Controversy in mechanistic modelling with Gaussian processes

Benn Macdonald B.MACDONALD.1@RESEARCH.GLA.AC.UK
School of Mathematics & Statistics, University of Glasgow

Catherine Higham CATHERINE.HIGHAM@GLASGOW.AC.UK
School of Mathematics & Statistics, University of Glasgow

Dirk Husmeier DIRK.HUSMEIER@GLASGOW.AC.UK
School of Mathematics & Statistics, University of Glasgow

Abstract

Parameter inference in mechanistic models based on non-affine differential equations is computationally onerous, and various faster alternatives based on gradient matching have been proposed. A particularly promising approach is based on nonparametric Bayesian modelling with Gaussian processes, which exploits the fact that a Gaussian process is closed under differentiation. However, two alternative paradigms have been proposed. The first paradigm, proposed at NIPS 2008 and AISTATS 2013, is based on a product of experts approach and a marginalization over the derivatives of the state variables. The second paradigm, proposed at ICML 2014, is based on a probabilistic generative model and a marginalization over the state variables. The claim has been made that this leads to better inference results. In the present article, we offer a new interpretation of the second paradigm, which highlights the underlying assumptions, approximations and limitations. In particular, we show that the second paradigm suffers from an intrinsic identifiability problem, which the first paradigm is not affected by.

1. Introduction

Many processes in science and engineering can be described by dynamical systems models based on ordinary differential equations (ODEs). Examples range from simple models of predator-prey interactions in ecosystems (Lotka, 1932) or activation/deactivation dynamics of spiking neurons (Nagumo et al., 1962) to increasingly complex mathematical descriptions of biopathways that aim to predict the time-varying concentrations of different molecular species, like mRNAs and proteins, inside the living cell (Pokhilko et al., 2012). ODEs are typically constructed from well understood scientific principles and include clearly interpretable parameters that define the kinetics of the processes and the interactions between the species. However, these parameters are often unknown and not directly measurable. In principle, the task of statistically inferring them from data is not different from statistical inference in more conventional models. For given initial concentrations and under fairly mild regularity conditions, the solution of the ODEs is uniquely defined; hence, the kinetic parameters could be inferred e.g. by minimizing the mismatch between the data and the ODE solutions in a maximum likelihood sense. In practice, a closed-form solution for non-linear ODEs usually does not exist. Any variation of the kinetic parameters thus requires a numerical integration of the ODEs, which is computationally expensive and imposes severe limitations on the number of parameter adaptation steps that are practically feasible.

To circumvent the high computational complexity of numerically integrating the ODEs, several authors have explored approximate inference based on gradient matching. The details vary from method to method, but they all have in common the combination of a *data interpolation (DI)* and a *parameter adaptation (PA)* step. In the DI step, an established statistical model or procedure is applied to obtain a set of smooth interpolants from noisy (measured or observed) concentration time series (for each species). In the PA step, the time derivatives obtained from the time-varying slopes of the tangents to the interpolants are compared with the time derivatives predicted by the ODEs, and the kinetic parameters are adjusted so as to minimize some measure of mismatch. More advanced methods (see overleaf) allow the ODEs to regularize the interpolation, and

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

the two steps are thus interconnected and are iterated until some convergence criterion is met. The reduction of the computational complexity, compared to the direct approach, results from the fact that the ODEs never have to be solved explicitly, and the typically unknown initial conditions are effectively profiled over. Representative examples of this paradigm are the papers by (Ramsay et al., 2007) and (Liang & Wu, 2008) (using P-splines for interpolation), (González et al., 2013) (proposing an approach based on reproducing kernel Hilbert spaces), and (Campbell & Steele, 2012) (exploring inference with parallel tempering).

The present paper focuses on a particular approach to gradient matching based on nonparametric Bayesian modelling with Gaussian processes (GPs). The key insight, first discussed in (Solak et al., 2003) and (Graepel, 2003), and more recently exploited in (Holsclaw et al., 2013), is that for a differentiable kernel, the time derivative of a GP is also a GP. Hence a GP in data space imposes a conjugate GP in derivative space and thereby provides a natural framework for gradient matching. This idea has been exploited in recent high-profile publications, like (Babtie et al., 2014). The limitation of (Babtie et al., 2014) is that the interpolant obtained from the GP is kept fixed, and all subsequent inference critically depends on how accurately this initial interpolant matches the unknown true process. The implication is that the noise tolerance is typically low, as seen e.g. from Figure 4A in (Babtie et al., 2014), and that reliable inference requires tight prior constraints on the ODE parameters; see p.2 of the supplementary material in (Babtie et al., 2014). To improve the robustness of inference, more advanced methods aim to regularize the GP by the ODEs themselves. Two alternative conceptual approaches to this end have been proposed in the recent machine learning literature. The first paradigm, originally published in (Calderhead et al., 2008) and more recently extended in (Dondelinger et al., 2013), where it was called AGM (for 'adaptive gradient matching'), is based on a product-of-experts approach and a marginalization over the derivatives of the state variables. A competing approach, proposed in (Wang & Barber, 2014) and called GPODE by the authors, formulates gradient matching with GPs in terms of a probabilistic generative model by marginalizing over the state variables and conditioning on the state derivatives. (Wang & Barber, 2014) claim that their proposed paradigm shift achieves an improvement over the first paradigm in three respects: model simplification, tractable inference, and better predictions.

In the present paper, we offer an alternative interpretation of the GPODE model, which leads to deeper insight into intrinsic approximations that were not apparent from the original publication. We discuss that the GPODE model suffers from an inherent identifiability problem, which models of the first paradigm are not affected by. We complement our theoretical analysis with empirical demonstrations on simulated data, using the same model systems as in the original publications, (Wang & Barber, 2014) and (Dondelinger et al., 2013).

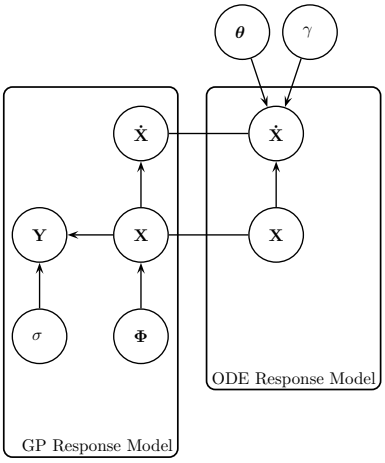

Figure 1. Gradient matching with Gaussian processes, as proposed in (Calderhead et al., 2008) and (Dondelinger et al., 2013).

