^2, \gamma)$ back to $P(\boldsymbol{\theta})$. This essentially cuts $\boldsymbol{\theta}$ off from the data, preventing useful information from trickling through to guide the motion of $\boldsymbol{\theta}$ towards its true value. While $\boldsymbol{\theta}$ is then less restricted, the algorithms convergence depends heavily on luck to randomly improve its value. If $\boldsymbol{\theta}$ happens to move to a more useful region, PEN will be reduced and γ will increase, feeding more information into $\boldsymbol{\theta}$.

One way of potentially encouraging the data to guide the movement of $\boldsymbol{\theta}$ more strongly is to place a strong prior on γ. However, forcing γ to be too large hinders movement across the

parameter space inducing similar problems to those of the method in section 2.2. Alternatively, constraining γ too small effectively eliminates the useful information in $P(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{c}, \sigma^2, \gamma)$. It is therefore not clear that this model can be improved by choice of prior on γ. Fixing γ instead of obtaining a posterior density for it will maintain constant pressure on $\boldsymbol{\theta}$ to move towards a reasonable set of values but even if $\boldsymbol{\theta}$ converges properly, however $\text{var}(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{c}, \sigma^2, \gamma)$ will still be larger than necessary.

4.2 Bayesian Collocation Method: A Second Method.

An alternative Bayesian collocation model was produced, building on the method in the previous section, to improve by accessibility to nonlinear ODE models. The model, induces a prior on the functional form of the smooth $X(t)$ rather than directly placing a prior on the basis expansion coefficients:

$$\begin{aligned} Y(t) \mid X(t), \sigma^2, \gamma &\sim N\left(X(t), \sigma^2\right) \\ \pi(\boldsymbol{\theta}) &\propto \exp\left(-\lambda \texttt{PEN}\right) P(\boldsymbol{\theta}) \\ \pi(\boldsymbol{\theta}) &\propto \exp\left(-\lambda \int_t \left[DX(s) - f(X(s), \boldsymbol{\theta}, \mathbf{u}(s), s)\right]^2 ds\right) P(\boldsymbol{\theta}) \\ \lambda &\sim G(A_\lambda, B_\lambda). \end{aligned} \tag{4.8}$$

The induced prior on $X(t)$ increases as the smooth moves towards a solution to the ODE model. This method is similar to the model in the previous section in that it is a collocation method producing a posterior density for $\boldsymbol{\theta}$ and λ, but differs because, given $\boldsymbol{\theta}$ and λ, the smooth $X(t)$ is a deterministic function which is further refined by the data $\mathbf{y}$. The induced prior on $X(t)$ reduces the fine tuning problem of dealing from the posteriors of hundreds of basis coefficients.

While the method of section 4.1 induced a prior on $X(t) = \mathbf{c}\boldsymbol{\phi}(t)$ through λ and $\boldsymbol{\theta}$, a prior was also placed directly on $\mathbf{c}$ and hence on $\mathbf{X}(t)$. If these direct and induced priors arise from independent information sources, they may produce an improper prior or place very low prior density on a region of the parameter space considered important by the likelihood. Discrepancies between the information from the direct and induced priors can therefore reduce, eliminate or otherwise isolate important regions of the posterior space, reducing the mixing of the Markov chain. Consequently, removing the prior directly on $\mathbf{c}$ and instead depending only on the induced

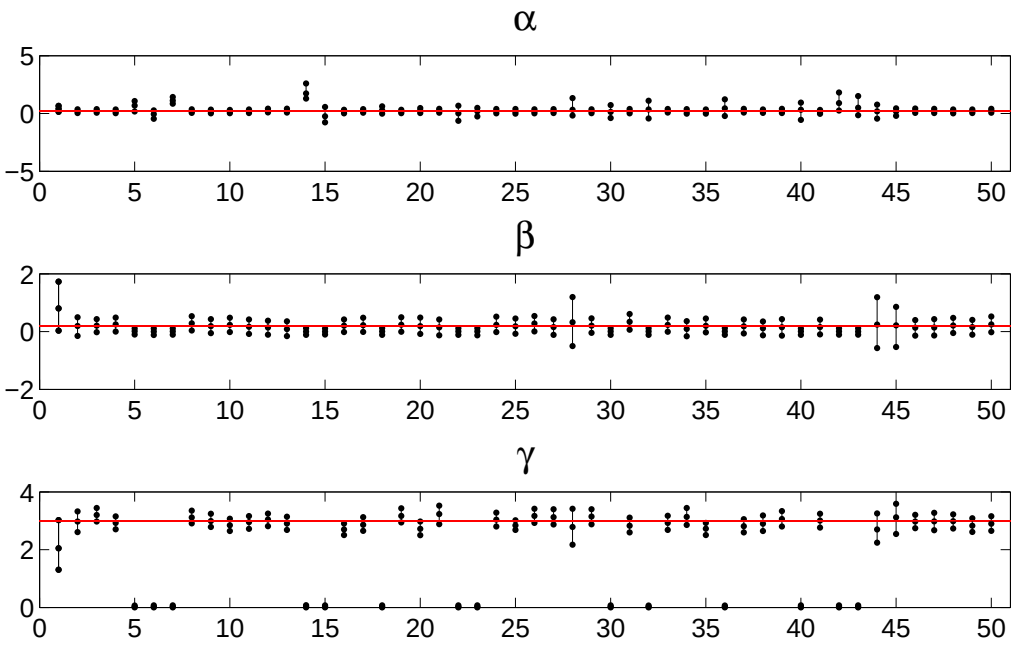

Figure 4–2: The 95% highest posterior densities of the Bayesian Collocation ODE model of section 4.2.

prior, avoids the need for Bayesian Melding (Poole and Raftery 2000) to combine multiple sources of prior information about $\mathbf{X}(t)$. The resulting model (4.8) is like having the generalized profile smoothing step nested within an MCMC collocation model.

The model of (4.8) was applied to the 50 FitzHugh-Nagumo simulated data sets using a fifth order b-spline basis with 89 interior knots, hyper-parameters $A_\lambda = B_\lambda = 2$ and log uniform prior densities on σ_V^2 and σ_R^2. Prior densities on $\boldsymbol{\theta}$ were the same as those in section 2.2:

$$\begin{aligned} \alpha \sim \beta &\sim N(0, .4^2), \\ \gamma &\sim \chi^2_2. \end{aligned} \tag{4.9}$$

The 95% highest posterior density intervals using the last half of the 100,000 posterior draws appear in figure 4–2. While some of the more difficult sets of initial parameter estimates to estimate with the standard MCMC method of section 2.2, which included large values of γ, are no longer a problem for this model, there are several new problems. For example simulated data sets number 2, 25 and 45 were not able to converge to the neighbourhood of the true

parameter values in the standard MCMC method or through NLS. With the model (4.8), these simulated data sets did converge to the neighbourhood of the true values. Many other simulated data sets however, did were not able to converge. Furthermore, even when parameters were attained, the uncertainty in λ and hence $X(t)$ feeds into $\boldsymbol{\theta}$ giving extremely wide intervals especially in simulated data set numbers 28 and 44. Consequently, if this model does move to the neighbourhood of the true parameters, the uncertainty in the posterior due to the hierarchical nature of the model, is considerably wider than it was in the standard MCMC model. A variation of this model dealing with these problems is revisited in section 4.4 building on insights from the next section.

4.3 Parallel Tempering

Although it was developed for sampling from multi-modal densities, parallel tempering (Geyer 1991) allows easier movement around the parameter space by building a sequence of approximations to the posterior density, $\{P_m : m = 1, \ldots, M\}$. Posterior P_M is the desired posterior but densities P_m, $m < M$ are approximations which have been smoothed towards the uniform density. The degree of the posterior approximation is determined by the temperature parameter λ_m in

$$P_i(\boldsymbol{\theta} \mid \mathbf{y}) = P(\boldsymbol{\theta} \mid \mathbf{y})^{\lambda_i/(1+\lambda_i)}, \quad 0 \leq \lambda_1 < \lambda_2 \ldots < \lambda_M = \infty. \tag{4.10}$$

As λ_m decreases, the posterior modes are less sharply peaked, and the near zero probability valleys separating modes become easier to cross as these regions fill in towards uniformity. When temperature parameter $\lambda_1 = 0$, $P_1 \sim U$, a uniform density on the potentially unbounded domain of the parameters.

The M approximations are run as parallel MCMC chains and parameters are allowed to swap between chains taking advantage of the mobility of smaller λ chains using the following algorithm:

1. For draw n=1 initialize the parameter values $\boldsymbol{\theta}_1^{(n)}, \ldots, \boldsymbol{\theta}_M^{(n)}$ where each of the M chains may have the same set of parameter values.

2. With probability p sample i and j independently from the discrete uniform density on the interval (1,M).

3. If $i \neq j$ then propose to swap parameter values between these chains. Sample u from a continuous uniform on $(0, 1)$ and accept the swap by setting $\boldsymbol{\theta}_i^{(n+1)} = \boldsymbol{\theta}_j^{(n)}$ and $\boldsymbol{\theta}_j^{(n+1)} = \boldsymbol{\theta}_i^{(n)}$ if $u < R_{i,j}$ where $R_{i,j}$ is given by

$$R_{i,j} = min\left(1, \frac{P_i(\boldsymbol{\theta}_j^{(n)} \mid \mathbf{y})P_j(\boldsymbol{\theta}_i^{(n)} \mid \mathbf{y})}{P_i(\boldsymbol{\theta}_i^{(n)} \mid \mathbf{y})P_j(\boldsymbol{\theta}_j^{(n)} \mid \mathbf{y})}\right). \tag{4.11}$$

   If the swap is not accepted then retain the previous values $\boldsymbol{\theta}_i^{(n+1)} = \boldsymbol{\theta}_i^{(n)}$ and $\boldsymbol{\theta}_j^{(n+1)} = \boldsymbol{\theta}_j^{(n)}$

   .

4. For the remaining $M-2$ chains, update $\boldsymbol{\theta}_k^{(n+1)}$, $k \in \{1, \ldots, i-1, i+1, \ldots, j-1, j+1, \ldots, M\}$ omitting chains i and j, with the usual MCMC step independently for each chain.

5. Repeat steps 2 to 4 many times.

Parallel tempering (Geyer 1991) and simulated tempering (Marinari and Parisi 1992), a single chain version, are commonly used to sample from multi-modal posterior densities. With differential equation models, this methodology is attractive as smoothing the posterior density should enable easier movement around the posterior space.

4.3.1 Parallel Tempering and the FitzHugh-Nagumo Model

Figure 4–3 shows several tempered approximations to the un-normalized log posterior of γ in a simulated data set from the FitzHugh-Nagumo system of (1.16). In this figure, the remaining model parameters are held at their true values. The highest posterior mode is centered on the true value of $\gamma = 3$, however a smaller mode at $\gamma = 9$ is also present. This corresponds to the modal location in figure 2–9 which was trapping γ larger than its true value using the standard MCMC method of section 2.2. In figure 2–9 this modal value appeared shifted towards $\gamma = 5$ because the other model parameters were not held fixed at their true values. While the mode around $\gamma = 9$ is very small (due to it only partially fitting the data), the valley that separates this mode from the higher mode is very wide and deep. The smaller λ parallel tempered posterior approximations fill in this valley.

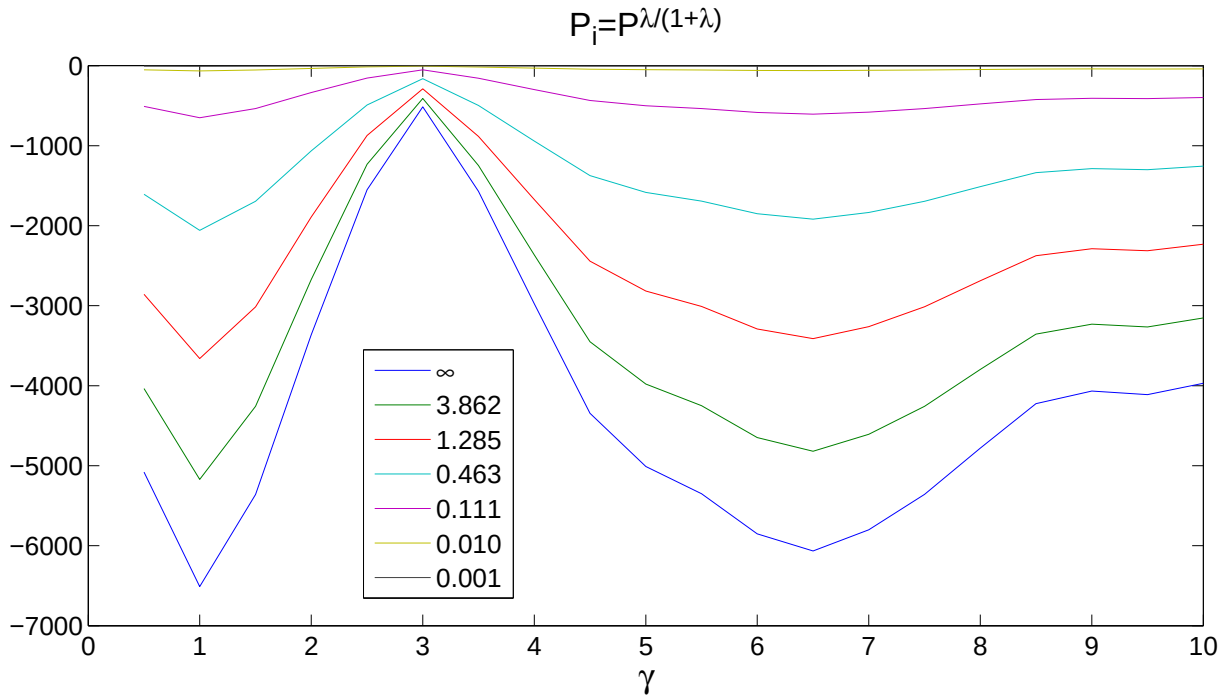

Figure 4–3: The un-normalized log posterior, $P_i = P^{\lambda_1/(1+\lambda_i)}$, changing with λ and γ in the FitzHugh-Nagumo system holding all other parameters at their true values.

