Statistics and Computing (2019) 29:853–867
https://doi.org/10.1007/s11222-018-9840-4

Model selection via marginal likelihood estimation by combining thermodynamic integration and gradient matching

Benn Macdonald[1] · **Dirk Husmeier**[1]

Received: 28 April 2018 / Accepted: 8 November 2018 / Published online: 14 December 2018

Abstract
Conducting statistical inference on systems described by ordinary differential equations (ODEs) is a challenging problem. Repeatedly numerically solving the system of equations incurs a high computational cost, making many methods based on explicitly solving the ODEs unsuitable in practice. Gradient matching methods were introduced in order to deal with the computational burden. These methods involve minimising the discrepancy between predicted gradients from the ODEs and those from a smooth interpolant. Work until now on gradient matching methods has focused on parameter inference. This paper considers the problem of model selection. We combine the method of thermodynamic integration to compute the log marginal likelihood with adaptive gradient matching using Gaussian processes, demonstrating that the method is robust and able to outperform BIC and WAIC.

Keywords Ordinary differential equations · Model selection · Thermodynamic integration · Gradient matching

1 Introduction

Ordinary differential equations (ODEs) are a powerful way of providing an observed system with a mathematical description. The system can be expressed as

$$\dot{x}_s(u) = \frac{dx_s(u)}{du} = f_s(\mathbf{x}(u), \boldsymbol{\theta}_s, u), \tag{1}$$

where $s \in \{1, \ldots, N\}$ denotes one of N components (referred to throughout as "species"), $x_s(u)$ denotes the concentration of species s as a function of u (typically time or space), $\boldsymbol{\theta}_s$ is the vector of ODE parameters for species s and $\mathbf{x}(u)$ the vector of concentrations of all species of the variable u.

Examples of biological systems described by ODEs include predator–prey interactions in ecology (Lotka 1932), autocatalysis in chemical kinetics (Atkins 1986), signalling pathways in a biochemical cascade (Vyshemirsky and Girolami 2008), activation and deactivation of spiking neurons (FitzHugh 1961), cardiac excitation (Biktashev et al. 2008; Adon et al. 2015) and kinetics of enzyme reactions (Gratie et al. 2013).

Parameter inference can be carried out by solving the system of equations for a given parameter set and minimising the discrepancy between the predicted signals from the ODEs and the data. Since solutions to the ODEs typically do not exist in closed form, explicit solutions of the ODEs need to be computed numerically. Robinson (2004) contains a background on methods used for obtaining numerical solutions for ODEs and amongst other topics, discusses the use of Euler's method and the Runge–Kutta method as ways to do so.

The main drawback to inference using an explicit solution of the system is the computational burden. Every time the parameters are changed, for example at each step of a minimisation algorithm or sampling scheme (such as MCMC), the system needs to be re-solved rendering the approach too time consuming for practical use in many applications.

In recent years, gradient matching methods have been developed in order to deal with this computational burden. The methods start by smoothing data in order to obtain an interpolant. Gradient matching is then carried out by minimising the differences between the predicted gradients from

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11222-018-9840-4) contains supplementary material, which is available to authorized users.

✉ Benn Macdonald
Benn.Macdonald@glasgow.ac.uk

Dirk Husmeier
Dirk.Husmeier@glasgow.ac.uk

[1] School of Mathematics and Statistics, University of Glasgow, Glasgow G12 8SQ, UK

the ODEs (for a given parameter set) and slopes of the tangents to the interpolants. In this fashion, the ODEs never need to be explicitly solved, making gradient matching a computationally attractive approach. Different gradient matching methods differ by their choice of interpolation scheme and metric for penalising the difference between the gradients.

Wu et al. (2014) propose a multi-step approach based on penalised splines. Ramsay et al. (2007) conduct parameter inference by using a penalised likelihood approach with hierarchical regularisation to tune the parameter controlling the mismatch between the gradients from the interpolant and the predictions from the ODEs and the parameters controlling the spline interpolant. González et al. (2013) use a penalised likelihood method where the penalty incorporates the information of the ODEs, and then by using the properties of reproducing kernel Hilbert spaces (RKHS), they perform parameter inference in a computationally fast manner.

