ODE parameter inference using adaptive gradient matching with Gaussian processes

F. Dondelinger
The Netherlands Cancer Institute
f.dondelinger@nki.nl

M. Filippone and S. Rogers
School of Computing Science
University of Glasgow

D. Husmeier
School of Mathematics and Statistics
University of Glasgow

Abstract

Parameter inference in mechanistic models based on systems of coupled differential equations is a topical yet computationally challenging problem, due to the need to follow each parameter adaptation with a numerical integration of the differential equations. Techniques based on gradient matching, which aim to minimize the discrepancy between the slope of a data interpolant and the derivatives predicted from the differential equations, offer a computationally appealing shortcut to the inference problem. The present paper discusses a method based on nonparametric Bayesian statistics with Gaussian processes due to Calderhead et al. (2008), and shows how inference in this model can be substantially improved by consistently sampling from the joint distribution of the ODE parameters and GP hyperparameters. We demonstrate the efficiency of our adaptive gradient matching technique on three benchmark systems, and perform a detailed comparison with the method in Calderhead et al. (2008) and the explicit ODE integration approach, both in terms of parameter inference accuracy and in terms of computational efficiency.

1 INTRODUCTION

In many domains of applications, ordinary differential equations (ODEs) are a useful tool for modeling the behaviour of a system. Systems where they have been applied range from physics and engineering to ecology (Lotka, 1932), and recently, systems biology (see e.g. De Jong, 2002). In systems biology, ODEs have been used to describe the dynamics of pathways and gene regulatory interactions in the cell (Pokhilko et al., 2010). Frequently, molecular biologists will have sufficient knowledge about a system to define the equations that govern its behaviour, but there will be uncertainty about the kinetic or thermodynamic parameters. A common way to resolve this uncertainty is to use some form of parameter inference based on the available experimental data (Ashyraliyev et al., 2009). Previous approaches to parameter inference in ODEs have ranged from maximum likelihood over variational approximations and Markov Chain Monte Carlo (MCMC) to Hamiltonian Monte Carlo (Girolami and Calderhead, 2011). Generally, all of these approaches involve explicitly solving the ODE system at each inference step to evaluate how well the inferred parameter values match the data. As this incurs a computational cost at each step, which grows linearly with the dataset size and size of the system, alternatives have been developed that avoid explicitly solving the system of differential equations (Varah, 1982; Poyton et al., 2006; Ramsay et al., 2007; Calderhead et al., 2008). These alternatives work by interpolating the signal from the observed experimental data and calculating the gradients, to which the ODE system can then be fitted directly.

One recent approach is described in Calderhead et al. (2008). This approach uses Gaussian Processes (GPs) to model the experimental data, which has the advantage that all the parameters can be inferred from the data. A disadvantage of the method proposed in Calderhead et al. (2008) is that the hyperparameters of the Gaussian process are inferred based on the data alone, without any rectifying feedback mechanism from the ODE system. This falls short of related previous approaches, like Ramsay et al. (2007). While the approach in Calderhead et al. (2008) generally works well for the limiting case of zero noise, we have observed that it tends to lead to rather poor parameter estimation from data subject to noise. In the present paper, we propose an improved inference

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31.

scheme, which we call adaptive gradient matching (AGM). In this scheme, both the hyperparameters of the Gaussian process as well as the ODE parameters are jointly and consistently inferred from the posterior distribution, leading to an essential information coupling between both, by taking account of their correlation. The scheme is adaptive, in that unlike in Calderhead et al. (2008), the GP is adapted during the inference based on information from the ODE system. We demonstrate that this leads to a significant improvement in the robustness with respect to noise.

2 METHOD

2.1 Proposal by Calderhead et al. (2008)

Consider a set of T arbitrary time points $t_1 < \ldots < t_T$, and a sequence of noisy observations $\mathbf{Y} = (\mathbf{y}(t_1), ..., \mathbf{y}(t_T))$,

$$\mathbf{y}(t) = \mathbf{x}(t) + \boldsymbol{\varepsilon}(t) \tag{1}$$

of a K-dimensional process $\mathbf{X} = (\mathbf{x}(t_1), ..., \mathbf{x}(t_T))$, $\dim[\mathbf{x}(t)] = \dim[\mathbf{y}(t)] = \dim[\boldsymbol{\varepsilon}(t)] = K$, whose evolution is defined by a system of K ordinary differential equations (ODEs):

$$\dot{\mathbf{x}}(t) \;=\; \frac{d\mathbf{x}(t)}{dt} \;=\; \mathbf{f}(\mathbf{x}(t), \boldsymbol{\theta}); \quad \mathbf{x}(t_1) \;=\; \mathbf{x}_1 \tag{2}$$

with parameter vector $\boldsymbol{\theta}$ of length P, and ε is a multivariate Gaussian noise process $\varepsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$, where $D_{ik} = \sigma_k^2 \delta_{ik}$, i.e. for simplicity we assume the covariance matrix $\mathbf{D}$ to be diagonal:

$$\begin{aligned} P(\mathbf{Y}|\mathbf{X}, \boldsymbol{\sigma}) &= \prod_k \prod_t P(y_k(t)|x_k(t), \sigma_k) \\ &= \prod_k \prod_t \mathcal{N}(y_k(t)|x_k(t), \sigma_k^2) \end{aligned} \tag{3}$$

The matrices $\mathbf{X}$ and $\mathbf{Y}$ are of dimension K-by-T. Let $\mathbf{x}_k$ and $\mathbf{y}_k$ denote T-dimensional column vectors that contain the kth row of the matrices $\mathbf{X}$ and $\mathbf{Y}$, respectively. Hence, $\mathbf{x}_k$ and $\mathbf{y}_k$ represent the respective time series of the kth state.

Given that inference based on an explicit numerical integration of the differential equations, as pursued in Vyshemirsky and Girolami (2008), tends to incur high computational costs, an alternative approach based on non-parametric Bayesian modelling with Gaussian processes was proposed in Calderhead et al. (2008). They put a Gaussian process prior on $\mathbf{x}_k$,

$$p(\mathbf{x}_k|\boldsymbol{\phi}_k) \;=\; \mathcal{N}(\mathbf{x}_k|\mathbf{0}, \mathbf{C}_{\phi_\mathbf{k}}) \tag{4}$$

where $\mathbf{C}_{\phi_\mathbf{k}}$ denotes a positive definite matrix of covariance functions with hyperparameters $\boldsymbol{\phi}_k$. Assuming additive Gaussian noise with a state-specific error variance σ_k^2, we get:

$$p(\mathbf{y}_k|\mathbf{x}_k, \sigma_k) \;=\; \mathcal{N}(\mathbf{y}_k|\mathbf{x}_k, \sigma_k^2\mathbf{I}) \tag{5}$$

$$\begin{aligned} p(\mathbf{y}_k|\boldsymbol{\phi}_k, \sigma_k) &= \int p(\mathbf{y}_k|\mathbf{x}_k, \sigma_k) p(\mathbf{x}_k|\boldsymbol{\phi}_k) d\mathbf{x}_k \\ &= \int \mathcal{N}(\mathbf{y}_k|\mathbf{x}_k, \sigma_k^2\mathbf{I}) \mathcal{N}(\mathbf{x}_k|\mathbf{0}, \mathbf{C}_{\phi_\mathbf{k}}) d\mathbf{x}_k \\ &= \mathcal{N}(\mathbf{y}_k|\mathbf{0}, \mathbf{C}_{\phi_\mathbf{k}} + \sigma_k^2\mathbf{I}) \end{aligned} \tag{6}$$

The conditional distribution for the state derivatives is given by

$$p(\dot{\mathbf{x}}_k|\mathbf{x}_k, \boldsymbol{\phi}) \;=\; \mathcal{N}(\mathbf{m}_k, \mathbf{A}_k) \tag{7}$$

where

$$\mathbf{m}_k = {}'\mathbf{C}_{\phi_\mathbf{k}} {\mathbf{C}_{\phi_\mathbf{k}}}^{-1} \mathbf{x}_k; \quad \mathbf{A}_k = \mathbf{C}''_{\phi_\mathbf{k}} - {}'\mathbf{C}_{\phi_\mathbf{k}} {\mathbf{C}_{\phi_\mathbf{k}}}^{-1} \mathbf{C}'_{\phi_\mathbf{k}} \tag{8}$$

