Stat Comput
DOI 10.1007/s11222-011-9234-3

Smooth functional tempering for nonlinear differential equation models

David Campbell · Russell J. Steele

Received: 27 April 2010 / Accepted: 20 January 2011

Abstract Differential equations are used in modeling diverse system behaviors in a wide variety of sciences. Methods for estimating the differential equation parameters traditionally depend on the inclusion of initial system states and numerically solving the equations. This paper presents Smooth Functional Tempering, a new population Markov Chain Monte Carlo 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. The tempered approximations depend on relaxations of the solution to the differential equation model, reducing the need for estimating the initial system states and obtaining 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 data features. Using our proposed approach, we observed faster convergence and robustness to both initial values and 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.

Keywords Dynamic systems · Parallel tempering · Model based smoothing · Functional data analysis · Population Markov chain Monte Carlo · Multi-grid Markov chain Monte Carlo

D. Campbell
Department of Statistics and Actuarial Science, Simon Fraser University, 13450 102nd avenue, Surrey, BC, Canada V3T 0A3
e-mail: dac5@stat.sfu.ca

R.J. Steele (✉)
Department of Mathematics and Statistics, McGill University, 805 Sherbrooke Ouest, Montreal, QC, Canada H3A 2K6
e-mail: steele@math.mcgill.ca

Published online: 30 March 2011

1 Introduction

Differential equations (DEs) are used to model complex phenomena in pharmacokinetics, neuro-physiology, chemical engineering, systems biology, and other sciences. They are typically constructed from well understood scientific principles such as conservation of mass, energy and momentum, providing an easily interpretable parameter vector $\boldsymbol{\theta}$ that is often unavailable with other classes of models. Through changes in initial states ($\mathbf{x}(0)$) and parameters ($\boldsymbol{\theta}$), a DE model can describe a wide variety of complex behaviors including oscillations, steady states, and exponential growth (or decay) with a small number of parameters. However, the flexibility of a DE to succinctly model these behaviors comes at a heavy price.

DEs describe the rate of change of a vector of functional system states $\mathbf{x}(t)$ with respect to an argument, such as time $t \in [0, T]$,

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

In the presence of additive measurement noise, one observes

$$\mathbf{y}(t) = \mathbf{x}(t) + \epsilon(t), \tag{2}$$

where only a subset of states may be observed and $\epsilon(t)$ is an error process resulting from some distribution. 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, $\mathbf{x}(t) = S(\boldsymbol{\theta}, \mathbf{x}_0, t)$. Due to the properties of $S(\boldsymbol{\theta}, \mathbf{x}_0, t)$, the

likelihood for $\mathbf{y}(t)$ in (2) may be rife with undesirable topography such as local maxima, ridges, ripples and/or large flat segments (Esposito and Floudas 2000). We will primarily focus on ordinary differential equations (ODEs) in our work.

There is a rich literature in the biological sciences proposing solutions to the parameter estimation problem for models like (1). Varah (1982) and Voit and Sauvegeau (1982) first proposed the use of smoothing for the estimation of the parameters of an ODE. Ramsay and Silverman (2005) and Poyton et al. (2006) extended Varah's approach to iterate between smoothing the data and estimating the parameters of the underlying ODE. A recent approach to parameter estimation based on generalized profiling (GP) also aims to improve the likelihood topology by using a data smooth $\hat{\mathbf{x}}(t) \approx S(\boldsymbol{\theta}, \mathbf{x}_0, t)$ resulting 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). Data smoothing using GP accounts 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 a profile likelihood approach can perform poorly in the presence of multiple nuisance parameters (Walley and Moral 1999). Additionally, a purely frequentist approach does not allow for valuable prior information about the system to be incorporated into the modeling.

We present a new Bayesian sampling method for posterior estimation of $\boldsymbol{\theta}$ and $\mathbf{x}_0$ (if desired) from ODE models. The proposed smooth functional tempering (SFT) is a population Markov Chain Monte Carlo (MCMC) method that uses the GP model as a bridging auxiliary density in a parallel tempering algorithm (PT). Our approach employs a GP 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 or a bounded posterior space. Furthermore, unlike previous implementations of PT, SFT is robust to situations where prior information is inconsistent with the data. We propose two variations on our technique, one that incorporates the initial conditions in the estimation (SFT1) and a more computationally efficient alternative approach which profiles over the initial conditions, reducing the dimensionality of the parameter space (SFT2). Section 2 reviews background methods and leads into the description of both proposed variants of SFT in Sect. 3. A simulation study is given in Sect. 4 which examines the performance of our approaches in a canonical example from the statistical ODE literature. We conclude with a real data case study in Sect. 5 and a discussion of our results (Sect. 6).

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. Gradient based methods like non-linear least squares (NLS) do not typically perform well and practitioners are warned to expect a method based error level of the order of 25% (Bates and Watts 1988; Marlin 2000). New evolutionary approaches to maximization of the likelihood function are able to overcome many of these shortcomings, although inference depends on asymptotic approximations for standard errors or computationally expensive bootstrap procedures (Rodriguez-Fernandez et al. 2006; Miao et al. 2009; Liang and Wu 2008; Liang et al. 2010). Bayesian models provide an alternative to asymptotic frequentist analysis of ODE data. Typical Bayesian parameter estimation methods for ODEs (for two examples, see Gelman et al. 1996 or Huang and Wu 2006) use a model of the form:

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

A Bayesian approach for ODEs requires Monte Carlo simulation or numerical integration, and most implementations of Bayesian ODE models have used MCMC methods. For example, Barenco et al. (2006) and Rogers et al. (2007) both used traditional Metropolis sampling methods to obtain Bayesian posterior estimates for ODE parameters used for predicting gene transcription activity. Klinke (2009) implemented an adaptive MCMC approach for Bayesian estimation of a complex signaling network. The GNU MCSim software (Bois 2009) allows for Bayesian estimation of ODE models coded in Systems Biology Markup Language (SBML). However, current applications of MCMC require that the ODE is numerically solved at each proposed parameter value which makes exploration of the posterior surface under these topological difficulties challenging. Examples of these kinds of problems are shown in Sect. 4.2. Simulated annealing has been used to circumvent the topological difficulties (Gonzalez et al. 2007), but still requires a numerical solution to the ODE at each iteration. The dependence on $S(\boldsymbol{\theta}, \mathbf{x}_0, t)$ also increases the dimensionality of the parameter space with the inclusion of $\mathbf{x}_0$, a set of nuisance parameters that grows in dimension with additional experimental runs. The structural parameters, $\boldsymbol{\theta}$, are of primary interest because they define the ODE dynamics, yet current methods treat $\mathbf{x}_0, \boldsymbol{\theta}$ and σ^2 in (3) equally, despite their differing influence on the data-generating process and importance in estimation.

The primary challenges for Bayesian ODE estimation methods are that the topology of the posterior and location of the dominant mode are difficult to determine, the likelihood (and, thus, the un-normalized posterior distribution) generally does not have a closed form expression, and the param-

