Smooth Functional Tempering with application to Nonlinear Differential Equation Models

David Campbell *
Department of Statistics and Actuarial Science, Simon fraser University
and
Russell Steele
Department of Mathematics and Statistics, McGill University

Abstract

Differential Equations are used in modeling diverse behaviors in a wide variety of sciences. Traditionally methods for estimating the Differential Equation parameters $\boldsymbol{\theta}$ depend on augmenting the parameter space to include initial system states $\mathbf{x}_0$ and numerically solving the equations. This paper presents Smooth Functional Tempering a new population MCMC approach for posterior estimation of parameters. The proposed method borrows insights from parallel tempering and model based smoothing to define a sequence of approximations to the posterior with increased basins of attraction for the mode. The tempered approximations depend on relaxations of the solution to the differential equation model reducing or removing the need for $\mathbf{x}_0$ and a numerical differential equation solution. Rather than tempering via approximations to the posterior that are more heavily rooted in the prior, this new method tempers towards to data features, providing faster convergence, robustness to values used to initialize the algorithm and robustness of the algorithm to prior distributions that do not reflect the features of the data. Two variations of the method are proposed and their performance is examined through simulation studies and a real application to the chemical reaction dynamics of producing nylon. Matlab files are available online.

Keywords: Dynamic Systems, Parallel Tempering, Model Based Smoothing, Functional Data Analysis, Population MCMC, Multi-Grid MCMC

1 Introduction

Differential Equations (DEs) are used to model complex phenomena in pharmacokinetics, neurophysiology, chemical engineering, systems biology, climate models and other sciences. They are

*The authors gratefully acknowledge funding from NSERC

typically built from well understood scientific principles such as conservation of mass, energy and momentum, while providing an easily interpretable parameter vector $\boldsymbol{\theta}$ that is often unavailable with other classes of models. DEs describe the rate of change of a vector of functional system states $\mathbf{x}(t)$ with respect to an argument, such as time $t \in [0, T]$ through a functional regression model;

$$d\mathbf{x}(t)/dt - f(\mathbf{x}(t), \boldsymbol{\theta}, t) = 0, \tag{1}$$

where in the presence of measurement noise one observes

$$\mathbf{y}(t) = \mathbf{x}(t) + \epsilon(t), \quad \epsilon(t) \sim N(0, \sigma^2), \tag{2}$$

where only a subset of states may be observed. When there is no analytic solution for $\mathbf{x}(t)$, as often arises when $f(\cdot)$ is a nonlinear function, the initial system states $\mathbf{x}_0 = \mathbf{x}(0)$ are required to produce the numerical solution to the DE, $\mathbf{x}(t) = S(\boldsymbol{\theta}, \mathbf{x}_0, t)$ using, for example, a Runge-Kutta solver. Consequently, current Bayesian parameter estimation methods Gelman *et al.* (1996) and Huang and Wu (2006) use a model of the form:

$$\begin{aligned} P(\mathbf{y}(t) \mid \boldsymbol{\theta}, \mathbf{x}_0, \sigma^2) &\sim N\left(S(\boldsymbol{\theta}, \mathbf{x}_0, t), \sigma^2\right), \\ \boldsymbol{\theta}, \mathbf{x}_0, \sigma^2 &\sim P(\boldsymbol{\theta}, \mathbf{x}_0, \sigma^2). \end{aligned} \tag{3}$$

Through changes in $\mathbf{x}_0$, $\boldsymbol{\theta}$ and manipulated system inputs, a DE model can describe a wide variety of complex behaviors including oscillations, steady states, exponential growth and decay with a small number of parameters. However, the flexibility of a DE to succinctly model these behaviors comes at a heavy price. When centered on $S(\boldsymbol{\theta}, \mathbf{x}_0, t)$, the likelihood in (2) may be rife with undesirable topography such as local maxima, ridges, ripples and/or large flat segments Esposito and Floudas (2000). Gradient based methods like non-linear least squares (NLS)Bates and Watts (1988) do not typically perform well and practitioners are warned to expect an method based error level of the order of 25% Marlin (2000). Sampling based methods like Metropolis Hastings (MH) where the DE is numerically solved at each proposed parameter value may also have difficulty exploring the posterior surface under these topological difficulties. Examples of these problems are shown in section 4.2.

A recent frequentist approach to parameter estimation based on generalized profiling (GP) aims to improve the posterior topology by using a data smooth $\hat{\mathbf{x}}(t) \approx S(\boldsymbol{\theta}, \mathbf{x}_0, t)$ from a basis expansion. Estimates of $\boldsymbol{\theta}$ are determined by the profile likelihood marginalizing over the nuisance

parameters used to construct $\hat{\mathbf{x}}(t)$ Ramsay *et al.* (2007) . The data smoothing in GP is performed accounting for both the dynamics in (1) and the data features, providing an increased basin of attraction for the mode of $\boldsymbol{\theta}$. Smoothing, removes the dependence on the nuisance parameters $\mathbf{x}_0$ and improves stability of the estimate of $\boldsymbol{\theta}$. However it has been shown that sometimes the profile likelihood performs poorly for eliminating nuisance parameters Walley and Moral (1999). This paper describes a Bayesian version of this method that improves upon the frequentist counterpart by eliminating the need to depend on profiling, permitting inference about $\mathbf{x}_0$ and enabling inference in the presence of multi-modality.

We present a new Bayesian sampling method for posterior estimation of $\boldsymbol{\theta}$ and optionally $\mathbf{x}_0$ from DE models. The proposed smooth functional tempering (SFT) is a population MCMC method that borrows insights from parallel tempering (PT) and GP by using a model based data smooth to define a sequence of approximations to the posterior with increased basins of attraction for the modes. SFT does not require a-priori knowledge of the posterior topology, sequential MCMC or a bounded posterior space. Furthermore unlike PT, SFT is robust to situations where prior information is inconsistent with the data. Since SFT is a population MCMC method using model based smoothing, Section 2 reviews background methods and leads up to the description of two variants of SFT in Section 3. A simulation study is given in Section 4 followed by a real data case study in Section 5.

2 Background

The lack of an analytical form for $S(\boldsymbol{\theta}, \mathbf{x}_0, t)$ implies that there is no closed form for the likelihood. Furthermore, the challenging posterior topologies associated with DE models prevents a gradient based approach and parameter estimation therefore requires simulation-based methods such as MCMC (Gelman et al. 1996 and Huang et al. 2006) or simulated annealing Gonzales *et al.* (2007) with a numerical solution to the DE computed at each iteration. The dependence on $S(\boldsymbol{\theta}, \mathbf{x}_0, t)$ augments the parameter space with the inclusion of $\mathbf{x}_0$, a set of nuisance parameters that increase with additional experimental runs. While typically the structural parameters, $\boldsymbol{\theta}$, are of primary interest because they define the DE dynamics, current methods treat $\mathbf{x}_0, \boldsymbol{\theta}$ and σ^2 in (3) equally despite their differing influence on the data-generating process.

In the context of differential equation models, the problems with many sampling methods are

that the topology of the posterior and location of the dominant mode are difficult to determine, the posterior generally does not have a closed form expression and the parameter space may be unbounded and high dimensional. Furthermore, the posterior surface may have local maxima surrounded by deep and wide likelihood valleys making determining the global mode difficult. Figure 1 shows an example of a multimodal posterior surface where local modes associated with a partial fit to the data are of negligible posterior relevance.