Calderhead et al. (2008) use Gaussian processes (GPs) to interpolate the data. GPs are able to fit highly non-linear functions and have the added benefit that it is possible to obtain the derivative process in closed form (see Holsclaw et al. 2013). The authors penalise the difference between the gradients using a product of experts approach to link the distribution of the interpolant gradients with the distribution of the gradients predicted by the ODEs. Dondelinger et al. (2013) improve upon the work by Calderhead et al. (2008) by allowing the estimates of the ODE parameters to influence the reshaping of the GP interpolant, dubbing the method adaptive gradient matching (AGM). Macdonald et al. (2016) combine the method of Dondelinger et al. (2013) with a parallel tempering scheme of the parameter controlling the amount of mismatch allowed between the gradients, examining the effect of doing so across a range of simulation studies.

Wang and Barber (2014) focused on trying to represent gradient matching with Gaussian processes as a probabilistic generative model, in order to dispense with the heuristics involved with linking distributions via a product of experts. The work by Macdonald et al. (2015) shows that it is only possible to represent gradient matching with GPs as a probabilistic generative model by making restrictive independence assumptions that lead to a non-negligible deterioration in the accuracy of parameter estimation. Since the function to be matched using gradient matching is both the output of and input to the ODEs, gradient matching naturally leads to a directed cycle. In order to model gradient matching as a directed acyclical graph, one must assume independence between the input and output of the ODE. This assumption is not only implausible, but also causes the model state variables to no longer be directly associated with the data, leading to intrinsic identifiability issues.

All these approaches focus on parameter inference for ODEs and have not considered the problem of model selection. Model selection for ODEs aims at distinguishing between different hypotheses describing the structure of the systems. There are two main approaches to model selection—explanatory model selection and predictive model selection. Explanatory model selection is the method of integrating over the parameters and focussing on the marginal likelihood of the data i.e. the probability of the data given the model and not the probability of the data given some parameter set. The posterior probability of the candidate models is given by

$$p(M|\mathbf{Y}) = \frac{p(\mathbf{Y}|M)p(M)}{p(\mathbf{Y})}, \tag{2}$$

where $\mathbf{Y}$ denotes the data and M represents different models.

Assuming a uniform prior over the models, $p(M|\mathbf{Y})$ in Eq. 2 is maximised by the term $p(\mathbf{Y}|M)$ and therefore explanatory model selection can be conducted by focussing on this term. This term is known as the marginal likelihood (for a given model) and is equal to

$$p(\mathbf{Y}|M) = \int p(\mathbf{Y}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)d\boldsymbol{\theta}, \tag{3}$$

where $\boldsymbol{\theta}$ is the parameter vector.

The main difficulty in computing the marginal likelihood is that usually the integral in Eq. 3 is not available in closed form, and the techniques used to calculate it are computationally expensive. One could follow the approach of Newton and Raftery (1994) and use the harmonic mean estimator to estimate $1/p(\mathbf{Y}|M)$, however in practice this works poorly (see Murphy 2012). The reason for this is that the estimator is a posterior mean of the inverse of the likelihood

$$\int \frac{1}{p(\mathbf{Y}|\boldsymbol{\theta}, M)} p(\boldsymbol{\theta}|\mathbf{Y}, M)d\boldsymbol{\theta} = \frac{1}{p(\mathbf{Y}|M)}. \tag{4}$$

Typically, the prior is uninformative i.e. the posterior is similar to the likelihood. Hence, when the posterior is large, the likelihood tends to be large and therefore the inverse of the likelihood is small. Contrariwise, when the posterior is small, the likelihood tends to be small and the inverse of the likelihood is large. Therefore, most of the important contributions come from the tail of the posterior distribution. This means that in numerical MCMC approximations, the majority of posterior samples make small contributions and the occasional outlier has a large influence. The variance of the estimator, as a consequence, becomes unbounded. Alternatively, one could implement the method of Chib (1995), which calculates the marginal likelihood by representing it in the form $p(\mathbf{Y}|M) = p(\mathbf{Y}|\boldsymbol{\theta}, M)p(\boldsymbol{\theta}|M)/p(\boldsymbol{\theta}|\mathbf{Y}, M)$. However, this method assumes that the posterior has been marginalised over all modes, which is often not the case and therefore the method can produce inaccurate results in practice (see

