AMSI
AUSTRALIAN MATHEMATICAL
SCIENCES INSTITUTE

National
Collaboration
in the Mathematical
Sciences

Bayesian inference in nonlinear differential equation models

Cody McRae
Supervised by Dr Jonathan Keith
Monash University

1 Introduction

Many biological processes can be accurately modelled as systems of ordinary differential equations (ODEs). The free parameters which comprise such models are, however, rarely accessible for direct measurement, and must be inferred by fitting the model to a set of experimental data. In this report, we consider a Bayesian approach to the problem of parameter estimation in nonlinear ODE models. To illustrate our approach, we perform parameter estimation using a synthetic dataset derived from the FitzHugh-Nagumo (FHN) model [2, 7]. The FHN model originated as a simplification to the Hodgkin-Huxley model [6], which describes the voltage potential across the cell membrane of the axon of giant squid neurons. In the FHN model, the voltage V across an axon membrane depends on a recovery variable R:

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

where $\alpha, \beta, \gamma \in \mathbb{R}^+$ are the free model parameters. The choice of the FHN equations was motivated by the highly nonlinear dynamics which they exhibit, as shown in Figure 1. Indeed, many of the difficulties associated with parameter estimation in biological models is attributable to their characteristic nonlinear behaviour. The source of these difficulties can be understood geometrically [10] by recognising that the set of all possible parameter values for the model induce a manifold within the space of observables quantities. Parameter estimation may then be reframed as a minimisation problem, where the aim is to find the point on the manifold closest to the experimental data. Because nonlinear models may have multiple local minima, any optimisation algorithm that is purely local is unlikely to converge on a global best fit.

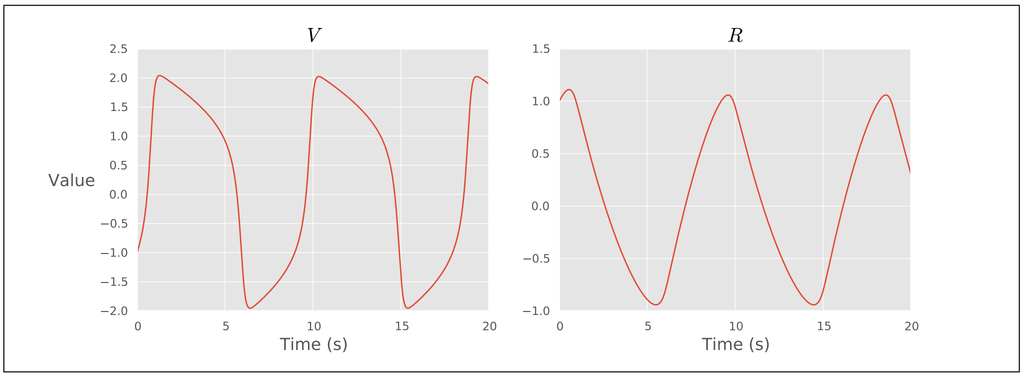

Figure 1: Solution to the FHN equations obtained using the parameter values $\alpha = 0.2$, $\beta = 0.2$, $\gamma = 3$ and initial conditions $V_0 = -1$, $R_0 = 1$.

2 Methods

2.1 Modelling

2.1.1 The dynamical system

Consider a dynamical system of N state variables in which the time-evolution of the state vector $\mathbf{x}(t)$ is described by a system of N differential equations:

$$\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t), \boldsymbol{\theta}) \tag{2}$$

which we assume, for illustrative purposes, have no closed form solution, but nevertheless can be simulated by numerical integration, given a set of parameters $\boldsymbol{\theta}$ and the initial conditions $\mathbf{x}_0$.