2. Paradigm A: the AGM model

We start by summarizing the AGM model of (Dondelinger et al., 2013), which is an extension of the model proposed in (Calderhead et al., 2008). Consider a continuous-time dynamical system in which the evolution of K states or 'species' $\mathbf{x}(t) = [x_1(t), x_2(t), ..., x_K(t)]^\mathsf{T}$ is represented by a set of K ordinary differential equations (ODEs) with parameter vector $\boldsymbol{\theta}$ and initial conditions $\mathbf{x}(0)$

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

We are typically interested in non-affine systems, for which $\mathbf{f}$ is nonlinear and a closed-form solution does not exist. We assume that we have noisy observations of the state variable $\mathbf{x}$ for N time points $t_1 < \ldots < t_N$:

$$\mathbf{y}(t) = \mathbf{x}(t) + \boldsymbol{\epsilon}(t). \quad (2)$$

For simplicity we assume the additive noise $\boldsymbol{\epsilon}(t)$ to follow a Normal distribution, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{D})$, with diagonal covariance matrix, $D_{ik} = \sigma_k^2 \delta_{ik}$. For notational convenience we introduce the K-by-N matrices

$$\mathbf{X} = [\mathbf{x}(t_1), \ldots, \mathbf{x}(t_N)] = [\mathbf{x}_1, \ldots, \mathbf{x}_K]^\mathsf{T} \quad (3)$$
$$\mathbf{Y} = [\mathbf{y}(t_1), \ldots, \mathbf{y}(t_N)] = [\mathbf{y}_1, \ldots, \mathbf{y}_K]^\mathsf{T} \quad (4)$$

where $\mathbf{x}_k = [x_k(t_1), \ldots, x_k(t_N)]^\mathsf{T}$ is the k^{th} state sequence, and $\mathbf{y}_k = [y_k(t_1), \ldots, y_k(t_N)]^\mathsf{T}$ are the corresponding noisy observations. Equation (2) can then be

rewritten as

$$P(\mathbf{Y}|\mathbf{X},\boldsymbol{\sigma}) = \prod_k \prod_t P(y_k(t)|x_k(t),\sigma_k) = \prod_k \prod_t \mathcal{N}(y_k(t)|x_k(t),\sigma_k^2). \quad (5)$$

Given that inference based on an explicit numerical solution of the differential equations tends to incur high computational costs, (Calderhead et al., 2008) proposed an alternative approach based on non-parametric Bayesian modelling with Gaussian processes. The idea is to put a Gaussian process prior on $\mathbf{x}_k$,

$$p(\mathbf{x}_k|\boldsymbol{\mu}_k,\boldsymbol{\phi}_k) = \mathcal{N}(\mathbf{x}_k|\boldsymbol{\mu}_k,\mathbf{C}_{\phi_\mathbf{k}}) \quad (6)$$

where $\mathbf{C}_{\phi_\mathbf{k}}$ denotes the covariance matrix, which is defined by some kernel with hyperparameters $\boldsymbol{\phi}_k$. In generalization of the expressions in (Calderhead et al., 2008) and (Dondelinger et al., 2013) we here explicitly include a potentially non-zero mean, $\boldsymbol{\mu}_k$, to allow for the fact that in many applications the state variables are non-negative (e.g. species concentrations). Since differentiation is a linear operation, a Gaussian process is closed under differentiation, and the joint distribution of the state variables $\mathbf{x}_k$ and their time derivatives $\dot{\mathbf{x}}_k$ is multivariate Gaussian with mean vector $(\boldsymbol{\mu}_k, \mathbf{0})^\mathsf{T}$ and covariance functions

$$cov[x_k(t), x_k(t')] = C_{\phi_k}(t,t') \quad (7)$$

$$cov[\dot{x}_k(t), x_k(t')] = \frac{\partial C_{\phi_k}(t,t')}{\partial t} := C'_{\phi_k}(t,t') \quad (8)$$

$$cov[x_k(t), \dot{x}_k(t')] = \frac{\partial C_{\phi_k}(t,t')}{\partial t'} := {}'C_{\phi_k}(t,t') \quad (9)$$

$$cov[\dot{x}_k(t), \dot{x}_k(t')] = \frac{\partial^2 C_{\phi_k}(t,t')}{\partial t \partial t'} := C''_{\phi_k}(t,t') \quad (10)$$

where $C_{\phi_k}(t,t')$ are the elements of the covariance matrix $\mathbf{C}_{\phi_\mathbf{k}}$ (Rasmussen & Williams, 2006). We introduce the definitions of the auto-covariance matrix of the kth state derivatives $\mathbf{C}''_{\phi_\mathbf{k}}$, which contains the elements defined in (10), and the cross-covariance matrices between the kth state and its derivatives, $\mathbf{C}'_{\phi_\mathbf{k}}$ and ${}'\mathbf{C}_{\phi_\mathbf{k}}$, which contain the elements defined in (8) and (9), respectively. From elementary transformations of Gaussian distributions, listed e.g. on p. 87 in (Bishop, 2006), the conditional distribution of the state derivatives is given by

$$p(\dot{\mathbf{x}}_k|\mathbf{x}_k,\boldsymbol{\phi}) = \mathcal{N}(\mathbf{m}_k,\mathbf{A}_k) \quad (11)$$

where

$$\mathbf{m}_k = {}'\mathbf{C}_{\phi_\mathbf{k}}{\mathbf{C}_{\phi_\mathbf{k}}}^{-1}(\mathbf{x}_k - \boldsymbol{\mu}_k); \mathbf{A}_k = \mathbf{C}''_{\phi_\mathbf{k}} - {}'\mathbf{C}_{\phi_\mathbf{k}}{\mathbf{C}_{\phi_\mathbf{k}}}^{-1}\mathbf{C}'_{\phi_\mathbf{k}}. \quad (12)$$

Assuming additive Gaussian noise with a state-specific error variance γ_k, one gets from (1):

$$p(\dot{\mathbf{x}}_k|\mathbf{X},\boldsymbol{\theta},\gamma_k) = \mathcal{N}(\mathbf{f}_k(\mathbf{X},\boldsymbol{\theta}),\gamma_k\mathbf{I}). \quad (13)$$

(Calderhead et al., 2008) and (Dondelinger et al., 2013) combine (11) and (13) with a product of experts approach:

$$p(\dot{\mathbf{x}}_k|\mathbf{X},\boldsymbol{\theta},\boldsymbol{\phi},\gamma_k) \propto p(\dot{\mathbf{x}}_k|\mathbf{x}_k,\boldsymbol{\phi})p(\dot{\mathbf{x}}_k|\mathbf{X},\boldsymbol{\theta},\gamma_k) = \mathcal{N}(\dot{\mathbf{x}}_k|\mathbf{m}_k,\mathbf{A}_k)\mathcal{N}(\dot{\mathbf{x}}_k|\mathbf{f}_k(\mathbf{X},\boldsymbol{\theta}),\gamma_k\mathbf{I}) \quad (14)$$

and obtain for the joint distribution:

$$p(\dot{\mathbf{X}},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\gamma}) = p(\dot{\mathbf{X}}|\mathbf{X},\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\gamma})p(\mathbf{X}|\boldsymbol{\phi})p(\boldsymbol{\theta})p(\boldsymbol{\phi})p(\boldsymbol{\gamma}) = p(\boldsymbol{\theta})p(\boldsymbol{\phi})p(\boldsymbol{\gamma})\prod_k p(\dot{\mathbf{x}}_k|\mathbf{X},\boldsymbol{\theta},\boldsymbol{\phi},\gamma_k)p(\mathbf{x}_k|\boldsymbol{\phi}_k) \quad (15)$$