Here, the matrix $\mathbf{C}''_{\phi_\mathbf{k}}$ denotes the auto-covariance for each state derivative, and the matrices $\mathbf{C}'_{\phi_\mathbf{k}}$ and ${}'\mathbf{C}_{\phi_\mathbf{k}}$ denote the cross-covariances between the kth state and its derivative. See supplementary material A.1 for details. Assuming additive Gaussian noise with a state-specific error variance γ_k, one gets from (2):

$$p(\dot{\mathbf{x}}_k|\mathbf{X}, \boldsymbol{\theta}, \gamma_k) \;=\; \mathcal{N}(\mathbf{f}_k(\mathbf{X}, \boldsymbol{\theta}), \gamma_k\mathbf{I}) \tag{9}$$

where $\mathbf{f}_k(\mathbf{X}, \boldsymbol{\theta}) = (\mathbf{f}_k(\mathbf{x}(t_1), \boldsymbol{\theta}), ..., \mathbf{f}_k(\mathbf{x}(t_T), \boldsymbol{\theta}))^\mathsf{T}$. Next, the approach taken in Calderhead et al. (2008) is to combine (7) and (9) with a product of experts approach:

$$\begin{aligned} p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbf{X}, \boldsymbol{\phi}) &= \int p(\dot{\mathbf{X}}, \boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbf{X}, \boldsymbol{\phi}) d\dot{\mathbf{X}} \\ &\propto p(\boldsymbol{\theta}) p(\boldsymbol{\gamma}) \int p(\dot{\mathbf{X}}, \mathbf{X}, \boldsymbol{\phi}|\boldsymbol{\theta}, \boldsymbol{\gamma}) d\dot{\mathbf{X}} \\ &\propto p(\boldsymbol{\theta}) p(\boldsymbol{\gamma}) \prod_k \int p(\dot{\mathbf{x}}_k|\mathbf{x}_k, \boldsymbol{\phi}) p(\dot{\mathbf{x}}_k|\mathbf{X}, \boldsymbol{\theta}, \gamma_k) d\dot{\mathbf{x}}_k \\ &= p(\boldsymbol{\theta}) p(\boldsymbol{\gamma}) \prod_k \int \mathcal{N}(\dot{\mathbf{x}}_k|\mathbf{m}_k, \mathbf{A}_k) \mathcal{N}(\dot{\mathbf{x}}_k|\mathbf{f}_k(\mathbf{X}, \boldsymbol{\theta}), \gamma_k\mathbf{I}) d\dot{\mathbf{x}}_k \\ &\propto \frac{p(\boldsymbol{\theta}) p(\boldsymbol{\gamma})}{\prod_k Z(\gamma_k)} \times \\ &\qquad \exp\left\{ -\frac{1}{2} \sum_k (\mathbf{f}_k - \mathbf{m}_k)^\mathsf{T} (\mathbf{A}_k + \gamma_k\mathbf{I})^{-1} (\mathbf{f}_k - \mathbf{m}_k) \right\} \end{aligned} \tag{10}$$

where $p(\boldsymbol{\theta})$ and $p(\boldsymbol{\gamma})$ are the prior distributions on $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$, $Z(\gamma_k) = |2\pi(\mathbf{A}_k + \gamma_k\mathbf{I})|^{1/2}$ and we have defined $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_K)$ and $\mathbf{f}_k = \mathbf{f}_k(\mathbf{X}, \boldsymbol{\theta})$. Inference is based on sampling the parameters of the ODEs $\boldsymbol{\theta}$, the hyperparameters of the Gaussian process $\boldsymbol{\phi}$, the noise variances $\boldsymbol{\gamma}, \boldsymbol{\sigma}$, and the state variables $\mathbf{X}$ from the posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\sigma}, \mathbf{X}|\mathbf{Y})$ with the following Gibbs sampling procedure:

$$\boldsymbol{\phi}, \boldsymbol{\sigma} \;\sim\; p^*(\boldsymbol{\phi}, \boldsymbol{\sigma}|\mathbf{Y}) \tag{11}$$

$$\mathbf{X} \;\sim\; p(\mathbf{X}|\mathbf{Y}, \boldsymbol{\sigma}, \boldsymbol{\phi}) \tag{12}$$

$$\boldsymbol{\theta}, \boldsymbol{\gamma} \;\sim\; p(\boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbf{X}, \boldsymbol{\phi}, \boldsymbol{\sigma}) \tag{13}$$

The distribution in the last sampling step, (13), is given by (10). This distribution does not have a standard form, and sampling from it directly is infeasible. Hence, MCMC with the Metropolis-Hastings algorithm (Hastings, 1970) is used. Note that $p(\boldsymbol{\phi}, \boldsymbol{\sigma}|\mathbf{Y}) = \int p(\dot{\mathbf{X}}, \boldsymbol{\phi}, \boldsymbol{\sigma}, \boldsymbol{\theta}, \boldsymbol{\gamma}|\mathbf{Y}) d\dot{\mathbf{X}} d\boldsymbol{\theta} d\boldsymbol{\gamma}$ is analytically intractable. Calderhead et al. (2008) approximate $p(\boldsymbol{\phi}, \boldsymbol{\sigma}|\mathbf{Y})$ by a distribution derived from a standard Gaussian process that is decoupled from the rest of the model. We call this $p^*(\boldsymbol{\phi}, \boldsymbol{\sigma}|\mathbf{Y})$. The

sampling steps (11) and (12) are broken up into the contributions from the individual states k:

$$\begin{aligned} \boldsymbol{\phi}_k, \sigma_k &\sim p^*(\boldsymbol{\phi}_k, \sigma_k|\mathbf{y}_k) \qquad (14)\\ &\propto p(\mathbf{y}_k|\boldsymbol{\phi}_k, \sigma_k)p(\boldsymbol{\phi}_k)p(\sigma_k)\\ &= \mathcal{N}(\mathbf{y}_k|\mathbf{0}, \sigma_k^2\mathbf{I} + \mathbf{C}_{\phi_\mathbf{k}})p(\boldsymbol{\phi}_k)p(\sigma_k)\\ \mathbf{x}_k &\sim p(\mathbf{x}_k|\mathbf{y}_k, \sigma_k, \boldsymbol{\phi}_k) = \mathcal{N}(\mathbf{x}_k|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \qquad (15)\end{aligned}$$

where $\boldsymbol{\mu}_k = \mathbf{C}_{\phi_\mathbf{k}}(\mathbf{C}_{\phi_\mathbf{k}}+\sigma_k^2\mathbf{I})^{-1}\mathbf{y}_k$ and $\boldsymbol{\Sigma}_k = \sigma_k^2\mathbf{C}_{\phi_\mathbf{k}}(\mathbf{C}_{\phi_\mathbf{k}}+\sigma_k^2\mathbf{I})^{-1}$. Equation (15) follows from $p(\mathbf{x}_k|\mathbf{y}_k, \sigma_k, \boldsymbol{\phi}_k) = p(\mathbf{y}_k|\mathbf{x}_k, \sigma_k)p(\mathbf{x}_k|\boldsymbol{\phi}_k)/p(\mathbf{y}_k|\sigma_k, \boldsymbol{\phi}_k)$, equations (4–6) are well-established results for Gaussian distributions. Sampling of the vector of latent variables $\mathbf{x}_k$ in (15) follows directly from a multivariate Gaussian distribution. For sampling $\boldsymbol{\phi}_k$ and σ_k in (14), one again has to resort to MCMC. The overall MCMC scheme then iteratively loops through the steps (11–13) until some convergence criterion has been met.[1] However, the approximation in equation (11) of the sampling scheme introduces a certain weakness: the parameters of the ODE systems, $\boldsymbol{\theta}, \boldsymbol{\gamma}$, which are sampled in the third step of the Gibbs sampling routine (13), never feed back into the first and second steps, (11–12). This implies that $\boldsymbol{\theta}, \boldsymbol{\gamma}$ have no bearing on the inference of the state variables $\mathbf{X}$; these state variables are solely inferred from the observed data via a standard Gaussian process interpolation, (11–12). Hence the method proposed in Calderhead et al. (2008) is a two-step procedure, in which first an interpolation problem is solved, and then the parameters of the ODEs are inferred by matching the derivatives of the interpolant with those predicted from the ODEs. This falls short of the method proposed in Ramsay et al. (2007), where the interpolation fits both the noisy data and the derivatives from the ODEs simultaneously, allowing the system of ODEs to feed back onto the interpolation.