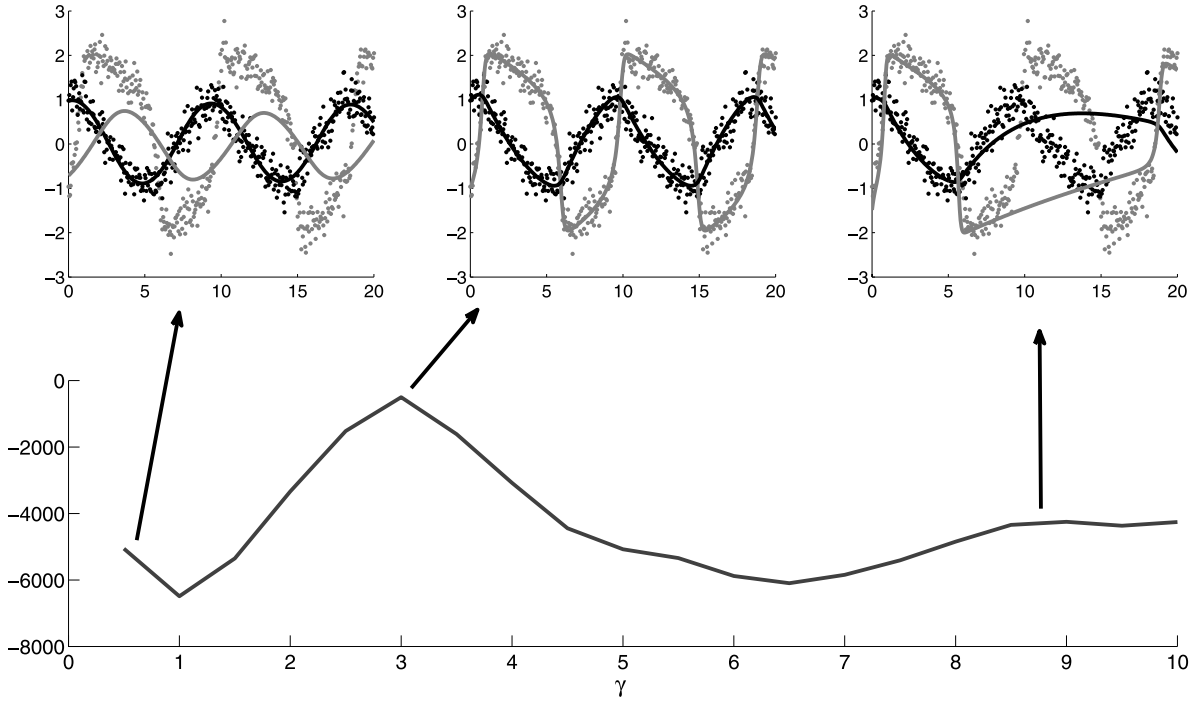

Fig. 1 A cross section of the FitzHugh-Nagumo log likelihood for γ (*bottom*) and the fits to the data for V (*grey*) and R (*black*) corresponding to the likelihood modes using the true parameter values (*top middle*), a small value (*top left*) and a large value (*top right*)

eter 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. Therefore, a Bayesian approach to ODE models requires a method that can adeptly manage these challenging features.

The nature of the posterior topology for difficult models is often referred to in the biological literature as "sloppiness". In statistics, the term normally used would be "non-identifiable" or, if not strictly not identifiable, then only *weakly* identifiable with respect to estimating parameters from the observed data. The wide extent of the sloppiness in biological systems problems is discussed in a paper by Gutenkunst et al. (2007b), where they identify sloppiness issues with parameters in a large number of models extracted from the BioModels database (Le Novère et al. 2006). There is a large amount of work that has been done to develop methods for identifying problems that are sloppy in nature and to choose parameterizations that remedy the problem (Gutenkunst et al. 2007a; Vilela et al. 2007; Raue et al. 2009). The body of work on identifying and remedying sloppy parameterizations is interesting and could be potentially used in conjunction with our approach to improve inference, but it is beyond the scope of our work here.

2.1 Population MCMC

Population based simulation methods are designed to improve mobility of MCMC samplers using information from parallel MCMC chains based on a sequence of approximations to the posterior density. Parallel tempering (PT), for example, approximates the posterior distribution 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 $0 \leq \xi_1 < \cdots < \xi_M = 1$ (Geyer 1991). The most commonly used series of approximations is

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

We have that for $\xi_1 = 0$, $P_1(\boldsymbol{\psi} \mid \mathbf{y}) = P(\boldsymbol{\psi})$ and for $\xi_M = 1$, $P_M(\boldsymbol{\psi} \mid \mathbf{y}) = P(\boldsymbol{\psi} \mid \mathbf{y})$, the posterior distribution of interest. The M posterior approximations are used as the target densities of M parallel MCMC chains. The posterior approximations from chains corresponding to smaller values of ξ_m are rooted more heavily in the prior, affording $\boldsymbol{\psi}_m$ greater mobility around the posterior parameter space compared to larger ξ_m-valued chains. Consequently, the smaller ξ_m-valued chains explore a larger area of the parameter surface while the larger ξ_m-valued chains remain trapped in the basin of attraction of a local posterior mode. Figure 2 shows the impact of changes in ξ on the posterior surface of γ in the FitzHugh-Nagumo model to be discussed in Sect. 4.

At the i^{th} iteration, each chain independently performs a Metropolis-Hastings (MH) step to update $\boldsymbol{\psi}^{(i)}$. However,

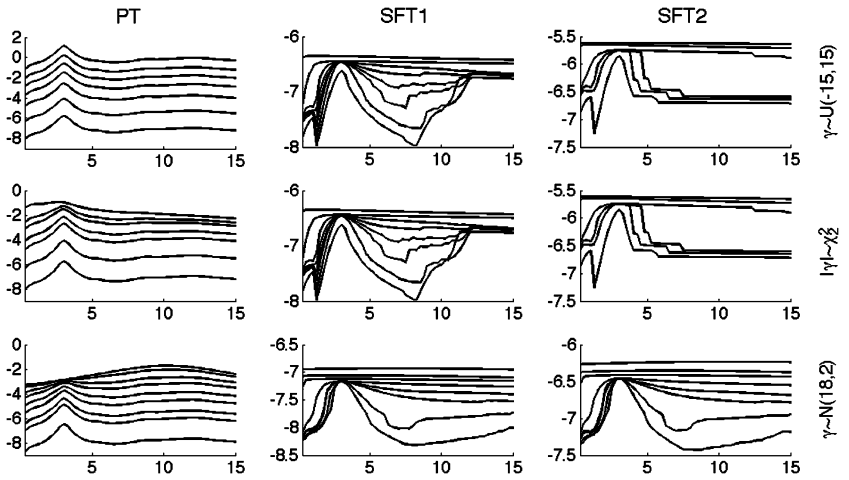

Fig. 2 The effect of changing the temperature gradient parameter on the -log(-log(non-normalized posterior)) for methods (*columns*) using different priors for γ (*rows*). Increasing values of the parameter (ξ for PT, λ for SFT1 and SFT2) gives lines appearing lower down within each plot

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 be exchanged between the chains rather than mutate independently. The exchange is accepted with probability

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

Over time, 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 for the chain sampling from the posterior of interest, $P(\boldsymbol{\psi} \mid \mathbf{y})$.

Parallel tempering and genetic algorithms (GA) share many conceptual similarities. The critical difference between the two approaches is that parallel tempering is used for generating samples from a distribution rather than for optimization of an objective function (Liang and Wong 2000). Parallel tempering allows for two kinds of updates to the model parameters, mutation steps (the MH updates performed within each chain) and exchange (the MH updates performed between chains). In their work, Liang and Wong (2000, 2001) also suggested a potential crossover (known as recombination in the GA literature in optimization research) move that would allow for only portions of the parameter vector to be exchanged between chains. However, this crossover move can be difficult to implement in practice and so we have not used it as part of our simulations. For a comprehensive and somewhat current review of the literature on these methods, we encourage the readers to see Jasra et al. (2007).

PT and variants (Marinari and Parisi 1992; Neal 1996; Calderhead et al. 2009) have been shown to work well for sampling from certain multi-modal densities. However, despite enabling the sampler to escape local posterior modes, 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 almost always done towards the prior, PT will fail when prior information does not agree with the features of the observed data (see Sect. 4.2).

2.2 Model-based smoothing

Model-based smoothing is a generalization of smoothing splines or penalized smoothing (Eilers and Marx 1996). The mean of the data is assumed to be a linear combination of basis functions ($\boldsymbol{\phi}(t)$) with coefficients ($\mathbf{c}$), i.e. $E[y(t)] = \mathbf{x}(t) = \mathbf{c}'\boldsymbol{\phi}(t)$. The shape of the smooth depends on the hyper-parameter λ and can be expressed as a distribution on $\mathbf{x}(t)$,

$$P(\mathbf{x}(t) \mid \boldsymbol{\theta}, \lambda) \propto \exp\left[-\frac{\lambda}{2}\mathtt{PEN}(x, \boldsymbol{\theta}, t)\right]$$

where

$$\mathtt{PEN}(x, \boldsymbol{\theta}, t) = \int_t \left[\frac{dx(s)}{ds} - f\left(x(s), \boldsymbol{\theta}, s\right)\right]^2 ds. \tag{5}$$

A Bayesian extension of this model would then assume a further prior distribution for $(\theta, \lambda, \sigma^2)$

A standard choice in the smoothing literature for the penalty term is $\mathtt{PEN} = \int_t (d^2x(s)/ds^2 - 0)^2\, ds$. This defines a model structure that anticipates a linear model, whereas in (5), the penalty is more generally based on the integrated square of the residual of (1). When used as a kernel for a prior on $\mathbf{x}(t)$, we see that the prior density increases as $\mathbf{x}(t)$ approaches the shape defined by the ODE model through PEN. Model parameters $\boldsymbol{\theta}$ from (1) can be considered as hyper-parameters of the prior on $\mathbf{x}(t)$.