where $p(\boldsymbol{\theta}), p(\boldsymbol{\phi}), p(\boldsymbol{\gamma})$ denote the prior distributions of the ODE parameters $\boldsymbol{\theta}$, the GP hyperparameters $\boldsymbol{\phi}$, and the slack hyperparameters $\boldsymbol{\gamma}$; the latter define the tightness of the gradient coupling. Inserting (6) and (14) into (15) gives:

$$p(\dot{\mathbf{X}},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\gamma}) \propto p(\boldsymbol{\theta})p(\boldsymbol{\phi})p(\boldsymbol{\gamma}) \prod_k \mathcal{N}(\dot{\mathbf{x}}_k|\mathbf{m}_k,\mathbf{A}_k)\mathcal{N}(\dot{\mathbf{x}}_k|\mathbf{f}_k(\mathbf{X},\boldsymbol{\theta}),\gamma_k\mathbf{I})\mathcal{N}(\mathbf{x}_k|\boldsymbol{\mu}_k,\mathbf{C}_{\phi_\mathbf{k}}). \quad (16)$$

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

$$p(\mathbf{X},\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\gamma}) = \int p(\dot{\mathbf{X}},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\gamma})d\dot{\mathbf{X}} \propto p(\boldsymbol{\theta})p(\boldsymbol{\phi})p(\boldsymbol{\gamma})\prod_k \mathcal{N}(\mathbf{x}_k|\boldsymbol{\mu}_k,\mathbf{C}_{\phi_\mathbf{k}}) \int \mathcal{N}(\dot{\mathbf{x}}_k|\mathbf{m}_k,\mathbf{A}_k)\mathcal{N}(\dot{\mathbf{x}}_k|\mathbf{f}_k(\mathbf{X},\boldsymbol{\theta}),\gamma_k\mathbf{I})d\dot{\mathbf{x}}_k \quad (17)$$

is analytically tractable and yields:

$$p(\mathbf{X},\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\gamma}) \propto p(\boldsymbol{\theta})p(\boldsymbol{\phi})p(\boldsymbol{\gamma})p(\mathbf{X}|\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\gamma}) \quad (18)$$

$$p(\mathbf{X}|\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\gamma}) \propto \prod_k \frac{\mathcal{N}(\mathbf{x}_k|\boldsymbol{\mu}_k,\mathbf{C}_{\phi_\mathbf{k}})}{Z(\gamma_k)} \exp\Big[-\frac{1}{2}(\mathbf{f}_k-\mathbf{m}_k)^\mathsf{T}(\mathbf{A}_k+\gamma_k\mathbf{I})^{-1}(\mathbf{f}_k-\mathbf{m}_k)\Big] \propto \exp\left[-\frac{1}{2}\sum_k\left(\mathbf{x}_k^\mathsf{T}\mathbf{C}_{\phi_k}^{-1}\mathbf{x}_k + (\mathbf{f}_k-\mathbf{m}_k)^\mathsf{T}(\mathbf{A}_k+\gamma_k\mathbf{I})^{-1}(\mathbf{f}_k-\mathbf{m}_k)\right)\right]\frac{1}{\prod_k Z(\gamma_k)} \quad (19)$$

where $Z(\gamma_k) = (2\pi)^k|\mathbf{A}_k + \gamma_k\mathbf{I}|$, $\mathbf{f}_k$ is shorthand notation for $\mathbf{f}_k(\mathbf{X},\boldsymbol{\theta},\mathbf{t})$, and $\mathbf{m}_k$ and $\mathbf{A}_k$ were defined in (12). Note that this distribution is a complicated function of the states $\mathbf{X}$, owing to the nonlinear dependence via $\mathbf{f}_k = \mathbf{f}_k(\mathbf{X},\boldsymbol{\theta},\mathbf{t})$. For the joint probability distribution of the whole system this gives:

$$p(\mathbf{Y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\gamma},\boldsymbol{\sigma}) = p(\mathbf{Y}|\mathbf{X},\boldsymbol{\sigma})p(\mathbf{X}|\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\gamma})p(\boldsymbol{\theta})p(\boldsymbol{\phi})p(\boldsymbol{\gamma})p(\boldsymbol{\sigma}) \quad (20)$$

where the first factor, $p(\mathbf{Y}|\mathbf{X},\boldsymbol{\sigma})$, was defined in (5), and the second factor is given by (18). A graphical representation of the model is given in Figure 1. Inference is analytically intractable. (Calderhead et al., 2008) introduced a modularization approximation to (20), which for space restrictions we cannot discuss here. (Dondelinger et al., 2013) have developed an effective MCMC scheme to sample $\mathbf{X},\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\gamma},\boldsymbol{\sigma}$ directly from the posterior distribution $p(\mathbf{X},\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\gamma},\boldsymbol{\sigma}|\mathbf{Y}) \propto p(\mathbf{Y},\mathbf{X},\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\gamma},\boldsymbol{\sigma})$. Due to space restrictions, we refer the reader to the original publications for the methodological details.

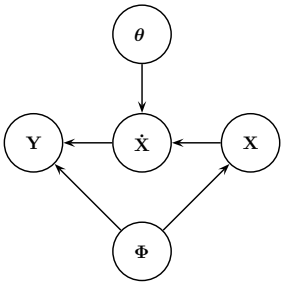

Figure 2. GPODE model, proposed in (Wang & Barber, 2014).

3. Paradigm B: the GPODE model

An alternative approach was proposed by (Wang & Barber, 2014) and termed the GPODE model. As for AGM, the starting point in (Wang & Barber, 2014) is to exploit the fact that the derivative of a Gaussian process is also a Gaussian process, and that the joint distribution of the state variables $\mathbf{X}$ and their time derivatives $\dot{\mathbf{X}}$ is multivariate Gaussian with covariance functions given by (7-10). Application of elementary transformations of Gaussian distributions, as shown e.g. on p. 93 in (Bishop, 2006), leads to the following conditional distribution of the states given the state derivatives:

$$p_{GP}(\mathbf{x}_k|\dot{\mathbf{x}}_k, \boldsymbol{\phi}) \;=\; \mathcal{N}(\mathbf{x}_k|\tilde{\mathbf{m}}_k, \tilde{\mathbf{A}}_k) \tag{21}$$

where for clarity we refer to the GP with a subscript, and

$$\tilde{\mathbf{m}}_k = \boldsymbol{\mu}_k + {}'\mathbf{C}_{\phi_\mathbf{k}}\mathbf{C}''^{\,-1}_{\phi_\mathbf{k}}\dot{\mathbf{x}}_k; \quad \tilde{\mathbf{A}}_k = \mathbf{C}_{\phi_\mathbf{k}} - {}'\mathbf{C}_{\phi_\mathbf{k}}\mathbf{C}''^{\,-1}_{\phi_\mathbf{k}}\mathbf{C}'_{\phi_\mathbf{k}}. \tag{22}$$