2.2 Adaptive Gradient Matching

We demonstrate that with a mathematically more consistent formulation of the inference procedure, we can close the desired feedback loop between interpolation and parameter estimation of the ODEs. Following Calderhead et al. (2008), we combine (7) and (9) with a product of experts approach:

$$\begin{aligned} p(\dot{\mathbf{x}}_k|\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\phi}, \gamma_k) &\propto p(\dot{\mathbf{x}}_k|\mathbf{x}_k, \boldsymbol{\phi})p(\dot{\mathbf{x}}_k|\mathbf{X}, \boldsymbol{\theta}, \gamma_k) \qquad (16)\\ &= \mathcal{N}(\dot{\mathbf{x}}_k|\mathbf{m}_k, \mathbf{A}_k)\mathcal{N}(\dot{\mathbf{x}}_k|\mathbf{f}_k(\mathbf{X}, \boldsymbol{\theta}), \gamma_k\mathbf{I})\end{aligned}$$

We obtain for the joint distribution:

$$\begin{aligned} &p(\dot{\mathbf{X}},\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\gamma})\\ &\quad= p(\dot{\mathbf{X}}|\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\gamma})p(\mathbf{X}|\boldsymbol{\phi})p(\boldsymbol{\theta})p(\boldsymbol{\phi})p(\boldsymbol{\gamma}) \qquad (17)\\ &\quad= p(\boldsymbol{\theta})p(\boldsymbol{\phi})p(\boldsymbol{\gamma})\prod_k p(\dot{\mathbf{x}}_k|\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\phi}, \gamma_k)p(\mathbf{x}_k|\boldsymbol{\phi}_k)\end{aligned}$$

Inserting (4) and (16), we get:

$$\begin{aligned} p(\dot{\mathbf{X}}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\gamma}) \propto\ & p(\boldsymbol{\theta})p(\boldsymbol{\phi})p(\boldsymbol{\gamma})\prod_k \mathcal{N}(\dot{\mathbf{x}}_k|\mathbf{m}_k, \mathbf{A}_k)\\ & \mathcal{N}(\dot{\mathbf{x}}_k|\mathbf{f}_k(\mathbf{X}, \boldsymbol{\theta}), \gamma_k\mathbf{I})\mathcal{N}(\mathbf{x}_k|\mathbf{0}, \mathbf{C}_{\phi_\mathbf{k}}) \qquad (18)\end{aligned}$$

The marginalization over the state derivatives $\dot{\mathbf{X}}$

$$\begin{aligned} p(\mathbf{X}, \boldsymbol{\theta},\boldsymbol{\phi}, \boldsymbol{\gamma}) &= \int p(\dot{\mathbf{X}}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\gamma})d\dot{\mathbf{X}}\\ &\propto p(\boldsymbol{\theta})p(\boldsymbol{\phi})p(\boldsymbol{\gamma})\prod_k \mathcal{N}(\mathbf{x}_k|\mathbf{0}, \mathbf{C}_{\phi_\mathbf{k}})\times \qquad (19)\\ &\quad\int \mathcal{N}(\dot{\mathbf{x}}_k|\mathbf{m}_k, \mathbf{A}_k)\mathcal{N}(\dot{\mathbf{x}}_k|\mathbf{f}_k(\mathbf{X}, \boldsymbol{\theta}), \gamma_k\mathbf{I})d\dot{\mathbf{x}}_k\end{aligned}$$

is analytically tractable and yields:

$$\begin{aligned} p(\mathbf{X}, \boldsymbol{\theta},\boldsymbol{\phi}, \boldsymbol{\gamma}) &\propto p(\boldsymbol{\theta})p(\boldsymbol{\phi})p(\boldsymbol{\gamma})p(\mathbf{X}|\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\gamma})\\ &\propto \prod_k \mathcal{N}(\mathbf{x}_k|\mathbf{0}, \mathbf{C}_{\phi_\mathbf{k}})\times\\ &\quad \exp\Big[-\frac{1}{2}(\mathbf{f}_k - \mathbf{m}_k)^\mathsf{T}(\mathbf{A}_k + \gamma_k\mathbf{I})^{-1}(\mathbf{f}_k - \mathbf{m}_k)\Big]\\ &\propto \exp[-\frac{1}{2}\sum_k(\mathbf{x}_k^\mathsf{T}\mathbf{C}_{\phi_k}^{-1}\mathbf{x}_k +\\ &\quad (\mathbf{f}_k - \mathbf{m}_k)^\mathsf{T}(\mathbf{A}_k + \gamma_k\mathbf{I})^{-1}(\mathbf{f}_k - \mathbf{m}_k))] \qquad (20)\end{aligned}$$

where $\mathbf{m}_k$ and $\mathbf{A}_k$ were defined in (8). Note that this distribution is a complicated function of the states $\mathbf{X}$, owing to the nonlinear dependence via $\mathbf{f}_k = \mathbf{f}_k(\mathbf{X}, \boldsymbol{\theta})$. For the joint probability distribution of the whole system we obtain:

$$\begin{aligned} &p(\mathbf{Y}, \mathbf{X},\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\sigma}) =\\ &\qquad p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\sigma})p(\mathbf{X}|\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\gamma})p(\boldsymbol{\theta})p(\boldsymbol{\phi})p(\boldsymbol{\gamma})p(\boldsymbol{\sigma}) \qquad (21)\end{aligned}$$

where the first factor, $p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\sigma})$, was defined in (3), and the second factor is given by (20). Note that the functional form of the second term is defined up to an unknown normalization constant. To bypass the problem of normalizing the distribution (20), we follow a Metropolis-Hastings scheme. Denote by $q_1(\boldsymbol{\sigma})$, $q_2(\boldsymbol{\phi})$, $q_3(\mathbf{x}_k)$, $q_4(\boldsymbol{\theta})$ and $q_5(\boldsymbol{\gamma})$ the proposal distributions for the inferred parameters. We propose new values from these distributions; q_1 and q_5 are sparse exponential distributions with $\lambda = 10$ to ensure small noise values and q_2, q_3 and q_4 are uniform distributions over the intervals $[0, 100]$, $[0, 10]$ and $[0, 20]$, respectively. These proposal distributions correspond to the prior distributions for the parameters in our model, except for $\boldsymbol{\sigma}$ where we use a sparse gamma prior $Ga(1, 1)$,

[1] Note that the method proposed in Calderhead et al. (2008) slightly deviates from the summary given here in that (8) is defined as follows: $\mathbf{m}_k = {}'\mathbf{C}_{\phi_\mathbf{k}}[\mathbf{C}_{\phi_\mathbf{k}} + \sigma_k^2\mathbf{I}]^{-1}\mathbf{x}_k$ and $\mathbf{A}_k = \mathbf{C}''_{\phi_\mathbf{k}} - {}'\mathbf{C}_{\phi_\mathbf{k}}[\mathbf{C}_{\phi_\mathbf{k}} + \sigma_k^2\mathbf{I}]^{-1}\mathbf{C}'_{\phi_\mathbf{k}}$, which leads to the dependence of (10) on $\boldsymbol{\sigma}$. However, this formulation, which is motivated by including information from the data $\mathbf{Y}$, is methodologically inconsistent.

and $\boldsymbol{\theta}$, where we have imposed a gamma distribution $Ga(4, 0.5)$ as a prior to encode our prior belief about parameter values, which is that most parameters will be > 0 and < 5.

We then accept or reject these proposal moves according to the standard Metropolis-Hastings criterion (Hastings, 1970). Define $\pi(\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\sigma}) = \frac{p(\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\sigma})}{q_1(\boldsymbol{\sigma}) q_2(\boldsymbol{\phi}) q_4(\boldsymbol{\theta}) q_5(\boldsymbol{\gamma}) \prod_k q_3(\mathbf{x}_k)}$, then:

$$P_{accept} = \min \left\{ 1, \frac{\pi(\mathbf{Y}, \tilde{\mathbf{X}}, \tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\phi}}, \tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\sigma}})}{\pi(\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\sigma})} \right\} \qquad (22)$$

For improved mixing and convergence, it is advisable to not propose all moves simultaneously, but to apply a blocking strategy and employ a Gibbs sampling scheme. We do not make that explicit in our notation, though. The effect of (22) is that the parameters $\boldsymbol{\theta}$ have an influence on the acceptance probabilities for $\mathbf{X}$. This mechanism closes the feedback loop, with the system of ODEs acting back in an adaptive manner on the interpolants $\mathbf{x}_k$ via the parameters $\boldsymbol{\theta}$. In this way, we address the main shortcoming of the method proposed in Calderhead et al. (2008).

3 SAMPLING SETUP

For running simulations with the model in Calderhead et al. (2008), we make use of the MATLAB code provided by the authors. Our adaptive gradient matching model was implemented in R, where we followed the sampling scheme from Calderhead et al. (2008) whenever possible. Like Calderhead et al., we used population MCMC (Jasra et al., 2007) to deal with the potentially rugged likelihood landscapes of the nonlinear ODE systems. For all MCMC simulations in this paper, we ran 10 chains at different temperatures, starting from an exponential scale which we tuned during the burn-in phase to achieve an acceptance rate of 0.25 for exchange moves.[2]. Similarly, proposal widths for all parameters and hyperparameters were tuned to achieve an acceptance rate of 0.25. The choice of 0.25 is motivated by analogy to Gelman (1997), where an acceptance rate of 0.234 was found to be asymptotically optimal for a random walk Metropolis algorithm. We initialised $\mathbf{X}$ and $\boldsymbol{\phi}$ using a GP regression fit with maximum likelihood to the data $\mathbf{Y}$; the same initial GP hyperparameters were used for the Calderhead et al. model and for our improved gradient matching model. All other parameters were initialised by drawing samples from the prior distributions defined in Section 2.2.