Murphy 2012; Neal 1998). Thermodynamic integration, successfully used in the field of Statistical Physics and more recently introduced into the wider Statistical community by Friel and Pettitt (2008), is a promising method for computing the log marginal likelihood. It uses the components that are already calculated in the parallel tempering scheme to be outlined in Sect. 2. This can be done for each of the competing models to find which one produces the highest marginal likelihood. Alternatively, one could use the computationally cheaper approach of BIC, which is asymptotically equivalent to minus two times the log marginal likelihood, and select the model that returns the smallest value.

The latter approach, predictive model selection, is a measure of out of sample predictive performance. However, approaches such as cross validation are computationally expensive and quite often information criteria are used as an approximation that is correct in the asymptotic limit. In contrast to explanatory model selection, these approaches do not integrate over the parameters. Instead they use the likelihood, which is the probability of the data given the parameters, and therefore model selection in this manner can be thought of as being conducted by means of predictive performance.

Information criteria tend to be estimates and approximations to some cross-validated fit (see Gelman et al. 2013). Cross-validation has been demonstrated to provide an accurate way of estimating a model's predictive performance (a survey of cross-validation results is presented in Arlot and Celisse 2010), however, these methods tend to be time-consuming. The natural step would then be to approximate the method of cross-validation to some degree, for example, AIC is asymptotically equivalent to cross-validation (see Fang 2011). WAIC on the other hand (which is an improvement over DIC, since DIC cannot deal with singular likelihood functions), is a recent method that is asymptotically equivalent to Bayesian leave-one-out cross-validation (see Spiegelhalter et al. 2002 for further details on DIC; Watanabe 2010 for further details on WAIC).

It should be noted that it can be difficult in practice to satisfy the asymptotic assumptions of information criteria, which can often lead to poor approximations of the quantity of interest.

This paper combines the method of calculating the log marginal likelihood using thermodynamic integration with that of gradient matching. This combination is not immediately straightforward since the general approach to parallel tempering, see Campbell and Steele (2012) for example, tempers the distribution of the data only and does not temper the distribution controlling the mismatch between the gradients (the general approach to parallel tempering is discussed in Sect. 2 and the consequences of using this approach are discussed in Sect. 3). It will be shown that the proposed way of combining thermodynamic integration and gradient matching results in accurate estimates of the log marginal likelihood and a robust way of performing model selection in systems described by ordinary differential equations. This new method will be compared to the results of WAIC and BIC on a set of benchmark problems.

2 Review of adaptive gradient matching

We begin by summarising the method of adaptive gradient matching from Dondelinger et al. (2013). This is included for the convenience of the reader and the methodology combining this method with thermodynamic integration is given in Sect. 3.

Consider a set of T arbitrary timepoints $t_1 < \cdots < t_i < \cdots < t_T$, and noisy observations $\mathbf{Y} = (\mathbf{y}(t_1), \ldots, \mathbf{y}(t_T))$, where

$$\mathbf{y}(t_i) = \mathbf{x}(t_i) + \boldsymbol{\epsilon}(t_i), \tag{5}$$

$N = \dim(\mathbf{x}(t_i))$ is equal to the number of species, $\mathbf{X} = (\mathbf{x}(t_1), \ldots, \mathbf{x}(t_T))$, $\mathbf{y}(t_i)$ is the data vector of the observations of all species concentrations at time t_i, $\mathbf{x}(t_i)$ is the vector of the concentrations of all species at time t_i, $\mathbf{y}_s$ is the data vector of the observations of species concentrations s at all timepoints, $\mathbf{x}_s$ is the vector of concentrations of species s at all timepoints, $y_s(t_i)$ is the observed datapoint of the concentration of species s at time t_i, $x_s(t_i)$ is the concentration of species s at time t_i and $\boldsymbol{\epsilon}$ is multivariate Gaussian noise, $\boldsymbol{\epsilon} \sim N(\boldsymbol{\epsilon}|\mathbf{0}, \sigma_s^2\mathbf{I})$. Note that here and throughout this paper the conditional Gaussian distribution is represented as $N(a|b, c)$, where the distribution describes variable a with mean b and variance c.