Parallel tempering was performed on the 50 simulated FitzHugh-Nagumo data sets of section 2.2. Four temperatures $\lambda \in \{0.0010, 0.0101, 0.1111, \infty\}$, defined the chains which were again initialized using the same starting parameter estimates consistent with the other simulation studies. The values of λ were chosen because they flatten out the posterior to a relatively high degree but still retain some of its shape. Since the $\lambda = \infty$ chain is identical to the standard MCMC model tested in section 2.2, parallel tempering was run for only 5,000 iterations with the probability of proposing a swap between chains $p = 1$. The reason for using so few posterior draws was to examine how the parallel chains and smoother posterior approximations improve movement around the posterior parameter space.

Simulated data set number 45, initialized with a particularly poor set of parameter values, became trapped in the higher γ smaller mode with the standard MCMC model of section 2.2. Figure 4–4 shows the first 200 draws of $\boldsymbol{\theta}_m$ from $P_m(\boldsymbol{\theta}_m \mid \mathbf{y}, \sigma^2, \mathbf{X}_0)$, $m = 1, \ldots, 4$ for this data set. The major breakthrough from parallel tempering is the fast propagation of the extremely large starting values of γ and poor values of α and β towards their true parameter values of

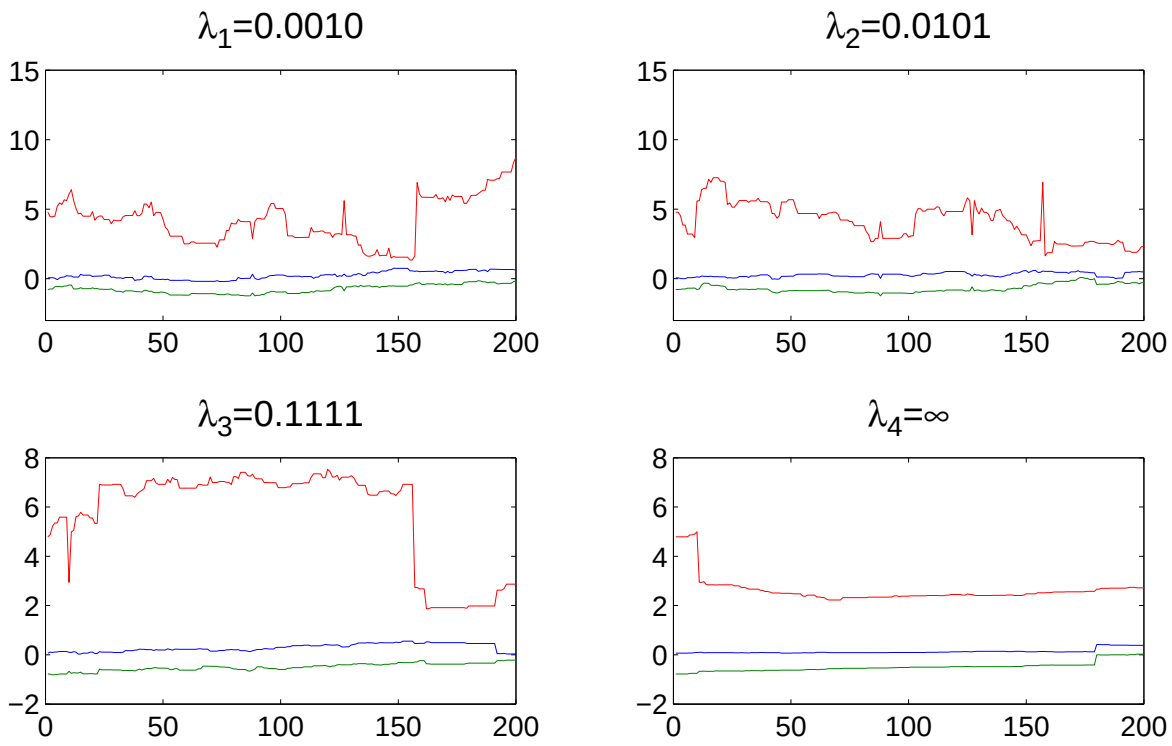

Figure 4–4: The first 200 MCMC iterations for α(blue),β(green) and γ(red) from all four Parallel Tempering chains of the FitzHugh-Nagumo simulated data set #45.

$\boldsymbol{\theta}_{true} = [\alpha, \beta, \gamma]_{true} = [.2, .2, 3]$. At iteration number 9, $\boldsymbol{\theta}_3$ swaps with $\boldsymbol{\theta}_2$ passing values of γ near 6 into the lower λ chain in exchange for values including γ near 3. In the next iteration $\boldsymbol{\theta}_3$ swaps with $\boldsymbol{\theta}_4$ passing the near true values to the top chain, and returning the poor parameter values (including γ near 6) back to the lower λ chain. While these near true values of $\boldsymbol{\theta}_4$ remain in the top $\lambda = \infty$ chain indefinitely, the poor values remain in the $\lambda_3 = 0.1111$ chain until iteration number 156, when another round of swapping pushes them into the λ_2 and then λ_1 chains. Parameters $\boldsymbol{\theta}_4, \boldsymbol{\theta}_3$ and $\boldsymbol{\theta}_2$ essentially remain close to $\boldsymbol{\theta}_{true}$ for the duration of the posterior draws.

Figure 4–5 shows the entire 5,000 posterior draws from this data set for $[\boldsymbol{\theta}, \mathbf{X}_0]$. In addition to simulated data set number 45 needing to fight against an initial estimate of $\gamma^{(initial)} = 5.1$, it also must overcome the initial system states including $\mathbf{X}_0^{(initial)} = [V_0, R_0]^{(initial)} = [.1, 1.8]$ when $\mathbf{X}_0^{(true)} = [-1, 1]$. Despite the fast movement of $\boldsymbol{\theta}$ towards its true value, improved initial system conditions swap up from $P_1 \rightarrow P_2 \rightarrow P_3 \rightarrow P_4$ after the 3000^{th} posterior draws.

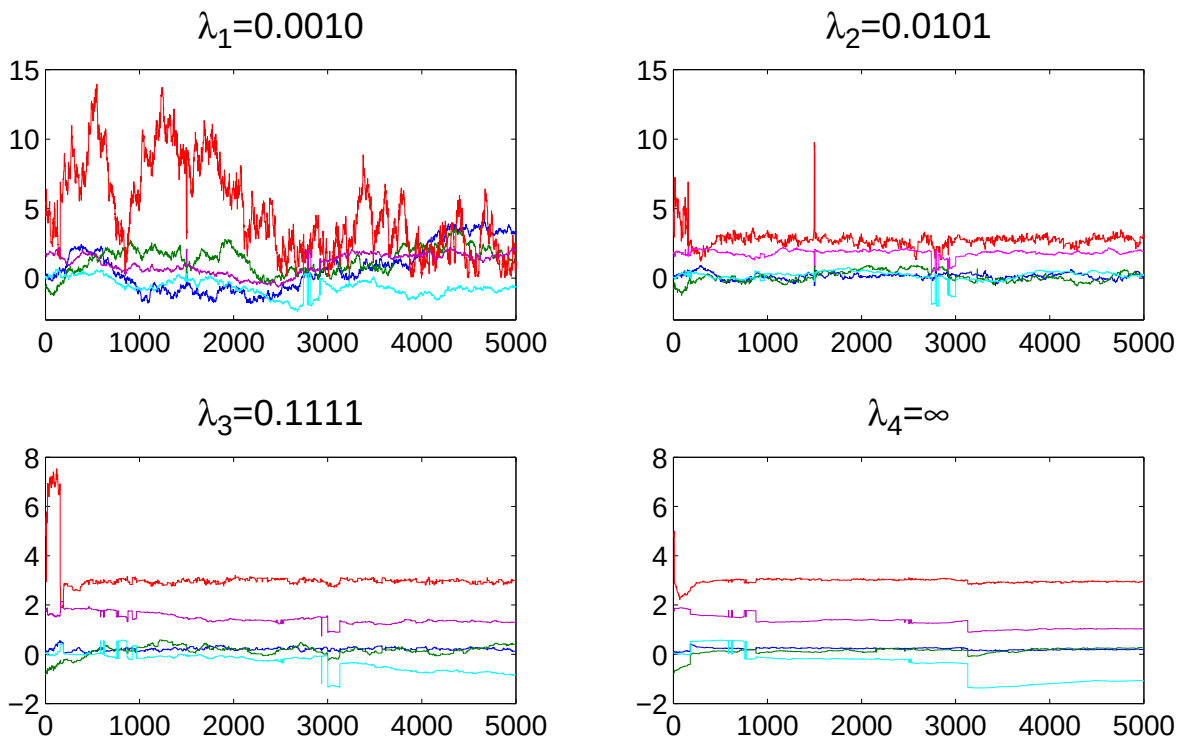

Figure 4–5: The MCMC iterations for α(blue),β(green),γ(red), V_0(cyan) and R_0(magenta) from all four Parallel Tempering chains of the FitzHugh-Nagumo simulated data set #45.

Figure 4–5 also shows the increased variability in the parameters as P_m is smoothed towards a uniform posterior. For example the λ_1 chain accepts many proposed values of $\gamma > 10$. This increased mobility eases movement towards higher posterior modes in the smaller λ chains. However, each decrease in λ also allows parameters to navigate into a larger region of the parameter space exhibiting unreasonable model dynamics and potentially chaotic behaviour. This was especially problematic when poor parameter values excessively slowed or altogether prevented the numerical ODE solver from producing a solution. This occurred quite often preventing the parallel tempering from progressing in several of the chains. When poor parameter values caused problems with the ODE solver, the algorithm was manually stopped and the proposed problematic parameter values were manually rejected before allowing the algorithm to continue. Often the MCMC would subsequently propose a somewhat better parameter set which did not stall the ODE solver. If the problem persisted the entire parallel tempered chain of posterior draws were discarded and the chains restarted from the first parameter values. Simulated data sets number 32 and 40 begin with exceptionally poor parameter values and consequently were unable to progress beyond a small handful of draws despite numerous attempts. By increasing λ, and restricting the movement of $\boldsymbol{\theta}$ and $\mathbf{X}_0$, this problem can be exchanged for slower mixing and more difficult navigation of the parameter space.

Figure 4–6 shows the number of posterior draws required for the M^{th} chain to reach within a tolerance of ± 0.25 of the true values of γ, V_0 and R_0. These three parameters seemed to be the most difficult to move towards the neighbourhood of the true values. This figure shows 27 of the 50 were able to meet the tolerance criteria within the 5,000 draws. Figure 4–7 shows a histogram of the time to reach within this same tolerance using the standard MCMC method of section 2.2. With the standard MCMC model only 22 of the 50 were able to meet this tolerance within the first 5,000 draws and 4 were not able to meet the tolerance within 200,000 draws.

Figures 4–8 and 4–9 show the number of draws to reach within $\pm .25$ of the true value of the single parameter γ using parallel tempering and the standard MCMC method respectively. In the standard MCMC model, γ is slow to travel to its proper neighbourhood and in four cases it

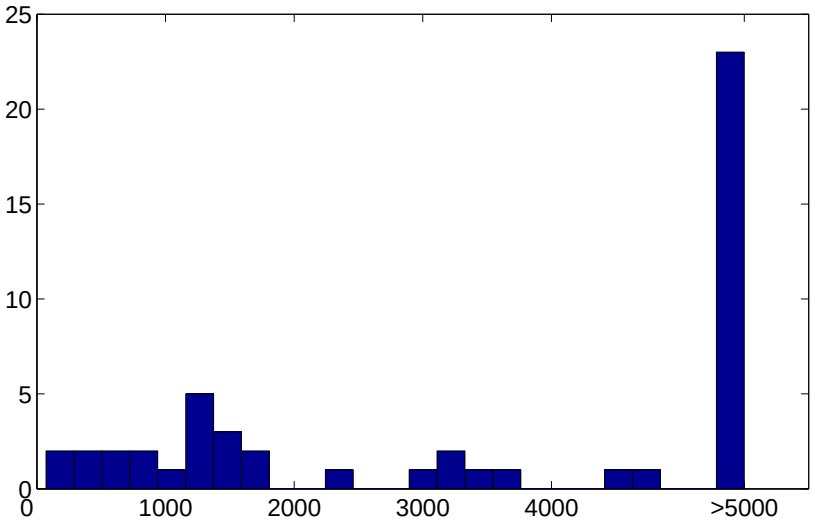

Figure 4–6: The number of draws to reach within $\pm.25$ of the true values of the parameters γ, V_0 and R_0 using parallel tempering for the 50 simulated FitxHugh-Nagumo simulated data sets.

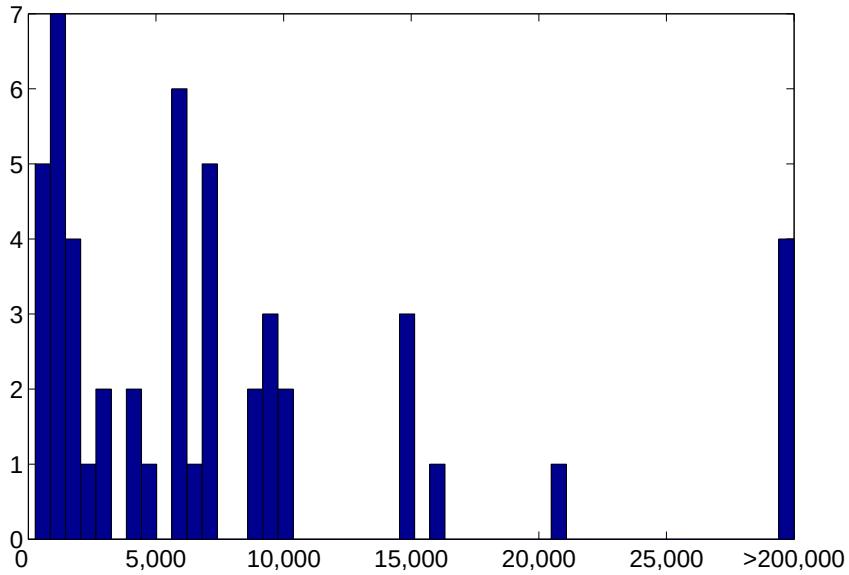

Figure 4–7: The number of posterior draws required for the 50 simulated FitzHugh-Nagumo data sets to move to within $\pm.25$ of the true values of γ, V_0 and R_0 using the standard MCMC model.