[2]We did not employ cross-over moves in this sampler, although implementing a cross-over scheme similar to the one in Jasra et al. (2007) could potentially speed up mixing and convergence.

The sampling of the hyperparameters $\boldsymbol{\phi}$ and the latent variables $\mathbf{X}$ warrants further explanation. Although we could in principle propose new values for $\mathbf{X}$ and $\boldsymbol{\phi}$ by sampling them alternately from the prior, or from some other distribution, e.g. via a random walk, this is highly inefficient due to the strong coupling between them. To avoid this problem, we apply a whitening of the prior, following Murray and Adams (2010). We introduce an independent Gaussian vector $\boldsymbol{\nu}$, and update the hyperparameters $\boldsymbol{\phi}$ for fixed $\boldsymbol{\nu}$ instead of fixed $\mathbf{X}$, by using the transformation $\mathbf{X} = \mathbf{L}_{\mathbf{C}_{\phi_\mathbf{k}}} \boldsymbol{\nu}$, where $\mathbf{L}_{\mathbf{C}_{\phi_\mathbf{k}}} \mathbf{L}_{\mathbf{C}_{\phi_\mathbf{k}}}^\top = \mathbf{C}_{\phi_\mathbf{k}}$. Since $\boldsymbol{\nu}$ and $\boldsymbol{\phi}$ are independent, this scheme removes the problems created by strong coupling. Furthermore, these updates will change both $\mathbf{X}$ and $\boldsymbol{\phi}$; in effect, we are now treating the latent variables as ancillary to the GP hyperparameters.

For the GP methods, the choice of covariance function can be important, as the GP needs to be able to fit the dynamics of the data. For the *PIF4/5* model and the Lotka-Volterra model described in Section 4, a radial basis function covariance function $k(t, t') = \sigma_{kern}^2 \exp(-0.5 * (t - t')^2 / l^2)$ with hyperparameters σ_{kern}^2 and l^2 (variance and characteristic lengthscale) was used, which provided a good fit. However, this covariance function does not provide a good fit for data from the model for the signal transduction cascade (also described in Section 4). We therefore switched to a sigmoid covariance function $k(t, t') = \sigma_{kern}^2 \arcsin\left(\frac{a + b*t*t'}{\sqrt{(a+b*t*t+1)(a+b*t'*t'+1)}} \right)$ with hyperparameters σ_{kern}^2, a and b. Note that in general the sigmoid covariance function gives good regression fits for all models. For a more in-depth treatment of GP covariance functions, see Chapter 4 in Rasmussen and Williams (2006).

In addition to the scheme described in Section 2.2, we also implemented a sampler which uses the explicit integration of the ODE system. This sampler is based on the same population MCMC setup as above, but samples from the distribution: $P(\mathbf{Y}, \boldsymbol{\theta}^*, \boldsymbol{\sigma}) = P(\mathbf{Y}|\boldsymbol{\theta}^*, \boldsymbol{\sigma}) P(\boldsymbol{\theta}^*) P(\boldsymbol{\sigma})$, where $\boldsymbol{\theta}^*$ is the parameter vector for the ODE system, augmented with the initial concentrations for each species, and $P(\boldsymbol{\theta}^*)$ and $P(\boldsymbol{\sigma})$ are the priors defined in Section 2.2. Then we have $P(\mathbf{Y}|\boldsymbol{\theta}^*, \boldsymbol{\sigma}) = \prod_k \prod_t P(y_k(t)|\boldsymbol{\theta}^*, \sigma_k)$, with $P(y_k(t)|\boldsymbol{\theta}^*, \sigma_k) = \mathcal{N}(y_k(t)|x_k(t, \boldsymbol{\theta}^*), \sigma_k^2)$ where $x_k(t, \boldsymbol{\theta}^*)$ is the solution of the ODE system for species k at time t, given $\boldsymbol{\theta}^*$. Parameters corresponding to the initial concentrations are initialised using the observed concentrations at time $t = 0$ for each species; however these are only starting values, and the actual initial concentrations need to be sampled from the joint distribution as part of the MCMC.

4 BENCHMARK ODE SYSTEMS

In this section, we present three small-to-medium-sized ODE models of biological systems that we will use to benchmark the parameter inference methods.

The *PIF4/5* model. We apply our GP parameter inference method to a model for gene regulation of genes *PIF4* and *PIF5* by *TOC1* in the circadian clock gene regulatory network of *Arabidopsis thaliana*. The overall network is represented by the Locke 2-loop model (Locke et al., 2005), with fixed parameters that were originally inferred following Pokhilko et al. (2010). Only the parameters involved in regulation of *PIF4* and *PIF5* are inferred from the data using the methods described in this paper. We simplify the model to represent genes *PIF4* and *PIF5* as a combined gene *PIF4/5*. We are interested in the promoter strength s, the rate constant K_d and Hill coefficient h of the regulation by $TOC1$, and the degradation rate d of the $PIF4/5$ mRNA. The regulation process is represented by the following ODE:

$$\frac{d[PIF4/5]}{dt} = s \cdot \frac{K_d^h}{K_d^h + [TOC1]^h} - d \cdot [PIF4/5] \quad (23)$$

where $[PIF4/5]$ and $[TOC1]$ represent the concentration of *PIF4/5* and *TOC1*, respectively.

For the experiments presented here, data were generated with parameters $\{s = 1, K_d = 0.46, h = 2, d = 1\}$, which generates concentrations that are close to real-life measurements of *PIF4/5*. For each dataset, we simulated data over the interval $[0, 24]$ with sampling intervals in $\{2, 4\}$. We use the *PIF4/5* concentration from a measurement of Arabidopsis gene expressions at the beginning of the day (0.386) as the concentration at time t=0 which is used to generate the data.

The Lotka-Volterra model. The Lotka-Volterra model is a 2-equation system that was originally developed for modelling predator-prey interaction in ecology (Lotka, 1932). There are two species, a prey species S (the 'sheep') and a predator species W (the 'wolves'). The dynamics of their interactions are described by a system of two ODEs, $\frac{d[S]}{dt} = [S] \cdot (\alpha - \beta \cdot [W])$ and $\frac{d[W]}{dt} = -[W] \cdot (\gamma - \delta \cdot [S])$. This system is of interest because it exhibits periodicity, and there are non-linear interactions between the two species.

For the experiments presented here, data were generated with parameters $\{\alpha = 2, \beta = 1, \gamma = 4, \delta = 1\}$, which generates stable oscillations. For each dataset, we simulated data over the interval $[0, 2]$ with sampling intervals of 0.25. The initial values for the prey species S and the predator species W were set at $[S] = 5$ and $[W] = 3$ to generate the data.

The signal transduction cascade. Our third and final model is a model of a signal transduction cascade that was described in Vyshemirsky and Girolami (2008) (Model 1). At the top of the cascade we have protein S, which can degrade into S_d. S activates protein R into active state Rpp by first binding to it to form RS, which is then activated to turn into Rpp. Rpp can degrade back into R, and RS can separate back into S and R. The model is described by the following system of five ODEs:

$$\begin{aligned}
\frac{d[S]}{dt} &= -k_1 \cdot [S] - k_2 \cdot [S] \cdot [R] + k_3 \cdot [RS] \\
\frac{d[S_d]}{dt} &= k_1 \cdot [S] \\
\frac{d[R]}{dt} &= -k_2 \cdot [S] \cdot [R] + k_3 \cdot [RS] + \frac{V \cdot [Rpp]}{Km + [Rpp]} \\
\frac{d[RS]}{dt} &= k_2 \cdot [S] \cdot [R] - k_3 \cdot [RS] - k_4 \cdot [RS] \\
\frac{d[Rpp]}{dt} &= k_4 \cdot [RS] - \frac{V \cdot [Rpp]}{Km + [Rpp]} \quad (24)
\end{aligned}$$

This system is of interest as it represents a realistic formulation of signal transduction as an ODE system, using mass action and Michaelis-Menten kinetics.

For the experiments presented here, data were generated with parameters $\{k_1 = 0.07, k_2 = 0.6, k_3 = 0.05, k_4 = 0.3, V = 0.017, Km = 0.3\}$, following Vyshemirsky and Girolami (2008). For each dataset, we simulated data over the interval $[0, 100]$ and took samples at time points $\{0, 1, 2, 4, 5, 7, 10, 15, 20, 30, 40, 50, 60, 80, 100\}$. This means that we sampled more timepoints during the earlier part of the timeseries, where the dynamics tend to be faster. We also followed Vyshemirsky and Girolami (2008) in setting the initial values for generating the timecourses of the 5 species: $\{[S] = 1, [S_d] = 0, [R] = 1, [RS] = 0, [Rpp] = 0\}$.