2.1 Population MCMC

Population based simulation methods are designed to improve mobility of the parameters using information from parallel MCMC chains based on a sequence of approximations to the posterior density (see Jasra, Stephens and Holmes 2007 for a recent overview.) Parallel tempering (PT), for example, approximates the posterior of $\boldsymbol{\psi} = [\boldsymbol{\theta}, \mathbf{x}_0]$ through a sequence of $m = 1, \ldots, M$ approximations; $P_m(\boldsymbol{\psi} \mid \mathbf{y}) \approx P(\boldsymbol{\psi} \mid \mathbf{y})$ defined by a temperature gradient $0 \leq \lambda_1 < \ldots < \lambda_M =$ 1Geyer (1991). The m^{th} such approximation is

$$P_m(\boldsymbol{\psi} \mid \mathbf{y}) \propto \Big(P(\mathbf{y} \mid \boldsymbol{\psi})\Big)^{\lambda_m} P(\boldsymbol{\psi}). \tag{4}$$

At $\lambda_1 = 0$, $P_1(\boldsymbol{\psi} \mid \mathbf{y}) = P(\boldsymbol{\psi})$ and at $\lambda_M = 1$, $P_M(\boldsymbol{\psi} \mid \mathbf{y}) = P(\boldsymbol{\psi} \mid \mathbf{y})$ the posterior of interest. The M posterior approximations are the target densities of parallel Metropolis Hastings (MH) MCMC chains. The posterior approximations from smaller λ_m are based more heavily on the prior affording $\boldsymbol{\psi}_m$ greater mobility around the posterior parameter space compared to larger m chains. Consequently, the smaller m chains explore a wider parameter surface while the larger m chains remain trapped in the basin of attraction of a local posterior mode. Figure 2 shows the impact of changes in λ on the posterior surface of γ in the FitzHugh-Nagumo model to be discussed in Section 4.

At the i^{th} iteration, each chain independently performs a MH step to update $\boldsymbol{\psi}^{(i)}$. The M chains are not generated entirely independently. With some probability, two chains k and ℓ are randomly selected and their parameters $\boldsymbol{\psi}_k^{(i)}$ and $\boldsymbol{\psi}_\ell^{(i)}$ are proposed to exchange between the chains rather than mutate independently. The exchange is accepted with probability

$$r_{swap} = min\left(1, \frac{P_k(\boldsymbol{\psi}_\ell^{(i)} \mid \mathbf{y}) P_\ell(\boldsymbol{\psi}_k^{(i)} \mid \mathbf{y})}{P_k(\boldsymbol{\psi}_k^{(i)} \mid \mathbf{y}) P_\ell(\boldsymbol{\psi}_\ell^{(i)} \mid \mathbf{y})}\right).$$

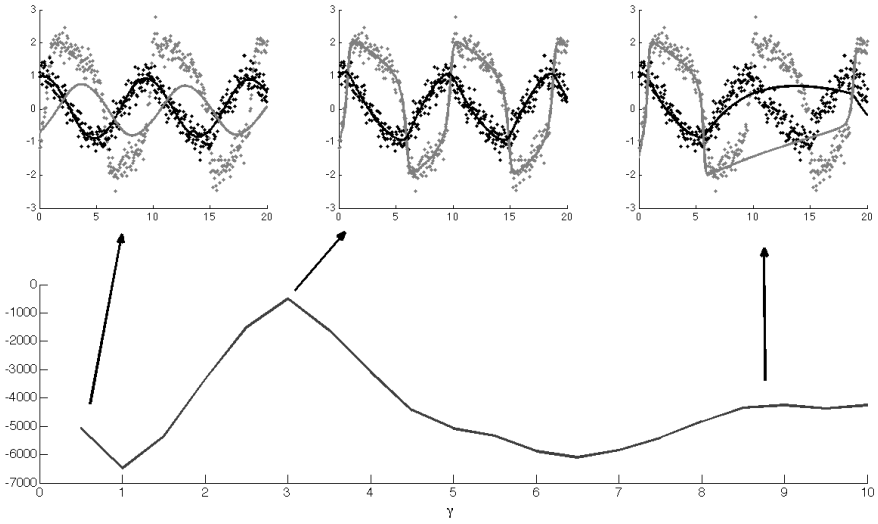

Figure 1: A cross section of the FitzHugh-Nagumo likelihood for γ (bottom) and the fits to the data for V (grey) and R (black) corresponding to the likelihood modes using the true parameter values (top middle), a small value (top left) and a large value (top right)

In the long run the proposed exchanges between neighboring chains should be accepted approximately 50% of the time to ensure reasonably smooth sequence of distributions Liu (2001). The exchange step enables multiple modes to be sampled and improves mixing.

PT and variants (such as Marinari and Parisi1992 or Neal 1996 for example or more specific to dynamic systems; Calderhead, Girolami and Lawrence 2009) have been shown to work well for sampling from multi-modal densities. Despite being less likely to get stuck in a local mode, the posterior flattening strategies that improve the mobility of some parameters may over-flatten parameter dimensions with less complex posterior topologies leading to slower mixing and burn-in in the target distribution Geyer and Thompson (1995). Additionally since tempering is towards the prior, PT will fail when prior information does not agree with the features of the observed data (see section 4.2).

2.2 Model Based Smoothing

Model based smoothing is a generalization of smoothing splines or penalized smoothing Eilers and Marx (1996). Using the vector of basis functions $\boldsymbol{\phi}(t)$ and coefficients $\mathbf{c}$, the smooth $\mathbf{x}(t) = \mathbf{c}'\boldsymbol{\phi}(t)$

is the location parameter in the likelihood, for example:

$$\mathbf{y}(t) \mid \mathbf{x}(t), \sigma^2 \sim N(\mathbf{x}(t), \sigma^2). \tag{5}$$

The prior on the shape of the smooth depends on the hyper-parameter γ and

$$\begin{aligned} P(\mathbf{x}(t) \mid \boldsymbol{\theta}, \gamma) &\propto exp(-\tfrac{\gamma}{2}\texttt{PEN}) \\ \text{where} \quad \texttt{PEN} &= \int_t \left[\frac{dx(s)}{ds} - f(x(s), \boldsymbol{\theta}, s) \right]^2 ds \\ \text{and} \quad \theta, \gamma, \sigma^2 &\sim P(\theta)P(\gamma)P(\sigma^2). \end{aligned} \tag{6}$$

Often in smoothing literature $\texttt{PEN} = \int_t \left(d^2x(s)/ds^2 - 0\right)^2 ds$ which defines a model where prior information anticipates a linear model, whereas in (6), the penalty is more generally based on the integrated square of the residual of (1). The prior on $\mathbf{x}(t)$ increases in density as $\mathbf{x}(t)$ approaches the shape defined by the DE model through PEN. Model parameters $\boldsymbol{\theta}$ from (1) are hyper-parameters to the prior on $\mathbf{x}(t)$.

The smoothing parameter $\lambda = \gamma\sigma^2$ defines the posterior balance between measurement error σ^2 and deviation from the model. As $\gamma \to 0$, the posterior mode of $\mathbf{x}(t) \mid \mathbf{y}, \boldsymbol{\theta}, \sigma^2, \gamma$ is the function space spanned by the basis that interpolates the data. As $\gamma \to \infty$, the posterior mode of $\mathbf{x}(t) \mid \mathbf{y}, \boldsymbol{\theta}, \sigma^2, \lambda$ occurs on the function space spanned by the DE solution.

Model based smoothing was not designed for optimal estimation of $\boldsymbol{\theta}$ when the parametric structure of (1) is assumed. To highlight this, note that γ controls the flow of information between $\mathbf{y}$ and $\boldsymbol{\theta}$ since $\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{x}(t), \gamma, \sigma^2 = \boldsymbol{\theta} \mid \mathbf{x}(t), \gamma, \sigma^2$. Consequently, using model based smoothing, the impact of changes in $\boldsymbol{\theta}$ on $\mathbf{x}(t)$ is reduced and the posterior variance for $\boldsymbol{\theta}$ is inflated compared to estimating $\boldsymbol{\theta}$ via (3) without the hierarchical layer of the data smooth.