The time-dependent signals of the system can be described by ordinary differential equations

$$\dot{x}_s(t) = \frac{dx_s(t)}{dt} = f_s(\mathbf{x}(t), \boldsymbol{\theta}_s, t). \tag{6}$$

We further define $\dot{\mathbf{x}}_s$ as the collection of the state derivatives across all timepoints for species s i.e. $\dot{\mathbf{x}}_s = (\dot{x}_s(t_1), \ldots, \dot{x}_s(t_T))^{\mathsf{T}}$ and $\mathbf{f}_s(\ldots, \mathbf{t})$ as the collection of ODE predictions across all timepoints for species s i.e. $\mathbf{f}_s(\ldots, \mathbf{t}) = (f_s(\ldots, t_1), \ldots, f_s(\ldots, t_T))^{\mathsf{T}}$. Hence,

$$\dot{\mathbf{x}}_s = \mathbf{f}_s(\mathbf{X}, \boldsymbol{\theta}_s, \mathbf{t}), \tag{7}$$

where $\mathbf{X}$ is defined below Eq. 5. Then,

$$p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\sigma}^2) = \prod_{s=1}^{N}\prod_{i=1}^{T} N(y_s(t_i)|x_s(t_i), \sigma_s^2), \tag{8}$$

where the dimension of the matrices $\mathbf{X}$ and $\mathbf{Y}$ are N by T. Following Calderhead et al. (2008), a Gaussian process (GP) prior is placed on $\mathbf{x}_s$,

$$p(\mathbf{x}_s|\boldsymbol{\phi}_s, \boldsymbol{\eta}) = N(\mathbf{x}_s|\boldsymbol{\phi}_s, \mathbf{K}_{\eta_s}), \tag{9}$$

where $\mathbf{K}_{\eta_s}$ is a positive definite matrix of covariance functions with hyperparameters η_s and $\boldsymbol{\phi}_s$ is a mean vector, which for simplicity we set as the mean of $\mathbf{Y}$.

Differentiation is a linear operation, and therefore a Gaussian process is closed under differentiation, provided the kernel is differentiable (see Solak et al. 2002; Holsclaw et al. 2013; Rasmussen and Williams 2006). Hence, the joint prior distribution of the concentrations of the species $\mathbf{x}_s$ and their time derivatives $\dot{\mathbf{x}}_s$ is multivariate Gaussian with mean $(\boldsymbol{\phi}_s, \mathbf{0})^\mathsf{T}$ and covariance functions

$$cov[x_s(t_i), x_s(t_j)] = K_{\eta_s}(t_i, t_j), \tag{10}$$

$$cov[\dot{x}_s(t_i), x_s(t_j)] = \frac{\partial K_{\eta_s}(t_i, t_j)}{\partial t_i} := K'_{\eta_s}(t_i, t_j), \tag{11}$$

$$cov[x_s(t_i), \dot{x}_i(t_j)] = \frac{\partial K_{\eta_s}(t_i, t_j)}{\partial t_j} := {}'K_{\eta_s}(t_i, t_j), \tag{12}$$

$$cov[\dot{x}_s(t_i), \dot{x}_s(t_j)] = \frac{\partial^2 K_{\eta_s}(t_i, t_j)}{\partial t_i \partial t_j} := K''_{\eta_s}(t_i, t_j), \tag{13}$$

where $K_{\eta_s}(t_i, t_j)$ are the components of the covariance matrix $\mathbf{K}_{\eta_s}$. The conditional distribution for the state derivatives is obtained using elementary transformations of Gaussian distributions (see p. 87 of Bishop 2006 for details), yielding

$$p(\dot{\mathbf{x}}_s|\mathbf{x}_s, \boldsymbol{\phi}_s, \boldsymbol{\eta}_s) = N(\dot{\mathbf{x}}_s|\boldsymbol{\mu}_s, \mathbf{A}_s), \tag{14}$$

where

$$\begin{aligned}\boldsymbol{\mu}_s &= {}'\mathbf{K}_{\eta_s}{\mathbf{K}_{\eta_s}}^{-1}(\mathbf{x}_s - \boldsymbol{\phi}_s) \quad \text{and} \\ \mathbf{A}_s &= \mathbf{K}''_{\eta_s} - {}'\mathbf{K}_{\eta_s}{\mathbf{K}_{\eta_s}}^{-1}\mathbf{K}'_{\eta_s}.\end{aligned} \tag{15}$$