The smoothing parameter λ defines a balance between measurement error σ^2 and deviation from the ODE model. As $\lambda \to 0$, the posterior mode of $\mathbf{x}(t) \mid \mathbf{y}, \boldsymbol{\theta}, \sigma^2, \lambda$ is the function space spanned by the basis that interpolates the data. As $\lambda \to \infty$, the posterior mode of $\mathbf{x}(t) \mid \mathbf{y}, \boldsymbol{\theta}, \sigma^2, \lambda$ occurs on the function space spanned by the ODE 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 λ controls the flow of information between $\mathbf{y}$ and $\boldsymbol{\theta}$ because $\boldsymbol{\theta}$ is conditionally independent of $\mathbf{y}$ given $\mathbf{x}(t), \lambda$, and σ^2. Consequently, model based smoothing reduces the impact of changes in $\boldsymbol{\theta}$ on $\mathbf{x}(t)$, inflating $var(\boldsymbol{\theta} \mid \mathbf{y})$ 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. These initial value problems generally 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. With respect to 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)

Our novel approach to Bayesian estimation of ODE models, Smooth Functional Tempering (SFT), is a particular form of parallel tempering, as it is defined by a sequence of M distributions towards the posterior of the measurement error model in (3). However, SFT is best seen as a collocation tempering method that uses the data-smooth as an auxiliary distribution. SFT depends on a basis expansion for the approximation $\mathbf{x}(t) = \mathbf{c}'\boldsymbol{\phi}(t) \approx \mathbf{S}(\boldsymbol{\theta}, \mathbf{x}_0, t)$ and tempers towards the posterior by varying the smoothing parameter. When using a B-spline basis, as the smoothing parameter increases and the ODE model is more rigorously enforced, $\mathbf{x}(t) \to \mathbf{S}(\boldsymbol{\theta}, \mathbf{x}_0, t)$, where $\mathbf{x}_0 = \mathbf{c}'\boldsymbol{\phi}(t = 0)$ and $\mathbf{S}(\boldsymbol{\theta}, \mathbf{x}_0, t)$ is numerically computed using an implicit Runge-Kutta method with stepping points at the knot locations (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 ODE solution. In this section, we outline two variations of this process. The first variation (SFT1) employs a smooth approximation to the initial value problem and utilizes a fixed point in the data smoothing step in conjunction with a numerical ODE solution. The second variation (SFT2) uses smooth approximations and does not depend on numerical ODE solutions or $\mathbf{x}_0$.

3.1 SFT1: parameter estimation with a smooth and a numerical ODE solution

We first assume that we are interested in making inference about $\mathbf{x}_0$ and/or the function space spanned by the possible ODE solutions as well as $\boldsymbol{\theta}$. SFT1 defines a tempering strategy towards model (3) based on the increasing sequence of fixed smoothing parameters $0 < \lambda_1 \leq \cdots \leq \lambda_M = \infty$:

$$P_m(\mathbf{y} \mid \mathbf{x}_m(t, \mathbf{x}_0), \sigma^2) \sim N\left(\mathbf{x}_m(t, \mathbf{x}_0), \sigma^2\right)$$

$$P_m(\boldsymbol{\theta}, \mathbf{x}_0, \sigma^2) \propto \exp\left(-\lambda_m \texttt{PEN}(\mathbf{x}, \mathbf{x}_0, \boldsymbol{\theta}, t)\right) P(\boldsymbol{\theta}, \mathbf{x}_0, \sigma^2)$$

where

$$\texttt{PEN}(\mathbf{x}, \mathbf{x}_0, \boldsymbol{\theta}, t) = \int_t \left[\frac{d}{ds}\mathbf{x}_m(s, \mathbf{x}_0) - f(\mathbf{x}_m(s, \mathbf{x}_0), \boldsymbol{\theta}, s)\right]^2 ds. \tag{6}$$

The parameters $(\lambda_1, \ldots, \lambda_M)$ in the SFT1 model assume the role of the temperature gradient parameters $(\xi_1, \ldots, \xi_M)$ in the PT approach.

Our proposed SFT1 approach differs from the model based smoothing approach described in Sect. 2.2 in that SFT1 removes one layer of the hierarchical model. SFT1 implicitly defines a distribution on the smooth and uses fixed values of λ_m. As with model based smoothing, as $\lambda_m \to 0$, the posterior mean for $\mathbf{y}$ under the SFT 1 model tends towards a data interpolant because the induced prior for $\mathbf{x}(t)$ is uniform over the function space spanned by the basis. Additionally, when $\lambda_m = 0$, $P_m(\boldsymbol{\theta} \mid \mathbf{y}) = P(\boldsymbol{\theta})$ since $\boldsymbol{\theta}$ is not used to define the shape of $\mathbf{x}(t, \mathbf{x}_0)$. If one assumes that $\lambda_M = \infty$ then

$$\exp\left(-\lambda_m \texttt{PEN}(\mathbf{x}, \mathbf{x}_0, \boldsymbol{\theta}, t)\right) = \begin{cases} 1 & \text{if } \mathbf{x}(t, \mathbf{x}_0) = S(\boldsymbol{\theta}, \mathbf{x}_0, t) \\ 0 & \text{otherwise.} \end{cases}$$

Consequently, the M^{th} chain produces samples from the ODE measurement error model in (3) but the $M - 1$ chains with $\lambda_m < \infty$ use a relaxation of the ODE solution enabling $\mathbf{x}(t, \mathbf{x}_0)$ to deviate from the dynamics in (1) to better match the features of the data. Thus, while the usual PT algorithm tempers towards the prior, SFT1 tempers towards the data features. Additionally, the impact of changes in $\boldsymbol{\theta}$ and $\mathbf{x}_0$ on $\mathbf{x}(\mathbf{x}_0, t)$ when using SFT1 are reduced with decreasing λ_m relative to the standard PT approach.

3.2 SFT2: parameter estimation without a numerical ODE 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 due to the reliance on $\mathbf{x}_0$. SFT2 avoids the potential liability of numerically solving the ODE and eliminates the need to explicitly model $\mathbf{x}_0$ by tempering via the sequence of distributions for $0 < \lambda_1 \leq \cdots \leq \lambda_M \leq \infty$:

$$\begin{aligned} P_m(\mathbf{y} \mid \boldsymbol{\theta}, \sigma^2) &= N(\mathbf{x}_m(t), \sigma^2) \\ P_m(\boldsymbol{\theta}, \sigma^2 \mid \mathbf{y}) &\propto \exp\left[-\lambda_m \texttt{PEN}(\mathbf{x}, \boldsymbol{\theta}, t)\right] P(\boldsymbol{\theta}, \sigma^2). \end{aligned} \tag{7}$$

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 $S(\boldsymbol{\theta}, \mathbf{x}_0, t)$ by changing λ_M (Qi and Zhao 2010). In practice, we choose $\lambda_M < \infty$ both 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 ODE solution. However, even at $\lambda_M = \infty$ model (7) is not equivalent to (3), because SFT2 effectively profiles over $\mathbf{x}_0$.

3.3 Impact of the smoothing parameter λ

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 Sect. 4) based on SFT1, SFT2 and PT under 3 different priors. The effectiveness of the tempering in PT changes drastically with the prior. Under the Uniform or χ^2_2 priors, the minor mode at $\gamma = 13$ eventually disappears with any tempering approach. 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.

Figure 2 also shows that SFT2, which does not use a fixed value of $\mathbf{x}_0$, has the necessary 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 suggest as a rule of thumb that the smallest value of λ should be able to approximate well the data dynamics, or if in doubt, it should nearly interpolate the data regardless of 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.

It is important to note that the choice of λ values will have a large effect on the performance of the algorithm. Finding an optimal number and sequence of temperatures for tempering strategies is currently a problem of open research (Jasra et al. 2007; Gramacy et al. 2010; Atchadé and Liu 2010; Liu 2001). The usual spacings for the temperature gradient tend to be either geometric or logarithmic, and the number of temperatures is typically selected based on the computational demands of generating samples for a single temperature (i.e. the smaller the demands for a single chain, the larger number of temperatures that can be used). It should also be noted that such choices tend to be problem dependent, so our approach (as with all tempering approaches) may require some initial exploration of these choices in order to ensure good performance.