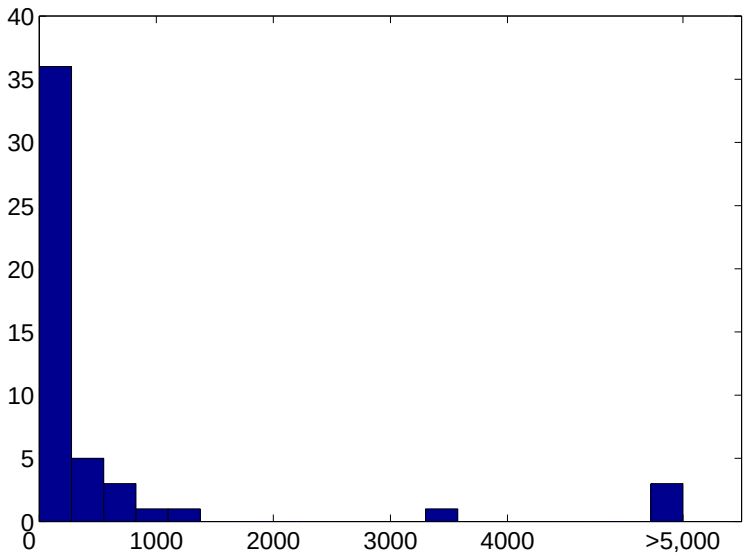

Figure 4–8: The number of draws to reach within $\pm.25$ of γ using parallel tempering.

does meet the criteria within 200,000 draws. By contrast in parallel tempering, γ moves much faster to its true neighbourhood, meeting this tolerance within the first 1,000 draws in all but 4 of the simulated data sets. With both of these methods, the initial system states are the slowest and most difficult parameters to move around the posterior space. While parallel tempering produces improved convergence overall, the dependency on the numerical ODE solution remains a major problem with this model formulation. While there is room for adaptive methods for adjusting the values of λ, it is perhaps more reasonable to listen to the abundant calls of section 2.3 for incorporating pseudo-orbits or collocation methods.

4.4 Bayesian Collocation Tempering for ODE Models

Like parallel tempering, collocation tempering uses a sequence of M parallel MCMC chains, $\{P_m, m = 1, \ldots, M\}$, each with a smoother approximation to the posterior of interest. However the smooth approximations are built using the likelihood smoothing principle behind generalized profile estimation. The m^{th} model uses a likelihood centered on a data smooth, $X(t) = g\{\mathbf{c}'\boldsymbol{\phi}\}$ for basis functions $\boldsymbol{\phi}$, coefficients $\mathbf{c}$ and constraint function $g\{\cdot\}$ giving the model

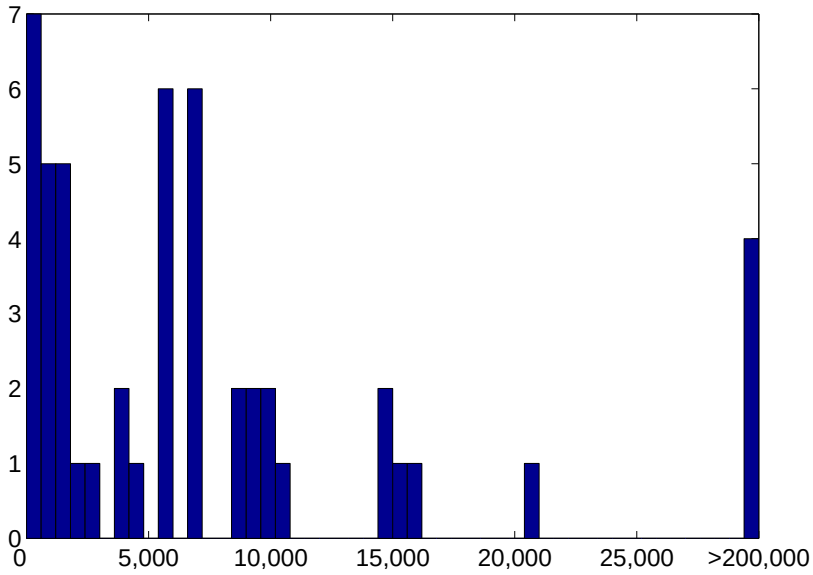

Figure 4–9: The number of draws to reach within $\pm.25$ of γ using the standard MCMC model.

$$\begin{aligned} P_m(Y(t) \mid X(t), \sigma^2) &= N\left(X(t), \sigma^2\right) \\ \pi(\boldsymbol{\theta}) &\propto \exp\left(-\lambda_m \texttt{PEN}\right) P(\boldsymbol{\theta}) \\ \pi(\boldsymbol{\theta}) &\propto \exp\left(-\lambda_m \textstyle\int_t \left[DX(s) - f(X(s), \boldsymbol{\theta}, \mathbf{u}(s), s)\right]^2 ds\right) P(\boldsymbol{\theta}) \\ \text{Where} \qquad & 0 < \lambda_1 < \ldots < \lambda_M = \infty. \end{aligned} \tag{4.12}$$

Centering the likelihood on the data smooth defines each of the parallel chains to be similar to the model in section 4.1 with fixed λ_m. However this model is improved in that $\pi(\boldsymbol{\theta})$ induces a prior on the shape of the functional mean $X(t)$ which increases in density as the smooth approaches the solution to the differential equation. This induced prior is such that $X(t)$ and hence $\mathbf{c}$ are deterministic functions given $\mathbf{y}$ and $\boldsymbol{\theta}$, through an optimization step similar to the generalized profile estimation smoothing in (3.1). The smoothing parameter λ_m acts as a temperature gradient influencing the flow of information between $\boldsymbol{\theta}$ and $\mathbf{c}$ and consequently influencing the flow of information from $\mathbf{y}$ to $\boldsymbol{\theta}$.

As λ_1 approaches zero, the prior for $X(t)$ is a functional uniform density and $P_1(X(t) \mid \mathbf{y}, \sigma^2)$ is a data interpolator whose shape between observations may be non-uniquely defined depending

on the basis functions $\boldsymbol{\phi}(t)$. Furthermore, when $\lambda_1 \to 0$, $P_1(\boldsymbol{\theta}, \mathbf{y}, X(t)\sigma^2) \to P(\boldsymbol{\theta})$, similar to one of the major problems with the model in section 4.1, except in this case λ_1 is fixed and defines only one of several parallel chains.

As λ_m increases, the induced prior for $X(t)$ becomes more sharply peaked, moving towards the solution to the differential equation. Furthermore, the posterior information flow from $\mathbf{y} \to \boldsymbol{\theta}$ is more abundant, producing an even sharper peaked $P_m(\boldsymbol{\theta} \mid \mathbf{y}, X(t), \sigma^2)$. In the M^{th} chain, with $\lambda_M = \infty$, the induced prior on $X(t)$ is again a uniform density but with zero probability on every functional shape other than those following a solution to the ODE $S(\boldsymbol{\theta}, \mathbf{X}_0, t)$:

$$P(\mathbf{X}(t) \mid \boldsymbol{\theta}) \propto \begin{cases} 1 & \text{if } X(t) = S(\boldsymbol{\theta}, \mathbf{X}_0, t) \text{ for any } X_0 \\ 0 & \text{Otherwise} \end{cases}$$

Consequently, for P_M, a reasonable set of basis functions are the ODE solutions, with coefficients $\mathbf{c} = \mathbf{X}_0$, the initial system states. If a prior is directly placed on $\mathbf{X}_0$ instead of induced, P_M is identical to the posterior in the standard MCMC model of section 2.2, $P^*(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{X}_0, \sigma^2)$.

Using the ODE solutions as basis functions for the M^{th} chain, $\mathbf{X}_0$ could be determined from the sensitivity equations in a nonlinear least squares (NLS) step. However, based on difficult experiences with NLS estimation, this is not a quick and simple optimization. Therefore using P^* in place of P_M is highly recommended.

Figure 4–10 shows a cross section of the un-normalized log posterior of γ in the FitzHugh-Nagumo model for several values of the temperature parameter with all other parameters held at their true values. As with figure 4–3, the $\lambda = \infty$ chain has an additional smaller model near $\gamma = 9$. While in parallel tempering the posterior approximations flattened out, but maintained the locations of the posterior modes, in collocation tempering the smaller mode flattens out and shifts towards the main and dominant mode. The shift in the mode occurs due to lack of restrictions on the initial system states from using a collocation method. This allows even the larger λ_m approximations to use more information from the data than would be available from the $\lambda_M = \infty$ chain with fixed initial system states. It is not clear whether a smooth enough set of λ values will combine the modes or if the second mode simply disappears as λ decreases.

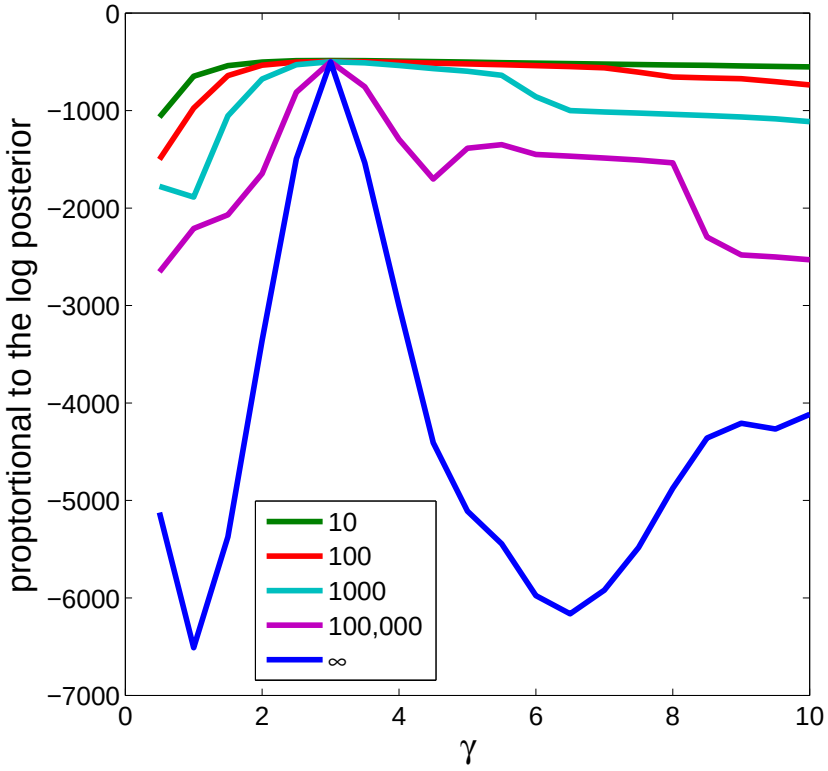

Figure 4–10: The un-normalized log posterior of the collocation tempered chains for γ in the FitzHugh-Nagumo system holding all other parameters at their true values.

4.4.1 Collocation Tempering Algorithm

The M collocation tempering chains are run in parallel using the following algorithm:

1. For iteration n=1 initialize the parameter values $\boldsymbol{\theta}_1^{(n)}, \ldots, \boldsymbol{\theta}_M^{(n)}$ where each of the M chains may have the same set of parameter values. For the M^{th} chain also initialize $\mathbf{X}_0^{(n)}$.
2. With probability p sample i and j independently from the discrete uniform density on the interval (1,M).
3. If $i \neq j$ and $i, j \neq M$, then propose to swap their parameter values. Accept the swap by setting $\boldsymbol{\theta}_i^{(n+1)} = \boldsymbol{\theta}_j^{(n)}$ and $\boldsymbol{\theta}_j^{(n+1)} = \boldsymbol{\theta}_i^{(n)}$ if $u < R_{i,j}$ where u is sampled from a continuous uniform on (0,1) and $R_{i,j}$ is given by
$$R_{i,j} = min\left(1, \frac{P_i(\boldsymbol{\theta}_j^{(n)} \mid \mathbf{y}) P_j(\boldsymbol{\theta}_i^{(n)} \mid \mathbf{y})}{P_i(\boldsymbol{\theta}_i^{(n)} \mid \mathbf{y}) P_j(\boldsymbol{\theta}_j^{(n)} \mid \mathbf{y})}\right). \tag{4.13}$$
If the swap is not accepted then retain the previous values $\boldsymbol{\theta}_i^{(n+1)} = \boldsymbol{\theta}_i^{(n)}$ and $\boldsymbol{\theta}_j^{(n+1)} = \boldsymbol{\theta}_j^{(n)}$.
4. If $i \neq j$ but either $i = M$ of $j = M$, then propose to swap their overlapping parameters. Without loss of generality let $j = M$, then $(\mathbf{X}_0^{(n)})_i$ is obtained directly from the data smooth $X_i^{(n)}(t = 0)$. Accept the swap by setting $\boldsymbol{\theta}_j^{(n+1)} = \boldsymbol{\theta}_i^{(n)}$, $(\mathbf{X}_0^{(n+1)})_j = (\mathbf{X}_0^{(n)})_i$ and $\boldsymbol{\theta}_i^{(n+1)} = \boldsymbol{\theta}_j^{(n)}$ if $u < R_{i,j}$ where u is sampled from a continuous uniform on (0,1) and $R_{i,j}$ is given by
$$R_{i,j} = min\left(1, \frac{P_i(\boldsymbol{\theta}_j^{(n)} \mid \mathbf{y}) P_j(\boldsymbol{\theta}_i^{(n)}, (\mathbf{X}_0^{(n)})_i \mid \mathbf{y})}{P_i(\boldsymbol{\theta}_i^{(n)} \mid \mathbf{y}) P_j(\boldsymbol{\theta}_j^{(n)}, (\mathbf{X}_0^{(n)})_j \mid \mathbf{y})}\right). \tag{4.14}$$
If the swap is not accepted then retain the previous values $\boldsymbol{\theta}_j^{(n+1)} = \boldsymbol{\theta}_j^{(n)}$, $(\mathbf{X}_0^{(n+1)})_j = (\mathbf{X}_0^{(n)})_j$ and $\boldsymbol{\theta}_i^{(n+1)} = \boldsymbol{\theta}_i^{(n)}$.
5. For the remaining $M-2$ chains, update $\boldsymbol{\theta}_k^{(n+1)}$, $k \in \{1, \ldots, i-1, i+1, \ldots, j-1, j+1, \ldots, M\}$ omitting chains i and j, with the usual MCMC step independently for each chain.
6. Repeat steps 2 to 5 many times.