Note the difference between AGM and GPODE, where for the former method we compute $p(\dot{\mathbf{x}}_k|\mathbf{x}_k, \boldsymbol{\phi})$, as expressed in (11-12), whereas for the latter model we compute $p(\mathbf{x}_k|\dot{\mathbf{x}}_k, \boldsymbol{\phi})$, as expressed in (21-22). Under the assumption that the observations $\mathbf{Y}$ are subject to additive iid Gaussian noise, (2,5), the marginalization over the state variables leads to a standard Gaussian convolution integral, which is analytically tractable with solution

$$\begin{aligned} p_\diamond(\mathbf{y}_k|\dot{\mathbf{x}}_k, \boldsymbol{\phi}) &= \int p(\mathbf{y}_k|\mathbf{x}_k)p_{GP}(\mathbf{x}_k|\dot{\mathbf{x}}_k, \boldsymbol{\phi})d\mathbf{x}_k \\ &= \int \mathcal{N}(\mathbf{y}_k|\mathbf{x}_k, \sigma_k^2\mathbf{I})\mathcal{N}(\mathbf{x}_k|\tilde{\mathbf{m}}_k, \tilde{\mathbf{A}}_k)d\mathbf{x}_k \\ &= \mathcal{N}(\mathbf{y}_k|\tilde{\mathbf{m}}_k, \tilde{\mathbf{A}}_k + \sigma_k^2\mathbf{I}). \end{aligned} \tag{23}$$

The authors factorize

$$p(\mathbf{Y}, \mathbf{X}|\boldsymbol{\phi}, \boldsymbol{\theta}) \;=\; p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\phi}, \boldsymbol{\theta})p_{GP}(\mathbf{X}|\boldsymbol{\phi}) \tag{24}$$

and obtain the first term by marginalization over the state derivatives $\dot{\mathbf{X}}$:

$$\begin{aligned} p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\phi}, \boldsymbol{\theta}) &= \int p(\mathbf{Y}, \dot{\mathbf{X}}|\mathbf{X}, \boldsymbol{\phi}, \boldsymbol{\theta})d\dot{\mathbf{X}} \\ &= \int p_\diamond(\mathbf{Y}|\dot{\mathbf{X}}, \boldsymbol{\phi})p_{ODE}(\dot{\mathbf{X}}|\mathbf{X}, \boldsymbol{\theta})d\dot{\mathbf{X}} \\ &= p_\diamond(\mathbf{Y}|f[\mathbf{X}, \boldsymbol{\theta}], \boldsymbol{\phi}) \end{aligned} \tag{25}$$

where $p_\diamond(\mathbf{Y}|\dot{\mathbf{X}}, \boldsymbol{\phi}) = \prod_k p_\diamond(\mathbf{y}_k|\dot{\mathbf{x}}_k, \boldsymbol{\phi})$, with $p_\diamond(\mathbf{y}_k|\dot{\mathbf{x}}_k, \boldsymbol{\phi})$ given in (23), and assuming that the state derivatives are deterministically defined by the ODEs:

$$p_{ODE}(\dot{\mathbf{X}}|\mathbf{X}, \boldsymbol{\theta}) \;=\; \delta(\dot{\mathbf{X}} - f[\mathbf{X}, \boldsymbol{\theta}]). \tag{26}$$

Inserting (25) into (24) gives:

$$p(\mathbf{Y}, \mathbf{X}|\boldsymbol{\phi}, \boldsymbol{\theta}) \;=\; p_\diamond(\mathbf{Y}|f[\mathbf{X}, \boldsymbol{\theta}], \boldsymbol{\phi})p_{GP}(\mathbf{X}|\boldsymbol{\phi}). \tag{27}$$

This is a deceptionally simple and elegant formulation, illustrated as a graphical model in Figure 2, with two advantages over the AGM model. Conceptually, the GPODE is a proper probabilistic generative model, which can be consistently represented by a directed acyclic graph (DAG). Practically, the normalization constant of the joint distribution in (27) is known, which facilitates inference.

4. Shortcomings of the GPODE model

The Achilles heel of the GPODE model is equation (23), which includes a marginalization over the state variables $\mathbf{x}_k$ to obtain $p_\diamond(\mathbf{y}_k|\dot{\mathbf{x}}_k)$. The derivations in (24) and (25) then treat $\mathbf{y}_k$ as independent of $\mathbf{x}_k$ given $\dot{\mathbf{x}}_k$: $p(\mathbf{y}_k|\dot{\mathbf{x}}_k, \mathbf{x}_k) = p_\diamond(\mathbf{y}_k|\dot{\mathbf{x}}_k)$, or $p(\mathbf{Y}|\mathbf{X}, \dot{\mathbf{X}}) = p_\diamond(\mathbf{Y}|\dot{\mathbf{X}})$; this is consistent with the graphical model in Figure 2. Having integrated the state variables $\mathbf{X}$ out in (23), the method subsequently conditions on them in (25). The underlying assumption the authors make is that the marginalization over the random variables $\mathbf{x}_k$ in (23) is equivalent to their elimination. However, marginalization merely means that for the purposes of inference, the variables that have been integrated out do not need to be taken into consideration explicitly. However, these variables remain in the model conceptually. In our particular model, the data $\mathbf{Y}$ consist of noisy observations of the state variables $\mathbf{X}$, not their derivatives $\dot{\mathbf{X}}$. Consider, for instance, the tracking of a set of exoplanets with a space telescope, where the state variables $\mathbf{X}$ are the positions of the planets. Given the knowledge of the initial conditions and the velocities of the planets, $\dot{\mathbf{X}}$, we can compute the positions of the planets $\mathbf{X}$ using established equations from classical mechanics. This procedure might dispense with the need to keep detailed records of the planets' positions. However, it does *not* imply that the positions of the planets have disappeared.

For methodological consistency, we need to reintroduce the state variables $\mathbf{X}$ into the model, as shown in Figure 3, left panel. However, this leads to the inconsistency that the same random variables, $\mathbf{X}$, are used in two different places of the graph. As a further correction, we therefore introduce a set of dummy variables $\tilde{\mathbf{X}}$, as shown in Figure 3, centre panel. This is a methodologically consistent representation of the model, but leaves open the question what the difference between $\mathbf{X}$ and $\tilde{\mathbf{X}}$ is. Ideally, there is no difference, which can be represented mathematically as