2.1.2 Observation model

We assume that noisy observations of the dynamical system in (2) are related to the 'true' state through a Gaussian error model, that is, measurements made at a given time t, conditional on the model parameters $\boldsymbol{\theta}$ and the initial conditions $\mathbf{x}_0$, are assumed to be iid normal random variables. Let $\mathbf{y}_n$ denote the time series of length T which contains T discrete time observations of the state variable x_n. To simplify our notation, we let $\mathbf{x}_n(\boldsymbol{\theta}, \mathbf{x}_0)$ be a vector of length T whose components are the corresponding numerical solution to the differential equations for each observation time. Then, given time series measurements for a set of state variables indexed by a set K, the joint observational likelihood may be written as:

$$p(\mathbf{y}|\mathbf{x}(\boldsymbol{\theta}, \mathbf{x}_0), \boldsymbol{\sigma}^2) = \prod_{k \in K} \mathcal{N}(\mathbf{x}_k(\boldsymbol{\theta}, \mathbf{x}_0), \sigma_k^2 \mathrm{I}_{\mathrm{T}_k}) \tag{3}$$

where $\boldsymbol{\sigma}^2 = (\sigma_1^2, \ldots, \sigma_k^2)$ are the observational variances, $\mathrm{I}_{\mathrm{T}_k}$ is the $\mathrm{T}_k \times \mathrm{T}_k$ identity matrix and T_k is the number of discrete time observations that are available for the state variable x_k

2.1.3 Inference

For the general case in which our knowledge regarding $\mathbf{x}_0$, $\boldsymbol{\theta}$, and $\boldsymbol{\sigma}^2$ is uncertain, we encode our current state of knowledge through an appropriate choice of priors. The model parameters $\boldsymbol{\theta}$ can then be inferred by marginalising the joint posterior shown below,

$$p(\mathbf{x}_0, \boldsymbol{\theta}, \boldsymbol{\sigma}^2|\mathbf{y}) \propto \pi(\boldsymbol{\theta})\pi(\boldsymbol{\sigma}^2)\pi(\mathbf{x}_0)p(\mathbf{y}|\mathbf{x}(\boldsymbol{\theta}, \mathbf{x_0}), \boldsymbol{\sigma^2})\,. \tag{4}$$

Since $\mathbf{x}(\boldsymbol{\theta}, \mathbf{x}_0)$ has no closed form solution, the posterior density (4) has no closed form, thus, Markov chain Monte Carlo (MCMC) methods are required for simulating from the posterior distribution. The choice of MCMC algorithm for inference is, however, non trivial. Nonlinearities inherit in the dynamical system can produce multiple modes in the likelihood surface, as is known to be the case for the FHN model [8]. Such complexities in the likelihoods topology can trap standard MCMC methods, such as Metropolis-Hastings, in deep local minima, making it unlikely for chain to fully explore the target distribution in any reasonable amount of time. Many techniques aimed at circumventing these issues exist. These include: simulated annealing, simulated tempering, parallel tempering etc. However, the the additional computational cost associated with these methods can become prohibitively large when applied to ODE inference problems, as every iteration of the sampler will require numerical integration of the system of ODEs. [1]

2.2 Illustrative example: FHN model

Samples were generated at 20 equally spaced time points on the interval $[0, 20]$ by solving the FHN equations (1) using the parameter values $\alpha = 0.2$, $\beta = 0.2$, $\gamma = 3$ and the initial conditions $V_0 = -1$, $R_0 = 1$. Random Gaussian noise with zero mean and 0.5 standard deviation was added to each individual sample to artificially simulate measurement errors. The synthetic data set is shown in Figure 2.

A wide gamma prior of $\Gamma(1, 3)$ was employed for each of the parameters α, β and γ. As a simplifying assumption, we assume that both the standard deviation of the noise process and the initial conditions are known.

[1] Unfortunately, time and computational constraints have prevented full exploration of the issues discussed in this section. We make no claims whatsoever regarding the suitability of our chosen algorithm for similar applications.