Although this produces a dimensional leap from P_M to $P_{m<M}$, in the smaller λ chains, it takes only $\boldsymbol{\theta}$ and $\mathbf{y}$ to uniquely define $X(t)$, whereas in this formulation for the M^{th} chain, $\mathbf{X}_0$ is also required. This removes the dependency of the smaller λ chains on the initial system states,

a benefit which is passed along to P_M in every swap. The ability to change dimension between parallel chains borrows from a tempering generalization called sintering (Liu and Sabatti 1998).

As λ_m increases, so does the importance of the discrepancy between the features that the basis can accommodate and the ODE model. This discrepancy produced bias in the generalized profile results when λ was too large for the basis. Using a wide spread of λ values and not restricting the pairs of chains which are permitted to swap, reduces the impact of this problem. A poorly chosen λ_m for would effectively isolate itself from the information sharing chain swapping process.

4.4.2 Bayesian Collocation Tempering Results for the FitzHugh-Nagumo System

Collocation tempering was performed on the 50 simulated FitzHugh-Nagumo data sets, using the initial parameter estimates described in 1.2. A third order b-spline basis was used with 79 equally spaced unique interior knots at the 4 temperatures $\lambda = [10, 100, 1000, \infty]$. The $M = 4^{th}$ chain used priors on the initial system state making this chain identical to the model used in 2.6 rather than using the ODE solution as the basis for this chain. The values of λ were chosen to be comparable in impact those used in the parallel tempering attempt of section 4.3.1. However, since the basis used here is coarser than the basis used in generalized profile estimation, the temperature values do not easily compare to the generalized profiling smoothing parameters. The parallel MCMC chains were run for 5,000 draws with $p = 1$ to examine the speed of convergence of the algorithm.

Figure 4–11 shows the first 200 iterations from all four temperatures for simulated data set number 45. This example chain began with a challenging set of parameter estimates that blocked both NLS and the typical MCMC model from converging to the neighbourhood of the true parameter values. Furthermore, with this simulated data set, parallel tempering moved $\boldsymbol{\theta}$ quickly to the true neighbourhood but V_0 and R_0 took an additional 3,000 draws to reach their intended location. With collocation tempering in the first 50 draws all parameters from all four chains moved to the neighbourhoods of the true parameters. Figure 4–12 shows the parameters

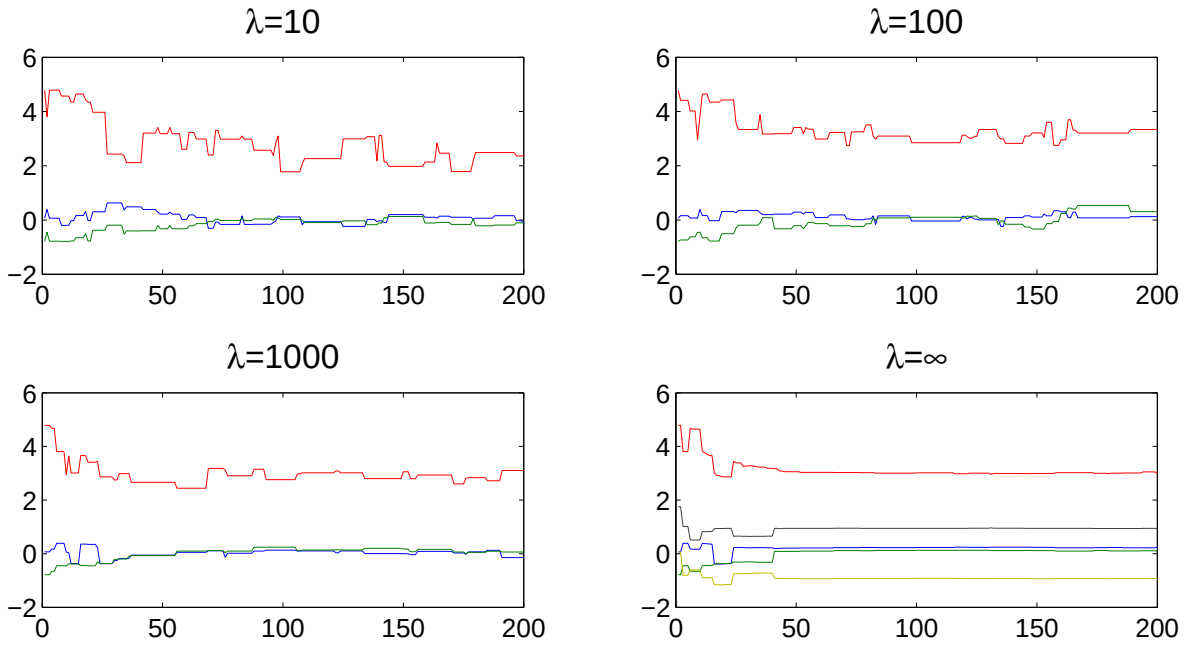

Figure 4–11: The first 200 MCMC draws of α(blue),β(green),γ(red), V_0(cyan) and R_0(magenta) from all four Collocation Tempering chains of simulated data set #45.

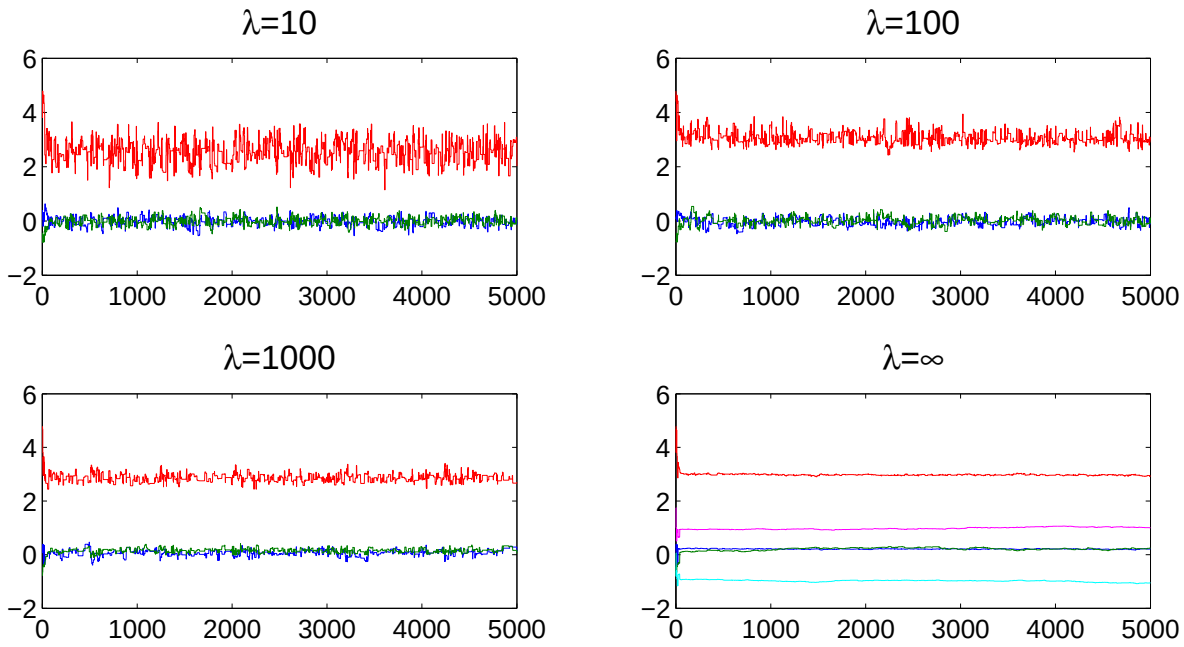

Figure 4–12: 5,000 MCMC draws for α(blue),β(green),γ(red), V_0(cyan) and R_0(magenta) from all four Collocation Tempering chains of simulated data set #45.

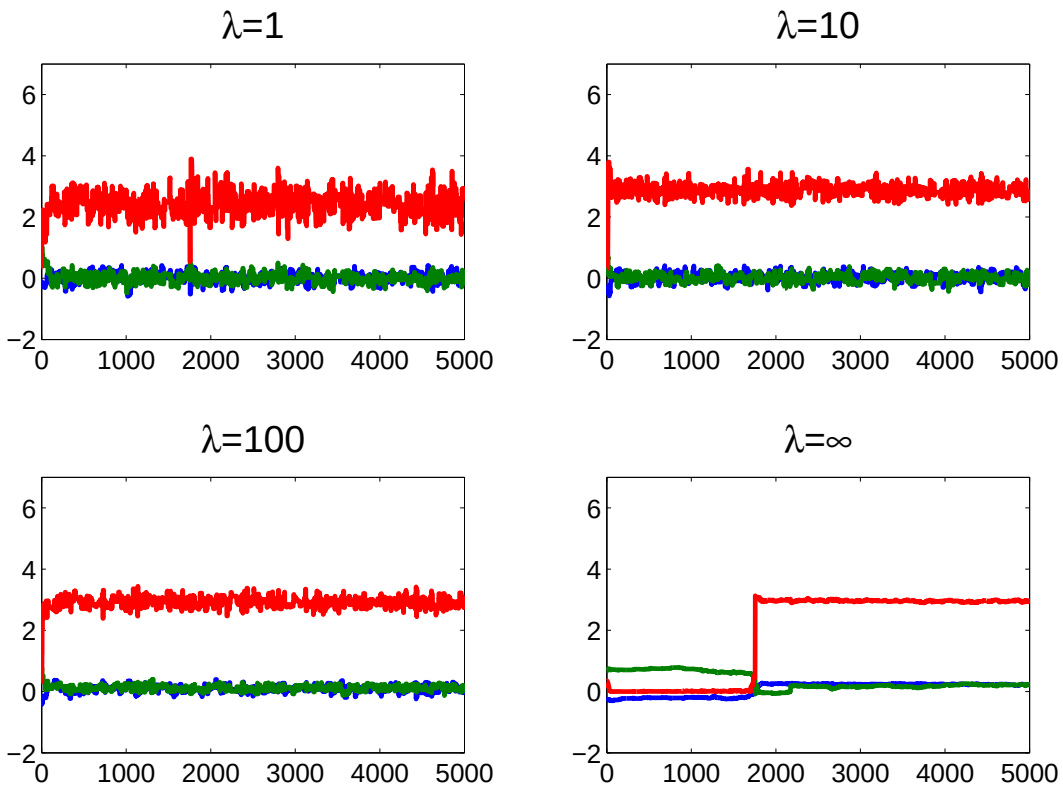

Figure 4–13: 5,000 MCMC draws of α(blue),β(green),γ(red), V_0(cyan) and R_0(magenta) from all four Collocation Tempering chains of simulated data set #18.

for all four chains. All parameters remain reasonably close to the true values for almost the entire 5,000 draws. From this figure, the increased variance of the smaller λ chains is evident.

If a finer basis is used, in exchange for increased computational load, large λ smooth chains can produce posterior densities with the variance asymptotically approaching that of P_M. Using a smoother chain instead of a numerical solution as the top chain of interest avoids the potential for the numerical solver to break down if particularly poor parameter values are proposed. Therefore Bayesian collocation tempering avoids one of the pitfalls of parallel tempering.

Figure 4–13 shows the slowest converging of the 50 collocation tempered simulations. With this simulated data set, the lower temperature chains move towards the true value almost immediately but the $\lambda = \infty$ chain is slow to swap into an improved parameter region. Eventually the $\lambda = 10$ chain swaps parameters with the $\lambda = \infty$ chain at close to iteration #2000. The

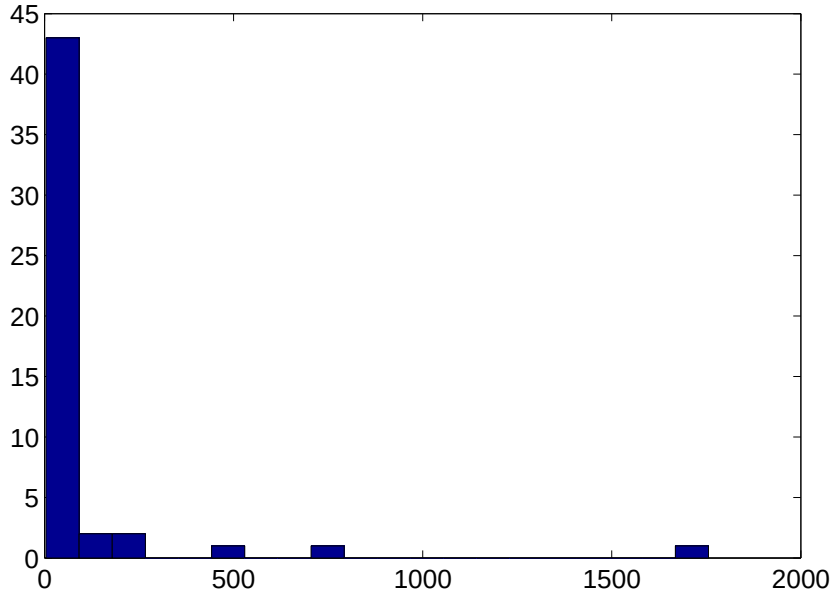

Figure 4–14: The number of posterior draws required for the 50 simulated FitzHugh-Nagumo data sets to move to within $\pm.25$ of the true values of γ, V_0 and R_0.

lower λ chain almost immediately moves back from this poor parameter location towards the true values. All of the chains in all of the simulated data sets moved to the neighbourhood of the true parameter value well before end of the 2,500 iterations discarded to burn in. Figure 4–14 shows a histogram of the number of draws required for the M^{th} chain for the 50 simulated data sets to reach within $\pm.25$ of the true values for γ, V_0 and R_0. The slowest converging chain took 1756 draws to reach within this tolerance but the vast majority converged in less than 200 draws. Recall from parallel tempering in figures 4–6 and 4–8, this speed of convergence was only possible when examining γ individually and not considering the initial system states as well. Furthermore, only 27 of the 50 parallel tempered data sets had met this tolerance criteria within the 5,000 posterior draws. The combination of posterior density smoothing and reducing the impact of initial system states dramatically improves convergence.