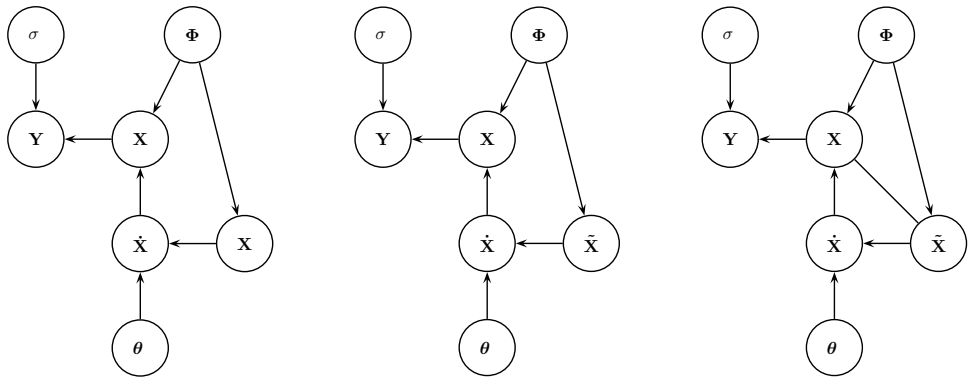

Figure 3. Left panel: GPODE model, as proposed in (Wang & Barber, 2014), but explicitly presenting all random variables included in the model. The graph is inconsistent, in that the same random variables, $\mathbf{X}$, have been assigned to two different nodes. *Centre panel:* Correcting the inconsistency in the notation of (Wang & Barber, 2014). The model distinguishes between the unknown true state variables $\mathbf{X}$, and their model approximation $\tilde{\mathbf{X}}$. *Right panel:* In the ideal GPODE model, the true state variables $\mathbf{X}$ and their model approximation $\tilde{\mathbf{X}}$ are coupled, ideally via an identity constraint. This introduces an undirected edge between $\mathbf{X}$ and $\tilde{\mathbf{X}}$, which is no longer a consistent probabilistic graphical model represented by a DAG. To reintroduce the DAG constraint, (Wang & Barber, 2014) have discarded this undirected edge, leading to the model shown in the centre panel. The disadvantage is that the model state variables $\tilde{\mathbf{X}}$ are no longer directly associated with the data. As we discuss in the main text, this leads to an intrinsic identifiability problem.

$p(\mathbf{X}, \tilde{\mathbf{X}}) = \delta(\mathbf{X} - \tilde{\mathbf{X}})$. However, in this way we have introduced an edge from the node $\tilde{\mathbf{X}}$ to $\mathbf{X}$, as shown in Figure 3, right panel. This causes methodological problems, in whatever definition we choose for that edge. If we treat it as an undirected edge, $p(\mathbf{X}, \tilde{\mathbf{X}}) = \delta(\mathbf{X} - \tilde{\mathbf{X}})$, as shown in the right panel of Figure 3, based on the symmetry of the identity relation between $\tilde{\mathbf{X}}$ and $\mathbf{X}$, then we get a chain graph. A chain graph is not a probabilistic generative model, and the main objective of (Wang & Barber, 2014) was to obtain the latter. If we introduce a directed edge from $\mathbf{X}$ to $\tilde{\mathbf{X}}$, based on $p(\tilde{\mathbf{X}}|\mathbf{X}) = \delta(\tilde{\mathbf{X}} - \mathbf{X})$, then we end up with a directed cycle that violates the DAG constraint. In order to get a valid probabilistic graphical model, we have to introduce a directed edge in the opposite direction, from $\tilde{\mathbf{X}}$ to $\mathbf{X}$, based on $p(\mathbf{X}|\tilde{\mathbf{X}}) = \delta(\tilde{\mathbf{X}} - \mathbf{X})$. However, this structure will require us to define the probability $p(\mathbf{X}|\dot{\mathbf{X}}, \tilde{\mathbf{X}})$, and it is not clear how to do that. For that reason, the approximation taken in (Wang & Barber, 2014) is to discard the edge between $\mathbf{X}$ and $\tilde{\mathbf{X}}$ altogether. This simplification leads to a probabilistic generative model that can be consistently represented by a DAG. However, the disadvantage is that the true state variables $\mathbf{X}$ and their approximation $\tilde{\mathbf{X}}$ are only weakly coupled, via their common hyperparameters Φ. We will discuss the consequences below.

The upshot of what has been explained so far is that, by not properly distinguishing between $\mathbf{X}$ and $\tilde{\mathbf{X}}$, equation (27) introduced in (Wang & Barber, 2014) is misleading. The correct form is

$$p(\mathbf{Y}, \tilde{\mathbf{X}}|\boldsymbol{\phi}, \boldsymbol{\theta}) \;=\; p_\diamond(\mathbf{Y}|f[\tilde{\mathbf{X}}, \boldsymbol{\theta}], \boldsymbol{\phi}) p_{GP}(\tilde{\mathbf{X}}|\boldsymbol{\phi}) \qquad (28)$$

where $\tilde{\mathbf{X}}$ are *not* the unknown true state variables $\mathbf{X}$, but some model approximation. This subtle difference has non-negligible consequences. As an illustration, consider the simple second-order ODE (using $\ddot{x} = d^2x/dt^2$)

$$\ddot{x} + \theta^2 x \;=\; 0 \qquad (29)$$

which, with the standard substitution $(x_1, x_2) := (x, \dot{x})$, leads to the linear system of first-order ODEs:

$$\dot{x}_1 \;=\; x_2; \qquad \dot{x}_2 \;=\; -\theta^2 x_1. \qquad (30)$$

These ODEs have the closed-form solution:

$$x_1(t) \;=\; A\sin(\theta t + \phi); \quad x_2(t) \;=\; A\theta\cos(\theta t + \phi) \quad (31)$$

where A and ϕ are constants, which are determined by the initial conditions. Now, according to the GPODE paradigm, illustrated in the centre panel of Figure 3, x_1 and x_2 in (30) have to be replaced by separate variables:

$$\dot{x}_1(t) \;=\; \tilde{x}_2(t); \qquad \dot{x}_2(t) \;=\; -\theta^2 \tilde{x}_1(t) \qquad (32)$$

where $\tilde{x}_1(t)$ and $\tilde{x}_2(t)$ are modelled with a GP. Recalling that $\mathbf{x}_k = [x_k(t_1), \ldots, x_k(t_N)]^\mathsf{T}$, we rewrite (32) as:

$$\dot{\mathbf{x}}_1 = \mathbf{f}_1(\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2; \theta) = \tilde{\mathbf{x}}_2; \;\; \dot{\mathbf{x}}_2 = \mathbf{f}_2(\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2; \theta) = -\theta^2 \tilde{\mathbf{x}}_1.$$

Inserting these expressions into (28), we get:

$$\begin{aligned} &p(\mathbf{y}_1, \mathbf{y}_2, \tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2|\boldsymbol{\phi}, \theta) = \qquad (33)\\ &p_\diamond(\mathbf{y}_1, \mathbf{y}_2|f_1[\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \theta], f_2[\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \theta], \boldsymbol{\phi}) p(\tilde{\mathbf{x}}_1|\boldsymbol{\phi}) p(\tilde{\mathbf{x}}_2|\boldsymbol{\phi}) = \\ &p_\diamond(\mathbf{y}_1|f_1[\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \theta], \boldsymbol{\phi}) p_\diamond(\mathbf{y}_2|f_2[\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \theta], \boldsymbol{\phi}) p(\tilde{\mathbf{x}}_1|\boldsymbol{\phi}) \\ &p(\tilde{\mathbf{x}}_2|\boldsymbol{\phi}) = p_\diamond(\mathbf{y}_1|\tilde{\mathbf{x}}_2, \boldsymbol{\phi}) p_\diamond(\mathbf{y}_2| - \theta^2 \tilde{\mathbf{x}}_1, \boldsymbol{\phi}) p(\tilde{\mathbf{x}}_1|\boldsymbol{\phi}) p(\tilde{\mathbf{x}}_2|\boldsymbol{\phi}). \end{aligned}$$

