Inference in complex biological systems with Gaussian processes and parallel tempering.

Benn Macdonald[1], Frank Dondelinger[2], Dirk Husmeier[1]

[1] University of Glasgow, Department of Mathematics and Statistics, Scotland
[2] The Netherlands Cancer Institute, Netherlands

E-mail for correspondence: b.macdonald.1@research.gla.ac.uk

Abstract: Parameter inference in mathematical models of complex biological systems, expressed as coupled ordinary differential equations (ODEs), is a challenging problem. These depend on kinetic parameters, which cannot all be measured and have to be ascertained a different way. However, the computational costs associated with repeatedly solving the ODEs are often staggering, making many techniques impractical. Therefore, aimed at reducing this cost, new concepts using gradient matching have been proposed. This paper combines current adaptive gradient matching approaches, using Gaussian processes, with a parallel tempering scheme, in order to compare 2 different paradigms using the same nonlinear regression method. We use 2 ODE systems to assess our technique, showing an improvement over the recent method in Calderhead et al. (2008).

Keywords: Parameter inference; Ordinary differential equations; Adaptive gradient matching; Gaussian processes; Parallel tempering.

1 Introduction

Ordinary differential equations (ODEs) have many applications in modelling the behaviours of systems, from fluid mechanics to systems biology. Often, there is enough knowledge of a system to model it through mathematical equations, but there is intrinsic uncertainty in the kinetic parameters governing these. Conventional methods involving Markov Chain Monte Carlo (MCMC) tend to involve integrating the system of ODEs at each iterative step, to compare how well the sampled parameters match the data. However, the computational cost can be overbearing, making these methods impractical for larger systems, and more modern methods have sought an alternative to the explicit solution. The work by Calderhead et al. (2008), Campbell and Steele (2012) and Dondelinger et al. (2013), involves fitting an interpolant to the data, then comparing the gradients from the interpolant to those from the ODEs (known as gradient matching). The original method proposed by Calderhead et al. (2008) uses a methodological simplification, which effectively ignores the posterior correlation between the

ODE parameters and the Gaussian process (GP) hyperparameters in the sampling scheme, whereas Dondelinger et al. (2013) sample all the parameters from the posterior distribution (adaptive gradient matching (AGM)). Both Dondelinger et al. (2013) and Campbell and Steele (2012) temper towards the posterior (β-tempering, Section 2.), but Campbell and Steele (2012) differs with regards to the mismatch parameter (the difference between the gradients). Whereas Dondelinger et al. (2013) infer the mismatch parameter, Campbell and Steele (2012) temper this mismatch towards zero (γ-tempering). Since this is gradual, it avoids convergence problems. We combine both methods to create an adaptive gradient matching technique, using Gaussian processes and parallel tempering (both the β and γ variety).

2 Methodology

Consider a set of T arbitrary time points $t_1 < ... < t_T$, and a set of noisy observations $\mathbf{Y} = (\mathbf{y}(t_1), ..., \mathbf{y}(t_T))$, where $\mathbf{y}(t) = \mathbf{x}(t) + \boldsymbol{\epsilon}(t)$, N = $\dim(\mathbf{x}(t))$, $\mathbf{X} = (\mathbf{x}(t_1), ..., \mathbf{x}(t_T))$. The signals of the system are described by ordinary differential equations (ODEs), of the form

$$\mathbf{x}' = \frac{d\mathbf{x}(t)}{dt} = \mathbf{f}(\mathbf{x}(t), \boldsymbol{\theta}, t); \qquad \mathbf{x}(t_1) = \mathbf{x}_1 \tag{1}$$

where $\boldsymbol{\theta}$ is a parameter vector of length p, and $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma_n^2 \mathbf{I})$. Then,

$$P(\mathbf{Y}|\mathbf{X}, \boldsymbol{\sigma}) = \prod_n \prod_t P(y_n(t)|x_n(t), \sigma_n) = \prod_n \prod_t N(y_n(t)|x_n(t), \sigma_n) \tag{2}$$

Now let $\mathbf{x}_n$ and $\mathbf{y}_n$ be T dimensional column vectors containing the n^{th} row of $\mathbf{X}$ and $\mathbf{Y}$. Following Calderhead et al. (2008), we place a GP prior on $\mathbf{x}_n$, $p(\mathbf{x}_n|\boldsymbol{\phi}) = N(\mathbf{x}_n|\mathbf{0}, \mathbf{C}_{\phi_n})$, where $\mathbf{C}_{\phi_n}$ is a positive definite matrix of covariance functions with hyperparamters ϕ_n. As the derivative of a GP is itself a GP, the conditional distribution for the state derivatives is

$$p(\mathbf{x}'|\mathbf{x}, \boldsymbol{\phi}) = N(\mathbf{m}_n, \mathbf{K}_n) \tag{3}$$

(analytical solutions to $\mathbf{m}_n$ and $\mathbf{K}_n$ in Dondelinger et al. (2013)). Assuming additive Gaussian noise with state-specific variance γ_n, from (1) we get

$$p(\mathbf{x'}_n|\mathbf{X}, \boldsymbol{\theta}, \gamma_n) = N(\mathbf{f}_n(\mathbf{X}, \boldsymbol{\theta}), \gamma_n \mathbf{I}) \tag{4}$$