When using SFT2 the value of $\texttt{PEN}(\mathbf{x}, \boldsymbol{\theta}, t)$ 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 prior

B-splines permit considerable flexibility in shape, allowing high order smooth or discontinuous derivatives where needed which makes them a convenient choice for SFT. However, other bases such as Fourier, wavelet or (truncated) polynomial bases can also used for smoothing and producing solutions to ODEs and may provide additional advantages in some problems. The type and number of basis functions used must permit $\mathbf{x}(t)$ to accommodate the ODE model dynamics and deviations for a wide range of values of $\boldsymbol{\theta}$. Use of a model based smoothing strategy suggest that the optimal number of B-spline bases is of order $O(n)$ (Olhede 2008), however with small n there may be a need for far more basis functions than observations. We anticipate this would occur more often if the dynamics of the ODE model are complex. Note that neither SFT1 nor SFT2 explicitly sample $\mathbf{c}$, so employing a large number of basis functions does not complicate the convergence or tuning of the chains.

The integral terms in (6) and (7) to compute $\texttt{PEN}(\cdot)$ can be computed through numerical quadrature. Using B-spline basis with quadrature points at the unique knot locations produces a computationally fast result. Some relevant discussion about quadrature and calculating $\texttt{PEN}(\cdot)$ in model based smoothing can be found in the discussion of Ramsay et al. (2007).

As an additional practical note, we emphasize here that care must be taken in real applications in producing a prior on $\boldsymbol{\theta}$ in ODE systems. A prior should be placed on the

shape of $S(\boldsymbol{\theta}, \mathbf{x}_0, t)$ and transformed to the parameter space $[\boldsymbol{\theta}, \mathbf{x}_0]$. Non-informative or loosely informative (vague) priors on the parameter space may lead to prior distributions that are quite informative on the function space (see, for example, Salway and Wakefield 2008; Wakefield 1996; Wakefield and Bennett 1996; and Bates and Watts 1988). One could, for example, place a bounded, piece-wise constant set of uniform priors over regions of the function space where there is confidence the function values must lie. Then one can use a (potentially non-linear) transformation to transform this vague prior distribution on the function space to a prior over the parameters of the model (Salway and Wakefield 2008). In practice, working entirely with a prior distribution on the function space (for example, using a Gaussian process prior) is often easier to specify and manage computationally than transforming distributions on function spaces into priors on $\boldsymbol{\theta}$. However, the challenge then shifts to relating posterior inference on the function space to the underlying parameters of the ODE (Gao et al. 2008) A formal comparison between these two approaches is beyond the scope of this paper.

4 Simulated examples from the FitzHugh-Nagumo model

The FitzHugh-Nagumo differential equations (FitzHugh 1961; Nagumo et al. 1962) comprise a simple model for the voltage potential across the cell membrane of the axon of giant squid neurons. These equations are used in neurophysiology 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:

$$\begin{aligned} \frac{dV}{dt} &= \gamma\left(V - \frac{V^3}{3} + R\right), \\ \frac{dR}{dt} &= -\frac{1}{\gamma}\left(V - \alpha + \beta R\right). \end{aligned} \tag{8}$$

An example of a simulated data set and the true underlying process when $\gamma = 3$ appears in Fig. 1. Figure 1 also includes a cross section of the log likelihood and additional ODE solutions using parameter values corresponding to minor modes of the cross section. The mode corresponding to values of $\gamma \approx 0.5$ produces a ODE solution with the correct period but the shape is too sinusoidal to represent the dynamics of V. The likelihood mode corresponding to values of $\gamma \approx 9$ produces approximately the correct shape but does not match the period. Traversing the likelihood surface in either direction from the local modes causes a deterioration in the data fit before it can be improved. Any sampling or optimization algorithm would encounter wide regions of prohibitively deep posterior topology of approximately 4000 units deep on the log scale. We consider a particular one-parameter version of this model, where all parameters other than γ are held fixed, in order to highlight the ability to accurately estimate the posterior in Sect. 4.1. We have set the fixed parameters to values that yield a bimodal likelihood surface that is representative of the types of irregularities that are observed in problems involving ODEs. We compare Bayesian and frequentist methods using the one dimensional version of this model in Sect. 4.2. The full FitzHugh-Nagumo model, where all unknown parameters are estimated, is explored in Sect. 4.3. A simulation examining computational effort in a real data setting is saved for Sect. 5.1.

4.1 One dimensional bimodal example

In this section we alter (8) to produce a symmetric, bimodal posterior for γ;

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

Due to the computational intensity of working with differential equations, we restricted ourselves to ten simulated data sets, obtained from the numerical solution to (9) 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 ($\alpha = 0.2$, $\beta = 0.2$, $\sigma_V^2 = 0.25$, $\sigma_R^2 = 0.16$, $V_0 = -1$, and $R_0 = 1$) so that the posterior density can be evaluated numerically and compared with results from SFT1, SFT2 and PT under two different prior distributions:

$$\begin{aligned} P(\gamma) &= \frac{1}{2}\chi_2^2, \quad \gamma > 0 \\ P(-\gamma) &= \frac{1}{2}\chi_2^2, \quad \gamma < 0 \end{aligned} \tag{10}$$

and

$$P(\gamma) = \mathit{Uniform}(-15, 15). \tag{11}$$

Neither of the two prior distributions had much influence on the posterior, whose most important features could be reasonably approximated by 2 identical Gaussians whose means are separated by 312 standard deviations. SFT1 and SFT2 algorithms were constructed with 101 evenly spaced knots from a 5^{th} order B-spline basis. From the largest λ_M for each method, parallel chains were added in sequence by tuning the next value of λ_m so that the between-swap acceptance rate approached 50%.

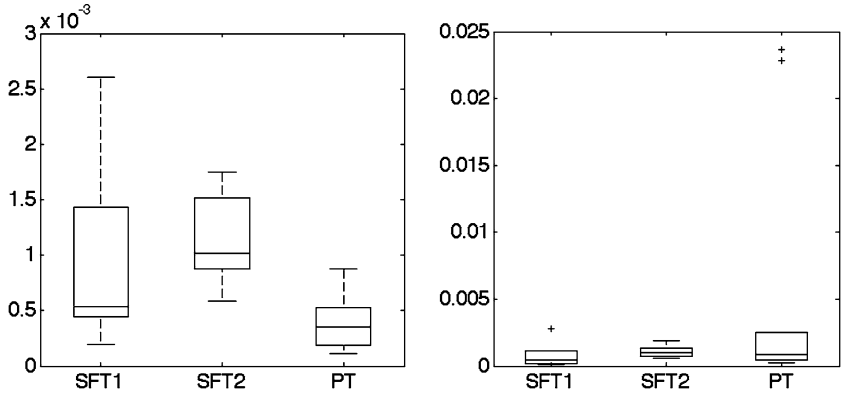

Fig. 3 Discrepancy between sampled and numerical posterior estimates using different prior distributions and sampling methods. Boxplots show $D(\hat{P})$ using the uniform prior (*left*) and the χ^2 based prior (*right*)

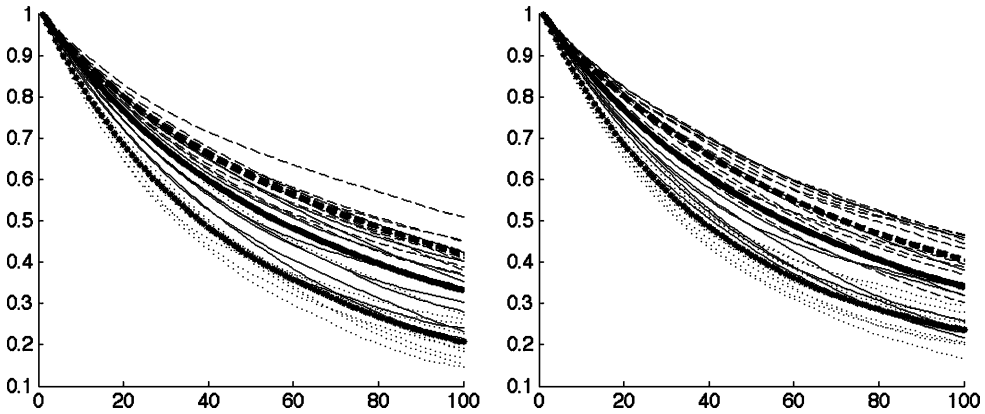

Fig. 4 The autocorrelation functions for the SFT1 (*solid line*), SFT2 (*dotted line*) and PT (*dashed line*) for the bimodal problem of Sect. 4.1, with the uniform prior (*left*) and the χ^2 based prior (*right*). The heavy lines are the point-wise mean autocorrelation functions

For each of the two prior distributions for γ, the numerically evaluated posterior (P_{num}) was compared with the results of the sampling based methods using the Integrated Squared Error (ISE):

$$D(\hat{P}_{sampled}) = \int \left[P_{num}(\gamma \mid \mathbf{y}) - \hat{P}_{sampled}(\gamma \mid \mathbf{y}) \right]^2 d\gamma. \tag{12}$$

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

For comparison, note that the M^{th} chains of SFT1 and PT use the same target distribution. In our simulations, PT performed somewhat better than SFT1 when using a uniform prior on γ but somewhat worse with a χ^2 based prior, although both performed well based on the ISE. SFT1 and PT both use the true value of $\mathbf{x}_0$, but SFT2 estimates γ without this additional knowledge. Consequently, SFT2 uses less information than the other two approaches, leading to a posterior variance around the modes (at $\gamma = \pm 3$), which is approximately 7 times wider than that using SFT1 or PT. In order to make for a fair comparison, $D(\hat{P}_{SFT2})$ was computed comparing the sampled density with the numerical estimate of its 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. SFT2 performs the best with respect to this criterion, in part due to the lack of dependence on the initial conditions. SFT1 generally ranks second, likely due to the reduced impact of initial conditions in the finite λ_m parallel chains. The ACFs for PT are slowest to decay. We do not observe an impact of the choice of prior distribution on the ordering of the ACFs in this example.