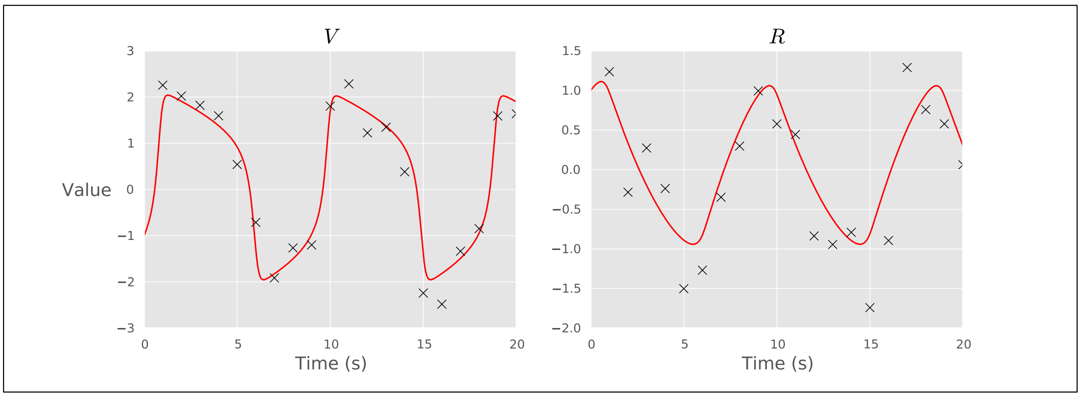

Figure 2: Plot of the synthetic data set (black crosses) used for inferring the FHN model parameters. The ‘true’ solution to the equations is shown in red.

2.2.1 MCMC Simulation

Simulations from the posterior distribution were performed using the robust adaptive Metropolis (RAM) algorithm [11]. The RAM algorithm is an adaptive random walk Metropolis algorithm. It can be viewed as a generalisation of adaptive scaling Metropolis (ASM) methods, in that the scaling factor used to adjust the size of the proposal jumps is replaced by a scaling matrix $\mathrm{S} \in \mathbb{R}^{d \times d}$, where d is the dimension of the target density. Like ASM methods, the RAM algorithm attempts to coerce the mean acceptance probability to a prefixed value through adaption of the scaling matrix. The adaptive step is performed by either a rank-1 Cholesky update or downdate, which has the same $O(d^2)$ computational complexity as that of the seminal Adaptive Metropolis (AM) algorithm [4]. Further implementation details and pseudocode are provided in Appendix A.

2.2.2 Computational methods

The RAM algorithm was implemented using the ANSI C programming language. Numerical integration of the differential equations was performed using the SUNDIALS CVode package [5] with both the absolute and relative tolerances set to 10^{-6}. Random number generation required for the Metropolis accept-reject step and sampling of the proposal distribution was performed using the GNU Scientific Library [3]. The algorithm was run on a 2011 Macbook Air (1.7 Ghz Intel Core i5 with 4GB RAM), requiring approximately 18 seconds to complete 5×10^4 iterations of the algorithm.

3 Results and Discussion

3.1 MCMC output analysis

5,000 posterior samples were generated from 51,000 iterations of the RAM algorithm, where the first 1,000 values were used as burn-in and the remaining samples were thinned by a factor of 10. Trace plots of the MCMC simulation are shown in Figure 3. The trace plots show that all parameters have good mixing, which is further justified by the autocorrelation plots shown Figure 4. Similar traces were observed over multiple MCMC runs (not shown), where the initial position of the chain was selected randomly from the prior distributions, suggesting the sampler has converged to the stationary distribution.

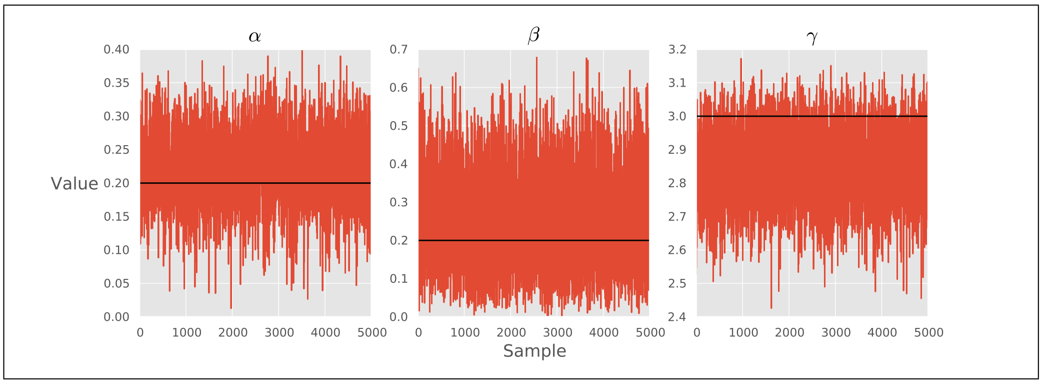

Figure 3: Trace plots for 5,000 posterior samples. The true parameters values are depicted by the black line.