Assuming the model for the gradients has additive Gaussian error, with a state-specific variance γ_s, using Eq. 6 gives

$$p(\dot{\mathbf{x}}_s|\mathbf{X}, \boldsymbol{\theta}_s, \gamma_s) = N(\dot{\mathbf{x}}_s|\mathbf{f}_s(\mathbf{X}, \boldsymbol{\theta}_s, \mathbf{t}), \gamma_s\mathbf{I}). \tag{16}$$

Using a product of experts approach, Calderhead et al. (2008) and Dondelinger et al. (2013) link the distribution of the interpolant in Eq. 14 with the distribution of the ODE derivatives in Eq. 16, obtaining the following distribution

$$\begin{aligned}&p(\dot{\mathbf{x}}_s|\mathbf{X}, \boldsymbol{\theta}_s, \boldsymbol{\phi}_s, \boldsymbol{\eta}_s, \gamma_s) \\ &\quad\propto p(\dot{\mathbf{x}}_s|\mathbf{x}_s, \boldsymbol{\phi}_s, \boldsymbol{\eta}_s)p(\dot{\mathbf{x}}_s|\mathbf{X}, \boldsymbol{\theta}_s, \gamma_s) \\ &\quad= N(\dot{\mathbf{x}}_s|\boldsymbol{\mu}_s, \mathbf{A}_s)N(\dot{\mathbf{x}}_s|\mathbf{f}_s(\mathbf{X}, \boldsymbol{\theta}_s, \mathbf{t}), \gamma_s\mathbf{I}).\end{aligned} \tag{17}$$

The joint distribution is given by

$$\begin{aligned}p(\dot{\mathbf{X}}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\gamma}|\boldsymbol{\phi}) &= p(\boldsymbol{\theta})p(\boldsymbol{\eta})p(\boldsymbol{\gamma}) \\ &\prod_{s=1}^{N} p(\dot{\mathbf{x}}_s|\mathbf{X}, \boldsymbol{\theta}_s, \boldsymbol{\phi}_s, \boldsymbol{\eta}_s, \gamma_s)p(\mathbf{x}_s|\boldsymbol{\eta}_s),\end{aligned} \tag{18}$$

where $\boldsymbol{\gamma}$ is the vector which contains all the gradient mismatch parameters and $p(\boldsymbol{\theta})$, $p(\boldsymbol{\eta})$, $p(\boldsymbol{\gamma})$ are the prior distributions over the respective parameters. Dondelinger et al. (2013) show that the marginalisation over the state derivatives yields a closed form solution

$$\begin{aligned}&p(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\gamma}|\boldsymbol{\phi}) \\ &\quad= \int p(\dot{\mathbf{X}}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\gamma}|\boldsymbol{\phi})d\dot{\mathbf{X}} \\ &\propto p(\boldsymbol{\theta})p(\boldsymbol{\eta})p(\boldsymbol{\gamma}) \prod_{s=1}^{N} N(\mathbf{x}_s|\mathbf{0}, \mathbf{K}_{\eta_s}) \\ &\quad\int N(\dot{\mathbf{x}}_s|\boldsymbol{\mu}_s, \mathbf{A}_s)N(\dot{\mathbf{x}}_s|\mathbf{f}_s(\mathbf{X}, \boldsymbol{\theta}_s, \mathbf{t}), \gamma_s\mathbf{I})d\dot{\mathbf{x}}_s \\ &\propto p(\boldsymbol{\theta})p(\boldsymbol{\eta})p(\boldsymbol{\gamma}) \prod_{s=1}^{N} N(\mathbf{x}_s|\mathbf{0}, \mathbf{K}_{\eta_s}) \\ &\quad\exp\left[-\frac{1}{2}(\mathbf{f}_s - \boldsymbol{\mu}_s)^T(\mathbf{A}_s + \gamma_s\mathbf{I})^{-1}(\mathbf{f}_s - \boldsymbol{\mu}_s)\right],\end{aligned} \tag{19}$$