Figure 4–15 shows the 95% highest posterior density intervals using the last 2,500 draws from each of the 50 simulated FitzHugh-Nagumo data sets described in section 1.2. Although the number of iterations is potentially too small for reasonable inference on some quantiles,

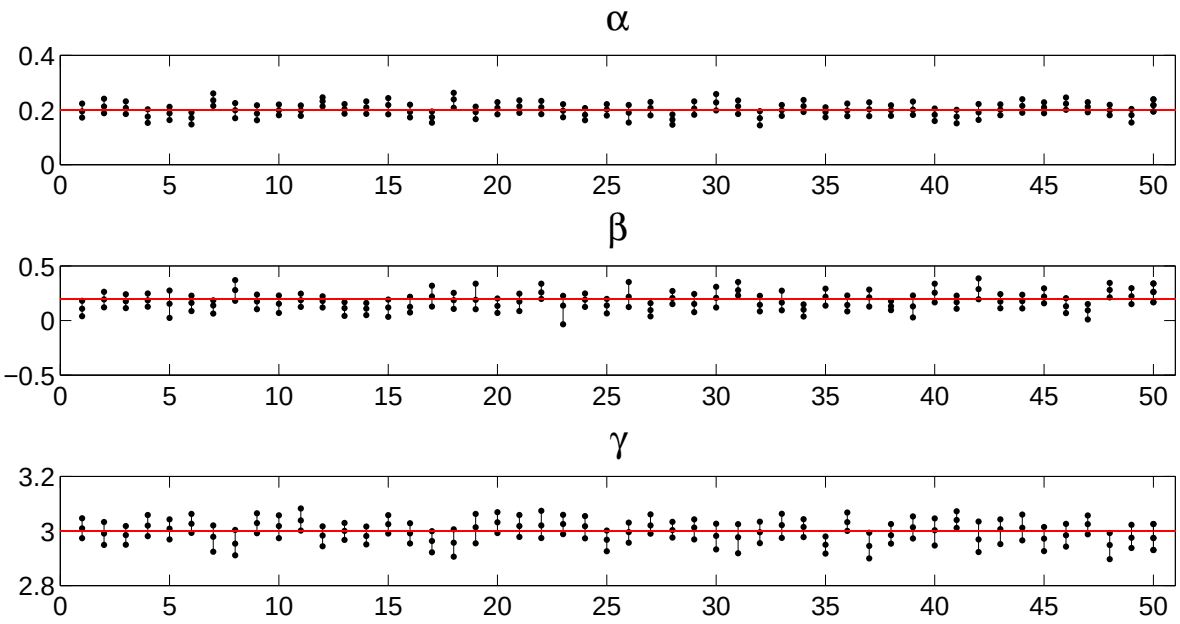

Figure 4–15: The 95% highest posterior density intervals from Bayesian collocation tempering for the 50 simulated FitzHugh-Nagumo data sets.

all of the chains managed to converge to a close neighbourhood of the true parameter values. Furthermore since the $\lambda = \infty$ chain is the same model as the typical one chain MCMC model of section 2.2, the posterior density estimates are asymptotically the same. However the model of section 2.2 may require a massive number of draws to eventually cross the low probability valley preventing some parameter values from moving to the neighbourhood of the true values.

In attempting extremely poor choices of initial parameter estimates, Bayesian Collocation Tempering required additional tuning in the choice of λ. However if λ_1 was made small enough, the method was able to converge to the neighbourhood of the true parameter values relatively quickly, because if a λ_1 is small enough the smooth will interpolate the data. This provides the opportunity for $\boldsymbol{\theta}$ to move parameters governing large scale ODE features to the correct neighbourhood. Fine tuning of these parameters is then performed by larger λ chains.

4.5 Overcoming Challenges of the Nylon Data

This section considers the special challenges of the nylon data set and describes how they might be overcome using Bayesian collocation tempering.

4.5.1 Multiple Experimental Runs

When there are multiple experimental runs, the collocation chains smooth the data using the same process as profile estimation; experimental runs can be smoothed in parallel, but all components within an experimental run must smoothed together. The M^{th} chain in collocation tempering is based on the numerical solution to the ODE, where solutions for each experimental run can be computed in parallel.

4.5.2 Step Function System Inputs

The collocation chains require removing a small interval from the integral in (4.12) or defining the right or left hand derivatives at the points of discontinuity as described in section 3.4.2 in the context of profile estimation. The M^{th} chain, based on the numerical solution to the ODE, must be divided into intervals bounded by the times of step changes. The data fit at the step change is determined by the numerical solution from the interval ending at this point. The fitted value at this endpoint is then used to initialize the numeric solution across the next interval to ensure a continuous ODE solution.

4.5.3 Outputs Measured With Different Precision

In the profile estimation process it is necessary to account for different levels of precision, scales and units of measurement in order to optimize parameters and produce a reasonable fit to the data. Having different levels of precision is easily accommodated by a Bayesian model where it is natural to include parameters σ_j^2 for each of the j observed system outputs.

4.5.4 Unobserved Outputs

The M^{th} chain inherently suffers from the same difficult estimation process as NLS and the basic MCMC model for unobserved outputs. With these methods, the estimation of the unobserved components is based on the initial system state with only indirect observations through the observed system outputs. As with generalized profile estimation, the smaller λ chains benefit from the reduced impact of the initial system state and the potentially chaotic behaviour which can be modelled by an ODE. The chain using only the ODE solutions reaps these benefits by swapping values from smaller λ chains.

4.6 Nylon Bayesian Collocation Tempering Results

Using the 4 parameter nylon system from (2.4) Bayesian Collocation Tempering was attempted using $\lambda \in \{5, 500, \infty\}$. Only three chains were attempted because the smoothing step is relatively computationally heavy when multiple experimental runs are involved. The $\lambda = \infty$ chain used the solution to the ODE and a prior on the initial system states making the $M = 3$ chain equivalent to the Bayesian model of section 2.2.2. Since the M^{th} chain is the same as the model in 2.2.2 the results were similarly inconclusive, offering little new information beyond the assumptions made in the prior density.

CHAPTER 5
Conclusion and Future Work

Throughout this work, two efficient and effective methods are developed to estimate parameters from nonlinear ODE models: generalized profile estimation (GPE) and Bayesian collocation tempering (BCT). They are based on smoothing the likelihood surface to increase the basin of attraction and to allow parameters to cross deep and wide caverns of unlikely parameter values. In previous chapters, it was shown that these methods produce reasonable point and interval estimates in situations where the benchmark methods of nonlinear least squares and the standard MCMC model failed. Furthermore, GPE and BCT induce additional robustness to choice of initial parameter estimates and converge more quickly than these benchmark methods.

While the performance of GPE and BCT have been compared to the benchmark methods throughout this work, the performance of GPE is compared to BCT in section 5.1. This section highlights the differences produced by these philosophically complimentary methods. While GPE and BCT represent new efficient and effective means to estimate parameters from ODE models, statistical research in ODE models is far from complete. Section 5.2 describes some directions for future research, a few of which already in progress.

5.1 Comparing Generalized Profile Estimation and Bayesian Collocation Tempering

In comparing GPE and BCT, their different strategies produce subtle differences in results. Consequently, the performance of GPE and BCT are compared on two levels. Section 5.1.1 compares the fit to the data produced by these two estimation strategies. Section 5.1.2 compares the performance of these two methods in terms of their point and interval estimates for the ODE parameters.

5.1.1 Comparison of Estimates of the Underlying Dynamic System

Using a data set simulated from the FitzHugh-Nagumo system with dynamics,

$$\begin{aligned} DV &= \gamma(V - V^3/3 + R), \\ DR &= -\tfrac{1}{\gamma}(V - \alpha + \beta R), \end{aligned} \tag{5.1}$$

with $V(t)$ and $R(t)$ subject to Gaussian noise with variance $.5^2$, the estimated fit to the true underlying process is examined in this section without being subject to model mis-specification.

Figure 5–1 shows the functional difference between the true underlying process and the fit to the data using the ODE solution from parameters $\hat{\boldsymbol{\theta}} = [\hat{\alpha}, \hat{\beta}, \hat{\gamma}]$ and initial system states $\hat{\mathbf{X}}_0 = [\hat{V}_0, \hat{R}_0]$, from GPE and BCT:

$$S(\boldsymbol{\theta}^{(true)}, \mathbf{X}_0^{(true)}, t) - S(\hat{\boldsymbol{\theta}}, \hat{\mathbf{X}}_0, t). \tag{5.2}$$

The green lines in figure 5–1 show (5.2), computed using BCT as outlined in section 4.4.2, where parameters $\hat{\boldsymbol{\theta}}$ and $\hat{\mathbf{X}}_0$ are the posterior means from 5,000 posterior draws.

As an alternative to (5.2) for GPE, the ODE solution could be approximated using the data smooth at the optimal λ, but with a sufficiently dense basis this is equivalent to using the numerical solution to the ODE with initial system states estimated from the data smooth. The basis used to perform GPE, as outlined in section 3.6, includes 399 interior cubic b-spline knots for each of V and R, and $\hat{\lambda} = 10^8$ producing strong agreement with the solution to the ODE. Consequently, the red lines in figure 5–1 show (5.2) computed using the ODE solution based on initial system states estimated from the data smooth in GPE.

The spikes in figure 5–1 in (5.2) for component V show that the most difficult regions of the true underlying process to estimate are the short lived large magnitude slope segments near times $\{1, 6, 10, 15, 19\}$. These large magnitude slope regions of V are difficult to estimate because they contain few observations and are short lived. The regions of sharp changes towards positive values in V are steeper and therefore more difficult to estimate accurately at times $\{1, 10, 19\}$, than the counterpart regions at times $\{6, 15\}$. This behaviour in V is induced by the tendency towards positive values in R, which is caused by the positive value of α in (5.1).

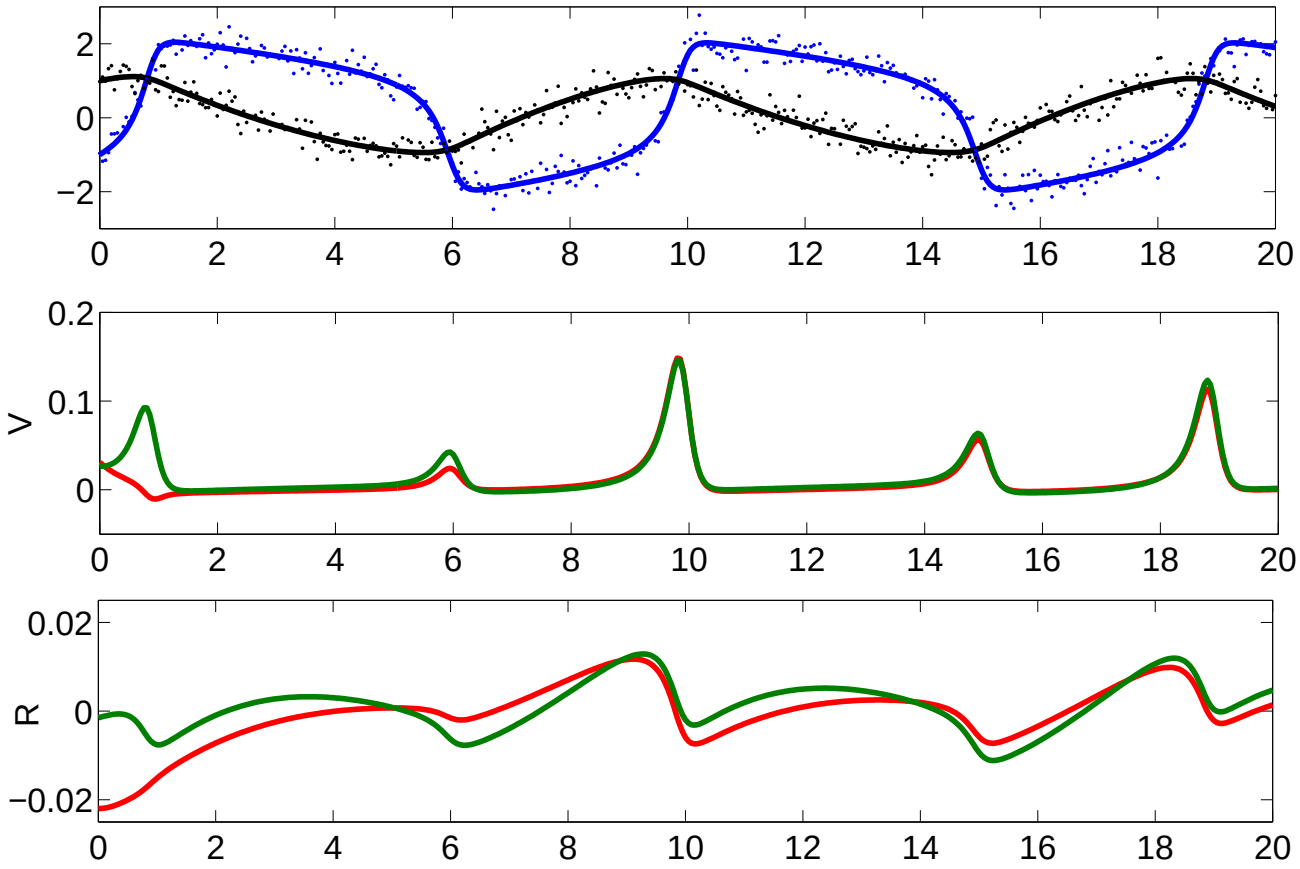

Figure 5–1: The top panel shows the fit to the data for one of the simulated FitzHugh-Nagumo data sets. The blue line is component V and the black line is component R. The bottom two panels show $S(\boldsymbol{\theta}^{(true)}, \mathbf{X}_0^{(true)}, t) - S(\hat{\boldsymbol{\theta}}, \hat{\mathbf{X}}_0, t)$, the difference between the true underlying process and the estimated fit to the data, using the parameter estimates from Bayesian and generalized profiling methods in green and red respectively. This difference is shown for component V in the middle panel, and component R in the bottom panel.