In some cases $\mathbf{x}_0$ may be known to high precision, but remaining trajectory $\mathbf{x}(t, \mathbf{x}_0)$ must be estimated. In general these initial value problems could be computed using constrained optimization, however the computation is simplified using a B-spline basis since there is only one basis function taking a non-zero value at each of the time interval boundaries. In terms of parameter estimation, if $\mathbf{x}_0$ is known, this additional information can improve reliability in the estimation of $\boldsymbol{\theta}$, especially when the model is sensitive to initial conditions Wu *et al.* (2008).

3 Smooth Functional Tempering (SFT)

SFT, similar to PT, is a population based algorithm using parallel MCMC chains defined by a sequence of M distributions approximating the posterior of the measurement error model in (3). However, SFT is a collocation tempering method. That is, it depends on a basis expansion for the approximation $\mathbf{x}(t) = \mathbf{c}'\boldsymbol{\phi}(t) \approx \mathbf{S}(\boldsymbol{\theta}, \mathbf{x}_0, t)$, where the tempering is defined by the smoothing parameter. When using a B-spline basis, as the smoothing parameter increases and the model is more rigorously enforced, $\mathbf{x}(t) \to \mathbf{S}(\boldsymbol{\theta}, \mathbf{x}_0, t)$, where $\mathbf{S}(\boldsymbol{\theta}, \mathbf{x}_0, t)$ is computed using an implicit Runge-Kutta method with stepping points at the knot locations and $\mathbf{x}_0 = \mathbf{c}'\boldsymbol{\phi}(t = 0)$ Deuflhard and Bornemann (2000). Consequently basing the tempering process on a collocation method is equivalent to basing the tempered chains on a relaxation to the DE solution.

In this section we outline two variations of this process, the first (SFT1) uses a smooth approximation to the initial value problem using a fixed point in the data smoothing step in conjunction with a numerical DE solution. The second (SFT2) uses smooth approximations and does not depend on numerical DE solutions or $\mathbf{x}_0$.

3.1 SFT1: Parameter Estimation with a Smooth and a Numerical DE Solution

In some cases we are interested in $\mathbf{x}_0$ and/or the function space spanned by the possible DE solutions is of inferential interest such that we would like to use model (3). SFT1 defines a tempering strategy towards model (3) based on the increasing sequence of fixed smoothing parameters $0 < \lambda_1 \leq \ldots \leq \lambda_M = \infty$:

$$\begin{aligned} &P_m(\mathbf{y} \mid \mathbf{x}_m(t, \mathbf{x}_0), \sigma^2) \sim N\left(\mathbf{x}_m(t, \mathbf{x}_0), \sigma^2\right) \\ &P_m(\boldsymbol{\theta}, \mathbf{x}_0, \sigma^2) \propto \exp\left(-\lambda_m \int_t \left[\tfrac{d}{dt}\mathbf{x}_m\Big(s, \mathbf{x}_0\Big) - f(\mathbf{x}_m\Big(s, \mathbf{x}_0\Big), \boldsymbol{\theta}, s)\right]^2 ds\right) P(\boldsymbol{\theta})P(\mathbf{x}_0)P(\sigma^2) \end{aligned} \tag{7}$$

The innovation in this model compared to model based smoothing with a fixed initial value in section 2.2 is that SFT1 removes one layer of the hierarchical model by implicitly defining a distribution on the smooth and using fixed values of λ_m. As with model based smoothing, in SFT1 as $\lambda_m \to 0$, the posterior mean for $\mathbf{y}$ tends towards a data interpolant because the induced prior for $\mathbf{x}(t)$ is uniform over the function space spanned by the basis. In addition when $\lambda_m = 0$,

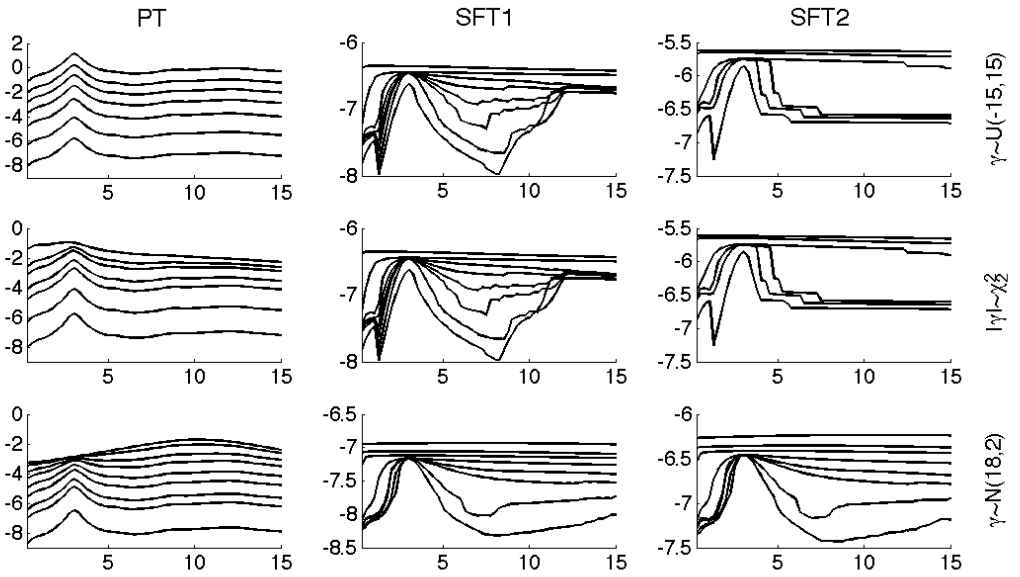

Figure 2: The effect of changing λ on the -log(-log(non-normalized posterior)) for parameter γ in the FitzHugh-Nagumo model using the model in PT, SFT1 and SFT2 in the left, middle and right columns respectively. Models are compared using uniform, χ^2 and Gaussian based priors in the top, middle and bottom rows. Increasing values of λ give lines generally appearing lower down within each plot.

$P_m(\boldsymbol{\theta} \mid \mathbf{y}) = P(\boldsymbol{\theta})$ since $\boldsymbol{\theta}$ is not used to define the shape of $\mathbf{x}(t, \mathbf{x}_0)$. Ensuring $\lambda_M = \infty$ means that

$$\exp\left(-\lambda_m \int_t \left[\frac{d}{dt}\mathbf{x}_m\Big(s, \mathbf{x}_0\Big) - f(\mathbf{x}_m\Big(s, \mathbf{x}_0\Big), \boldsymbol{\theta}, s)\right]^2 ds\right) = \begin{cases} 1 & \text{if } \mathbf{x}(t, \mathbf{x}_0) = S(\boldsymbol{\theta}, \mathbf{x}_0, t) \\ 0 & \text{otherwise} \end{cases} .$$

Consequently, the M^{th} chain is the DE measurement error model in (3) but the $M-1$ chains having $\lambda_m < \infty$, use a relaxation of the DE solution enabling $\mathbf{x}(t, \mathbf{x}_0)$ to deviate from the dynamics in (1) to better match the features of the data. While PT tempers towards the prior, SFT1 tempers towards the data features. Additionally, in SFT1, the impact of changes in $\boldsymbol{\theta}$ and $\mathbf{x}_0$ on $\mathbf{x}(\mathbf{x}_0, t)$ are reduced with decreasing λ_m.