4.2 Inconsistent prior information

In this section, we focus on the one dimensional problem of estimating $P(\gamma \mid \mathbf{y})$ using the FitzHugh-Nagumo model (8) with a prior that is inconsistent with the observed data: $\gamma \sim N(14, 2^2)$. The bottom row of Fig. 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 Sect. 4. SFT1 and SFT2 were performed with 6 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 the Raftery-Lewis criterion (Raftery and Lewis 1992) was less than 125 in all cases from this starting point. After discarding 1,000 iterations, both the Raftery-Lewis and

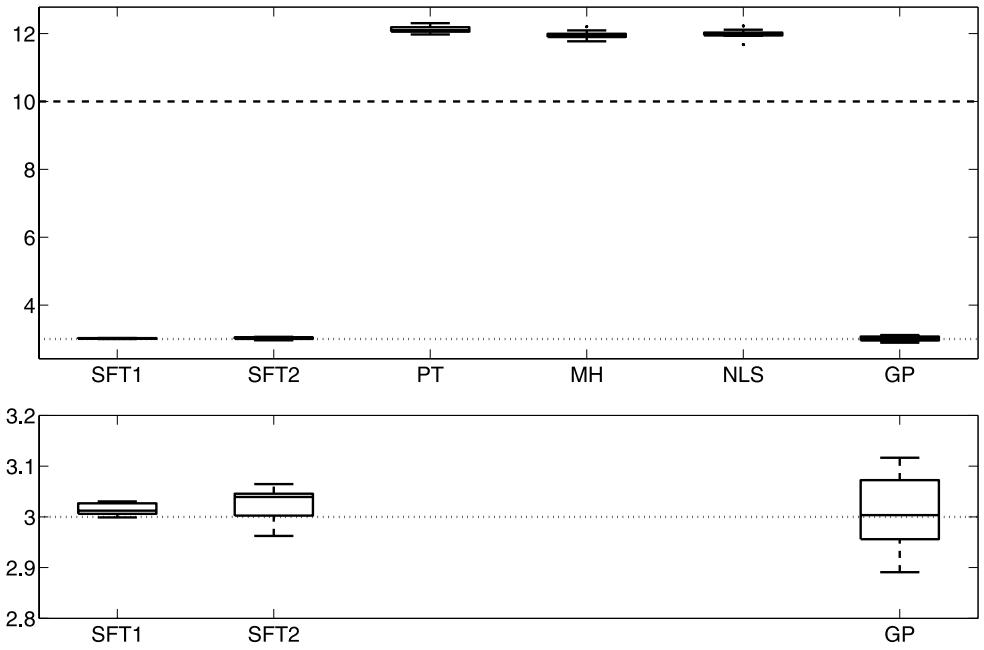

Fig. 5 Boxplots of the estimates of γ in Sect. 4.2, the *dashed line* is where the methods where initialized, the true parameter value is 3. *Top*: Estimates for all 6 methods, *bottom*, rescaled to show detail

Geweke convergence diagnostics (Geweke 1992) indicate convergence from all of the independent chains from all the sampling methods.

Figure 5 shows a boxplot of the final parameter estimates. MH and NLS are not able to 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-valued 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 negatively impacted because it did not use information about the initial conditions. In this section, we use model (8) with simulated data from the more realistic scenario where no parameters are known. Prior distributions for $\boldsymbol{\theta} = (\gamma, \alpha, \beta)$ were determined by numerically solving the ODE over a coarse grid of values of $\boldsymbol{\theta}$ and placing approximately 95% of the prior mass over the values that produce oscillatory dynamics giving:

$$\gamma \sim \chi^2_2, \qquad P(\alpha) = P(\beta) = N(0, 0.4^2) \tag{13}$$

Priors for $\mathbf{x}_0 = (V_0, R_0)$ were chosen based on the observed data, where the priors for both were chosen to be independent Gaussian densities centered on the first observed value with variance equal to the observed data variance, which places most of the prior mass in regions where data for V and R were actually observed. The priors for the variance parameters σ^2_V and σ^2_R were chosen to be Jeffreys, i.e. $P(\sigma^2_{V,R}) \propto 1/\sigma^2_{V,R}$. 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 a multivariate Normal distribution, making the Delta method interval estimates of Ramsay et al. (2007) good approximations.

For these simulations we focus on SFT1, SFT2 and GP because of the bad performance of PT for the inconsistent prior distribution in Sect. 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 Fig. 6 based on 30,000 posterior draws after burn-in. The observed magnitude of the bias is small and there are no significant differences in performance amongst the methods in this example.

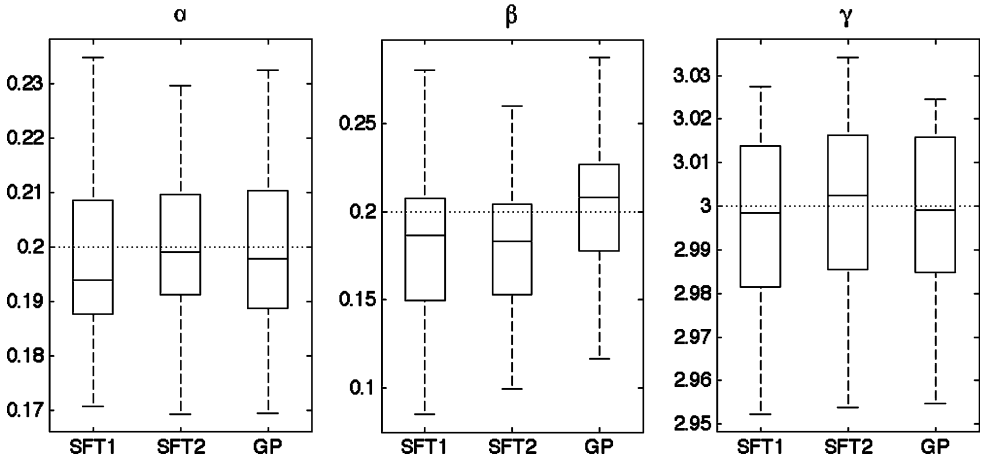

Fig. 6 Bias in point estimates for the FitzHugh-Nagumo parameters α (*top*), β (*middle*) and γ (*bottom*) of Sect. 4.3

5 Nylon example

In this section, we model the production of nylon in a heated reactor where its constituents, amine (A) and carboxyl (C), combine to produce the polymer, nylon (L), and water (W), which escapes as steam. At the same time, before escaping the system as steam, W decomposes L into A and C in the molten nylon mixture, giving the symbolic competing reactions $A + C \rightleftharpoons L + W$. In the experiment of Zheng et al. (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}, and then reduced until time τ_{i2}, at which point 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 of the model are described with the following system of differential equations:

$$-\frac{dL}{dt} = \frac{dA}{dt} = \frac{dC}{dt} = -k_p(CA - LW/K_a)10^{-3}, \tag{14}$$

$$\frac{dW}{dt} = k_p 10^{-3}(CA - LW/K_a) - 24.3(W - W_{eq}). \tag{15}$$

The reaction rate K_a is allowed to change with T_i and W_{eq} through relationships depending on the reference temperature $T_0 = 549.15$ giving four ODE parameters: $\boldsymbol{\theta} = [k_p, \gamma, K_{a0}, \Delta H]$ by the following expansion of K_a:

$$K_a = \left\{1 + W_{eq}\gamma 10^{-3}\right\} K_T[K_{a0}]\ell\left(\frac{\Delta H}{8.314}\right), \tag{16}$$

$$\ell(m) = \exp\left(-m10^3\left\{\frac{1}{T_i} - \frac{1}{T_0}\right\}\right), \tag{17}$$

$$K_T = 20.97\exp\left(-9.624 + \frac{3613}{T_i}\right). \tag{18}$$

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 exactly. Because 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; $[\boldsymbol{\theta}, \mathbf{x}_0, \sigma^2]$.