The functional residual for component R in figure 5–1 follows a wave pattern with somewhat of a sawtooth shape. The sharpening the residual slope coincides with the residual spikes in V. While the behaviour of R itself does not undergo sharp changes of growth like component V, the distinct functional residual shape to component R arises from the strong influence of V, and its functional residual on DR.

The residual functions of the estimated underlying process are nearly identical from GPE and BCT, but the GPE method produces a slightly improved estimated $\hat{V}(t)$ in the first two functional residual spikes at times $\{1, 6\}$ while BCT produces a slightly improved to R until time 5. The discrepancy in fit appears to be due to the estimated initial system states and their treatment in the two methods.

5.1.2 Comparison of Point Estimates

To compare parameter estimates from GPE and BCT, the 50 simulated FitzHugh-Nagumo data sets were modelled with both methods without model mis-specification error. Figure 5–2 shows the 95% confidence intervals from iteratively re-weighted GPE and the 95% highest posterior intervals from BCT for the 50 simulated FitzHugh-Nagumo data sets as described in sections 3.6.1 and 4.4.2 respectively. Both of these methods permit the model to account for the true error structure of the data, where the residual variances of V and R are unequal. The point and interval estimates for parameter α using these two methods are nearly identical. Interval estimates for β have essentially the same width using the two methods, although the point and interval estimates for β from GPE are shifted slightly above those of BCT. The role of β in DR from (5.1) is to guide the rate of exponential decay of R. However, in (5.1), β is multiplied by $1/\gamma$, muffling the impact in data fit due to changes in β, and producing the widest intervals of any of the three parameters in figure 5–2. While left for future work, it seems plausible that a re-parametrization of the FitzHugh-Nagumo equations would improve accuracy of the parameter estimate intervals.

In figure 5–2, BCT produces estimates for γ that are narrower and frequently centered closer to the true parameter values compared to those of GPE. While the iteratively re-weighted

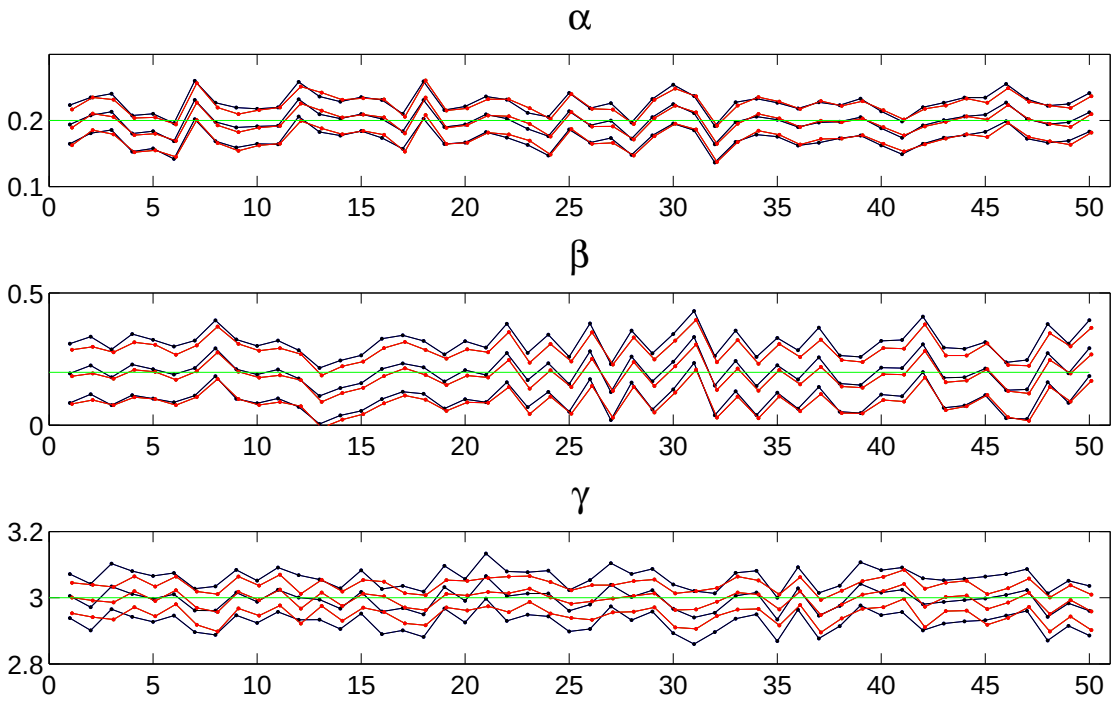

Figure 5–2: The 95% confidence intervals from the generalized profiling estimation method (in black) and the 95% highest posterior density estimates from the Bayesian Collocation Tempering method for the 50 simulated FitzHugh-Nagumo data sets. The true values are shown in green.

GPE method in this comparison is based on $\lambda = 10^4$, a larger value of λ may be enough to account for the differences in point estimates between these two methods. The wider interval estimates of GPE may also be affected by this choice of λ or they may be due to the delta method approximation. A thorough exploration of the reason behind differences in estimates using these two methods is left for future work.

Overall the GPE and BCT methods align very closely, giving essentially the same results for estimates of the underlying process and the parameter point estimates. This suggests two complimentary tools for parameter estimation from ODE models. There is still however, room for additional statistical research with ODE models. Some potential areas are outlined in the next sections.

5.2 Current and Future Areas of Research

The under-representation of methods for ODE models in statistical literature relative to the popularity of their use in a variety of scientific disciplines suggests that there remain many open research problems. Some of which are outlined in this section.

5.2.1 Extensions to Iteratively Re-Weighted Profile Estimation

When using iteratively re-weighted GPE on the nylon system in section 3.5.4, weights were determined by pooling information across all experiments (method 2) or partitioned further to each system output (method 3). Weights for individual observations could be also be estimated using Huber, Andrews or other residual weighting functions (Huber 1981) to provide additional robustness to outlier observation points. This may be a useful extension in the case of influential observation points producing a poor overall model based smooth fit to the data. In section 3.5.4, a few exceptionally large observations were enough to down-weigh an entire experimental run of the nylon system. Using observation specific weights might help to extract more information from the 'well behaved' observations, instead of nearly eliminating the entire experimental run through small weights. Furthermore, the iterative re-weighting process may prove to be important in estimating the auto-correlation structure of the data. While it is expected that the model based

smoothing process inherently builds in some robustness to outliers, a deeper exploration of this is left for future work.

5.2.2 Bayesian Collocation Tempering and Model Mis-Specification

In the BCT model, the temperature parameter λ_m controls the smooth approximation to the ODE solution. Essentially the smaller λ_m chains allow for model mis-specification. When multiple experimental runs are available, such as in the case of the nylon system, a Bayesian collocation model could be reformulated so that the temperature parameter λ controls a functional random effects term, producing insights into model mis-specification of the system. Consider the system with $i = 1, \ldots, I$ experimental runs, white noise measurement error $\epsilon(t)$ at time t, and smooth functional model mis-specification error $\xi_i(t)$:

$$\begin{aligned} Y_i(t) &= X_i(t) + \epsilon(t), \\ X_i(t) &= S\left(\boldsymbol{\theta}, X_{0i}, t\right) + \xi_i(t). \end{aligned} \tag{5.3}$$

To estimate $\xi_i(t)$, it's functional form will require a roughness or model deviation penalty and smoothing parameter $\lambda_{\xi i}$ to uniquely define its functional form in the presence of observation errors $\epsilon(t)$. It might be useful to constrain the form of $\hat{\xi}(t)$ to the function space orthogonal to the solution space of the ODE, possibly further constrained to be 'smooth' with an additional roughness penalty. Alternatively, the shape of could be related to the ODE model itself under alternative parameter values, like a random effects term. The posterior shape of this mis-specification term could be very useful in model diagnostics and model building. It could potentially be incorporated into a dimension jumping BCT MCMC model designed to select an optimal model by reducing constraints on the functional form of $\xi_i(t)$. Much work needs to be done to ensure that the model parameters remain identifiable without increasing the variance of $P(\boldsymbol{\theta} \mid \mathbf{y})$, as was problematic in the methods of section 4.2.

5.2.3 Bayesian Collocation Tempering for Multi-Modal ODE Posteriors

Due to the large changes in the functional behaviour of ODE models that may arise from small changes in parameters, it is likely that multi-modal posterior densities are an important

class of problems where Bayesian methods can contribute. BCT has the potential to accommodate and efficiently sample from multi-modal densities because of it's similarity to parallel tempering. Furthermore, BCT may be modified to include a model selection jump to determine the posterior probabilities of competing models. Testing under these conditions is left for future work and is anticipated to produce even more research problems.

5.2.4 Experimental Design and Selection of Optimal Observation Times

In dynamic systems, such as the FitzHugh-Nagumo and nylon systems, the behaviour exhibited by the system is determined by the inputs, experimental conditions and external forcing functions. These are often the main components of interest when determining the experimental design. When model building or during parameter estimation from an established model, the features of the model which will undergo data driven scrutiny, are only those occurring at observation times. Consequently, the system may exhibit behaviours which were entirely missed by the observation times. For example, if the FitzHugh-Nagumo system used in the simulation study were observed for only the first time unit, it would appear that V is increasing exponentially without bound. Alternatively, certain behaviours may be suggested by the model but there may not be adequate data to determine their validity. For example, in the nylon system including input W_{eq} in the definition of K_a in

$$\begin{aligned} -DL = DA = DC &= -\tfrac{k_{p0}}{1000}(CA - LW/K_a), \\ DW &= \tfrac{k_{p0}}{1000}(CA - LW/K_a) - 24.3(W - W_{eq}), \\ K_a &= \left\{1 + W_{eq}\tfrac{\gamma}{1000}\right\} K_T K_{a0} \ell\left(\tfrac{\Delta H}{8.314}\right), \\ \ell(m) &= \exp\left(-m10^3\left\{\tfrac{1}{T} - \tfrac{1}{T_0}\right\}\right), \\ \text{and} \quad K_T &= 20.97\exp\left(-9.624 + \tfrac{3613}{T}\right), \end{aligned} \tag{5.4}$$

induces a bump in the levels of A and C after step changes in input W_{eq}. However, there is insufficient data to assess the validity of the bump.

As a first step towards assessing the quality of the observation times from the nylon experiment, 100 simulated data sets from each of three different observation time sampling schemes were used under the experimental conditions of the nylon experiment, outlined in section 1.1.

Random Gaussian noise was added to the solution to the four parameter ODE model in 5.4 with variance equal to the assumed measurement error variance in Zheng et al. (2005), $\sigma_A^2 = .6^2$ and $\sigma_C^2 = 2.4^2$. These measurement error variances were assumed to be known and were consequently used as inverse weights in the profile estimation process following the estimation details of section 3.5. In this simulation study there is no model mis-specification or weight mis-specification to complicate the results. The total number of observations for each component within each experimental run is held constant however the observation times were determined through these three sampling schemes:

- Scheme X) The observation times are identical to those used in the original nylon experiment of Zheng et al. (2005), such that the n_{Ai} observations of A in the i^{th} experimental run were taken at times $\mathbf{t}_{Ai}^{(X)}$ and the n_{Ci} observations of C were taken at $\mathbf{t}_{Ci}^{(X)} = \mathbf{t}_{Ai}^{(X)}\mathbf{I}_i^*$. The n_{Ai} by n_{Ci} indicator matrix $\mathbf{I}_i^*$ accounts for the chronological ordering of missing and available observations for C.
- Scheme Y) The sampling times $\mathbf{t}_{Ai}^{(Y)}$ are equally spaced throughout the experimental duration for each of the $i = 1, \ldots, 6$ experimental runs. Observations for C were taken at times $\mathbf{t}_{Ci}^{(Y)} = \mathbf{t}_{Ai}^{(Y)}\mathbf{I}^*$.
- Scheme Z) The total number of observations n_{Ai} are divided into four groups. Three of those groups, each having $floor(n_{Ai}/4)$ observations, were equally spaced in the one hour time intervals beginning 0.05 hours after a step change in input or the start of the experiment. The remaining $\{n_{Ai} - 3 \times floor(n_{Ai}/4)\}$ observations were equally spaced across the experimental duration. In other words this produced the sampling times for an experiment running until time T_i with step input changes at times τ_{i1} and τ_{i2} by sorting