3.2 SFT2: Parameter Estimation Without a Numerical DE Solution

In many situations, $\mathbf{x}_0$ is not directly of interest but is required to numerically produce $S(\boldsymbol{\theta}, \mathbf{x}_0, t)$. Furthermore the numerical solution may be difficult to produce or may be subject to propagating numerical errors. SFT2 avoids the potential liability of numerically solving the DE and eliminates the reliance on $\mathbf{x}_0$ by tempering via the sequence of distributions for $0 < \lambda_1 \leq \ldots \leq \lambda_M \leq \infty$;

$$\begin{aligned} P_m(\mathbf{y} \mid \boldsymbol{\theta}, \sigma^2) &= N\Big(\mathbf{x}_m(t), \sigma^2\Big) \\ P_m(\boldsymbol{\theta}, \sigma^2 \mid \mathbf{y}) &\propto \exp\left(-\lambda_m \int_t \left[\tfrac{d}{dt}\mathbf{x}_m(s) - f\Big(\mathbf{x}_m(s), \boldsymbol{\theta}, s\Big)\right]^2 ds\right) P(\boldsymbol{\theta}) P(\sigma^2). \end{aligned} \tag{8}$$

As with SFT1, SFT2 uses fixed values of λ_m and induces a distribution on $\mathbf{x}(t)$. However SFT2 no longer requires $S(\boldsymbol{\theta}, \mathbf{x}_0, t)$ because $\mathbf{x}_M(t)$ can be made arbitrarily close to the $S(\boldsymbol{\theta}, \mathbf{x}_0, t)$ as determined by the value of λ_M. In practice $\lambda_M < \infty$ for computational reasons and because large values of λ_M ensure that the induced posterior on $\mathbf{x}_M(t)$ decays rapidly towards zero as $\mathbf{x}_M(t)$ deviates from the function space of the DE solution. However, even at $\lambda_M = \infty$ model (8) is not equivalent to (3), because SFT2 effectively profiles over $\mathbf{x}_0$.

3.3 Impact of λ and Choosing Values

Figure 2 shows the impact of changes in λ on a cross section of the posterior surface of the γ parameter in the FitzHugh-Nagumo model (discussed in section 4) based on SFT1, SFT2 and PT using 3 different priors. The effectiveness of the tempering in PT changes drastically with the prior. Using the uniform distribution or χ^2_2 based priors, the minor mode at $\gamma = 13$ eventually disappears with any tempering. However the minor mode becomes relatively more important in PT using the $N(14, 2)$ prior as λ_m decreases. Instead of tempering $P(\gamma \mid \mathbf{y})$ towards the prior, SFT methods temper the posterior function space of $\mathbf{x}(t)$ towards the data. Consequently the effectiveness of tempering is not as adversely impacted by changes in the prior.

Since SFT2 does not use a fixed value of $\mathbf{x}_0$, it has flexibility to induce additional smoothness into the topology of the tempered posterior. Consequently, the posterior modes for large λ values around $\gamma = 13$ in the PT and SFT1 posterior plots are avoided when using SFT2.

If λ_m is small then the m^{th} posterior approximation will have a larger posterior variance for $\boldsymbol{\psi}$ due to its reduced impact on $\mathbf{x}_m(t)$ or $\mathbf{x}(t, \mathbf{x}_0)$. This provides considerable robustness to parameter values used to initialize the algorithm and produces a wide basin of attraction for the

target posterior modes. To exploit this benefit, we propose the rule of thumb that the smallest value of λ should be able to approximate the data dynamics, or if in doubt, it should nearly interpolate the data despite the values used to initialize the algorithm. The other values of λ can be determined by increasing λ on the log scale until the discrepancy between neighboring chains permits an adequate exchange acceptance rate.

When using SFT2 the value of `PEN` should be examined to ensure that it is sufficiently small compared to the sum of squared residuals to enforce adequate fidelity to the model at λ_M. If λ_M is further increased the computation time will increase with negligible improvement in the approximation $\mathbf{x}(t) \approx S(\boldsymbol{\theta}, \mathbf{x}_0, t)$

3.4 Choice of Basis and Computation of Smoothing Criteria

In general B-splines permit considerable flexibility in shape allowing high order smooth or discontinuous derivatives where needed making them a convenient choice for SFT. However, other bases such as Fourier, wavelet or (truncated) polynomial bases are also used for smoothing and producing solutions to DEs and may provide additional advantages in some problems. The type and number of basis functions used must permit $\mathbf{x}(t)$ to accommodate the DE model dynamics and deviations thereof for a wide range of values of $\boldsymbol{\theta}$, so there may be a need for far more basis functions than observations, especially if the dynamics are complex. Neither SFT1 nor SFT2 explicitly sample $\mathbf{c}$ so having a large number of basis functions does not complicate the convergence or tuning of the chains.

The integral terms in (7) and (8) can be computed through numerical quadrature. When using a B-spline basis having quadrature points at the unique knot locations produces a computationally fast result. Some relevant discussion about quadrature and `PEN` in model based smoothing can be found in the discussion of Ramsay et al. 2007.

3.5 Prior Specification on $\boldsymbol{\theta}$

As an additional practical note, we emphasize here that care must be taken in producing a prior on $\boldsymbol{\theta}$ in DE systems. A prior should be placed on the shape of $S(\boldsymbol{\theta}, \mathbf{x}_0, t)$ and transformed to the parameter space $[\boldsymbol{\theta}, \mathbf{x}_0]$. Placing a proper prior on $[\boldsymbol{\theta}, \mathbf{x}_0]$ directly may lead to improper posterior distributions and increase the topological difficulties Bates and Watts (1988). In practice, a prior

on the function space is often easier to deal with than transforming distributions on function spaces into priors on $\boldsymbol{\theta}$.

4 Simulated examples from the FitzHugh-Nagumo model

The FitzHugh-Nagumo differential equations (FitzHugh 1961 and Nagumo, Arimoto and Yoshizawa 1962) are a simple model for the voltage potential across the cell membrane of the axon of giant squid neurons. These equations are used in neuro-physiology as an approximation of the observed spike potential. The voltage V moving across the cell membrane depends on the recovery variable R through the relationship:

$$\frac{dV}{dt} = \gamma\left(V - \frac{V^3}{3} + R\right), \qquad \frac{dR}{dt} = -\frac{1}{\gamma}\left(V - \alpha + \beta R\right). \tag{9}$$

An example of a simulated data set and the true underlying process based on $\gamma = 3$ appears in figure 1 along with a cross section of the log likelihood. Figure 1 also includes additional DE solutions using parameter values corresponding to minor modes of the cross section. The mode corresponding to values of $\gamma \approx .5$ produces a DE solution with the correct period but the shape is too sinusoidal to represent the dynamics of V. The mode corresponding to values of $\gamma \approx 9$ produces approximately the correct shape but does not match the period. To move from the local modes, γ causes a deterioration in the data fit before it can be improved, thereby producing wide regions of prohibitively deep posterior topology of approximately exp(4000) units deep on the log scale. We consider a bimodal version of this model to highlight the ability to accurately estimate the posterior in section 4.1. We compare Bayesian and frequentist methods using a one dimensional version of this model in the presence of prior information that is inconsistent with the data features in section 4.2. The full FitzHugh-Nagumo model is explored in section 4.3.

4.1 One Dimensional Bimodal Example

In this section we alter (9) to produce a symmetric, bimodal posterior for γ;

$$\frac{dV}{dt} = \mid \gamma \mid \left(V - \frac{V^3}{3} + R\right), \qquad \frac{dR}{dt} = -\frac{1}{|\gamma|}\left(V - \alpha + \beta R\right). \tag{10}$$

Due to the computationally intensity of working with differential equations, ten simulated data sets were obtained from the numerical solution to (10) using the parameter $\gamma = 3$ at the 201 evenly

spaced time points $t = 0, 0.1, 0.2, \ldots, 20$ with added Gaussian white noise. Focusing attention on γ, all other parameters are held fixed at their true values $[\alpha, \beta, \sigma_V^2, \sigma_R^2, V_0, R_0] = [.2, .2, .25, .16, -1, 1]$ so that the posterior density can be evaluated numerically and compared with results from SFT1, SFT2 and PT under two prior distributions:

$$\begin{cases} P(\gamma) = \frac{1}{2}\chi_2^2, & \gamma > 0 \\ P(-\gamma) = \frac{1}{2}\chi_2^2, & \gamma < 0 \end{cases} \tag{11}$$

and

$$P(\gamma) = Uniform(-15, 15) \tag{12}$$

Priors had little influence on the posterior, which could be reasonably approximated by 2 identical Gaussians whose means are separated by 312 standard deviations. SFT1 and SFT2 were set up with 101 evenly spaced knots from a 5^{th} order B-spline basis with one quadrature point at the unique knot locations. From the largest λ_M for each method, parallel chains were added by tuning the next value of λ_m to approach the swap acceptance rate of 50%.