We use the subscript in $p_\diamond$ to indicate that the functional form of this probability distribution is given by (23), but drop the subscript 'GP' used in the previous section. Now, recall that the variable x_2 represents the time derivative of x_1 and was introduced as an auxiliary variable to transform the second-order ODE from (29) into a system of first-order ODEs: equation (30). In most applications, only the variables themselves rather than their derivatives can be measured or observed, i.e. y_2 is systematically missing. From (33) we obtain for missing variables y_2:

$$\begin{aligned} p(\mathbf{y}_1, \tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2|\boldsymbol{\phi}, \theta) &= \int p(\mathbf{y}_1, \mathbf{y}_2, \tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2|\boldsymbol{\phi}, \theta) d\mathbf{y}_2 \\ &= p_\diamond(\mathbf{y}_1|\tilde{\mathbf{x}}_2, \boldsymbol{\phi}) p(\tilde{\mathbf{x}}_1|\boldsymbol{\phi}) p(\tilde{\mathbf{x}}_2|\boldsymbol{\phi}) \\ &\quad \int p_\diamond(\mathbf{y}_2| - \theta^2 \tilde{\mathbf{x}}_1, \boldsymbol{\phi}) d\mathbf{y}_2 \\ &= p_\diamond(\mathbf{y}_1|\tilde{\mathbf{x}}_2, \boldsymbol{\phi}) p(\tilde{\mathbf{x}}_1|\boldsymbol{\phi}) p(\tilde{\mathbf{x}}_2|\boldsymbol{\phi}) \quad (34) \end{aligned}$$

and

$$\begin{aligned} p(\mathbf{y}_1|\boldsymbol{\phi}, \theta) &= \int p(\mathbf{y}_1, \tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2|\boldsymbol{\phi}, \theta) d\tilde{\mathbf{x}}_1 d\tilde{\mathbf{x}}_2 \quad (35) \\ &= \int p_\diamond(\mathbf{y}_1|\tilde{\mathbf{x}}_2, \boldsymbol{\phi}) p(\tilde{\mathbf{x}}_2|\boldsymbol{\phi}) d\tilde{\mathbf{x}}_2 \int p(\tilde{\mathbf{x}}_1|\boldsymbol{\phi}) d\tilde{\mathbf{x}}_1 \\ &= \int p_\diamond(\mathbf{y}_1|\tilde{\mathbf{x}}_2, \boldsymbol{\phi}) p(\tilde{\mathbf{x}}_2|\boldsymbol{\phi}) d\tilde{\mathbf{x}}_2 = p(\mathbf{y}_1|\boldsymbol{\phi}). \end{aligned}$$

This implies that the likelihood, i.e. the probability of a set of observations $\mathbf{y}_1 = [y_1(t_1), \ldots, y_1(t_N)]^\mathsf{T}$, is independent of the ODE parameter θ. Consequently, in the GPODE model, the parameter of interest – the ODE parameter θ – is unidentifiable, i.e. it can *not* be inferred from the data. Note that this problem is intrinsic to the GPODE model, *not* the ODE itself. Equation (29) is a very simple ODE with a closed form solution for $x(t) = x_1(t)$, stated in (31). If this solution is known, the inference task reduces to inferring the frequency from noisy observations of a sine function. Hence, it is straightforward to infer θ from noisy observations $y_1(t) = x_1(t) + \varepsilon(t)$ alone, where $\varepsilon(t)$ is iid noise, and no observations of the derivative $x_2 = \frac{dx}{dt}$ are required. Even if the explicit solution were not known, it could be obtained by numerical integration of the ODEs, again rendering the inference of the ODE parameter θ a straightforward task. How do missing observations affect the AGM model? When y_2 is systematically missing, we need to marginalize over $\mathbf{y}_2$ in (20). This will only affect the first term on the right-hand side of (20), which as a consequence of the marginalization will reduce from $p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\sigma}) = p(\mathbf{y}_1, \mathbf{y}_2|\mathbf{X}, \boldsymbol{\sigma})$ to $p(\mathbf{y}_1|\mathbf{X}, \boldsymbol{\sigma})$. However, this term does not explicitly depend on the ODE parameters $\boldsymbol{\theta}$. Hence, as opposed to the GPODE model, missing observations do not systematically eliminate ODE parameters from the likelihood. In fact, an inspection of equation (30) provides an intuitive explanation of how inference in the AGM can work despite systematically missing values: noisy observations of x_1 provide information about the missing species x_2 via (30), left, using the very principle of gradient matching. Inference of x_2 then enables inference of the ODE parameter θ via (30), right. We will demonstrate, in Section 5, that AGM indeed can successfully infer the ODE parameter θ when observations for species y_2 are missing, whereas GPODE systematically fails on this task.

5. Empirical findings

The empirical analysis presented in (Wang & Barber, 2014) suggests that the GPODE model achieves very accurate parameter estimates. However, a closer inspection of the authors' study reveals that they used rather informative priors with relatively tight uncertainty intervals centred on the (known) true parameter values. In the present study, we have repeated the authors' simulations with less informative priors; all GPODE results were obtained with the original software from (Wang & Barber, 2014). We have also integrated the inference for the AGM model into their software, for a fair comparison between the two paradigms. Our code can be downloaded from http://tinyurl.com/otus5xq.

Computational inference. The objective of inference is to obtain the marginal posterior distributions of the quantities of interest, which are usually the ODE parameters. This is analytically intractable, and previous authors have used sampling methods based on MCMC. (Dondelinger et al., 2013) and (Calderhead et al., 2008) used MCMC schemes for continuous values, based on Metropolis-Hastings with appropriate proposal moves. (Wang & Barber, 2014) used Gibbs sampling as a faster alternative, based on a discretization of the latent variables, parameters and hyperparameters. For a fair comparison between the model paradigms (AGM versus GPODE), which is not confounded by the different convergence characteristics and potential discretization artefacts of the two MCMC schemes (Metropolis-Hastings versus Gibbs sampling), we have implemented the AGM model in the software of (Wang & Barber, 2014) to infer all quantities of interest with the same Gibbs sampling scheme. The basic idea is that due to the discretization, all quantities can be marginalized over in the joint probability density, and this allows the conditional probabilities needed for the Gibbs sampler to be easily computed. Due to space restrictions, we refer the reader to Section 3 of (Wang & Barber, 2014) for the methodological details. For the prior distribution over the latent variables, the software of (Wang & Barber, 2014) fits a standard GP to the data and chooses, for each time point, a uniform distribution with a 3-standard-deviation width centred on the GP interpolant. For faster convergence of the MCMC simulations, we set the noise variance σ_k^2 equal to the true noise variance, and the mean $\boldsymbol{\mu}_k$ equal to the sample mean. The parameters that had to be inferred (in addition to the latent state variables) were the ODE parameters, the kernel parameters of the GP, and the slack hyperparameter γ for the AGM. For all simula-