5 PARAMETER INFERENCE RESULTS

We use the three benchmark systems described in Section 4 to analyse the performance of our adaptive gradient matching, and to provide a thorough comparison with both the method in Calderhead et al. (2008), and the sampler which explicitly solves the ODE system, as described in Section 3.[3] We generated data from each system using the R package `deSolve` (Soetaert et al., 2010) for numerically integrating the systems of differential equations. See Section 4 for the parameter and initial concentration settings. We then added white Gaussian observation noise to the datasets. For the *PIF4/5* system and the signal transduction cascade, we added noise with standard deviation $\in \{0, 0.1\}$, and for the Lotka-Volterra system we added noise with

[3]Note that due to the higher computational cost involved, we could only apply the explicit ODE integration method to the Lotka-Volterra model and the signal transduction cascade. Applying it to the *PIF4/5* system would have required solving the entire 14-equation system of the Locke 2-loop model at each step, which was not feasible with the time and resources at our disposal.

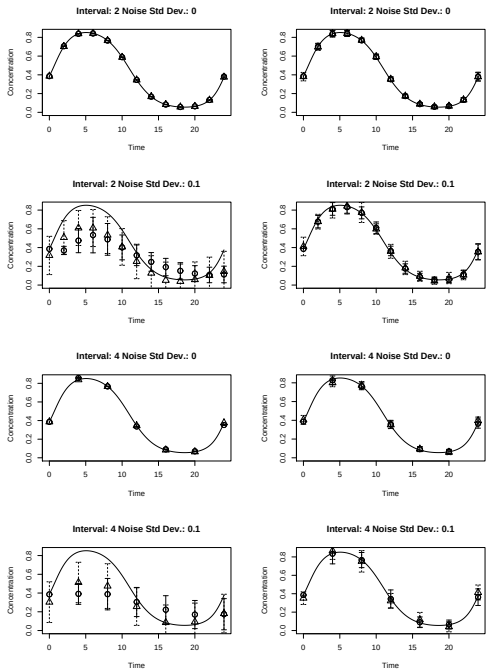

Figure 1: PIF4/5 expression levels with varying sampling intervals and noise. We show the true (noiseless) expression values, the sampled latent variables (triangles) and the expression profile simulated from the ODE using the sampled $\boldsymbol{\theta}$ values (circles). Error bars show one standard deviation. Left: Calderhead et al. model. Right: Adaptive gradient matching.

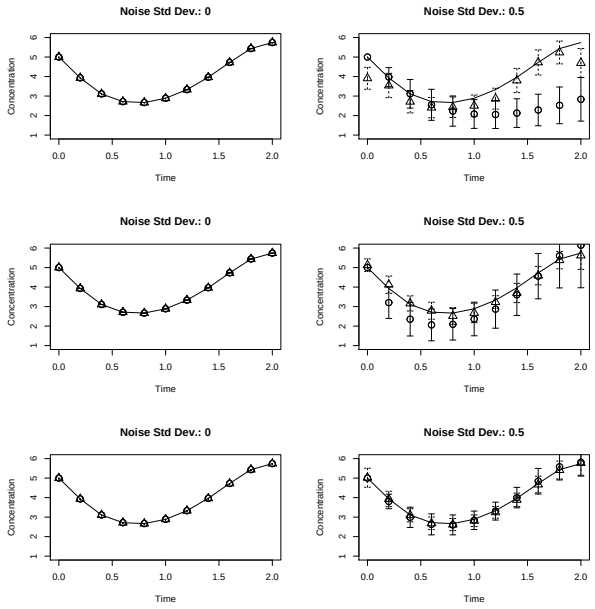

Figure 2: Lotka-Volterra concentrations for the prey species with varying observational noise. We show the true (noiseless) expression values, the sampled latent variables (triangles) and the expression profile simulated from the ODE using the sampled $\boldsymbol{\theta}$ values (circles). Error bars show one standard deviation. Top: Calderhead et al. model. Middle: Adaptive gradient matching. Bottom: Explicit ODE Integration.

standard deviation $\in \{0, 0.5\}$. The higher noise level for the Lotka-Volterra system reflects the higher amplitude of the signal in this system.

We generated 10 datasets for each noise level and system. Convergence was monitored via diagnostic plots and the potential scale reduction factor (PSRF) (Gelman and Rubin, 1992). A PSRF < 1.1 for all ODE parameters in $\boldsymbol{\theta}$ was taken as an indication of sufficient convergence. We collected 1000 samples at intervals of 100 steps from the converged chains. Samples from all 10 independent datasets were pooled to obtain the final predictions. Note that we were unable to obtain a PSRF < 1.1 for the Calderhead et al. model in the presence of non-zero Gaussian observation noise; in this case, we resorted to running the MCMC chains for 200,000 steps, which corresponds to roughly twice the number of steps that it took adaptive gradient matching to reach convergence, before taking samples as described above.

Figure 1 shows the results for the *PIF4/5* system. The data used for the parameter inference was sampled at intervals 2 and 4 timesteps, where 4 is a realistic sampling interval for actual measurements. We compare the method in Calderhead et al. (2008) with our adaptive gradient matching technique. We see that when there is no noise, the two methods perform equally well, but as soon as we introduce noise into the system, the predictions by the Calderhead et al. method become unreliable due to non-convergence.

Figures 2 show the results for the prey species in the Lotka-Volterra system. Results for the predator species were similar, and can be found in the supplementary material. The data used for the parameter inference was sampled at intervals of 0.25 timesteps. Again the method by Calderhead et al. showed good performance in the noiseless case, but a deteriorated performance in the presence of noise. For noise level 0, adaptive gradient matching performed as well as the explicit ODE integration, and for the high noise level of 0.5, the performance of adaptive gradient matching is still competitive.

Finally, Figure 3 shows the results for the signal transduction cascade. Figure 3 only shows the predictions for *Rpp*, which represents the activated protein complex, and is arguably the central species in this system. Predictions for the other species can be found in the supplementary material. Figure 3 also includes

boxplots for the sampled parameters. For the last two parameters, we present the ratio V/Km, as this is the crucial quantity that determines reconstruction accuracy. Once again, our adaptive gradient matching performs well, and remains competitive with explicit ODE integration in the presence of noise. Note that even though the ratio V/Km is overestimated by our method for noise level 0.1, the sampled parameters still result in a good fit to the observed data.

6 SPEED AND COMPUTATIONAL COMPLEXITY

In Calderhead et al. (2008), the authors demonstrate that the moves of their sampler scale with $\mathrm{O}(NT^3)$, due to the requirement of inverting a $T\times T$ data matrix N times (where T is the length of the input time series and N is the number of species in the system). We can make a similar argument for adaptive gradient matching. The dominant computational cost for each sampling step comes from Equation (20), which requires inverting two $T\times T$ data matrices. Thus the complexity of each sampling step is $\mathrm{O}(2NT^3)=\mathrm{O}(NT^3)$ when the sampling is done for all N species[4]. Hence each MCMC move using adaptive gradient matching has the same computational complexity as a move in Calderhead et al. (2008).

What will matter most in practice is how long each method takes to converge. Although it is difficult to prove convergence, we can get an indication by using the potential scale reduction factor (PSRF) as a convergence diagnostic, as described in Section 5. For convenience, we will refer to an MCMC run as converged if the PSRF is ≤ 1.1 for all parameters in $\boldsymbol{\theta}$. Figure 4 compares the explicit ODE integration with the model by Calderhead et al. (2008), and with adaptive gradient matching in terms of computational time for 10^5 iterations (in seconds)[5] and number of MCMC iterations before reaching convergence. We used the signal transduction cascade described in Section 4 as the test model. Each method was run 10 times using 10 different data instantiations (adding Gaussian observation noise with standard deviation 0.1). We see that, as expected, adaptive gradient matching and the method in Calderhead et al. (2008) are both faster than explicit ODE integration for a fixed number of iterations. Furthermore, adaptive gradient matching is only marginally slower than the method in Calderhead et al. (2008). We see that the method in Calderhead et al. (2008) does not converge for any of the runs, confirming our observation from Section 5. Adaptive gradient matching, on the other hand, converges in fewer iterations than explicit ODE integration. This can be explained by the difference in the dimensionality of the parameter space; as we have pointed out in Section 3, to integrate the ODE system, we also need to infer the initial concentrations for each species, in effect increasing the number of parameters. Adaptive gradient matching avoids having to infer the initial concentrations by effectively profiling over them, which, along with the treatment of latent variables $\mathbf{X}$ as ancillary variables (see Section 3), leads to fast convergence.