Using both of the prior distributions for γ, the numerically evaluated posterior (P_{num}) was compared with the results of the sampling based methods through

$$D(\hat{P}_{sampled}) = \int \Big[P_{num}(\gamma \mid \mathbf{y}) - \hat{P}_{sampled}(\gamma \mid \mathbf{y})\Big]^2 d\gamma. \tag{13}$$

These values are shown in figure 3 for the M^{th} chains using the last 40,000 posterior draws after discarding burn-in.

The M^{th} chains of SFT1 and PT use the same target distribution but PT performed somewhat better than SFT1 with a uniform prior on γ but somewhat worse with a χ^2 based prior. SFT1 and PT assume that $\mathbf{x}_0$ were known exactly but SFT2 estimates γ without this additional knowledge. Consequently, SFT2 uses less information which translates into a posterior variance around the modes of $\gamma = \pm 3$ which approximately 7 times wider than that using SFT1 or PT. Consequently, $D(\hat{P}_{SFT2})$ was computed comparing the sampled density with the numerical estimate of it's smooth based density from the M^{th} chain.

Figure 4 shows the autocorrelation functions (ACFs) and their point-wise mean ACFs for the posterior samples of the λ_M chains. The main factor dominating the ACF is the exchange between the modes at ± 3. In general, the ACFs for SFT2 perform the best in part due to the lack of dependence on the initial conditions. SFT1 generally ranks second of these methods likely

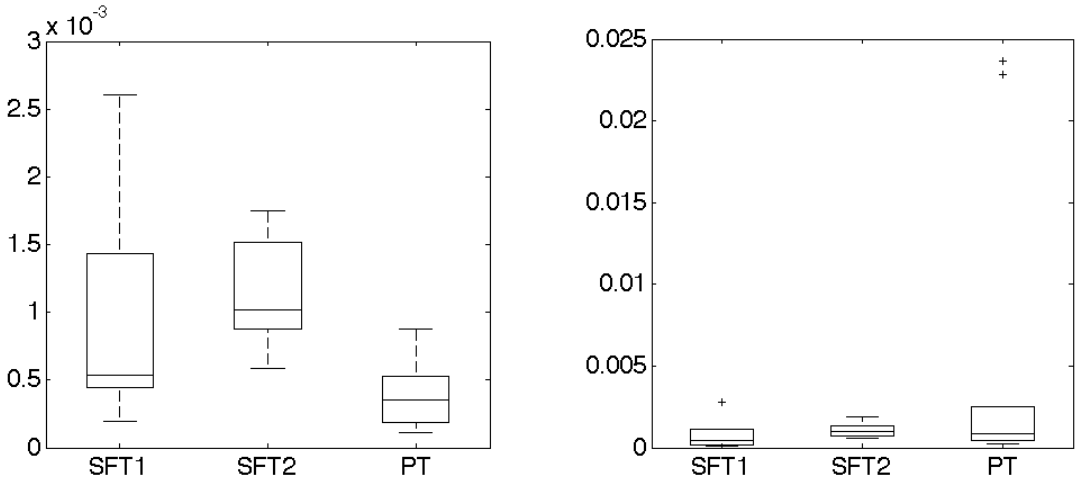

Figure 3: Discrepancy between sampled and numerical posterior estimates using different prior distributions and sampling methods. Boxplots show $D(\hat{P})$ using the uniform prior (left) and the χ^2 based prior (right).

due to the reduced impact of initial conditions in the finite λ_m parallel chains. The ACFs for PT are generally slowest to decay. The ordering of the ACFs are not impacted by the choice of prior distribution in this example.

4.2 Inconsistent Prior Information

In this section we focus on the one dimensional problem of estimating only $P(\gamma \mid \mathbf{y})$ using the FitzHugh-Nagumo model (9) with a prior that is inconsistent with the observed data: $\gamma \sim N(18, 2^2)$. The bottom row of figure 2 shows that the global mode of the target posterior at $\gamma = 3$ remains virtually unchanged by this change in prior. Parameter estimation was attempted using SFT1, SFT2, PT, single chain Metropolis Hastings (MH), NLS and GP on 10 data sets from the measurement error model from section 4. SFT1 and SFT2 were performed with 4 parallel chains each and PT was equipped with 10. All chains in all methods were initialized at $\gamma = 10$. The number of burn-in iterations determined by Raftery-Lewis Raftery and Lewis (1992) and was less than 125 in all cases from this starting point. After discarding 1,000 iterations, Raftery-Lewis and Geweke Geweke (1989) confirm convergence from all of the independent chains from all the sampling methods.

Figure 5 also shows a boxplot of the final parameter estimates. MH and NLS are not able to

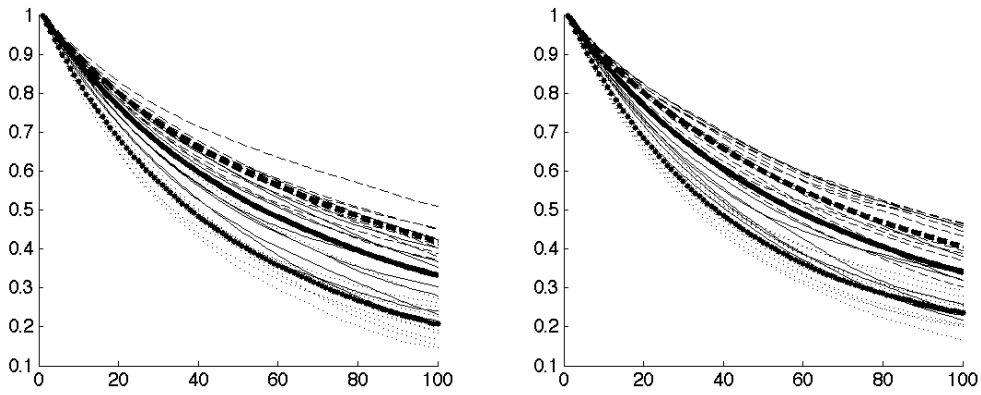

Figure 4: The autocorrelation functions for the SFT1 (solid line), SFT2 (dotted line) and PT (dashed line) for the bimodal problem of section 4.1, with the uniform prior (left) and the χ^2 based prior (right). The heavy lines are the point-wise mean autocorrelation functions for each model.

escape the strong gradient towards the local mode at $\gamma = 12$. The strategy of tempering towards the prior hindered any of the PT chains from finding the global mode because the smaller λ chains enforce behavior inconsistent with the data features and emphasize the local mode at $\gamma = 12$ within the alloted 100,000 iterations.

Both SFT1 and SFT2 used the increased basin of attraction of their smaller λ parallel chains and tempering towards data features to avoid the impact of the inconsistent prior information. GP also smoothes the likelihood towards the data features and the point estimate converged quickly close to the true value. Since $\mathbf{x}_0$ is assumed known, SFT1 uses this additional information to perform better than SFT2 and GP.