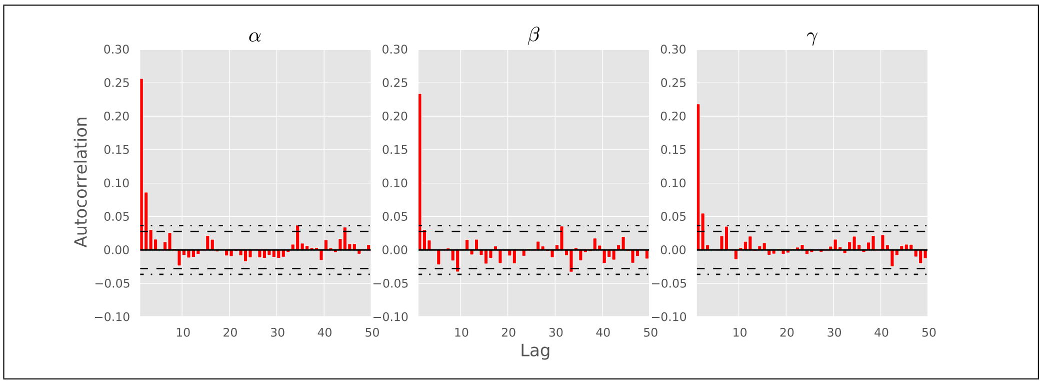

Figure 4: Autocorrelation plots for the posterior samples.

3.2 Posterior summary

The posterior means (see Table 1) appear to be in close agreement with the true values despite being inferred from a fairly sparse and noisy dataset. The β parameter exhibits a relatively large standard deviation in contrast to the other parameters (see also Figure 5). It is not clear why this is so and it may possibly be insignificant altogether. Further investigation into the role of this parameter in determining the equilibrium point of the oscillatory output is warranted. No striking correlations are apparent from the the pairwise scatterplots shown in Figure 6.

Parameter	**True Value**	**Posterior $\mu \pm \sigma$**
α	0.2	0.22 ± 0.05
β	0.2	0.26 ± 0.13
γ	3.0	2.85 ± 0.11

Table 1: Summary statistics for each of the inferred parameters of the FHN model

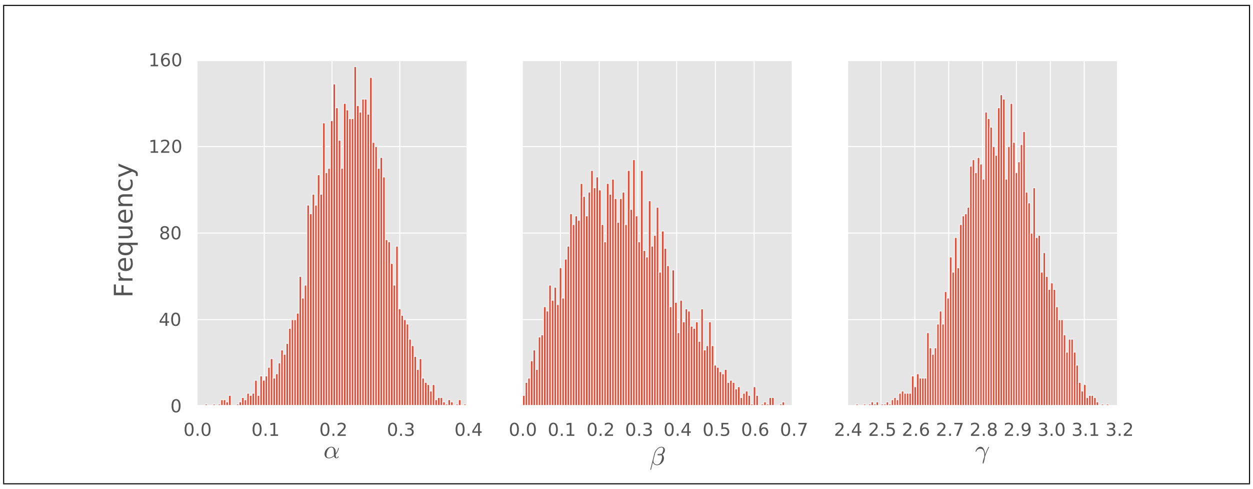

Figure 5: Posterior distributions (histograms) for each of the FHN model parameters.