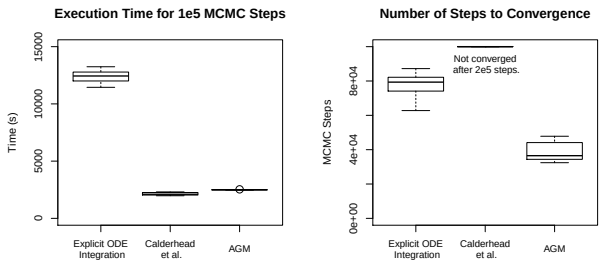

Figure 4: Computational efficiency of the different methods: Explicit ODE integration, Calderhead et al. (2008) and adaptive gradient matching (AGM). We use parameter inference for the signal transduction model as a test case. Left: Time taken for 10^5 MCMC iterations. Right: Number of MCMC iterations to convergence (PSRF $\leq$ 1.1). Note that Calderhead et al. (2008) did not achieve PSRF $\leq$ 1.1 in any of the runs. The horizontal bar of the boxplots shows the median, the box margins show the 25th and 75th percentiles, the whiskers indicate data within 2 times the interquartile range, and circles are outliers.

7 DISCUSSION

We have described an adaptive gradient matching approach for parameter inference in ODE systems based on Calderhead et al. (2008). Adaptive gradient matching avoids the need for explicitly solving the ODE system at each MCMC sampling step, which significantly reduces the computational burden. In the method of Calderhead et al., an adaptation of the ODE parameters has no influence on the inference of the GP hyperparameters. This corresponds to a unidirectional information flow from GP interpolation to parameter inference in the system of ODEs. We have developed a methodological improvement that infers both GP hyperparameters and ODE parameters jointly from the posterior distribution, and where due to conditional dependence between both groups, the latter may exert an influence on the former. This closes the inference

[4]Note that in practice the inverted matrices can be cached, so we only have to invert both matrices for MCMC moves that change the GP hyperparameters. Therefore we should not expect the computational costs to be double those of Calderhead et al. (2008).

[5]Although the simulations were run on the same machine, there may be implementation-dependent speed differences between Calderhead et al. (2008) (implemented in MATLAB) and AGM (implemented in R).

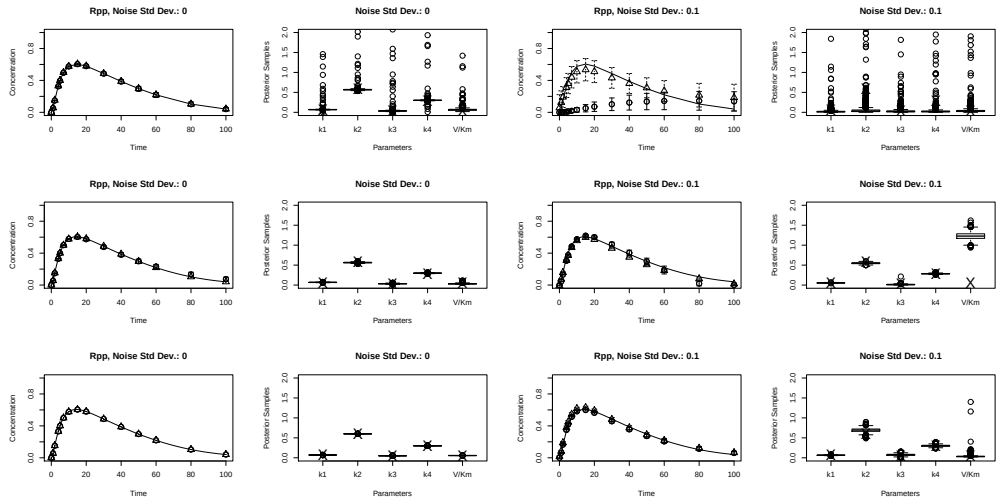

Figure 3: Expression levels of activated protein complex Rpp in the signal transduction pathway, with varying observational noise. Expression levels for other species in the system can be found in the supplementary material. We show the true (noiseless) expression values, the sampled latent variables (triangles) and the expression profile simulated from the ODE using the sampled $\boldsymbol{\theta}$ values (circles). Error bars show one standard deviation. The boxplots give an idea of the distribution of the sampled parameters, where the true parameter value is marked with an X. The horizontal bar shows the median, the box margins show the 25th and 75th percentiles, the whiskers indicate data within 2 times the interquartile range, and circles are outliers. Top Row: Calderhead et al. model. Middle Row: Adaptive gradient matching. Bottom Row: Explicit ODE integration.

procedure by effectively introducing an important information feedback loop from the ODE system back to the GP interpolation.

We have applied adaptive gradient matching to three model systems from ecology and systems biology, and have demonstrated that our method improves on Calderhead et al. (2008) and performs on a par with a sampler which explicitly solves the ODE system at each step. Regarding computational complexity, our method is marginally slower than the method of Calderhead et al. (2008) in terms of CPU time per iteration due to the fact that two matrix inversions rather than one are needed to calculate equation (20). However, both methods have the same asymptotic complexity of $O(NT^3)$, and caching techniques reduce the practical difference to much less than a factor of two (see Figure 4, left panel). Regarding the efficiency of the MCMC sampler, we found that the method of Calderhead et al. (2008) often fails to converge for non-zero noise variance, and that our new sampling approach substantially improves convergence and mixing (see Figure 4, right panel). In particular, our method improves both execution time (CPU time per iteration) and MCMC convergence (number of iterations) over explicit ODE integration. The former improvement, which is due to the gradient matching approach, was found to lead to an acceleration by a whole order of magnitude. In general, the improvement will depend on the size and stiffness of the ODE system. The latter improvement results from the fact that gradient matching does not require knowledge or inference of the initial conditions, which reduces the dimension of the parameter space by effectively profiling over the corresponding subdomain.

A close relative of our work is the recent method of functional tempering (Campbell and Steele, 2012), which is based on the same gradient matching paradigm as our approach, but uses B-splines instead of Gaussian processes for data interpolation. Their approach has one vector of regularization parameters, which corresponds to our hyperparameter vector $\boldsymbol{\gamma}$ and penalizes the mismatch between the gradients. Our model additionally profits from the hyperparameters of the Gaussian process, $\boldsymbol{\phi}$, which define the flexibility of the interpolant and are automatically inferred from the data, while in the model of Campbell and Steele (2012) this flexibility is defined by the B-splines basis and has to be set in advance. An interesting difference is the tempering scheme of Campbell and Steele (2012), which applied to our model corresponds to gradually forcing $\boldsymbol{\gamma}$ to zero rather than inferring it from the posterior distribution. A comparative evaluation is the subject of our future research.

Acknowledgements This work was partially funded by EPSRC and the EU (FP7 project "Timet").

References

Ashyraliyev, M., Fomekong-Nanfack, Y., Kaandorp, J., and Blom, J. (2009). Systems biology: parameter estimation for biochemical models. *FEBS Journal*, 276(4):886–902.

Calderhead, B., Girolami, M. A., and Lawrence, N. D. (2008). Accelerating Bayesian inference over nonlinear differential equations with Gaussian processes. In *Advances in Neural Information Processing Systems (NIPS)*, volume 22.

Campbell, D. and Steele, R. (2012). Smooth functional tempering for nonlinear differential equation models. *Statistics and Computing*, pages 1–15.

De Jong, H. (2002). Modeling and simulation of genetic regulatory systems: a literature review. *Journal of Computational Biology*, 9(1):67–103.

Gelman, A. (1997). Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1):110–120.

Gelman, A. and Rubin, D. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472.

Girolami, M. and Calderhead, B. (2011). Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.

Jasra, A., Stephens, D., and Holmes, C. (2007). On population-based simulation for static inference. *Statistics and Computing*, 17(3):263–279.

Locke, J., Southern, M., Kozma-Bognar, L., Hibberd, V., Brown, P., Turner, M., and Millar, A. (2005). Extension of a genetic network model by iterative experimentation and mathematical analysis. *Molecular Systems Biology*, 1:(online).

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):461–469.

Murray, I. and Adams, R. (2010). Slice sampling covariance hyperparameters of latent Gaussian models. In *Advances in Neural Information Processing Systems (NIPS)*, volume 23.

Pokhilko, A., Hodge, S., Stratford, K., Knox, K., Edwards, K., Thomson, A., Mizuno, T., and Millar, A. (2010). Data assimilation constrains new connections and components in a complex, eukaryotic circadian clock model. *Molecular Systems Biology*, 6(1).

Poyton, A., Varziri, M., McAuley, K., McLellan, P., and Ramsay, J. (2006). Parameter estimation in continuous-time dynamic models using principal differential analysis. *Computers & Chemical Engineering*, 30(4):698–708.