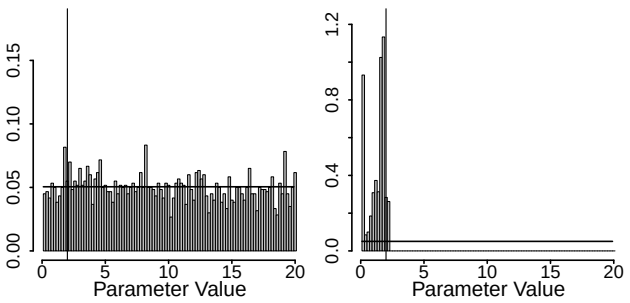

Figure 4. Inference results for the ODEs (30) with missing species. Vertical line: true parameter value. Horizontal line: uniform prior. Histogram: average posterior distribution obtained with Gibbs sampling, averaged over ten independent data instantiations. Left panel: GPODE model. Right panel: AGM model.

tions, we used a squared exponential kernel, and chose a $U(5, 50)$ prior for the length scale and a $U(0.1, 1)$ prior for the amplitude hyperparameters, respectively, as in the paper by (Wang & Barber, 2014). We tried different prior distributions of the ODE parameters, as specified in the figure captions; note that these priors are less informative than those used in (Wang & Barber, 2014). Observational noise was added in the same way as in (Wang & Barber, 2014). We monitored the convergence of the MCMC chains with the diagnostics proposed by (Gelman & Rubin, 1992), and terminated the burn-in phase when the potential scale reduction factor fell below a threshold of 1.1. All simulations were repeated on ten independent data instantiations.

Simple ODE with missing values. As a first study, we generated noisy data from the simple ODEs of (30), with species 2 missing, using a sample size of $N = 20$ and an average signal-to-noise ratio of $SNR = 10$. The results are shown in Figure 4. They confirm what was discussed below equation (35): paradigm B completely fails to infer the ODE parameter; in fact, the inferred posterior distribution is indistinguishable from the prior. Paradigm A succeeds in inferring the ODE parameter: the posterior distribution is significantly different from the prior and includes the true parameter.

The Lotka-Volterra system

$$\dot{x}_1 = \theta_1 x_1 - \theta_2 x_1 x_2; \quad \dot{x}_2 = -\theta_3 x_2 + \theta_4 x_1 x_2 \qquad (36)$$

is a simple model for prey-predator interactions in ecology (Lotka, 1932), and autocatalysis in chemical kinetics (Atkins, 1986). It has four kinetic parameters $\theta_1, \theta_2, \theta_3, \theta_4 > 0$, which we try to infer. This model was used for the evaluation of parameter inference in (Dondelinger et al., 2013) and (Wang & Barber, 2014), and we repeated the simulations with the same parameters as used in these studies. First, $N = 11$ data points were generated

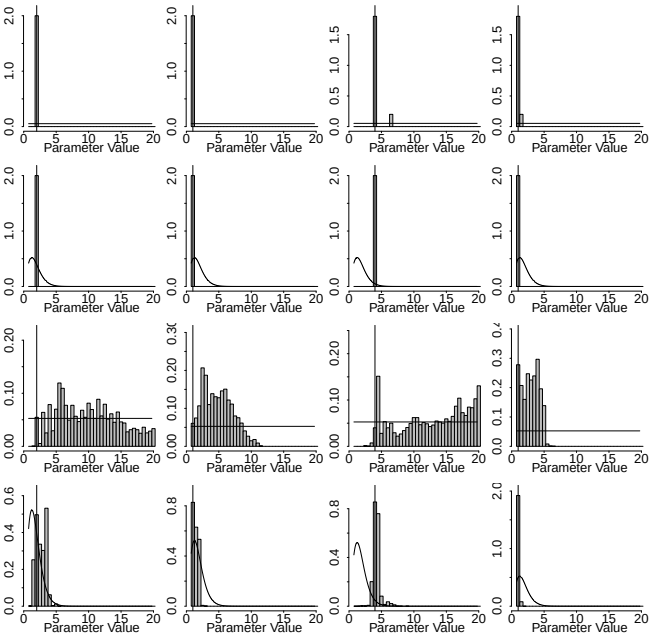

Figure 5. Inference results for the Lotka-Volterra system (36). Each column represents one of the four kinetic parameters of the system, and the histograms show the average posterior distributions of the respective parameter, averaged over ten data instantiations. Vertical line: true parameter value. Horizontal line or curve: prior distribution - uniform or $\Gamma(4, 0.5)$. The top two rows show the results for the AGM model (paradigm A). The bottom two rows show the results for the GPODE model (paradigm B).

with $\theta_1 = 2, \theta_2 = 1, \theta_3 = 4, \theta_4 = 1$. Next, iid Gaussian noise with an average signal-to-noise ratio $SNR = 4$ was added, and ten independent data sets were generated this way. The results are shown in Figure 5. The AGM model (paradigm A) shows a consistent performance over both parameter priors: the Gamma $\Gamma(4, 0.5)$ prior and the uniform prior. In both cases, the inferred posterior distributions are tightly concentrated on the true parameters. The GPODE model (paradigm B) sensitively depends on the prior. The inferred posterior distributions are always more diffuse than those obtained with paradigm A, and the performance is particularly poor for the uniform prior. Here, paradigm A clearly outperforms paradigm B.

The Fitz-Hugh Nagumo system

$$\frac{dV}{dt} = \psi(V - \frac{V^3}{3} + R); \quad \frac{dR}{dt} = -\frac{1}{\psi}(V - \alpha + \beta R) \qquad (37)$$

was introduced in (FitzHugh, 1961) and (Nagumo et al., 1962)) to model the voltage potential across the cell membrane of the axon of giant squid neurons. There are two species: Voltage (V) and Recovery variable (R), and 3 parameters; α, β and ψ. The model was used in (Campbell & Steele, 2012) to assess parameter inference in ODEs, using comparatively large sets of $N = 401$ observations. For