4.3 FitzHugh-Nagumo, full model

While the previous simulations showed the ability of the methods to produce reasonable results in a single dimension, the performance of SFT2 was limited because it used less information than methods relying on the ODE solution. In this section we use model (9) with simulated data from the more realistic scenario where all of the parameters must be estimated. Prior distributions for $\boldsymbol{\theta}$ were determined by numerically solving the DE over a coarse grid of values of $\boldsymbol{\theta}$ and placing

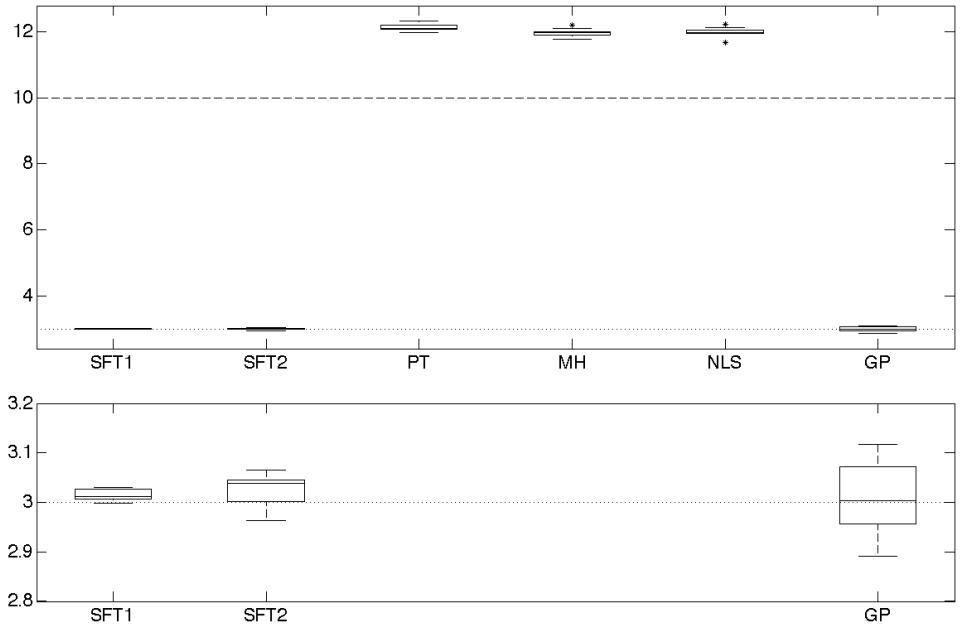

Figure 5: Boxplots of the estimates of γ in section 4.2, the dashed line is where the methods where initialized, the true parameter value is 3. Top: Estimates for all 6 methods, bottom, rescaled to show detail.

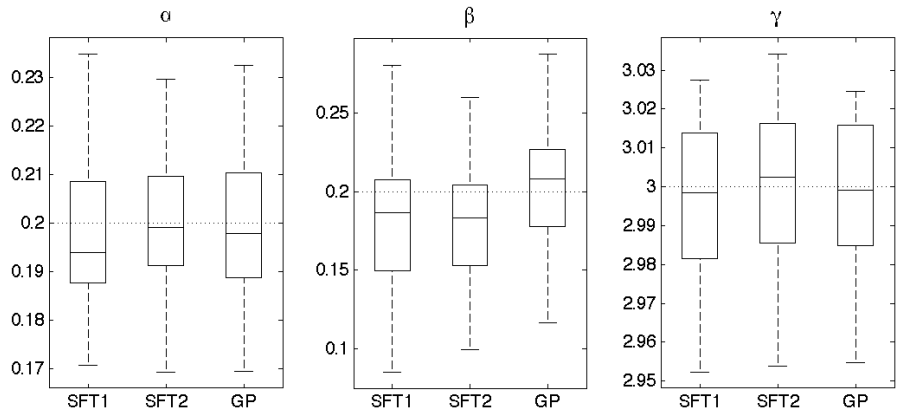

Figure 6: Bias in point estimates for the FitzHugh-Nagumo parameters α (left), β (middle) and γ (right) of section 4.3.

approximately 95% of the prior mass over the values that produce oscillatory dynamics giving:

$$\gamma \sim \chi^2_2, \qquad P(\alpha) = P(\beta) = N(0, .4^2). \tag{14}$$

Priors on V_0 and R_0 were empirically chosen Gaussian densities centered on the initial observations with variance equal to the observed data variance about its mean. Priors on σ^2 were assumed near zero but long tailed: $P(\sigma^2_{V,R}) \propto 1/\sigma^2_{V,R}$. In this simulation study we used 30 different data sets, each with 401 evenly spaced observations for each of V and R. This large amount of data ensured that the likelihood was well approximated by multivariate Normal distributions, making the delta method interval estimates of Ramsay et al. 2007 good approximations.

For these simulations we focus on SFT1, SFT2 and GP because other competing methods were already shown to be inadequate in section 4.2. Parameters were initialized with draws from the prior. All parallel chains (across all methods) were initialized with the same values. SFT1 and SFT2 used 4 parallel chains and GP was performed using an increasing sequence of λ values as suggested in Ramsay et al. 2007 such that SFT2 and GP have the same value of λ_M=10,000. The point estimates are shown in figure 6 based on 30,000 posterior draws after burn-in. Bias is small, approximately centered on zero and there is no significant differences between the performance of the methods in this example.

5 Nylon example

This section models the production of nylon in a heated reactor where it's constituents, amine (A) and carboxyl (C) combine producing the polymer nylon (L) and water (W) which escapes as steam. At the same time, before leaving as steam, W in the molten nylon mixture, decomposes L into A and C giving the symbolic competing reactions $A + C \rightleftharpoons L + W$. In the experiment of Zheng, McAuley, Marchildon and Zhen 2005 steam is bubbled through molten nylon to maintain an approximately constant concentration of W in the system causing A, C and L to move towards equilibrium concentrations with 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 it returned to its original level for the remainder of the experiment. Each experiment was performed at a constant temperature T_i which, along with the input water pressure, determines the equilibrium concentration of water in the molten nylon mixture, W_{eq}. Using reaction rates k_p and K_a, the dynamics are described with differential equations:

$$-\frac{dL}{dt} = \frac{dA}{dt} = \frac{dC}{dt} = -k_p(CA - LW/K_a)10^{-3} \quad \text{and} \tag{15}$$

$$\frac{dW}{dt} = k_p 10^{-3}(CA - LW/K_a) - 24.3(W - W_{eq}). \tag{16}$$

The reaction rates are allowed to change with T_i and W_{eq} through relationships depending on the reference temperature $T_0 = 549.15$ giving four DE parameters: $\boldsymbol{\theta} = [k_p, \gamma, K_{a0}, \Delta H]$;

$$K_a = \left\{1 + W_{eq}\gamma 10^{-3}\right\} K_T [K_{a0}] \ell \left(\frac{\Delta H}{8.314}\right) \tag{17}$$

$$\ell(m) = \exp\left(-m10^3 \left\{\frac{1}{T_i} - \frac{1}{T_0}\right\}\right) \tag{18}$$

$$K_T = 20.97 \exp\left(-9.624 + \frac{3613}{T_i}\right). \tag{19}$$

Figure 7 shows the data for each of the 6 experimental runs. The plot shows the observed components A and C as well as input W_{eq}. Due to the mass balance of this system, given any three components the fourth can be computed. Since only A and C are observed, we must estimate the unobserved $W(t)$ for each experimental run. Furthermore since the components are chemical reactions they are constrained to take on non-negative values. In the nylon system $\mathbf{x}_0$ increases the dimension of the parameter space from 6 parameters in $[\boldsymbol{\theta}, \sigma^2]$, to 24 parameters; $[\boldsymbol{\theta}, \mathbf{x}_0, \sigma^2]$.