We set our prior distribution to be uniform 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. We chose this interval to be wider than is likely to be necessary to place non-zero mass over realistic functions for this problem. It is likely that values of the unobserved W would remain close to the values of W_{eq} (all take on value less than 100) but the more conservative value of 250 was used throughout for the states A, C and W. Additionally, the prior distributions on $1/\sigma_A^2$ and $1/\sigma_C^2$ were chosen to be independent Gamma densities with mean 9 and variance 27. These are pessimistic priors relative to the measurement error variance estimates from additional experiments by Zheng et al. (2005).

We implemented SFT1 and SFT2 using evenly spaced knots placed 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 $\tau + [0.1, 0.2, \ldots, 0.9]$ 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 was implemented with values $\lambda_1 = 100$, and $\lambda_2 = 10,000$. SFT1 required four times the number of parameters of the SFT2 model and consequently $M = 3$ chains were used, with tempering values $\lambda_1 = 200, \lambda_2 = 500$.

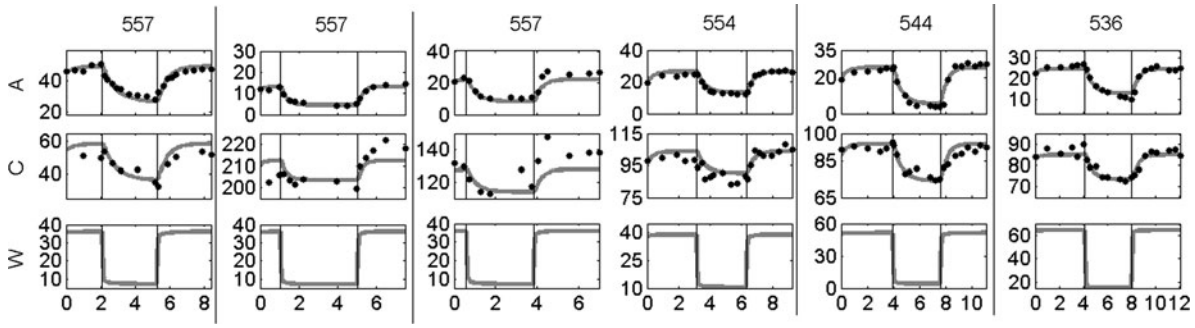

Fig. 7 The nylon observations along with the fit to the data. Temperatures of the experimental runs are given above component A in degrees Kelvin. *Vertical axes* are in concentration units and *horizontal axes* are in hours

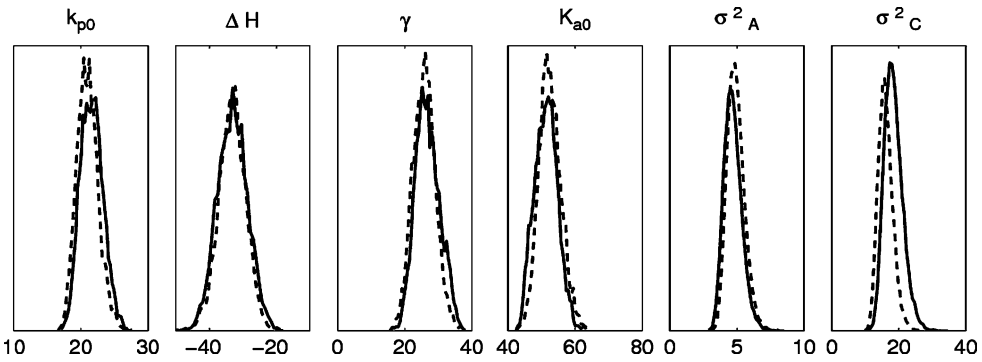

Fig. 8 A comparison of the posterior density estimates for the nylon parameters using SFT1 (*black line*) and SFT2 (*dashed line*)

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 Fig. 8. Estimates for the marginal posterior densities of $\boldsymbol{\theta}$ are similar between the methods, and the values for the integrated squared difference between marginal posteriors comparing SFT1 with SFT2 are 0.057, 0.016, 0.018 and 0.0073 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 equal to 0.11 and 0.15 respectively. The reason for this discrepancy may lie in the marginal posterior density estimates of $\mathbf{x}_0$, estimated by SFT1 and shown in Fig. 9. The dynamics of the system are quite fast, so 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 relatively flat (and uninformative) posterior distributions. 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 Fig. 8. SFT1 also allows for new insights into the vast uncertainty in W_0. The advantages of SFT2 are the reduced dimension of the problem and, in this case, a five fold computational time reduction.

5.1 Comparison of computational effort

The simulation studies of Sect. 4 were performed with high resolution, homoscedastic data from a fully observed system in order to highlight the algorithmic performance in the face of specific posterior topological difficulties. Data such as this are unlikely to be observed often in practice. In this section, we examine the variability in computation time of the SFT1, SFT2 and PT algorithms using more realistic replications of the same algorithmic process where the only difference between algorithmic trials is the random seed initializations. This is in contrast to the simulation studies of Sect. 4 where each trial used a different random data set. Variability in the computation time required for a single iteration of the sampler depends on the parameter values being proposed because they determine the stiffness (magnitude of the derivative) of the differential equation system. When numerically solving an ODE, a stiff system is considerably slower than a non-stiff system and the parameter values determine the stiffness (Deuflhard and Bornemann 2000). Consequently, we use repeated algorithmic simulations on a fixed real dataset to compare the algorithmic differences in methods and examine the practical advantages of the proposed methods.

The full nylon model in (14)–(18) requires multiple experimental runs to estimate the temperature dependency of k_p and K_a, however for this study we use a single experimental run of the nylon dataset, (shown in Fig. 7 with tem-

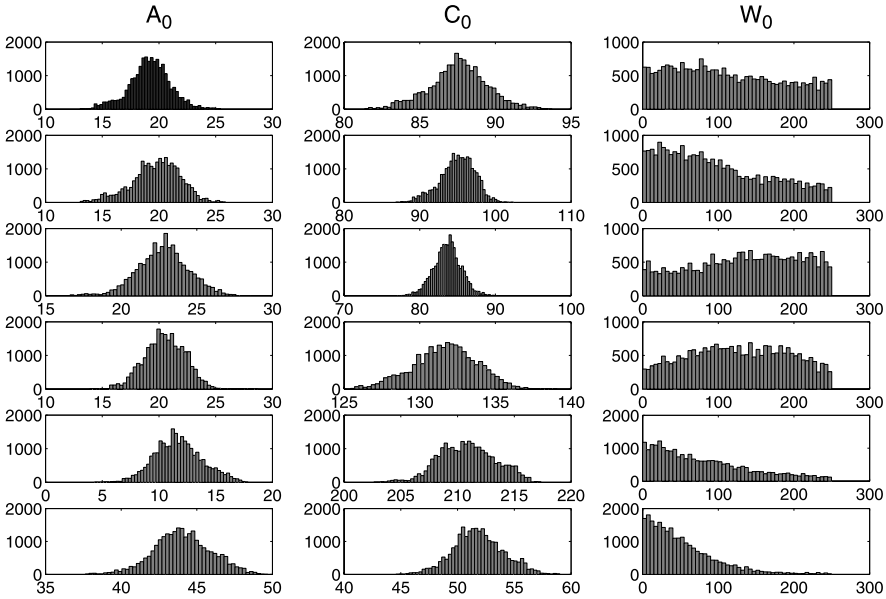

Fig. 9 Histograms of posterior draws $\mathbf{x}_0$ in the nylon system using SFT1. *Rows* are for the different experimental runs, while *columns* are (*left to right*) A_0, C_0 and W_0

perature of 544), performed at a fixed temperature. Consequently, we use the model (14) and (15) without further expansion of k_p and K_a. The priors from Sect. 5 were re-used to produce the statistical model.

The algorithmic study was run with this unevenly spaced partially observed data set 50 times for each of the three methods compared. Each of the 50 algorithmic simulations per method were run for 50,000 iterations from the same starting point and the computational time and effective sample size were measured after discarding the first 10,000 for burn in.