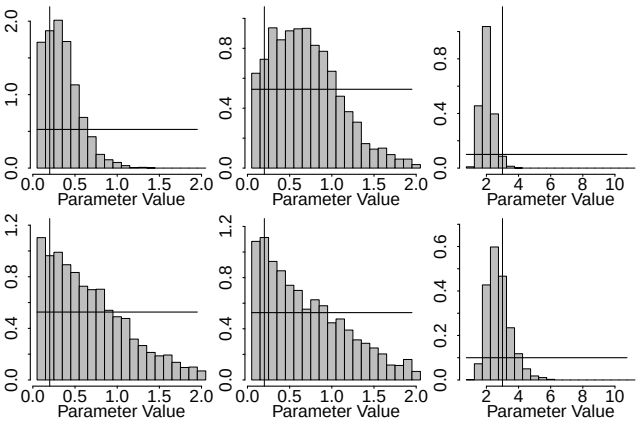

Figure 6. Inference results for the Fitz-Hugh Nagumo system (37). Each column represents one of the three kinetic parameters of the system, and the histograms show the average posterior distributions of the respective parameter, averaged over ten data instantiations. Vertical line: true parameter value. Horizontal line: prior distribution. The top row shows the results for the AGM model (paradigm A). The bottom row shows the results for the GPODE model (paradigm B). Due to space restrictions, only the results for the uniform prior are shown. The results for the priors used in (Campbell & Steele, 2012) – a non-negative truncated $N(0, 0.4)$ and a $\chi^2(2)$ distribution – were similar.

the present study, we generated data with the same parameters, $\alpha = 0.2$, $\beta = 0.2$ and $\psi = 3$, and same initial values, $V = -1, R = 1$, but making the inference problem harder by reducing the training set size to $N = 20$, covering the time interval $[0, 10]$. We emulated noisy measurements by adding iid Gaussian noise with an average signal-to-noise ratio $SNR = 10$, and generated ten independent data instantiations. The results are shown in Figure 6. Here, both paradigms show a similar performance. The GPODE model is slightly better than the AGM model in terms of reduced bias for the third parameter, but slightly worse in terms of increased posterior variance for the first parameter. The results are, overall, worse than for the Lotka-Volterra system. Note that the Fitz-Hugh Nagumo system poses a challenging problem, though; see (Campbell & Steele, 2012) and recall that our data set is considerably smaller (5%) than the one used but the authors.

6. Conclusion

Inference in mechanistic models based on non-affine ODEs is challenging due to the high computational costs of the numerical integration of the ODEs, and approximate methods based on adaptive gradient matching have therefore gained much attention in the last few years. The application of nonparametric Bayesian methods based on GPs is particularly promising owing to the fact that a GP is closed under differentiation. A new paradigm termed GPODE was proposed in (Wang & Barber, 2014) at ICML 2014, which was purported to outperform state-of-the-art GP gradient matching methods in three respects: providing a simplified mathematical description, constituting a probabilistic generative model, and achieving better inference results. The purpose of the present paper has been to critically review these claims. It turns out that the simplicity of the model presented in (Wang & Barber, 2014), shown in Figure 2, results from equating the marginalization over a random variable with its elimination from the model. A proper representation of the GPODE model leads to a more complex form, shown in Figure 3. We have shown that the GPODE model is turned into a probabilistic generative model at the expense of certain independence assumptions, which have not been made explicit in (Wang & Barber, 2014). We have further shown that as a consequence of these independence assumptions, the GPODE model is susceptible to identifiability problems when data are systematically missing. This problem is unique to the GPODE model, and is avoided when gradient matching with GPs follows the product of experts approach of (Calderhead et al., 2008) and (Dondelinger et al., 2013) (herein called paradigm A). Unlike (Wang & Barber, 2014), our empirical comparison has not shown any performance improvement over paradigm A. On the contrary, for two data sets (simple ODE with missing values, and the Lotka-Volterra system), paradigm A achieves significantly better results. For a third data set (Fitz-Hugh Nagumo system), both approaches are on a par, with different bias-variance characteristics.

The right-hand panel of Figure 3 demonstrates that gradient matching for inference in ODEs intrinsically violates the DAG constraint. This is because the function to be matched is both the output of and the input to the ODEs, leading to a directed cycle. The endeavour to model gradient matching with GPs as a probabilistic generative model based on a DAG at the expense of implausible dummy variables and independence assumptions (Figure 3, centre panel) is at the heart of the problems with the GPODE model, as discussed previously. We have demonstrated that these problems can be avoided with gradient matching paradigm A. Our study suggests that for practical applications, paradigm A is to be preferred over paradigm B. (Wang & Barber, 2014) argue that a principled shortcoming of paradigm A is the fact that the underlying product of experts approach cannot be formulated in terms of a probabilistic generative model. However, as we have just discussed, this is of little relevance, given that gradient matching cannot be consistently conceptualized as a probabilistic generative model *per se*. This methodological limitation is the price that has to be paid for the substantial computational advantages over the explicit solution of the ODEs that gradient matching yields.

Acknowledgements

This work was supported by EPSRC (EP/L020319/1). We thank Y. Wang and D. Barber for sharing their software.

References

Atkins, P. W. *Physical Chemistry*. Oxford University Press, Oxford, 3rd edition, 1986.

Babtie, A.C, Kirk, P., and Stumpf, M.P.H. Topological sensitivity analysis for systems biology. *PNAS*, 111(51): 18507–18512, December 2014.

Bishop, C.M. *Pattern Recognition and Machine Learning*. Springer, Singapore, 2006. ISBN 978-0387-31073-2.

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

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

Dondelinger, F., Filippone, M., Rogers, S, and Husmeier, D. ODE parameter inference using adaptive gradient matching with Gaussian processes. *Journal of Machine Learning Research - Workshop and Conference Proceedings: The 16th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 31:216–228, 2013.

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

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

González, J., Vujačić, I., and Wit, E. Inferring latent gene regulatory network kinetics. *Statistical Applications in Genetics and Molecular Biology*, 12(1):109–127, 2013.

Graepel, T. Solving noisy linear operator equations by Gaussian processes: Application to ordinary and partial differential equations. In *Machine Learning, Proceedings of the Twentieth International Conference (ICML 2003), August 21-24, 2003, Washington, DC, USA*, pp. 234–241, 2003.

Holsclaw, T., Sanso, B., Lee, H. K. H., Heitmann, K., Habib, S., Higdon, D., and Alam, U. Gaussian process modeling of derivative curves. *Technometrics*, 55:57–67, 2013.

Liang, H. and Wu, H. Parameter estimation for differential equation models using a framework of measurement error in regression models. *Journal of the American Statistical Association*, 103(484):1570–1583, 2008.

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

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

Pokhilko, A., Fernandez, A.P., Edwards, K.D, Southern, M.M., Halliday, K.J., and Millar, A.J. The clock gene circuit in arabidopsis includes a repressilator with additional feedback loops. *Molecular Systems Biology*, 8 (574), 2012.

Ramsay, J.O., Hooker, G., Campbell, D., and Cao, J. Parameter estimation for differential equations: a generalized smoothing approach. *J. R. Statist*, pp. 741–796, 2007.

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

Solak, E., Murray-Smith, R., Leithead, W.E., Leith, D.J., and Rasmussen, C.E. Derivative observations in Gaussian process models of dynamic systems. *Advances in Neural Information Processing Systems*, pp. 9–14, 2003.

Wang, Y. and Barber, D. Gaussian Processes for Bayesian estimation in ordinary differential equations. *Journal of Machine Learning Research - Workshop and Conference Proceedings (ICML)*, 32:1485–1493, 2014.