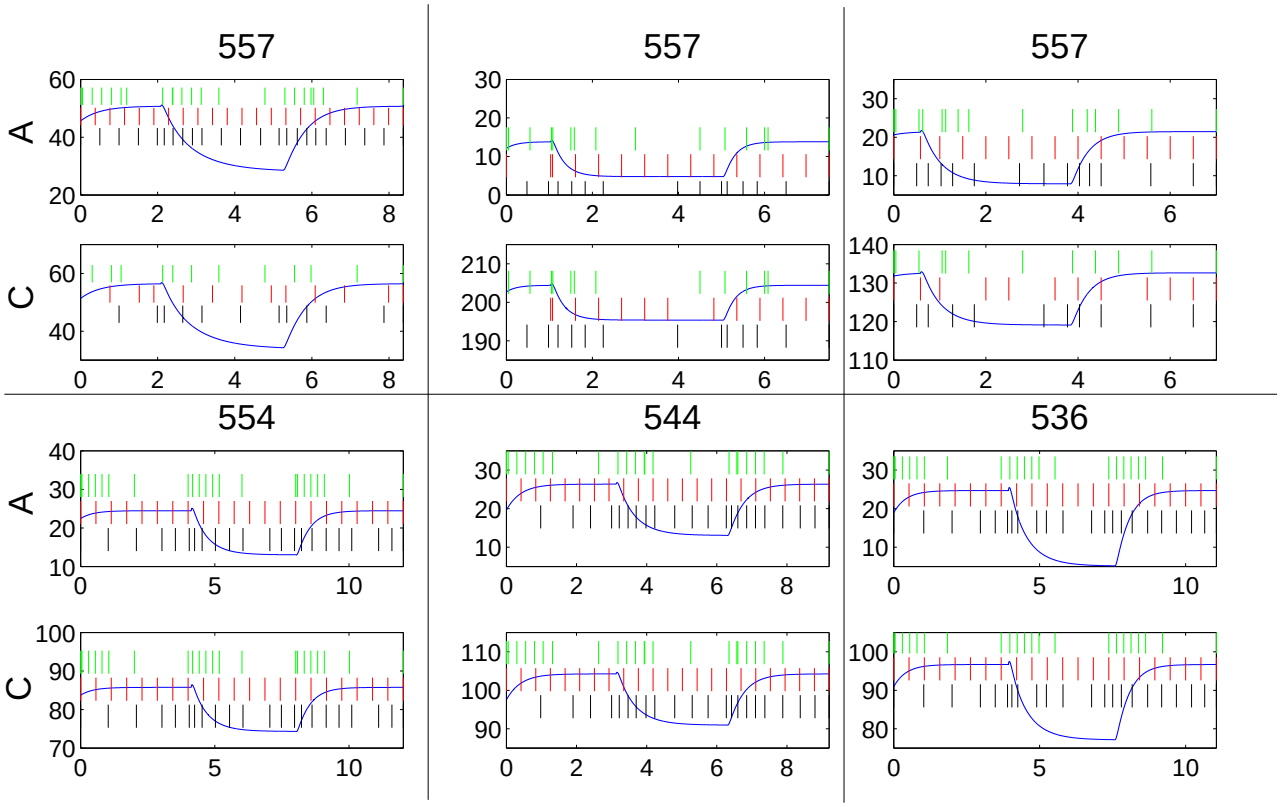

Figure 5–3: A comparison of the observation times from a nylon simulation study. Black marks denote the observation times from scheme X, the observation times from the original experiment. Red marks denote the equally spaced observation times from scheme Y. Green marks represent the observation times using scheme Z, placing additional emphasis on taking observations immediately after the step changes in input W_{eq}. The numerical solution to the ODE is shown in blue.

and combining these four sequences:

$$\begin{gathered}
[0, T_i/\left\{n_{Ai} - 3 \times floor(n_{Ai}/4) - 1\right\}, 2T_i/\left\{n_{Ai} - 3 \times floor(n_{Ai}/4) - 1\right\}, \ldots, T_i] \text{ and} \\
.05 + [floor(n_{Ai}/4), 2 \times floor(n_{Ai}/4), \ldots, 1.05] \text{ and} \\
\tau_{i1} + .05 + [floor(n_{Ai}/4), 2 \times floor(n_{Ai}/4), \ldots, 1.05] \text{ and} \\
\tau_{i2} + .05 + [floor(n_{Ai}/4), 2 \times floor(n_{Ai}/4), \ldots, 1.05].
\end{gathered} \tag{5.5}$$

Observations $\mathbf{t}_{Ci}^{(Z)} = \mathbf{t}_{Ai}^{(Z)} \mathbf{I}_i^*$

Figure 5–3 shows the observation times from the three sampling schemes. Table 5–1 compares the average parameter estimate bias, variance and MSE from GPE using the three different

parameter	True	Model X			Model Y			Model Z		
	value	bias	var	MSE	bias	var	MSE	bias	var	MSE
k_{p0}	20.587	.000	.215	.215	-.005	0.282	.282	-.812	0.182	.842
γ	26.859	-.027	1.064	1.065	.013	1.100	1.100	-.015	1.536	1.536
K_{a0}	50.222	.015	.931	.931	.020	.997	.997	-.018	1.291	1.291
ΔH	-36.462	-.038	2.026	2.026	-.080	2.034	2.034	-.255	1.314	1.364

Table 5–1: A comparison of point estimates and average 95% confidence interval widths for alternative observation time schemes.

schemes for choosing observation times. The parameter bias induced by the observation schemes is negligible except in scheme Z for parameter k_{p0}. From figure 5–3, the observations using scheme Z are highly concentrated in the times of sharp changes in the ODE solution. While k_{p0} plays an important role in determining the rate of change in these steep regions, k_{p0} is also an important parameter in determining the asymptotic equilibrium level of the system components. Observations taken from scheme Z are sparse in the segments where the components are near their steady state equilibrium levels, consequently the available information for k_{p0} is limited producing bias point estimates. Using scheme Z, despite the lack of available observations as the system approaches equilibrium levels, point estimates for γ and K_{a0} are approximately unbiased. However estimates for γ and K_{a0} have larger variance and MSE compared to their estimates using the alternative observation schemes. Effectively, observation scheme Z exchanges reduced accuracy and precision in the estimation of γ, K_{a0} and k_{p0} for substantial gains in the MSE of ΔH.

The equally spaced observations of method Y produced a slightly worse MSE for all four parameters compared to the estimates using scheme X. This suggests that the mix of taking observations in the rapidly changing segments immediately after step changes in input W_{eq}, and the near steady states used in scheme X, produced a good overall balance in the MSE of all four parameters.

While far from a comprehensive review of possible sampling strategies, the observation time scheme can have a substantial impact in estimating parameters from ODE models. Furthermore,

given that the real nylon data observations cost $30,000 to obtain, carefully planned sampling times are a vital stage in the experimental design.

5.2.5 Conclusion

Parameter estimation methods for ODE models are under-represented in statistical literature relative to their popularity as modelling tools. Furthermore, the most commonly used methods, nonlinear least squares and the standard MCMC model, produce results which are not to be fully trusted, as they may be highly dependent on initial parameter guesses, as was shown in chapter 2. However, GPE and BCT provide reliable parameter estimates by improving movement around the parameter space, and providing robustness to initial parameter estimates. Furthermore by comparing the fit to the data from GPE or BCT with large and small values of λ, these methods could provide useful insights into model mis-specification.

GPE and BCT provide two accurate and reliable methods for parameter estimation from ODE models spanning philosophically complementary approaches. Furthermore, the improved convergence rates, reduced dependence on initial system states and ability to include constraints on the structure of the data smooth improves accessibility of statistics to the modelers of ODEs.

APPENDIX A
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.

A.1 $\partial^2\mathbf{c}/\partial\boldsymbol{\theta}\partial\theta_k$

The implicit function theorem is required to define $\partial^2\mathbf{c}/\partial\boldsymbol{\theta}\partial\theta_k$ in (3.10). The term $\partial^2\mathbf{c}/\partial\boldsymbol{\theta}\partial\theta_k$ comes from the fact that $\partial J/\partial\mathbf{c} = 0$ at the optimal choice of $\mathbf{c} = \hat{\mathbf{c}}$. Then differentiating twice with respect to $\boldsymbol{\theta}$ and θ_k, equivalent to differentiating (3.5) with respect to θ_k produces (A.1) which is then rearranged to give the derivative in (A.2).

$$\begin{aligned}
\frac{\partial}{\partial\theta_k}\left(\frac{\partial^2 J}{\partial\hat{\mathbf{c}}\partial\boldsymbol{\theta}}\right) &= \frac{\partial}{\partial\theta_k}\left(\frac{\partial^2 J}{\partial g\partial\boldsymbol{\theta}}\frac{dg}{d\hat{\mathbf{c}}} + \left\{\left(\frac{dg}{d\hat{\mathbf{c}}}\right)'\frac{\partial^2 J}{\partial g^2}\frac{dg}{d\hat{\mathbf{c}}} + \frac{\partial J}{\partial g}\frac{d^2 g}{d\hat{\mathbf{c}}^2}\right\}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}}\right)\\
&= \frac{\partial^3 J}{\partial g\partial\boldsymbol{\theta}\partial\theta_k}\frac{dg}{d\hat{\mathbf{c}}} + \left(\frac{dg_\ell}{d\hat{\mathbf{c}}_\ell}\frac{d\hat{\mathbf{c}}_\ell}{d\boldsymbol{\theta}_k}\right)'\frac{\partial^3 J}{\partial g\partial\boldsymbol{\theta}\partial g_\ell}\frac{dg}{d\hat{\mathbf{c}}} + \frac{\partial^2 J}{\partial g\partial\boldsymbol{\theta}}\frac{d^2 g}{d\hat{\mathbf{c}}d\hat{\mathbf{c}}_\ell}\frac{d\hat{\mathbf{c}}_\ell}{d\theta_k}\\
&\quad + \left\{\left(\frac{d^2 g}{d\hat{\mathbf{c}}d\hat{\mathbf{c}}_\ell}\frac{d\hat{\mathbf{c}}_\ell}{d\theta_k}\right)'\frac{\partial^2 J}{\partial g^2}\frac{dg}{d\hat{\mathbf{c}}} + \left(\frac{dg}{d\hat{\mathbf{c}}}\right)'\frac{\partial^3 J}{\partial g\partial\boldsymbol{\theta}_k\partial g}\frac{dg}{d\hat{\mathbf{c}}} + \left(\frac{dg}{d\hat{\mathbf{c}}}\right)'\frac{\partial^3 J}{\partial g^2 dg_\ell}\frac{dg}{d\hat{\mathbf{c}}}\frac{dg_\ell}{d\hat{\mathbf{c}}_\ell}\frac{d\hat{\mathbf{c}}_\ell}{d\theta_k}\right.\\
&\quad \left. + \left(\frac{dg}{d\hat{\mathbf{c}}}\right)'\frac{\partial^2 J}{\partial g^2}\frac{d^2 g}{d\hat{\mathbf{c}}d\hat{\mathbf{c}}_\ell}\frac{d\hat{\mathbf{c}}_\ell}{d\theta_k} + \frac{\partial^2 J}{\partial g\partial\theta_k}\frac{d^2 g}{d\hat{\mathbf{c}}^2} + \frac{\partial^2 J}{\partial g\partial g_\ell}\frac{d^2 g}{d\hat{\mathbf{c}}^2}\frac{dg_\ell}{d\hat{\mathbf{c}}_\ell}\frac{d\hat{\mathbf{c}}_\ell}{d\theta_k} + \frac{\partial J}{\partial g}\frac{d^3 g}{d\hat{\mathbf{c}}^2 d\hat{\mathbf{c}}_\ell}\frac{d\hat{\mathbf{c}}_\ell}{d\theta_k}\right\}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}}\\
&\quad + \left\{\left(\frac{dg}{d\hat{\mathbf{c}}}\right)'\frac{\partial^2 J}{\partial g^2}\frac{dg}{d\hat{\mathbf{c}}} + \frac{\partial J}{\partial g}\frac{d^2 g}{d\hat{\mathbf{c}}^2}\right\}\frac{\partial^2\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}\partial\theta_k}
\end{aligned} \tag{A.1}$$

$$\begin{aligned}
\frac{\partial^2\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}\partial\theta_k} &= -\left[\left(\frac{dg}{d\hat{\mathbf{c}}}\right)'\frac{\partial^2 J}{\partial g^2}\frac{dg}{d\hat{\mathbf{c}}} + \frac{\partial J}{\partial g}\frac{d^2 g}{d\hat{\mathbf{c}}^2}\right]^{-1}\left[\frac{\partial^3 J}{\partial g\partial\boldsymbol{\theta}\partial\theta_k}\frac{dg}{d\hat{\mathbf{c}}} + \left(\frac{dg_\ell}{d\hat{\mathbf{c}}_\ell}\frac{d\hat{\mathbf{c}}_\ell}{d\boldsymbol{\theta}_k}\right)'\frac{\partial^3 J}{\partial g\partial\boldsymbol{\theta}\partial g_\ell}\frac{dg}{d\hat{\mathbf{c}}} + \frac{\partial^2 J}{\partial g\partial\boldsymbol{\theta}}\frac{d^2 g}{d\hat{\mathbf{c}}d\hat{\mathbf{c}}_\ell}\frac{d\hat{\mathbf{c}}_\ell}{d\theta_k}\right.\\
&\quad + \left\{\left(\frac{d^2 g}{d\hat{\mathbf{c}}d\hat{\mathbf{c}}_\ell}\frac{d\hat{\mathbf{c}}_\ell}{d\theta_k}\right)'\frac{\partial^2 J}{\partial g^2}\frac{dg}{d\hat{\mathbf{c}}} + \left(\frac{dg}{d\hat{\mathbf{c}}}\right)'\frac{\partial^3 J}{\partial g\partial\boldsymbol{\theta}_k\partial g}\frac{dg}{d\hat{\mathbf{c}}} + \left(\frac{dg}{d\hat{\mathbf{c}}}\right)'\frac{\partial^3 J}{\partial g^2 dg_\ell}\frac{dg}{d\hat{\mathbf{c}}}\frac{dg_\ell}{d\hat{\mathbf{c}}_\ell}\frac{d\hat{\mathbf{c}}_\ell}{d\theta_k}\right.\\
&\quad \left.\left. + \left(\frac{dg}{d\hat{\mathbf{c}}}\right)'\frac{\partial^2 J}{\partial g^2}\frac{d^2 g}{d\hat{\mathbf{c}}d\hat{\mathbf{c}}_\ell}\frac{d\hat{\mathbf{c}}_\ell}{d\theta_k} + \frac{\partial^2 J}{\partial g\partial\theta_k}\frac{d^2 g}{d\hat{\mathbf{c}}^2} + \frac{\partial^2 J}{\partial g\partial g_\ell}\frac{d^2 g}{d\hat{\mathbf{c}}^2}\frac{dg_\ell}{d\hat{\mathbf{c}}_\ell}\frac{d\hat{\mathbf{c}}_\ell}{d\theta_k} + \frac{\partial J}{\partial g}\frac{d^3 g}{d\hat{\mathbf{c}}^2 d\hat{\mathbf{c}}_\ell}\frac{d\hat{\mathbf{c}}_\ell}{d\theta_k}\right\}\frac{\partial\hat{\mathbf{c}}}{\partial\boldsymbol{\theta}}\right]
\end{aligned} \tag{A.2}$$

A.2 $\partial\hat{\mathbf{c}}/\partial\mathbf{y}$

The implicit function theorem is required to define $\partial\hat{\mathbf{c}}/\partial\mathbf{y}$ in (3.11). This derivative again uses the fact that $\partial J/\partial\mathbf{c} = 0$ at the optimal choice of $\mathbf{c} = \hat{\mathbf{c}}$. Then differentiating twice with respect to $\mathbf{y}$ produces (A.3) which is then rearranged to give the derivative in (A.4).