where $\mathbf{f}_s$ is the vector containing the ODE predicted gradients for species s. Using Eq. 19 and the noise model in Eq. 8, the full joint distribution becomes

$$\begin{aligned}&p(\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \sigma^2|\boldsymbol{\phi}) \\ &\quad= p(\mathbf{Y}|\mathbf{X}, \sigma^2)p(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\gamma}|\boldsymbol{\phi})p(\sigma^2) \\ &\quad= p(\mathbf{Y}|\mathbf{X}, \sigma^2)p(\mathbf{X}|\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\eta}, \boldsymbol{\gamma})p(\boldsymbol{\theta})p(\boldsymbol{\eta})p(\boldsymbol{\gamma})p(\sigma^2),\end{aligned} \tag{20}$$

where $p(\sigma^2)$ is the prior over the variance of the observational error and from Eq. 19

$$\begin{aligned}p(\mathbf{X}|\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\eta}, \boldsymbol{\gamma}) \propto \frac{1}{C}\exp\Bigg[-\frac{1}{2} \\ \sum_{s=1}^{N}\left(\mathbf{x}_s^T\mathbf{K}_{\eta_s}^{-1}\mathbf{x}_s + (\mathbf{f}_s - \boldsymbol{\mu}_s)^T(\mathbf{A}_s + \gamma_s\mathbf{I})^{-1}(\mathbf{f}_s - \boldsymbol{\mu}_s)\right)\Bigg],\end{aligned} \tag{21}$$

where $C = \prod_{s=1}^{N} |2\pi(\mathbf{A}_s + \gamma_s\mathbf{I})|^{\frac{1}{2}}$.

Dondelinger et al. (2013) want to perform Bayesian inference using Eq. 20, but ODE models tend to produce multi-modal likelihood landscapes. In order to efficiently sample from the posterior distribution, and avoid getting stuck in local optima for example, the authors temper the likelihood.

Consider a series of "temperatures",[1] $0 = \alpha_1 < \cdots < \alpha_i < \cdots < \alpha_M = 1$ and a power posterior distribution (see

[1] Note that our "temperatures" are equivalent to inverse-temperatures in Statistical Physics. It is however common in the likelihood tempering literature (see Friel and Pettitt 2008; Campbell and Steele 2012; Dondelinger et al. 2013; Macdonald et al. 2016 for examples) that the "temperatures" are scheduled according to our notation.

Friel and Pettitt 2008 for more details) of the parameters and latent variables in Eq. 20

$$p_\alpha(\boldsymbol{\theta}, \mathbf{X}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \sigma^2|\mathbf{Y}, \boldsymbol{\phi}) \propto p(\mathbf{Y}|\mathbf{X}, \sigma^2)^\alpha p(\mathbf{X}|\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\eta}, \boldsymbol{\gamma}) p(\boldsymbol{\theta}) p(\boldsymbol{\eta}) p(\boldsymbol{\gamma}) p(\sigma^2), \tag{22}$$

where the first term on the second line in Eq. 22 is an annealed likelihood. It is clear that Eq. 22 becomes the prior for $\alpha = 0$ and is the posterior when $\alpha = 1$. For $0 < \alpha < 1$ a distribution between the prior and posterior is created. The M α_is are "temperature" parameters that create annealed likelihoods that are used as the target densities of parallel MCMC chains (see Campbell and Steele 2012). Denote $\boldsymbol{\Omega}$ as shorthand for the parameters and latent variables in Eq. 22 i.e. $\boldsymbol{\Omega} = \{\boldsymbol{\theta}, \mathbf{X}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \sigma^2\}$. At each MCMC step, all "temperature" chains independently perform a Metropolis–Hastings step to update $\boldsymbol{\Omega}$, the parameters and latent variables associated with "temperature" α. This has acceptance probability

