Inference in complex systems using multi-phase MCMC sampling with gradient matching burn-in

Alan Lazarus[1], Dirk Husmeier[1], Theodore Papamarkou[1]

[1] School of Mathematics and Statistics, University of Glasgow, UK

E-mail for correspondence: a.lazarus.1@research.gla.ac.uk

Abstract: We propose a novel method for parameter inference that builds on the current research in gradient matching surrogate likelihood spaces. Adopting a three phase technique, we demonstrate that it is possible to obtain parameter estimates of limited bias whilst still adopting the paradigm of the computationally cheap surrogate approximation.

Keywords: Parameter inference; Delayed Rejection Adaptive Metropolis; Surrogate likelihood; Markov Chain Monte Carlo; Gradient matching

1 Introduction

Statistical inference in nonlinear differential equations (DE) is challenging. The log-likelihood landscape is typically multimodal and every parameter adaptation, e.g. in an MCMC simulation, requires a computationally expensive numerical integration of the DEs. Using numerical methods to solve the equations results in prohibitive computational cost; particularly when one adopts a Bayesian approach in sampling parameters from a posterior distribution. Alternatively, one can try to reduce this computational complexity by obtaining an interpolant to the data from which one can obtain a comparative objective function that matches the gradients of the interpolant and the DEs. By sampling on this cheap representative likelihood surface, bias is introduced to the modelling problem. Current research focuses on reducing this bias by introducing a regularising feedback mechanism from the DEs back to the interpolation scheme (e.g. Niu et al. 2016). The idea is to make the interpolant maximally consistent with the DEs. Although this paradigm has proved to improve performance over naïve gradient matching, the feedback loop fails to fully eradicate bias in the final estimate.

This paper was published as a part of the proceedings of the 32nd International Workshop on Statistical Modelling (IWSM), Johann Bernoulli Institute, Rijksuniversiteit Groningen, Netherlands, 3–7 July 2017.

For this reason, a natural progression would be to sample from the true likelihood space whilst reducing computational complexity in the discarded burnin steps. Assuming this hypothesis, we postulate the use of a surrogate likelihood in the burnin phase alone. Through an example possessing multimodal likelihood, we will show the ability of the algorithm to avoid any local entrapment whilst obtaining accurate parameter estimates.

2 Method

Consider time-dependent observations $\mathbf{y}(t) = \boldsymbol{x}(t) + \boldsymbol{\epsilon}$—where $\boldsymbol{x}(t)$ denotes the signal and $\boldsymbol{\epsilon}$ independent additive zero mean Gaussian noise with variance parameter σ^2—whose signals are governed by a system of differential equations:

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

dependent on some (partially) unknown parameters $\boldsymbol{\theta}$. Assuming Gaussian noise, we place a GP prior on the latent variable $\mathbf{x}$

$$\mathbf{x}(t) \sim \mathcal{GP}(0, k(\mathbf{t}, \mathbf{t}')), \tag{2}$$

leading us to a Gaussian distribution, $p(\mathbf{x}_i|\boldsymbol{\phi}_i) = \mathcal{N}(\mathbf{x}_i|\mathbf{0}, \mathbf{K}_i)$ for an arbitrary set of time points $\mathcal{T} = \{t_1, ..., t_n\}$ with entries of $\mathbf{K}_i$ given by evaluating kernel function k at each element of $\mathcal{T} \times \mathcal{T}$ (Rasmussen and Williams, 2006). Under our assumption of Gaussian noise, we consider the joint distribution, $p(\mathbf{y}, \mathbf{x}|\boldsymbol{\phi}, \sigma)$. Marginalising over latent variables $\mathbf{x}$ provides a zero mean distribution for the observations:

$$\mathbf{y} \sim \mathcal{N}(0, \mathbf{K} + \sigma^2\mathbf{I}) \tag{3}$$

(see Dondelinger et al. (2013) for details). Considering the joint distribution between our signal and observed values, we may implement an elementary transformation of a Gaussian distribution to obtain the posterior distribution for our signal with mean given by:

$$\mu(\tau) = k(\tau, \mathcal{T})(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y}, \tag{4}$$

where $k(\tau, \mathcal{T})$ denotes evaluation of the kernel function at τ over $\mathcal{T}$. Subsequently, firstly estimating the hyperparameters via ML, we adopt the mean of the posterior as a representation of our signal $\mathbf{x}$. This allows us to proceed under the supposition that we have a fixed interpolant for the true signal. Given that the derivative of eq. 4,

$$\frac{\partial\mu(\tau)}{\partial\tau} = k'(\tau, \mathcal{T})(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y}, \tag{5}$$

is the mean of the Gaussian distributed DE derivative (see section 7.5 of Vanhatalo et al., 2015), we may consider:

$$f(\hat{\mathbf{x}}(t), \boldsymbol{\theta}) \sim \mathcal{N}\Big(\frac{d\hat{\mathbf{x}}(t)}{dt}, \gamma^2\mathbf{I}\Big), \tag{6}$$