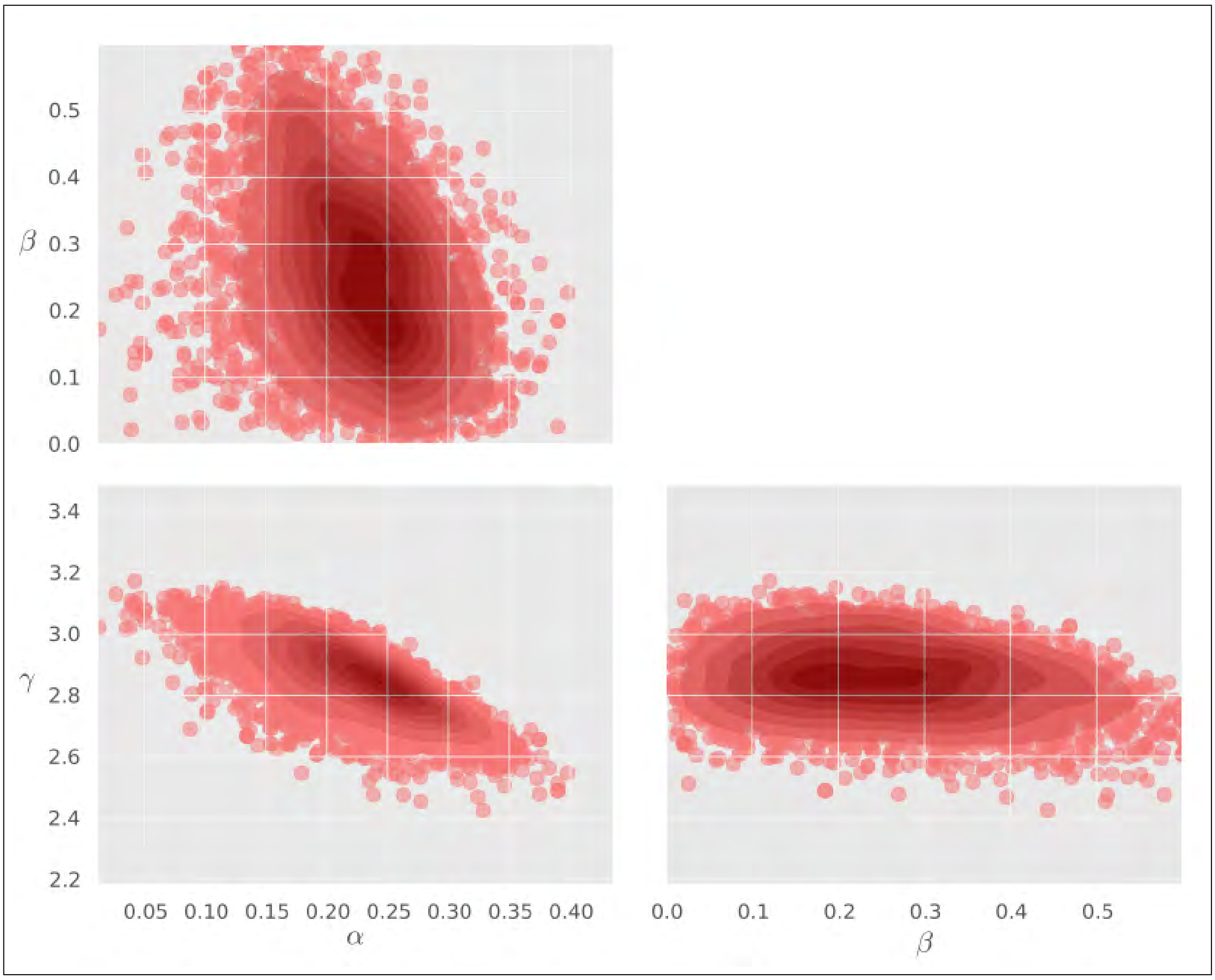

Figure 6: Pairwise scatter plots (red dots) and density estimates (contours) generated from the posterior samples. Density estimates were obtained using Gaussian kernel density estimation.