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

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

Soetaert, K., Petzoldt, T., and Setzer, R. (2010). Solving differential equations in R: Package deSolve. *Journal of Statistical Software*, 33(9):1–25.

Varah, J. (1982). A spline least squares method for numerical parameter estimation in differential equations. *SIAM Journal on Scientific and Statistical Computing*, 3:28.

Vyshemirsky, V. and Girolami, M. (2008). Bayesian ranking of biochemical system models. *Bioinformatics*, 24(6):833–839.

A SUPPLEMENTARY MATERIAL

Below, we present additional details and results that could not fit into the main paper: the explicit expressions of the cross-covariance matrices from Section 2.1, a comparison of the regression fits for the two alternative Gaussian process covariance functions, as well as further results for parameter inference comparison between adaptive gradient matching, the method from Calderhead et al. (2008), and explicit ODE integration.

A.1 Cross-Covariance Matrices

Below, we present the explicit expressions for the cross-covariance matrices. For a derivation of these results, see Rasmussen and Williams (2006). We obtain that:

$$\mathbf{C}'_{\phi_\mathbf{k}}(i,j) = \frac{d\mathcal{K}_{\phi_k}(t_i,t_j)}{dt_i} \quad (25)$$

$${}'\mathbf{C}_{\phi_\mathbf{k}}(i,j) = \frac{d\mathcal{K}_{\phi_k}(t_i,t_j)}{dt_j} \quad (26)$$

$$\mathbf{C}''_{\phi_\mathbf{k}}(i,j) = \frac{d^2\mathcal{K}_{\phi_k}(t_i,t_j)}{dt_i dt_j} \quad (27)$$

where $\mathcal{K}_{\phi_k}(t_i,t_j)$ is the chosen covariance function for the Gaussian process. For the RBF covariance function, we obtain:

$$\frac{d\mathcal{K}^{rbf}_{\phi_k}(t,t')}{dt} = -\frac{(t-t')}{l^2}\mathcal{K}^{rbf}_{\phi_k}(t,t') \quad (28)$$

$$\frac{d\mathcal{K}^{rbf}_{\phi_k}(t,t')}{dt'} = \frac{(t-t')}{l^2}\mathcal{K}^{rbf}_{\phi_k}(t,t') \quad (29)$$

$$\frac{d^2\mathcal{K}^{rbf}_{\phi_k}(t,t')}{dtdt'} = \left(\frac{1}{l^2} - \frac{(t-t')^2}{l^4}\right)\mathcal{K}^{rbf}_{\phi_k}(t,t') \quad (30)$$

For the sigmoid covariance function, we obtain:

$$\frac{d\mathcal{K}^{sig}_{\phi_k}(t,t')}{dt} = \frac{\sigma^2_{sig}}{\sqrt{1-Z^2}}\frac{dZ}{dt} \quad (31)$$

$$\frac{d\mathcal{K}^{sig}_{\phi_k}(t,t')}{dt'} = \frac{\sigma^2_{sig}}{\sqrt{1-Z^2}}\frac{dZ}{dt'}$$

$$\frac{d^2\mathcal{K}^{sig}_{\phi_k}(t,t')}{dtdt'} = \frac{\sigma^2_{sig}}{\sqrt{1-Z^2}} \times \quad (32)$$

$$\left(\frac{Z}{1-Z^2}\frac{dZ}{dt'}\frac{dZ}{dt} + \frac{d^2Z}{dtdt'}\right) \quad (33)$$

where:

$$Z = \frac{a+b*t*t'}{Z_{norm}} \quad (34)$$

with $Z_{norm} = \sqrt{(a+b*t*t+1)(a+b*t'*t'+1)}$, and we have:

$$\frac{dZ}{dt} = b\left(\frac{t'}{Z_{norm}} - \frac{tZ}{a+b*t*t+1}\right) \quad (35)$$

$$\frac{dZ}{dt'} = b\left(\frac{t}{Z_{norm}} - \frac{t'Z}{a+b*t'*t'+1}\right) \quad (36)$$

$$\frac{d^2Z}{dtdt'} = b\left(\frac{1}{Z_{norm}} - \frac{bt't'}{(a+b*t'*t'+1)Z_{norm}}\right) - \frac{bt}{(a+b*t*t+1)}\frac{dZ}{dt'} \quad (37)$$

A.2 GP Covariance Function Comparison

The two covariance functions used in this work are the RBF (radial basis function) covariance function $k(t,t') = \sigma^2_{kern}\exp(-0.5*(t-t')^2/l^2)$ with parameters σ^2_{kern} and l^2 (variance and characteristic lengthscale), and the sigmoid covariance function $k(t,t') = \sigma^2_{kern}\arcsin\left(\frac{a+b*t*t'}{\sqrt{(a+b*t*t+1)(a+b*t'*t'+1)}}\right)$ with parameters σ^2_{kern}, a and b. Figures 5 - 7 show a comparison of the GP regression fits (using maximum likelihood) to data from the different model systems. We see that the sigmoid covariance function always provides a good fit, while the RBF covariance function breaks down for some of the species in the signal transduction cascade. This is due to the fact that the RBF covariance function assumes stationarity, with a fixed lengthscale l^2. That assumption is not true for the signal transduction cascade. The sigmoid covariance function, on the other hand, is non-stationary and can deal with varying lengthscales. Note that the in the signal transduction example we applied added the Gaussian noise on log scale (in effect adding multiplicative noise), to avoid getting negative values for concentrations close to zero; this leads to slight distortion for the sigmoid covariance function as the noise model assumes additive Gaussian noise.

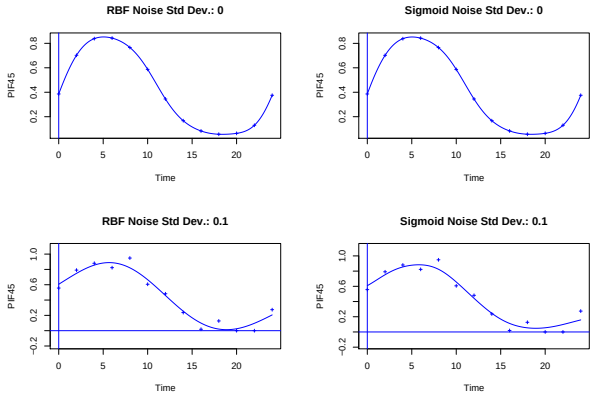

Figure 5: GP Regression fits to PIF4/5 expression levels, using the RBF and the sigmoidal covariance function. The crosses represent the data points, the solid line is the GP mean. Top Row: Gaussian noise with standard deviation 0. Bottom Row: Gaussian noise with standard deviation 0.1.

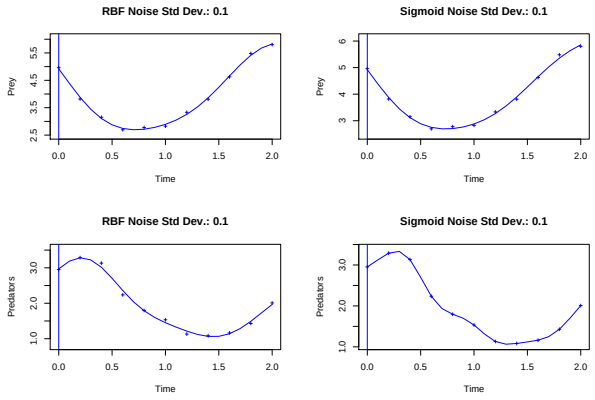

Figure 6: GP Regression fits to predator and prey concentrations in the Lotka-Volterra model, using the RBF and the sigmoidal covariance function. The crosses represent the data points, the solid line is the GP mean. Gaussian noise with standard deviation 0.1 was applied. Top Row: Prey species. Bottom Row: Predator Species.