where γ^2 is a fictitious noise term (assumed equal for both gradients) and $\hat{\mathbf{x}}$ is given by eq. 4. Contrary to the work done by Dondelinger et al. (2013), fixing the GP hyperparmeters $\boldsymbol{\phi}$ and the interpolant $\hat{\mathbf{x}}$ allows us to abandon the Gibbs sampling routine at this stage of the algorithm, further reducing the overall computational burden. The corresponding negative log-likelihood term provides a gradient matching objective function:

$$\pi(\boldsymbol{\theta}) = n \log\gamma^2 + \frac{1}{2\gamma^2}\left|\left|\frac{d\hat{\mathbf{x}}(t)}{dt} - f(\hat{\mathbf{x}}(t), \boldsymbol{\theta})\right|\right|^2 \tag{7}$$

which gives a representative computationally tractable surface as a surrogate for the log-likelihood. This involves the term γ^2 representing the mismatch between the gradient obtained from the differential equation and that from explicit differentiation of the GP posterior mean. This parameter will be sampled throughout the surrogate burnin phase of the algorithm. The proposed sampling scheme involves three phases. In the initial burnin phase, samples are drawn from the surrogate distribution in eq. 7 using a Delayed Rejection Adaptive Metropolis[1] (DRAM) scheme (Haario et al. (2006)). Assuming a degree of similarity between the surrogate and true likelihood surfaces, this drives the sampler towards the global minimum of the true likelihood function until a PSRF[2] value of 1.1 has been achieved. From here, we initialise a corrective phase in the true likelihood space, correcting for any bias introduced by the inconsistencies between the surrogate and true likelihood spaces. Sampling with DRAM, this phase is concluded upon obtainment of a PSRF equal to 1.1. The proceeding sampling phase replicates this corrective phase with sampling steps recorded until we achieve a PSRF value of 1.05. The stepwise decrease in target PSRF values allows time for the adaptive component of AM to learn the topology and adjust the covariance accordingly. We adopt an uninformative Inv-Gamma(0.001,0.001) prior for σ^2 and γ^2 and a $Ga(4, 0.5)$ prior for the parameters of the DE. All parameters are sampled on the log scale to account for the positivity constraint.

3 Results

We assess performance on the following DE model of circadian oscillation[3]:

$$\frac{dp_1}{dt} = \frac{k_1}{36 + k_2 p_2} - k_3, \quad \frac{dp_2}{dt} = k_4 p_1 - k_5 \tag{8}$$

This is a notoriously challenging problem due to the extreme multimodality of the likelihood. Following Girolami et al. (2010), we focus on the inference of two parameters (k_3 and k_4), setting the other parameters and initial conditions to

[1] Obtained using the adaptive Metropolis component of DRAM with *modMCMC* function in the *FME* package in R.

[2] Obtained at intervals of 20 steps using the *gelman.diag* function from the *coda* package in R.

[3] We used the same differential equations as in Girolami et al 2010. The actual Goodwin oscillator is of a slightly different form, where the terms k_3 and k_5 are replaced by $k_3 p_1$ and $k_5 p_2$, respectively.

the same fixed values as in Girolami et al. (2010). Five sets of initial parameter values for k_3 and k_4 were obtained using a Sobol sequence over the domain $[0, 5]^2$. Figure **??** shows the chain moving through the k_3-k_4 parameter domain. Comparing with the traditional method, we observe the ability of the proposed method to evade the various local minima. PSRF values of 17.1, 10.4 and 12.3 were obtained for k_3, k_4 and σ^2 simulations respectively after 10000 steps using the traditional DRAM method in true likelihood space. Comparatively, the proposed method required 1690 steps in surrogate space, 1690 in the corrective phase and 1010 in the sampling phase to achieve a PSRF of 1.05. The number of numerical integration steps required are given in Table **??** for each of the ten DRAM chains. In Figure **??**, boxplots are given that provide the distribution of

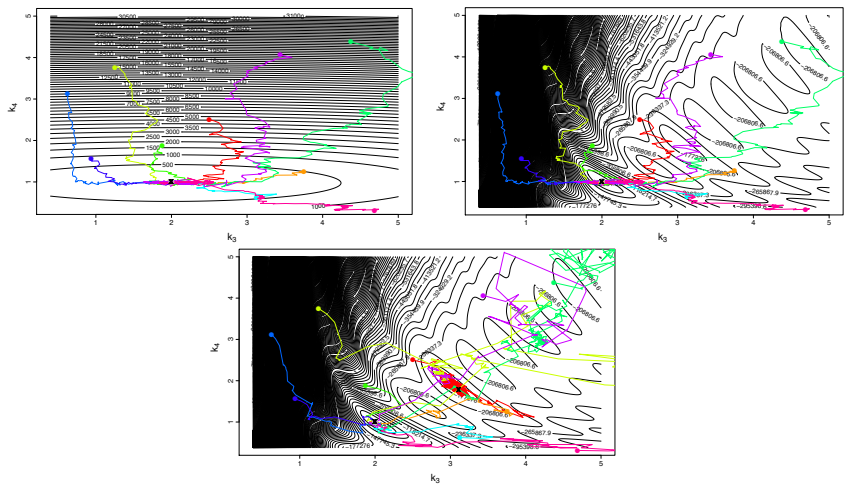

FIGURE 1: Ten chains generated with the proposed method shown in the parameter domain of the negative surrogate log-likelihood space (topleft) and negative log-likelihood space (topright). The bottom plot shows simulations generated using the traditional method. The true parameter value is given by a point at (2,1). The black crosses denote the final point of each chain.