$$\frac{\partial}{\partial\mathbf{y}}\left(\frac{\partial J}{\partial\hat{\mathbf{c}}}\right) = \frac{\partial^2 J}{\partial g\partial\mathbf{y}}\frac{dg}{d\hat{\mathbf{c}}} + \left(\frac{dg}{d\hat{\mathbf{c}}}\right)'\frac{\partial^2 J}{\partial g^2}\frac{dg}{d\hat{\mathbf{c}}}\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}} + \frac{\partial J}{\partial g}\frac{d^2 g}{d\hat{\mathbf{c}}^2}\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}} \tag{A.3}$$

$$\frac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}} = \left\{\left(\frac{dg}{d\hat{\mathbf{c}}}\right)'\frac{\partial^2 J}{\partial g^2}\frac{dg}{d\hat{\mathbf{c}}} + \frac{\partial J}{\partial g}\frac{d^2 g}{d\hat{\mathbf{c}}^2}\right\}^{-1}\left\{\frac{\partial^2 J}{dgd\mathbf{y}}\frac{dg}{d\hat{\mathbf{c}}}\right\} \tag{A.4}$$

A.3 $\partial^2\hat{\mathbf{c}}/\partial\mathbf{y}\partial\boldsymbol{\theta}$

We obtain this derivative by differentiating (A.3) with respect to θ_k to produce equation (A.5). Solving for $\partial^2\hat{\mathbf{c}}/\partial\mathbf{y}\partial\boldsymbol{\theta}$ gives us the results in (A.6).

$$\begin{aligned}
\tfrac{\partial}{\partial\theta_k}\left(\tfrac{\partial^2 J}{\partial\hat{\mathbf{c}}\partial\mathbf{y}}\right) &= \tfrac{\partial}{\partial\theta_k}\left(\tfrac{\partial^2 J}{dgd\mathbf{y}}\tfrac{dg}{d\hat{\mathbf{c}}} + \left(\tfrac{dg}{d\hat{\mathbf{c}}}\right)'\tfrac{\partial^2 J}{\partial g^2}\tfrac{dg}{d\hat{\mathbf{c}}}\tfrac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}} + \tfrac{\partial J}{\partial g}\tfrac{d^2 g}{d\hat{\mathbf{c}}^2}\tfrac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}\right)\\
&= \tfrac{\partial^3 J}{\partial g\partial\mathbf{y}\partial\theta_k}\tfrac{dg}{d\hat{\mathbf{c}}} + \tfrac{\partial^3 J}{\partial g\partial\mathbf{y}\partial g}\tfrac{dg}{d\hat{\mathbf{c}}}\tfrac{dg}{d\hat{\mathbf{c}}_\ell}\tfrac{d\hat{\mathbf{c}}_\ell}{d\theta_k} + \tfrac{\partial^2 J}{\partial g\partial\mathbf{y}}\tfrac{d^2 g}{d\hat{\mathbf{c}}d\hat{\mathbf{c}}_\ell}\tfrac{d\hat{\mathbf{c}}_\ell}{d\theta_k}\\
&\quad + \left(\tfrac{d^2 g}{d\hat{\mathbf{c}}d\hat{\mathbf{c}}_\ell}\tfrac{d\hat{\mathbf{c}}_\ell}{d\theta_k}\right)'\tfrac{\partial^2 J}{\partial g^2}\tfrac{dg}{d\hat{\mathbf{c}}}\tfrac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}} + \left(\tfrac{dg}{d\hat{\mathbf{c}}}\right)'\tfrac{\partial^3 J}{\partial g^2\partial\theta_k}\tfrac{dg}{d\hat{\mathbf{c}}}\tfrac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}} + \left(\tfrac{dg}{d\hat{\mathbf{c}}}\right)'\tfrac{\partial^3 J}{\partial g^3}\tfrac{dg}{d\hat{\mathbf{c}}}\tfrac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}\tfrac{dg}{d\hat{\mathbf{c}}_\ell}\tfrac{\partial\hat{\mathbf{c}}_\ell}{\partial\theta_k}\\
&\quad + \left(\tfrac{dg}{d\hat{\mathbf{c}}}\right)'\tfrac{\partial^2 J}{\partial g^2}\tfrac{d^2 g}{d\hat{\mathbf{c}}d\hat{\mathbf{c}}_\ell}\tfrac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}\tfrac{\partial\hat{\mathbf{c}}_\ell}{\partial\theta_k} + \left(\tfrac{dg}{d\hat{\mathbf{c}}}\right)'\tfrac{\partial^2 J}{\partial g^2}\tfrac{dg}{d\hat{\mathbf{c}}}\tfrac{\partial^2\hat{\mathbf{c}}}{\partial\mathbf{y}\partial\theta_k}\\
&\quad + \tfrac{\partial^2 J}{\partial g\partial\theta_k}\tfrac{d^2 g}{d\hat{\mathbf{c}}^2}\tfrac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}} + \left(\tfrac{dg}{d\hat{\mathbf{c}}_\ell}\tfrac{\partial\hat{\mathbf{c}}_\ell}{\partial\theta_k}\right)'\tfrac{\partial^2 J}{\partial g^2}\tfrac{d^2 g}{d\hat{\mathbf{c}}^2}\tfrac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}} + \tfrac{\partial J}{\partial g}\tfrac{d^3 g}{d\hat{\mathbf{c}}^2 d\hat{\mathbf{c}}_\ell}\tfrac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}\tfrac{\partial\hat{\mathbf{c}}_\ell}{\partial\theta_k} + \tfrac{\partial J}{\partial g}\tfrac{d^2 g}{d\hat{\mathbf{c}}^2}\tfrac{\partial^2\hat{\mathbf{c}}}{\partial\mathbf{y}\partial\theta_k}
\end{aligned} \tag{A.5}$$

$$\begin{aligned}
\tfrac{\partial^2\hat{\mathbf{c}}}{\partial\mathbf{y}\partial\boldsymbol{\theta}} &= -\left\{\left(\tfrac{dg}{d\hat{\mathbf{c}}}\right)'\tfrac{\partial^2 J}{\partial g^2}\tfrac{dg}{d\hat{\mathbf{c}}} + \tfrac{\partial J}{\partial g}\tfrac{d^2 g}{d\hat{\mathbf{c}}^2}\right\}^{-1}\left\{\tfrac{\partial^3 J}{\partial g\partial\mathbf{y}\partial\theta_k}\tfrac{dg}{d\hat{\mathbf{c}}} + \tfrac{\partial^3 J}{\partial g\partial\mathbf{y}\partial g}\tfrac{dg}{d\hat{\mathbf{c}}}\tfrac{dg}{d\hat{\mathbf{c}}_\ell}\tfrac{d\hat{\mathbf{c}}_\ell}{d\theta_k} + \tfrac{\partial^2 J}{\partial g\partial\mathbf{y}}\tfrac{d^2 g}{d\hat{\mathbf{c}}d\hat{\mathbf{c}}_\ell}\tfrac{d\hat{\mathbf{c}}_\ell}{d\theta_k}\right.\\
&\quad + \left(\tfrac{d^2 g}{d\hat{\mathbf{c}}d\hat{\mathbf{c}}_\ell}\tfrac{d\hat{\mathbf{c}}_\ell}{d\theta_k}\right)'\tfrac{\partial^2 J}{\partial g^2}\tfrac{dg}{d\hat{\mathbf{c}}}\tfrac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}} + \left(\tfrac{dg}{d\hat{\mathbf{c}}}\right)'\tfrac{\partial^3 J}{\partial g^2\partial\theta_k}\tfrac{dg}{d\hat{\mathbf{c}}}\tfrac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}} + \left(\tfrac{dg}{d\hat{\mathbf{c}}}\right)'\tfrac{\partial^3 J}{\partial g^3}\tfrac{dg}{d\hat{\mathbf{c}}}\tfrac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}\tfrac{dg}{d\hat{\mathbf{c}}_\ell}\tfrac{\partial\hat{\mathbf{c}}_\ell}{\partial\theta_k}\\
&\quad \left. + \left(\tfrac{dg}{d\hat{\mathbf{c}}}\right)'\tfrac{\partial^2 J}{\partial g^2}\tfrac{d^2 g}{d\hat{\mathbf{c}}d\hat{\mathbf{c}}_\ell}\tfrac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}\tfrac{\partial\hat{\mathbf{c}}_\ell}{\partial\theta_k} + \tfrac{\partial^2 J}{\partial g\partial\theta_k}\tfrac{d^2 g}{d\hat{\mathbf{c}}^2}\tfrac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}} + \left(\tfrac{dg}{d\hat{\mathbf{c}}_\ell}\tfrac{\partial\hat{\mathbf{c}}_\ell}{\partial\theta_k}\right)'\tfrac{\partial^2 J}{\partial g^2}\tfrac{d^2 g}{d\hat{\mathbf{c}}^2}\tfrac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}} + \tfrac{\partial J}{\partial g}\tfrac{d^3 g}{d\hat{\mathbf{c}}^2 d\hat{\mathbf{c}}_\ell}\tfrac{\partial\hat{\mathbf{c}}}{\partial\mathbf{y}}\tfrac{\partial\hat{\mathbf{c}}_\ell}{\partial\theta_k}\right\}
\end{aligned} \tag{A.6}$$

Bibliography

Arora, N. and L. Biegler (2004). A trust region sqp algorithm for equality constrained parameter estimation with simple parameteric bounds. *Computational Optimization and Applications 28*, 51–86.

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

Berry, S., R. J. Carroll, and D. Ruppert (2002). Bayesian smoothing and regression splilnes for measurement error problems. *journal of the American Statistical Association 457*(97), 160–169.

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

Brockwell, P. J. and R. A. Davis (1991). *Time Series: Theory and Methods*. New York: Springer-Verlag.

Campbell, D. and J. Cao (2006). Estimating differential equations with bayesian smoothing. Technical report, McGill University, Department of Mathematics and Statistics.

Cao, J. (2006). *Generalized Profiling Method and the Applications to Adaptive Smoothing, Generalized Semiparametric Additive Models and Estimating Differential Equations*. Phd thesis, McGill University.

Cao, J. and J. Ramsay (2007). Parameter cascades and profiling in functional data analysis. *journal of Computational Statistics (in press)*.

Dowd, M. (2007). Bayesian statistical data assimilation for ecosystem models using markov chain monte carlo. *Journal of Marine Research (in press)*.

Elton, C. and M. Nocholson (1914). The ten year cycle in numbers of the lynx in canada. *Journal of Animal Ecology 11*, 215–244.

Esposito, W. R. and C. Floudas (2000). Deterministic global optimization in nonlinear optimal control problems. *Journal of Global Optimization 17*, 97–126.

Gelman, A., F. Y. Bois, and J. Jiang (1996). Physiological pharmacokinetic analysis using population modeling and informative prior distributions. *journal of the American Statistical Association 91*(436), 1400–1412.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2004). *Bayesian data analysis* (2nd ed.). Texts in statistical science. Boca Raton, Fla.: Chapman & Hall.

Gelman, A. and D. B. Rubin (1992). Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science 7*, 457–472.

Geyer, C. (1991). Markov chain monte carlo maximum likelihood. In *Computing Science and Statistics: Proceedings of the 23rd Symposium on the interface*, pp. 156–163.

Hastie, T. and R. Tibshirani (2000). Bayesian backfitting. *Statistical Science 15*(3), 196–223.

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

Huang, Y. and H. Wu (2006). A bayesian approach for estimating antiviral efficacy in hiv dynamic models. *Journal of Applied Statistics 33*(2), 155–174.

Huber, P. J. (1981). *Robust Statistics*. New Jersey: John Wiley and Sons.

Ionides, E., C. Bretó, and A. King (2006). Inference for nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America 103*(49), 18438–18443.

Judd, K. (2003). Chaotic-time-series reconstruction by the bayesian paradigm: Right results by wrong methods. *Physical Review E 67*, 026212.

Kou, S. C., Q. Zhou, and W. H. Wong (2006). Equi-energy sampler with applications in statistical inference and statistical mechanics. *Annals of Statistics 34*, 1581.

Liu, J. S. and C. Sabatti (1998). Simulated sintering: Markov chain monte carlo with spaces of varying dimensions. In B. J.M., J. Berger, A. Dawid, and S. A.F.M. (Eds.), *Bayesian Statistics 6*. Oxford University Press.

Marinari, E. and G. Parisi (1992). Simulated tempering: A new monte carlo scheme. *Europhysics Letters 19*, 451–458.

Meyer, R. and N. Christensen (2000). Bayesian reconstruction of chaotic dynamical systems. *Physical Review E 62*, 3535.

Mukhin, D., A. Feigin, E. Loskutov, and Y. I. Molkov (2006). Modified bayesian approach for the reconstruction of dynamical systems from time series. *Physical Review E 73*, 036211.

Neyman, J. and E. L. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrika 16*, 1–32.

Poole, D. and A. E. Raftery (2000). Inference for deterministic simulation models: The bayesian melding approach. *Journal of the American Statistical Association 452*(95), 1244–1255.

Ramsay, J. O., G. Hooker, D. Campbell, and J. Cao (2007). Parameter estimation for differential equations: A generalized smoothing approach. *Journal of the Royal Statistical Society, Series B (in press)*.

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

Schaffer, M., K. B. McAuley, M. Cunningham, and E. K. Marchildon (2003). Experimental study and modeling of nylon polycondensation in the melt phase. *Ind. Eng. Chem. Res. 42*, 2946.

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

Varah, J. (1982). A spline least squares method for numerical parameter estimation in differential equations. *SIAM Journal on Scientific and Statistical Computing 3*(1), 28–46.

West, M. and J. Harrison (1997). *Bayesian Forecasting and Dynamic Models*. New York: Springer-Verlag.

Wilson, H. R. (1999). *Spikes, Decisions And Actions, The Dynamical Foundations Of Neuroscience*. Oxford University Press.

Zheng, W., K. B. McAuley, E. K. Marchildon, and K. Zhen Yao (2005). Effects of end-group balance on melt-phase nylon 612 polycondensation: Experimental study and mathematical model. *Ind. Eng. Chem. Res. 44*, 2675–2686.