The within chain ODE parameters, $[k_p, K_a, A_0, C_0, W_0]$ for PT and SFT1 or $[k_p, K_a]$ for SFT2 were updated in a single step using Metropolis Hastings tuned with the optimal normal jumping kernel. Variance terms were updated using Gibbs steps. The 3 methods were attempted with 3 parallel chains each where λ and ξ values were chosen to obtain an acceptance rate of 50% between neighboring chains. Each method required the same number of evaluations of $\texttt{PEN}(\cdot)$ per iteration. Although distributed computing is natural when dealing with population MCMC, all runs were allotted only a single 3 GHz processor core with 1 Gb RAM.

The compute time, effective sample size and ratio thereof are shown in Fig. 10. While PT was about 30% faster on average, 13 of the 50 PT trials proposed at least one set of stiff parameters such that the stiff Runge-Kutta solver, `ode15s` in Matlab® (The MathWorks 2010), failed to solve the ODE within the permitted numerical tolerance bounds. Breaking the solver in this way was much faster than actually solving the system at these numerical limits.

The autocorrelation differences in Fig. 4 were primarily due to the SFT algorithms' ability to jump between distant modes, whereas here the target posterior is unimodal, reducing the advantage of SFT1 over PT. Figure 10 shows that SFT1 and PT effective sample sizes are less than half that of SFT2. Consequently, SFT2 results in a large advantage in terms of the compute time per independent draw, shown in the right of Fig. 10 and defined as the average over relevant parameters $A, C, k_p, K_a, A_0, C_0, W_0$ of:

$$\frac{\text{Total compute time for 40,000 iterations}}{\text{Effective Sample Size from those 40,000 iterations}}.$$

6 Discussion

Parameter estimation for nonlinear differential equations presents challenges for both frequentist and Bayesian modeling where, despite the appearance of convergence, the likelihood topology may not permit either convergence to or sampling around the global optima. Our proposed SFT approaches utilize model based smoothing to construct auxiliary densities for PT to match the features of the data with the dynamics of the model and improve estimation. 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. 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.

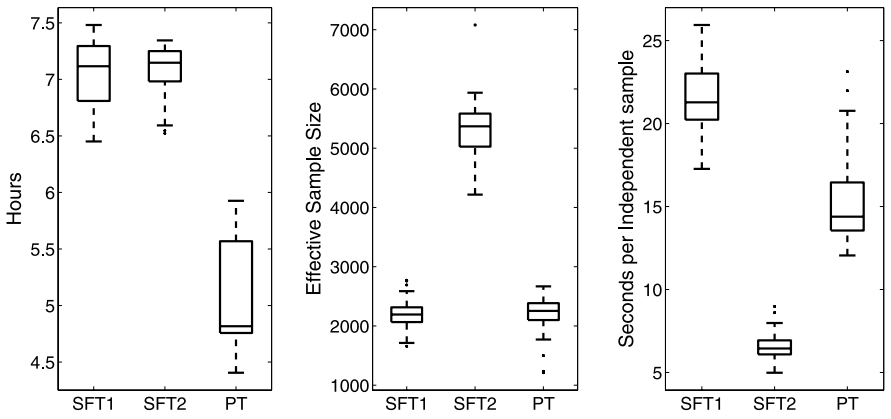

Fig. 10 Summary of computational effort for SFT1, SFT2, and PT. *Left to right*: total compute time, effective sample size, compute time per approximate independent posterior sample for the 2 parameter nylon simulation

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

When the likelihood and posterior were unimodal, SFT1, SFT2 and GP produced similar point estimates. 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 with only 10 parallel chains in an example where the posterior consisted of two widely separated modes.

It is rarely the case that a computational approach will universally provide improvement, as one can always construct examples where a particular method can fail. Our work has shown that the SFT approaches can significantly improve upon the usual implementation of the PT algorithm in certain situations without yielding significantly worse performance in situations where the standard PT approach performs well. The SFT2 approach, which avoids dependence on the initial system states ($\mathbf{x}_0$), performs very well and is significantly computationally less intensive per independent sample than the other approaches, especially when the ODE is computationally slow to solve.

Our objective in this paper was to propose a better computational approach to *Bayesian* analysis of nonlinear differential equation models. There exist many promising non-Bayesian solutions to this problem, particularly in the field of biology. Varah (1982) and Voit and Sauvegeau (1982) both proposed a data-smoothing approach to estimate the parameters of an ODE. More modern work in this area includes the use of artificial neural networks to estimate parameters in a non-parametric way (Voit and Almeida 2004). Chou and Voit (2009) provide a fairly comprehensive overview of not only different potential solutions for estimation of model parameters, but also for different approaches to specification of the model and general issues with these types of non-linear biological systems. Brunel (2008) explores the asymptotics of combining non-parametric smoothing with parameter estimation for data generated from non-linear ODE's and provides a new two-step approach that seems promising.

Similarly, although we have chosen a particular type of smoother, i.e. the generalized profiling method (Ramsay et al. 2007), as an auxiliary density, one would not be restricted to using only this kind of smoother. Other potential smoothing auxiliary models could be used, such as the perfect smoother of Eilers (2003) or the adaptation of Whitaker's smoother proposed by Vilela et al. (2007). We have used the generalized profiling method because of our familiarity with the smoother and the ease with which one can interpret the single smoothing parameter as a temperature in the tempering process. Other choices for the smoother could perform either better or worse (likely depending on the particular data problem) and we hope to explore this in future work. It would be interesting to see whether an ensemble of different choices of smoother would be computationally feasible as well.

Producing a data smooth to the ODE is not necessarily a computational improvement compared to producing a numerical solution to a ODE model. When the ODE is stiff, however, computing a numerical solution can already be extremely computationally intensive (Huang et al. 2006; Li et al. 2002) and using a relaxation of the numerical solution can accelerate iterations and convergence. The use of parallel processing reduces the total computational time of the population MCMC method and ensures minimal additional computational time from adding additional 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 only occasionally during mutation step, while updating always at the exchange steps. The success of this modification to our algorithm depends strongly on the quality of the smooth approximation to the ODE model and the

sensitivity of the dynamics to the parameters. This modification will certainly alter the effectiveness of the posterior sampling, although it is not clear exactly how. We thus leave the subject to future investigation.

There may be some interest in a mixed dimension approach that implements 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 X_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 $S(\boldsymbol{\theta}, \mathbf{x}_0, t)$, whereas in finite λ SFT, the smooth permits deviation from the function space of $S(\boldsymbol{\theta}, \mathbf{x}_0, t)$. For example, in Fig. 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 ODE 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 to account un-modeled deviations in period or other model inadequacies. SFT has the additional advantage that the 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, although more quantitative summaries could be derived in future work.

Finally, our work here focused only on generating samples from the posterior distribution of the model parameters. However, as one reviewer pointed out, the problem in many situations is not one of parameter estimation, but of model selection. Bayes factors (Kass and Raftery 1995) can be used to rank competing models on the basis of their posterior probability. The primary practical difficulty in using Bayes factors is integrating the unnormalized posterior over the space of parameters. The parallel tempering methods that have been used here to generate samples from the posterior can also be used to estimate the normalizing constants necessary for computation of posterior model probabilities (see, for example, Friel and Pettitt 2008). Calderhead and Girolami (2009), in particular, used population MCMC methods in conjunction with thermodynamic integration to compute normalizing constants for nonlinear ODE models. It would be quite interesting to observe whether our proposed SFT approach would also be useful for estimating normalizing constants in these problems, however this assessment is outside of the scope of our paper.

Acknowledgements The authors would like to thank the two reviewers and editor for their helpful comments on this manuscript. Both authors receive funding from Discovery Grants from the National Sciences and Engineering Research Council of Canada (NSERC) and David Campbell's work was also partially funded by Le Fonds québécois de la recherche sur la nature et les technologies (FQRNT).