TABLE 1: Number of numerical integration steps (N) for the traditional method. The number required in the proposed scheme is 2700.

Chain Index	1	2	3	4	5	6	7	8	9	10
N	25583	29778	29835	28452	29708	29672	29768	29530	29859	29847

bias in our sampled parameter estimates for each of the five chains. Figure **??**

provides RMS deviation in function space obtained using eq. 9,

$$RMS_{function} = \sqrt{\frac{1}{n}||\mathbf{x} - \hat{\mathbf{x}}||^2} \tag{9}$$

where $\mathbf{x}$ denotes the true signal and $\hat{\mathbf{x}}$ denotes the numerical solution of the DE for one parameter sample from the sample phase of the multiphase approach and the post burnin period of the traditional method. This provides a measure of the predictive accuracy of the MCMC samples.

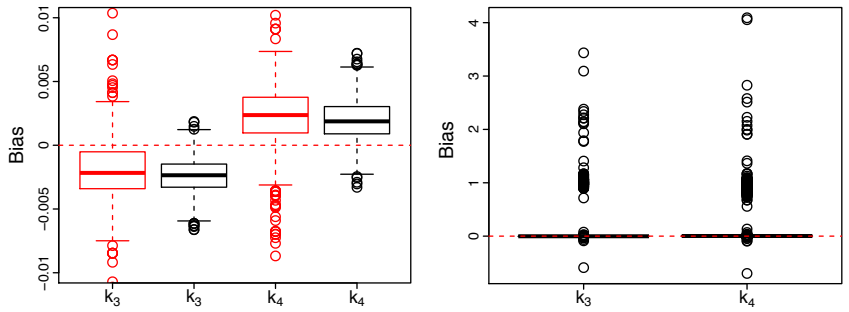

FIGURE 2: Boxplots showing the distribution of bias for both methods (left) where red boxes give the bias of the standard DRAM samples (without outliers) and the black boxes give the bias of the proposed method. The plot on the right gives the bias in both parameters for the DRAM method with outliers included.

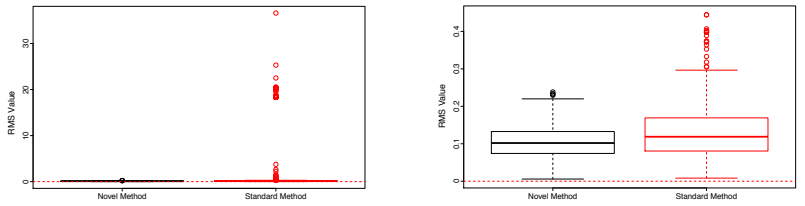

FIGURE 3: Functional RMS comparison between the proposed method (black) and DRAM (red). On the left, we include outliers (from DRAM) and, on the right, these are removed to enable better scalability of the plots for comparison. The red dotted line denotes a functional RMS equal to zero.

4 Conclusion

Our work considers the sampling in DE parameter inference as a computationally efficient three-phase scheme that achieves low levels of bias in sampled parameter estimates (Figure **??**). Achieving a PSRF of 1.05, we observe the ability of the

algorithm to converge in the parameter space of the circadian oscillator system of equations; a model for which the standard DRAM procedure fails to replicate this success (bottom of Figure **??**). Considering the results in function space, we observe the superior performance of the proposed method compared with the traditional DRAM method, showing that the performance improvement is witnessed in both domains of study. These features, along with the vast improvement in computational efficiency, demonstrate improved parameter inference compared with the traditional method.

Acknowledgments: This project was supported by a grant from the Engineering and Physical Sciences Research Council (EPSRC) of the UK, grant reference number EP/L020319/1.

References

Dondelinger, F. et al. (2013). *ODE parameter inference using adaptive gradient matching with Gaussian processes*. Proceedings of Machine Learning Research, Volume 31, pp 216-228.

Girolami, M. et al. (2010). *System identification and model ranking: the Bayesian perspective*. Learning and Inference in Computational Systems Biology. MIT Press, pp. 201-230.

Haario, H. et al. (2006). *DRAM: Efficient adaptive MCMC*. Statistics and Computing, Volume 16, Issue 4, pp 339-354.

Niu, M. et al. (2016). *Fast Parameter Inference in Nonlinear Dynamical Systems using Iterative Gradient Matching*. Proceedings of Machine Learning Research, Volume 48, pp 1699-1707.

Rasmussen, C.E. and Williams, C.K.I. (2006). *Gaussian Processes for Machine Learning*. MIT Press.

Vanhatalo, J. et al. (2015). *Bayesian Modeling with Gaussian Processes using the GPstuff Toolbox*. MIT Press.