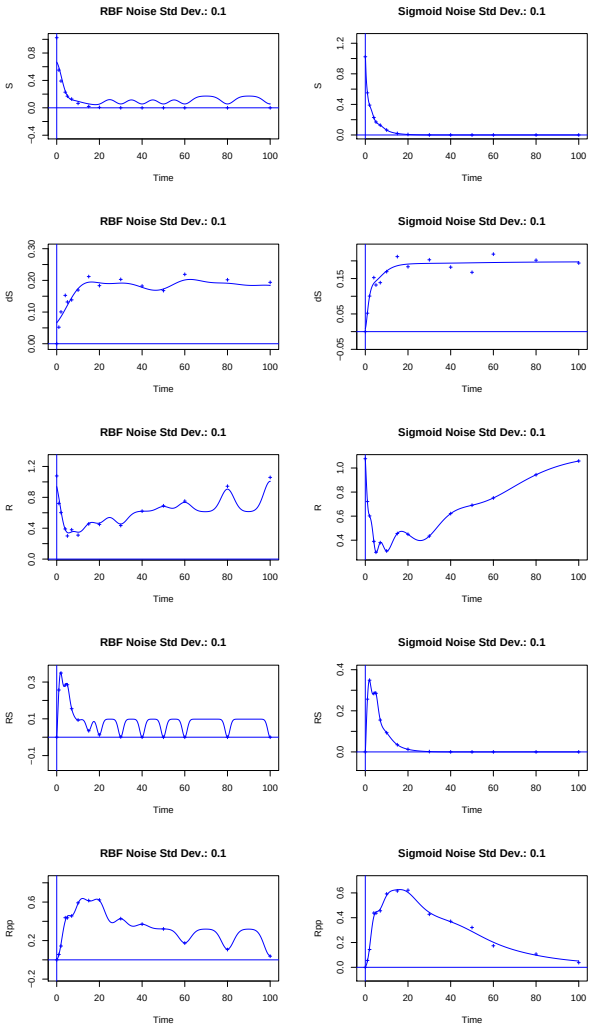

Figure 7: GP Regression fits to species concentrations in the signal transduction pathway, , using the RBF and the sigmoidal covariance function. The crosses represent the data points, the solid line is the GP mean. Gaussian noise with standard deviation 0.1 was applied. From top to bottom, the rows show species S, dS, R, RS and Rpp.

A.3 Additional Parameter Inference Results

We present some additional parameter inference results that we had to omit from the main paper due to space restriction. Figure 8 shows the results for the *PIF4/5* system with sampling interval 1. Figure 9 shows the results for the predator species in the Lotka-Volterra model. Figures 10 and 11 show the results for species S, S_d, R and RS in the signal transduction pathway, for Gaussian noise with standard deviation 0 and 0.1, respectively.

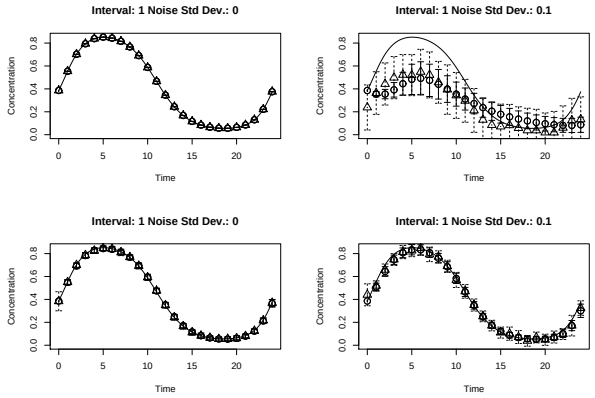

Figure 8: PIF4/5 expression levels with sampling interval 1 and varying observational noise. We show the true (noiseless) expression values, the sampled latent variables (triangles) and the expression profile simulated from the ODEs using the sampled $\boldsymbol{\theta}$ values (circles). Error bars show one standard deviation. Top Row: Calderhead et al. (2008) model. Bottom Row: Adaptive gradient matching.

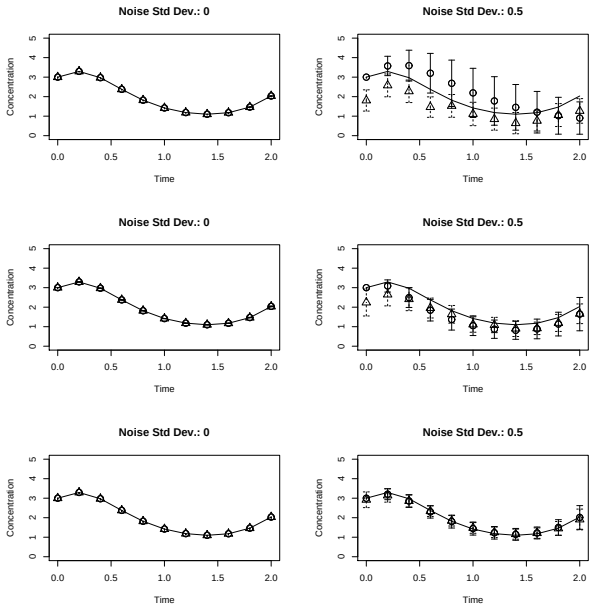

Figure 9: Lotka-Volterra concentrations for the predator species with varying observational noise. We show the true (noiseless) expression values, the sampled latent variables (triangles) and the expression profile simulated from the ODE using the sampled $\boldsymbol{\theta}$ values (circles). Error bars show one standard deviation. Top Row: Calderhead et al. (2008) model. Middle Row: Adaptive gradient matching. Bottom Row: Explicit ODE integration.

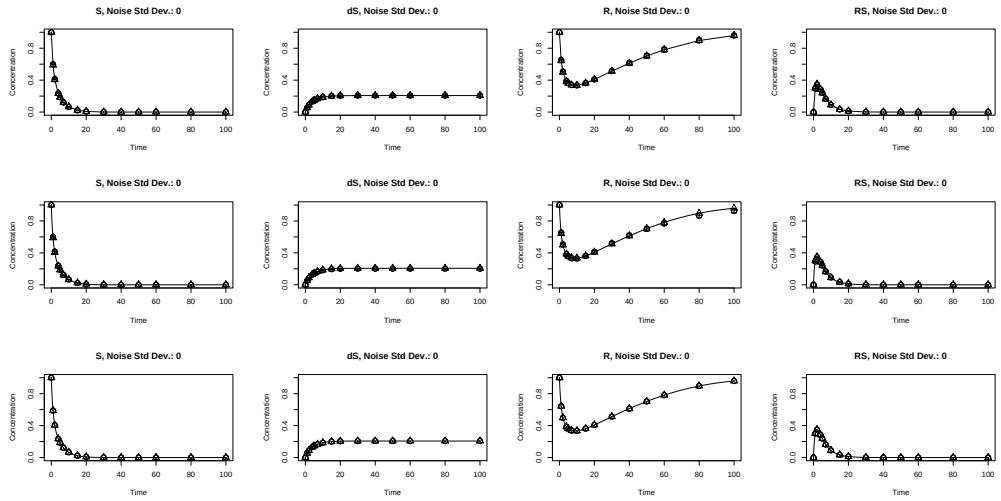

Figure 10: Expression levels of species other than Rpp in the signal transduction pathway, with no observational noise. We show the true (noiseless) expression values, the sampled latent variables (triangles) and the expression profile simulated from the ODE using the sampled $\boldsymbol{\theta}$ values (circles). Error bars show one standard deviation. Top Row: Calderhead et al. (2008) model. Middle Row: Adaptive gradient matching. Bottom Row: Explicit ODE integration.

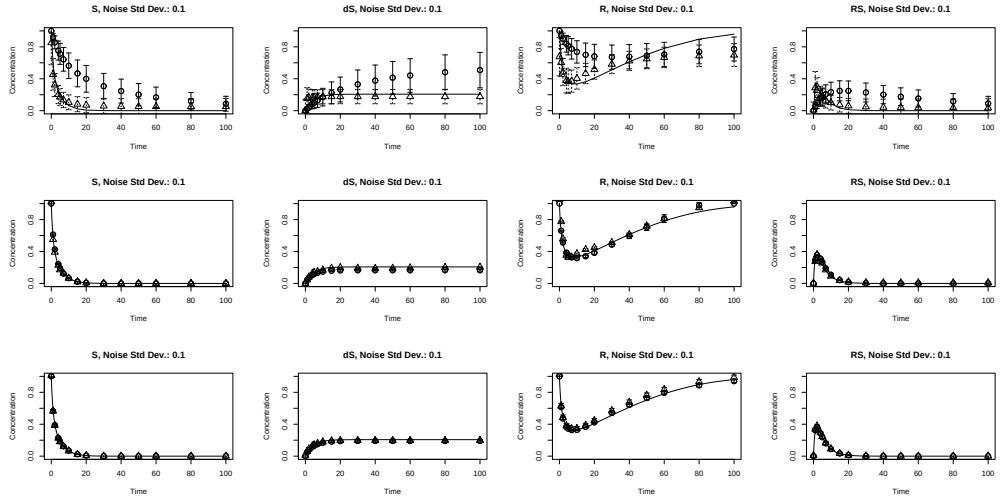

Figure 11: Expression levels of species other than Rpp in the signal transduction pathway, with observational noise with standard deviation 0.1. We show the true (noiseless) expression values, the sampled latent variables (triangles) and the expression profile simulated from the ODE using the sampled $\boldsymbol{\theta}$ values (circles). Error bars show one standard deviation. Top Row: Calderhead et al. (2008) model. Middle Row: Adaptive gradient matching. Bottom Row: Explicit ODE integration.