Dondelinger et al. (2013) link the interpolant in (3) with the ODE model in (4) using a products of experts approach, obtaining a joint distribution for $p(\mathbf{X}', \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\gamma})$. This can then be marginalised over in closed form (see Dondelinger et al. (2013) for details), to obtain $p(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\gamma})$.
Following Dondelinger et al. (2013), we sampled $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ from the posterior distribution with MCMC. However, we did not sample $\boldsymbol{\gamma}$ directly, but instead followed Campbell and Steele (2012) to set up a ladder of fixed values associated with the "temperatures" of a parallel tempering scheme, choosing a Log_{10} scale. For details see the online supplementary material at http://www.stats.gla.ac.uk/~dhusmeier/MyPapers/IWSM2013Macd.pdf

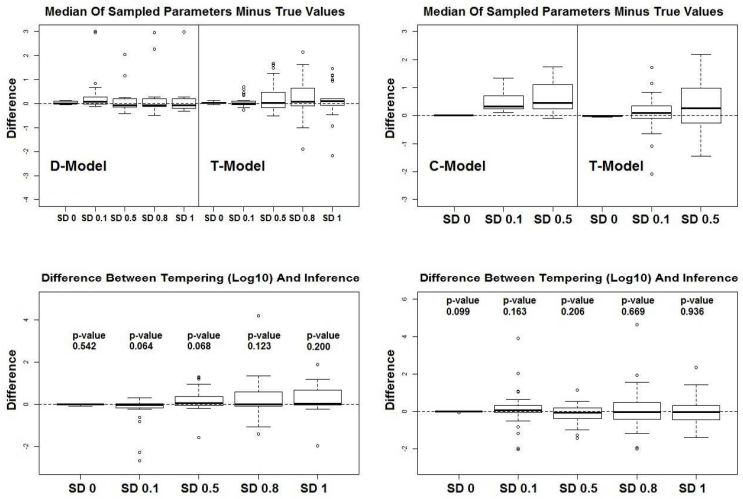

FIGURE 1. Parameter estimation accuracy of $\boldsymbol{\theta}$ over noise instantiations, for the Fitz-Hugh Nagumo (left) and Lotka-Volterra (right) systems. Some outliers in the plots have been removed for scalability. The dashed lines show zero difference. Top Row: Boxplots, over the 10 datasets, of differences between the median of sampled parameters and true values. The solid line splits the D-Model/C-Model (left) from the T-Model (right). Bottom Row: Boxplots, over the 10 datasets, of the differences in parameter estimation accuracy for the D-Model and T-Model. The p-values for a paired t-test are shown above the corresponding boxplot.

3 Results

We tested our method on the Fitz-Hugh Nagumo (FitzHugh (1961) and Nagumo et al. (1962)) and Lotka-Volterra (Lotka (1932)) ODE models. For space restrictions, details of the equations and parameters have been relegated to the online supplementary material.

We introduce the abridged notation used in this section: The method described in Calderhead et al. (2008) shall be denoted, C-Model, the method described in Dondelinger et al. (2013), D-Model, and the new combined method proposed in this paper, T-Model. For each system, method and added observational noise level, 10 datasets were generated. By averaging over these, we are able to remove specific characteristics of a dataset and observe more clearly our method's performance. The median was used as an estimator of the parameters and the true values were subtracted from the sampled parameter estimates. The distributions (of estimate minus true value) over the 10 datasets were compared.

The first row of FIGURE 1. shows the distribution of the estimate to the true parameter for the D-Model, C-Model and T-Model (Log_{10}), for the

FhN and LV systems. For zero noise, both the C-Model and T-Model have boxplots centred very close to zero, displaying good performance. However, when increasing the noise, the C-Model no longer has a distribution centred around zero (no part of the distribution for noise = 0.1 and only a small part of the lower tail for noise = 0.5). For all noise instantiations, the T-Model (and D-Model) has most of its mass centred around zero. Therefore, if averaging over all datasets, for the T-Model, the true parameters are close to the estimates i.e. this technique is unbiased. The second row of FIGURE 1. allows us to check how robust our technique is. The plots show the distributions of the differences between the absolute distance of the estimator to the true parameter for the T-Model and D-Model. These distributions are centred around zero, indicating that there is no noticeable difference between the parameter estimation accuracy of these two techniques. We can therefore see that our technique is robust to noise.

4 Conclusion

We have carried out a comparative evaluation of two schemes for adaptive gradient matching: posterior inference vs. parallel tempering of the gradient mismatch hyperparameter. The tempering scheme was originally proposed in the context of splines-based regression, which we have adapted to non-parametric Bayesian modelling, with Gaussian processes. An application to data, generated from two different systems of ODEs, shows no significant difference between the parallel tempering and posterior inference. We found that both methods outperform a related method by Calderhead et al. (2008), considered the current state of the art.

References

Calderhead, B. et al. (2008). Accelerating Bayesian inference over nonlinear differential equations with Gaussian processes. *NIPS*, **22**,

Campbell, D. and Steele., R.J. (2012). Smooth functional tempering for nonlinear differential equation models. *Stat Comput*, **22**, 429 – 443.

Dondelinger, F. et al. (2013). ODE parameter inference using adaptive gradient matching with Gaussian processes. *In Press: 16th AISTATS*

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

Lotka, A. (1932). The growth of mixed populations: two species competing for a common food supply. *Journal of the Washington Academy of Sciences*, **22**, 461 – 469.

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