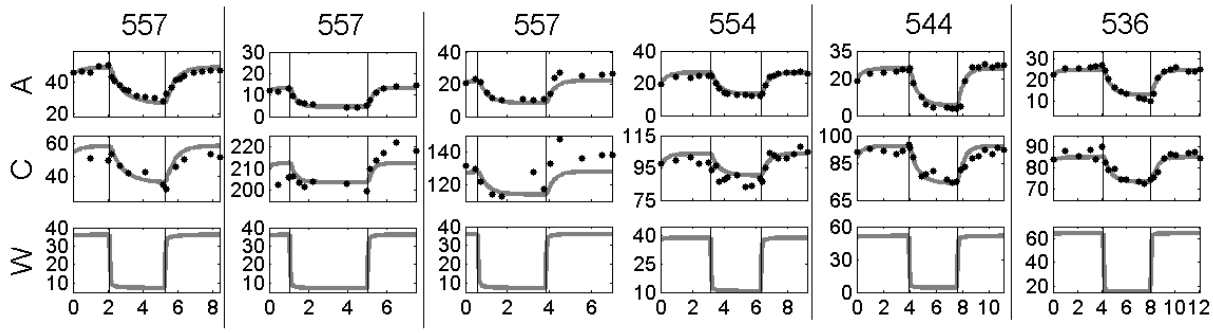

Figure 7: The nylon observations along with the fit to the data. Temperatures of the experimental runs are given above component A in degrees Kelvin. Vertical axes are in concentration units and horizontal axes are in hours.

Priors were indirectly set by placing a uniform distribution on the set of functions taking values between 0 and 250, where 250 was selected because it its about 10% larger than the largest observation and the interval is wider than expected to be necessary. Values of the unobserved W are expected to remain close to the values of W_{eq} which are all less than 100 but the more conservative value of 250 was used consistently for the states A, C and W. Additionally, the prior distributions on $1/\sigma_A^2$ and $1/\sigma_C^2$ were independent Gamma densities with mean 9 and variance 27, chosen to more pessimistic than the measurement error variance estimates from additional experiments by Zheng et al. 2005.

SFT1 and SFT2 used evenly spaced knots at a rate of 3 per hour of experimental duration. In anticipation of sharp dynamics after the step change in input W_{eq}, an additional 9 knots were evenly spaced at times $\tau+[0.1, 0.2, ..., 0.9]$ after the input change. The discontinuous first derivative induced by the step input change was accommodated by the addition of knots at the time of the step change. SFT2 used $\lambda_1 = 100$, and $\lambda_2 = 10,000$. SFT1 had four times the parameters of the SFT2 model and consequently $M = 3$ chains were used, with values $\lambda_1 = 200, \lambda_2 = 500$ and $\lambda_3 = \infty$.

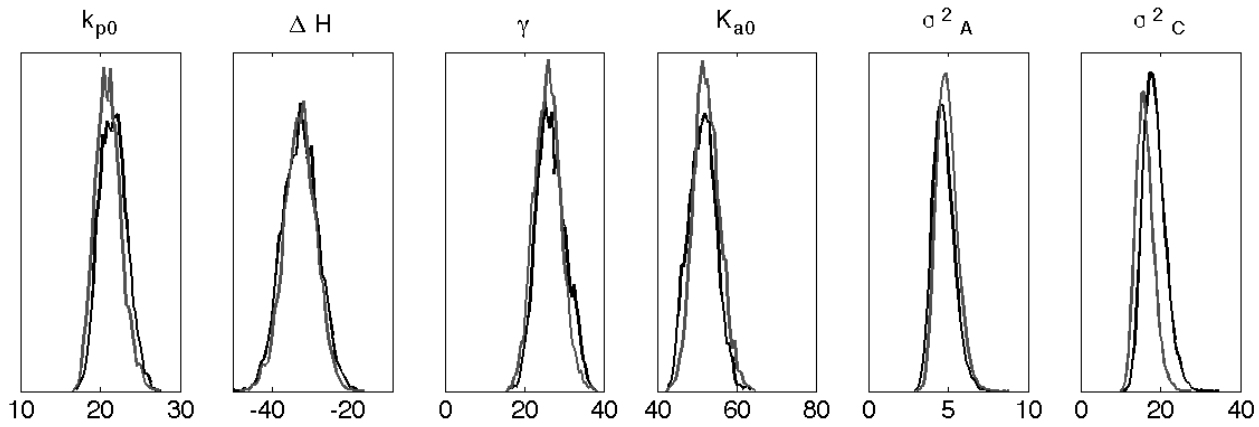

Figure 8: A comparison of the posterior density estimates for the nylon parameters using SFT1 (black line) and SFT2 (grey line)

5.1 Results

The small values of λ_1 in both methods produced considerable robustness with respect to values used to initialize the Markov Chains. The kernel density estimate of 40,000 posterior draws from the M^{th} chain of SFT1 and SFT2 for $\boldsymbol{\theta}, \sigma_A^2$ and σ_C^2 (after discarding burn-in) are shown in figure 8. Estimates for the marginal posterior densities of $\boldsymbol{\theta}$ are close between the methods, and the values for $D(P_{SFT1}(\boldsymbol{\theta} \mid, \mathbf{y}, \sigma^2), P_{SFT2}(\boldsymbol{\theta} \mid \mathbf{y}, \sigma^2))$ are .057, .016, .018 and .0073 for k_{p0}, γ, K_{a0} and ΔH respectively. The squared discrepancy between the marginal posterior density estimates deviates slightly more for σ_A^2 and σ_C^2 giving values of $D(P_{SFT1}(\sigma^2 \mid, \mathbf{y}, \boldsymbol{\theta}), P_{SFT2}(\sigma^2 \mid, \mathbf{y}, \boldsymbol{\theta}))$ equal to .11 and .15 respectively. The reason for this discrepancy may lie in the marginal posterior density estimates of $\mathbf{x}_0$, estimated by SFT1 and shown in figure 9. The dynamics of the system are fast such that the impact for some of the experimental runs on moving W_0 from near 0 to near 250 only affects the fit to the first few data points leading to some flat posteriors. SFT1 explores the distribution of X_0 and in the process finds more values that allow a better fit to A in exchange for a decrease in fit to C giving the shifted densities for σ_C^2 and σ_A^2 shown in figure 8.

The advantages of SFT2 are the reduced dimension of the problem along with, in this case, a five fold computational time reduction. However SFT1 produces new insights into the vast uncertainty in W_0.

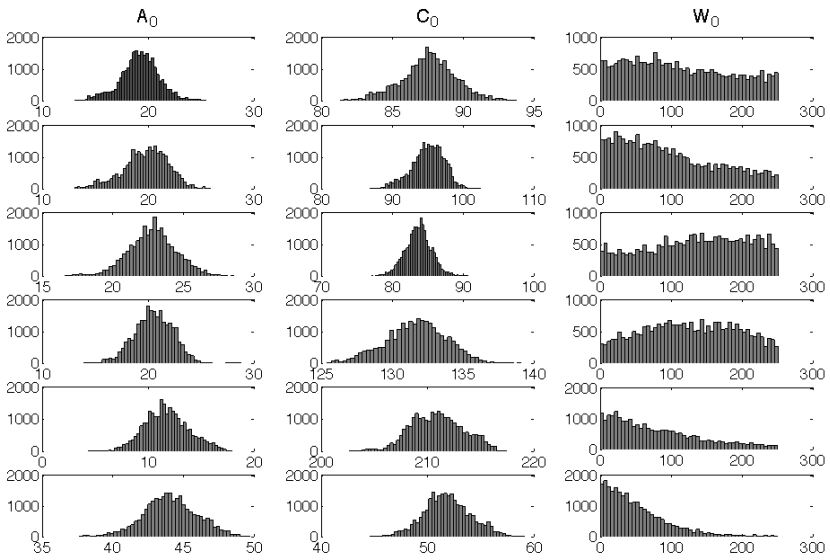

Figure 9: Histograms of posterior draws $\mathbf{x}_0$ in the nylon system using SFT1. Rows are for the different experimental runs, while columns are (left to right) A_0, C_0 and W_0.