3.3 Posterior predictive distribution

Samples from the posterior predictive distribution were generated by simulating the FHN equations for each of the posterior samples at 1,000 equally spaced time points on the interval [0,20]. Gaussian noise having standard deviation equal to 0.5 was then added to each sample, as per the observation model. The resulting samples are shown below in Figure 7. Interestingly, the mean of the posterior predictive distribution is in close agreement with the true solution to the FHN equations, despite uncertainties in the inferred parameters.

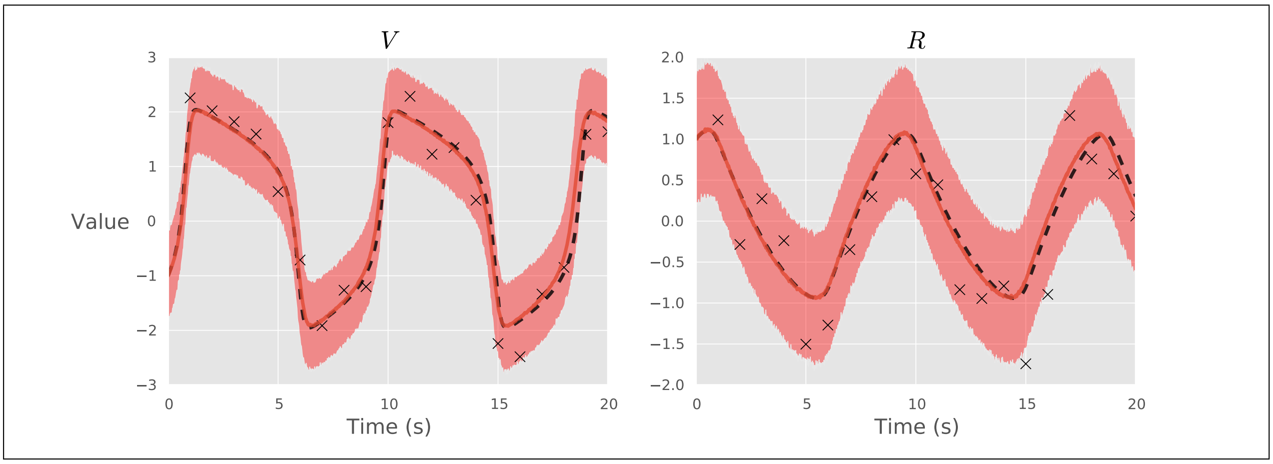

Figure 7: Comparison of the mean posterior predictive distribution (red line) to the true solution (black dashed line) and the synthetic data (black crosses). The red envelope corresponds to the smoothed 95% HPD intervals obtained at each sample point.

4 Conclusions and future work

Bayesian statistical methods are particularly well suited to the problem of parameter estimation in that they provide a coherent framework with which to systematically characterise and propagate uncertainties. This information can then be utilised at later stages in the modelling process to aid in the identification of tightly constrained predictions made by the model. The key difficulty in the application of Bayesian MCMC methods to nonlinear differential equation models appears to be the high computational cost of repeated numerical integration. For this reason, we believe future work should be directed towards minimising this computational bottleneck. Recently developed methods for parameter estimation in nonlinear differentials models based on Gaussian processes [1], which attempt to avoid numerical integration all together, appear to be one promising avenue for future research.

References

[1] Ben Calderhead, Mark Girolami, and Neil D Lawrence. Accelerating bayesian inference over nonlinear differential equations with gaussian processes. 2009.

[2] Richard Fitzhugh. Impulses and physiological states in theoretical models of nerve membrane. *Biophysical journal*, 1(6):445–466, 1961.