References

Atchadé, Y., Liu, J.: The Wang-Landau algorithm in general state spaces: applications and convergence analysis. Stat. Sin. **20**, 209–233 (2010)

Barenco, M., Tomescu, D., Brewer, D., Callard, R., Stark, J., Hubank, M.: Ranked prediction of p53 targets using hidden variable dynamic modeling. Genome Biol. **7**(3), R25 (2006)

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

Bois, F.Y.: GNU MCSim: Bayesian statistical inference for SBML-coded systems biology models. Bioinformatics **25**(11), 1453–1454 (2009)

Brunel, N.J.B.: Parameter estimation of ODE's via nonparametric estimators. Electron. J. Stat. **2**, 1242–1267 (2008)

Calderhead, B., Girolami, M.: Estimating Bayes factors via thermodynamic integration and population MCMC. Comput. Stat. Data Anal. **53**(12), 4028–4045 (2009)

Calderhead, B., Girolami, M., Lawrence, N.D.: Accelerating Bayesian inference over nonlinear differential equations with Gaussian processes. In: Advances in Neural Information Processing Systems, pp. 217–224 (2009)

Chou, I.C., Voit, E.O.: Recent developments in parameter estimation and structure identification of biochemical and genomic systems. Math. Biosci. **219**, 57–83 (2009)

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

Eilers, P.: A perfect smoother. Anal. Chem. **75**, 3631–3636 (2003)

Eilers, P.H.C., Marx, B.D.: Flexible smoothing with B-splines and penalties (with discussion). Stat. Sci. **11**, 89–102 (1996)

Esposito, W.R., Floudas, C.: Deterministic global optimization in nonlinear optimal control problems. J. Glob. Optim. **17**, 97–126 (2000)

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

Friel, N., Pettitt, A.N.: Marginal likelihood estimation via power posteriors. J. R. Stat. Soc., Ser. B, Stat. Methodol. **70**(3), 589–607 (2008)

Gao, P., Honkela, A., Rattray, M., Lawrence, N.D.: Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. Bioinformatics **24**, 70–75 (2008)

Gelman, A., Bois, F.Y., Jiang, J.: Physiological pharmacokinetic analysis using population modeling and informative prior distributions. J. Am. Stat. Assoc. **91**, 1400–1412 (1996)

Geweke, J.: Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (eds.) Bayesian Statistics. Proceedings of the Fourth Valencia International Meeting, vol. 4, pp. 169–193. Clarendon Press, Oxford (1992)

Geyer, C.J.: Markov chain Monte Carlo maximum likelihood. In: Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface, pp. 156–163 (1991)

Geyer, C.J., Thompson, E.A.: Annealing Markov Chain Monte Carlo with applications to ancestral inference. J. Am. Stat. Assoc. **90**, 909–920 (1995)

Gonzalez, O., Küper, C., Jung, K., Naval Jr. P., Mendoza, E.: Parameter estimation using simulated annealing for S-system models of biochemical networks. Bioinformatics **23**, 480–486 (2007)

Gramacy, R., Samworth, R., King, R.: Importance tempering. Stat. Comput. **20**, 1–7 (2010)

Gutenkunst, R.N., Casey, F.P., Waterfall, J.J., Myers, C.R., Sethna, J.P.: Extracting falsifiable predictions from sloppy models. Ann. N.Y. Acad. Sci. **1115**, 203–211 (2007a)

Gutenkunst, R.N., Waterfall, J.J., Casey, F.P., Brown, K.S., Myers, C.R., Sethna, J.P.: Universally sloppy parameter sensitivities in systems biology models. PLoS Comput. Biol. **3**, 1871–1878 (2007b)

Huang, Y., Liu, D., Wu, H.: Hierarchical Bayesian methods for estimation of parameters in a longitudinal HIV dynamic system. Biometrics **62**, 413–423 (2006)

Huang, Y., Wu, H.: A bayesian approach for estimating antiviral efficacy in HIV dynamic models. J. Appl. Stat. **33**, 155–174 (2006)

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

Kass, R.E., Raftery, A.: Bayes factors. J. Am. Stat. Assoc. **90**(430), 773–795 (1995)

Klinke, D.J.: An empirical Bayesian approach for model-based inference of cellular signaling networks. BMC Bioinform. **10**, 371 (2009)

Le Novère, N., Bornstein, B., Broicher, A., Courtot, M., Donizelli, M., Dharuri, H., Li, L., Sauro, H., Schilstra, M., Shapiro, B., Snoep, J., Hucka, M.: BioModels database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic Acids Res. **34**(Suppl 1), D689–D691 (2006)

Li, L., Brown, M.B., Lee, K.H., Gupta, S.: Estimation and inference for a spline-enhanced population pharmacokinetic model. Biometrics **58**, 601–611 (2002)

Liang, F., Wong, W.: Evolutionary Monte Carlo sampling: applications to Cp model sampling and change-point problem. Stat. Sin. **10**, 317–342 (2000)

Liang, F., Wong, W.H.: Real-parameter evolutionary Monte Carlo with applications to Bayesian mixture models. J. Am. Stat. Assoc. **96**, 653–666 (2001)

Liang, H., Miao, H., Wu, H.: Estimation of constant and time-varying dynamic parameters of HIV infection in a nonlinear differential equation model. Ann. Appl. Stat. **4**, 460–483 (2010)

Liang, H., Wu, H.: Parameter estimation for differential equation models using a framework of measurement error in regression models. J. Am. Stat. Assoc. **103**, 1570–1583 (2008)

Liu, Jun S.: Monte Carlo strategies in Scientific Computing. Springer, New York (2001)

Marinari, E., Parisi, G.: Simulated tempering: a new Monte Carlo scheme. Europhys. Lett. **19**, 451–458 (1992)

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

The MathWorks: Matlab ®7 Mathematics. The Mathworks, Inc. Natick, MA (2010)

Miao, H., Dykes, C., Demeter, L.M., Wu, H., Avenue, E., York, N., York, N.: Differential equation modeling of HIV viral fitness experiments: model identification, model selection, and multimodel inference. Biometrics **65**, 292–300 (2009)

Nagumo, J.S., Arimoto, S., Yoshizawa, S.: An active pulse transmission line simulating a nerve axon. Proc. Inst. Radio Eng. **50**, 2061–2070 (1962)

Neal, R.M.: Sampling from multimodal distributions using tempered transitions. Stat. Comput. **4**, 353–366 (1996)

Olhede, S.: Discussion on the paper by Ramsay, Hooker, Campbell and Cao. J. R. Stat. Soc. B **69**, 772–779 (2008)

Poyton, A., Varziri, M., McAuley, K., McLellan, P., Ramsay, J.: Parameter estimation in continuous-time dynamic models using principal differential analysis. Comput. Chem. Eng. **30**, 698–708 (2006)

Qi, X., Zhao, H.: Asymptotic efficiency and finite-sample properties of the generalized profiling estimation of parameters in ordinary differential equations. Ann. Stat. **38**(1), 435–481 (2010)

Raftery, A., Lewis, S.: How many iterations in the Gibbs sampler. In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (eds.) Bayesian Statistics. Proceedings of the Fourth Valencia International Meeting, vol. 4, pp. 763–773. Clarendon Press, Oxford (1992)

Ramsay, J.O., Hooker, G., Campbell, D., Cao, J.: Parameter estimation for differential equations: a generalized smoothing approach (with discussion). J. R. Stat. Soc. B **69**, 741–796 (2007)

Ramsay, J.O., Silverman, B.W.: Functional Data Analysis. Springer, New York (2005)

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U., Timmer, J.: Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics **25**, 1923–1929 (2009)

Rodriguez-Fernandez, M., Mendes, P., Banga, J.R.: A hybrid approach for efficient and robust parameter estimation in biochemical pathways. Biosystems **83**, 248–65 (2006)

Rogers, S., Khanin, R., Girolami, M.: Bayesian model-based inference of transcription factor activity. BMC Bioinform. **8**, S2 (2007)

Salway, R., Wakefield, J.: Gamma generalized linear models for pharmacokinetic data. Biometrics **64**, 620–626 (2008)

Varah, J.: A spline least squares method for numerical parameter estimation in differential equations. SIAM J. Sci. Stat. Comput. **3**, 28–46 (1982)

Vilela, M., Borges, C.C.H., Vinga, S., Vasconcelos, A.T.R., Santos, H., Voit, E.O., Almeida, J.S.: Automated smoother for the numerical decoupling of dynamics models. BMC Bioinform. **8**, 305 (2007)

Voit, E.O., Almeida, J.: Decoupling dynamical systems for pathway identification from metabolic profiles. Bioinformatics **20**, 1670–1681 (2004)

Voit, E.O., Sauvegeau, M.: Power-law approach to modeling biological systems; III. Methods of analysis. J. Ferment. Technol. **60**, 233–241 (1982)

Walley, P., Moral, S.: Upper probabilities based only on the likelihood function. J. R. Stat. Soc. B **61**, 831–847 (1999)

Wakefield, J.: The Bayesian analysis of population pharmacokinetic models. J. Am. Stat. Assoc. **91**, 62–75 (1996)

Wakefield, J., Bennett, J.: The Bayesian modeling of covariates for population pharmacokinetic models. J. Am. Stat. Assoc. **91**, 917–927 (1996)

Wu, H., Zhu, H., Miao, H., Perelson, A.S.: Parameter identifiability and estimation of HIV/AIDS dynamic models. Bull. Math. Biol. **70**, 785–799 (2008)

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