6 Discussion

Parameter estimation for nonlinear differential equations is challenging for standard methods like nonlinear least squares and basic metropolis hastings, where, despite the resemblance of convergence, the likelihood topology may not permit convergence towards the global optima. The proposed SFT combines PT and model based smoothing to improve the topological burden by matching the features of the data with the dynamics of the model. This variation of tempering smooths out the posterior enabling faster convergence towards the dominant mode, and as such, represents an important new tool for population-based MCMC simulation. While the simulations and application presented feature nonlinear differential equation models, the methods are applicable to nonlinear regression in general, especially when the response surface is prohibitive.

SFT1 and SFT2 temper towards the data features to improve posterior mobility, whereas PT was shown to fail when tempering towards a prior that is inconsistent with the data in section 4.2. Using SFT, priors can therefore be used to describe knowledge about the system without needing to also account for it's utility in providing an adequate tempering strategy. Furthremore, SFT was able to traverse the difficult topology of section 4.2 using 4 parallel chains where PT was

unable to do so while using 10 chains.

In the presence of prior information consistent with the data features, SFT1, SFT2 and PT perform similarly at estimation as was measured by integrated squared error discrepancy between simulated and numerical posterior estimates. However SFT has the advantage of reduced dependence on initial conditions which in turn reduced the autocorrelation by improving the mixing.

When the likelihood and posterior were unimodal, SFT1, SFT2 and GP produced results that were consistent with each other. Given additional information in the form of $\mathbf{x}_0$, SFT1 was able to out-perform both of these methods, even with a prior that was inconsistent with the data. In the case of multi-modality, GP requires additional information to find additional modes, whereas SFT methods were shown to be successful in the FitzHugh-Nagumo bimodal example of section 4.1, where the target posterior is like a mixture of 2 identical Gaussians separated by 312 standard deviations.

Computationally, producing a numerical solution to the DE model can be slow Huang *et al.* (2006), Li *et al.* (2006) and producing a data smooth is not necessarily a computational improvement, but use of parallel processing reduces the total computational time of using a population MCMC method and ensures minimal additional computational time from adding more parallel chains. To further reduce the computational load in SFT, one could omit computing $\mathbf{x}(t)$ or $\mathbf{x}(t, \mathbf{x}_0)$ at each proposed value and instead update it's estimate only occasionally during mutation steps but constantly during the proposed exchange moves. However the success of this variation depends strongly on the quality of the smooth approximation to the DE model and the sensitivity of the dynamics to the parameters. This short cut will surely alter the posterior and the role of the shortcut on convergence is not clear making guidelines for when it might be useful, the subject of further investigation.

There may be some interest in a mixed dimension approach that uses SFT2 along with an additional parallel chain using the model (3), where for some $\ell < M$, an exchange move proposes to swap $(\mathbf{x}(t=0))_\ell$ with $(\mathbf{x}_0)_M$. However, the dimension jumping between chains with and without $\mathbf{x}_0$ is not guaranteed to produce the desired target distribution for (3). In the nylon example, the induced density $P_{SFT2}(\mathbf{x}(t = 0) \mid \mathbf{y}, \lambda_M)$ does not have the same distribution for W_0 as $P_{SFT1}(\mathbf{x}_0 \mid \mathbf{y}, \lambda_M)$ because the former is essentially a profile posterior mode, whereas the latter is intended to explore the distribution of X_0.

The variability in $\mathbf{x}_0$ and $\boldsymbol{\theta}$ translates into variability within the function space spanned by $S(\boldsymbol{\theta}, \mathbf{x}_0, t)$ when working with the ODE solution, whereas in finite λ SFT, the smooth permits deviation from $S(\boldsymbol{\theta}, \mathbf{x}_0, t)$. For example, in figure 1 the data exhibit rapid changes in the trajectory of component V near times $5.5, 10, 15$ and 18. By permitting small deviations from the DE model leading up to these times of rapid change the smoothing based methods have the advantage of allowing some flexibility in the timing of these steep changes in trajectory. SFT has the additional advantage that it's parallel chains can be used to provide qualitative diagnostics. The evolution of the induced posteriors of $\mathbf{x}_1(t), \ldots, \mathbf{x}_M(t)$ will show deviations from the model dynamics towards data. Large deviations between the data features and the model features provides a qualitative goodness of fit diagnostic.

SUPPLEMENTAL MATERIALS

Matlab code requires the functional data analysis package (fdaM) from `http:\\functionaldata.org` , the GP software from `http:\\www.bscb.cornell.edu\~hooker\profile_webpages\profiling.html` and the statistics and optimization toolboxes from Matlab.

SFT.zip: Matlab code and data for simulations and analysis of the nylon application. (GNU zipped tar file)

File_Info.rtf This file is contained within the zip file. It describes the included Matlab scripts for SFT nylon analysis and for all methods applied to the simulation of section 4.2.

References

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

Calderhead, B., Girolami, M., and Lawrence, N. (2009). Accelerating bayesian inference over nonlinear differential equations with gaussian processes. In *NIPS 2008*. MIT Press.

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

Eilers, P. and Marx, B. (1996). Flexible smoothing with b-splines and penalties (with discussion). *Statistical Science*, **11**, 89–102.

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

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

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

Geweke, J. (1989). Bayesian inference in econometric models using monte carlo integration. *Econometrica*, **57**(6), 1317–1339.

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

Geyer, C. and Thompson, E. (1995). Annealing markov chain monte carlo with applications to ancestral inference. *Journal of the American Statistical Association*, **90**, 909–920.

Gonzales, O., Kper, C., Jung, K., Naval Jr, P., and Mendoza, E. (2007). Parameter estimation using simulated annealing for s-system models of biochemical networks. *Bioinformatics*, **23**(4), 480–486.

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

Huang, Y., Liu, D., and Wu, H. (2006). Hierarchical bayesian methods for es- timation of parameters in a longitudinal hiv dynamic system. *Biometrics*, page 413423.

Jasra, A., Stephens, D. A., and Holmes, C. C. (2007). On population-based simulation for static inference. *Statistics and Computing*, **17**, 263–279.

Li, L., Brown, M. B., Lee, K., and Gupta, S. (2006). Estimation and inference for a spline-enhanced population pharmacokinetic model. *Biometrics*, page 601611.

Liu, J. S. (2001). *Monte Carlo Strategies in Scientific Computing*. Springer, New York.

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

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

Nagumo, J., Arimoto, S., and Yoshizawa, S. (1962). An active pulse transmission line simulating a nerve axon. *Proceedings of the Institute of Radio Engineers*, **50**, 2061–2070.

Neal, R. (1996). Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*, **4**, 353–366.

Raftery, A. E. and Lewis, S. (1992). How many iterations in the gibbs sampler? In J. M. Bernardo, J. Berger, A. Dawid, and A. Smith, editors, *Bayesian Statistics 4*, pages 763–773. Oxford University Press, Oxford, U.K.

Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. (2007). Parameter estimation for differential equations: A generalized smoothing approach (with discussion). *Journal of the Royal Statistical Society, Series B*, **69**(part 5), 1–30.

Walley, P. and Moral, S. (1999). Upper probabilities based only on the likelihood function. *Journal of the Royal Statistical Society, Series B*, **64**(4), 831–847.

Wu, H., Zhu, H., Miao, H., and Perelson, A. (2008). Identifiability and statistical estimation of dynamic parameters in hiv/aids dynamic models. *Bulletin of Mathematical Biology*, **70**, 785–799.

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