[3] Mark Galassi, Jim Davies, James Theiler, Brian Gough, Gerard Jungman, Michael Booth, and Fabrice Rossi. *GNU Scientific Library Reference Manual*, volume 954161734. Network Theory Ltd., 2009.

[4] Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive metropolis algorithm. *Bernoulli*, pages 223–242, 2001.

[5] Alan C. Hindmarsh, Peter N. Brown, Keith E. Grant, Steven L. Lee, Radu Serban, Dan E. Shumaker, and Carol S. Woodward. Sundials: Suite of nonlinear and differential/algebraic equation solvers. *ACM Trans. Math. Softw.*, 31(3):363–396, September 2005.

[6] Alan L Hodgkin and Andrew F Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, 117(4):500, 1952.

[7] Jinichi Nagumo, S Arimoto, and S Yoshizawa. An active pulse transmission line simulating nerve axon. *Proceedings of the IRE*, 50(10):2061–2070, 1962.

[8] Jim O Ramsay, G Hooker, D Campbell, and J Cao. Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5):741–796, 2007.

[9] Gareth O Roberts, Andrew Gelman, and Walter R Gilks. Weak convergence and optimal scaling of random walk metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120, 1997.

[10] Mark K Transtrum, Benjamin B Machta, and James P Sethna. Geometry of non-linear least squares with applications to sloppy models and optimization. *Physical Review E*, 83(3):036701, 2011.

[11] Matti Vihola. Robust adaptive metropolis algorithm with coerced acceptance rate. *Statistics and Computing*, 22(5):997–1008, 2012.

Appendices

Appendix A

Algorithm 1: Robust adaptive Metropolis

```
Input: $p(x)$: d-dimensional target density
       $q(x)$: d-dimensional symmetric proposal distribution
       $\{\gamma_n\}_{n \geq 1}$: Nonnegative adaption weight sequence which decays to zero
       $\alpha_*$: Target mean acceptance probability
       $N$: Number of iterations
Output: $\{x^{(n)}\}$: Chain of samples
begin
    Initialise $x^{(1)}$ such that $p(x^{(1)}) > 0$
    Initialise $S_1 \in \mathbb{R}^{d \times d}$ to a lower-diagonal matrix with positive diagonal
    elements
    for $n = 2 \ldots N$ do
        // Random walk Metropolis
        Compute $x^* = x^{(n-1)} + S_{n-1} u_n$ where $u_n \sim q$
        Let
            $x^{(n)} = \begin{cases} x^*, & \text{with probability } \alpha_n = \min\left\{1\,, p(x^*)/p(x^{(n-1)})\right\} \\ x^{(n-1)}, & \text{otherwise} \end{cases}$
        // Adaptive step
        Let $v = \left( \frac{\gamma_n |\alpha_n - \alpha_*|}{\|u_n\|^2} \right)^{1/2} S_{n-1} u_n$
        if $\alpha_n - \alpha_* > 0$ then
            Compute $S_n = S_{n-1} + vv^{\mathrm{T}}$ by a rank-1 Cholesky update
        else
            Compute $S_n = S_{n-1} - vv^{\mathrm{T}}$ by a rank-1 Cholesky downdate
```

For parameter inference in the FHN model, the adjustable parameters of the RAM algorithm were set to the following:

Adjustable parameter	**Value used**
Dimension d	3
Proposal density $p(x)$	$\mathcal{N}(0, \mathrm{I}_3)$
Adaptive weight sequence γ_n	$\min\{1, 3n^{-2/3}\}$ [1]
Target mean acceptance probability α_*	0.234 [2]
Number of iterations N	51000 [3]
Initial parameter values $x^{(1)}$	Random draw from priors
Initial scaling matrix S_1	I_3

Table 2: Adjustable parameter settings used for the FHN model.

[1] The factor of 3 is added to compensate for the expected growth or shrinkage in the scaling matrix eigenvalues which known to be of order d^{-1}, where d is the dimension [11].

[2] Theoretically optimal acceptance probability for a multidimensional proposal distribution [9].

[3] 1000 iterations were used as burn-in. Remaining samples were thinned by a factor of 10.