$$p_{\text{move}} = \min\left(1, \frac{p\left(\mathbf{Y}|\boldsymbol{\Omega}^{\text{prop}}\right)^\alpha p\left(\boldsymbol{\Omega}^{\text{prop}}\right) q\left(\boldsymbol{\Omega}^{\text{curr}}|\boldsymbol{\Omega}^{\text{prop}}\right)}{p\left(\mathbf{Y}|\boldsymbol{\Omega}^{\text{curr}}\right)^\alpha p\left(\boldsymbol{\Omega}^{\text{curr}}\right) q\left(\boldsymbol{\Omega}^{\text{prop}}|\boldsymbol{\Omega}^{\text{curr}}\right)}\right), \tag{23}$$

where $q(\,)$ represents the proposal distribution and the superscripts "prop" and "curr" indicate whether the algorithm is being evaluated at the proposed or current state, respectively. At each MCMC step, two "temperature" chains are randomly selected (uniformly) and the corresponding parameters are proposed to swap between them. This proposal has acceptance probability

$$p_{\text{swap}} = \min\left(1, \frac{p_{\alpha_j}(\boldsymbol{\Omega}^i|\mathbf{Y}) p_{\alpha_i}(\boldsymbol{\Omega}^j|\mathbf{Y})}{p_{\alpha_i}(\boldsymbol{\Omega}^i|\mathbf{Y}) p_{\alpha_j}(\boldsymbol{\Omega}^j|\mathbf{Y})}\right), \tag{24}$$

where $\boldsymbol{\Omega}^i$ are the parameters and latent variables associated with "temperature" α_i and $\boldsymbol{\Omega}^j$ are the parameters and latent variables associated with "temperature" α_j.

3 Methodology

A primary challenge of model selection using Bayes factors is to calculate the marginal likelihood of a model. The Bayes factor is then computed by calculating the ratio of the marginal likelihoods, or the difference of the log marginal likelihoods, of the competing models. Computing the log marginal likelihood is challenging, however, and thermodynamic integration is widely regarded as one of the most promising approaches to its calculation in practice (Friel and Pettitt 2008). This gives a framework for the computation of the integral in Eq. 3, which is one of the main difficulties in practically performing explanatory model selection. Note that for the remainder of this paper the dependency on the particular model is not made explicit in the notation, for ease of reading, i.e. $p(\mathbf{Y}) = p(\mathbf{Y}|M)$.

Friel and Pettitt (2008) show that the log marginal likelihood can be computed by adapting the thermodynamic integration method used in Statistical Physics for computing free energies (see Schlitter and Husmeier 1992) and taking the derivative of $\log p(\mathbf{Y}|\alpha)$ with respect to the "temperatures" and then integrating over the "temperatures". In what follows, we show how this scheme can be adapted to the gradient matching method discussed in the present paper.

Based on Eq. 19, it is possible to write the joint probability of the latent variables and parameters as

$$p(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\gamma}) = \frac{\zeta(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\gamma}) p(\mathbf{X}|\boldsymbol{\eta}) p(\boldsymbol{\theta}) p(\boldsymbol{\eta}) p(\boldsymbol{\gamma})}{C}, \tag{25}$$

where $\zeta(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\gamma})$ is a potential function (an un-normalised probability distribution), defined by the last line in Eq. 19 ($\zeta(\cdot)$ here is being used as shorthand for the solution to the integral of Eq. 19), $p(\mathbf{X}|\boldsymbol{\eta})$ is the distribution of the Gaussian process with hyperparameters $\boldsymbol{\eta}$ and the normalisation constant C is defined as

$$C = \int_{\mathbf{X}} \int_{\boldsymbol{\theta}} \int_{\boldsymbol{\eta}} \int_{\boldsymbol{\gamma}} \zeta(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\gamma}) p(\mathbf{X}|\boldsymbol{\eta}) p(\boldsymbol{\theta}) p(\boldsymbol{\eta}) p(\boldsymbol{\gamma}) d\mathbf{X} d\boldsymbol{\theta} d\boldsymbol{\eta} d\boldsymbol{\gamma}. \tag{26}$$

The joint probability of the whole system now becomes

$$\begin{aligned} p(\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \sigma^2) &= p(\mathbf{Y}|\mathbf{X}, \sigma^2) p(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\gamma}) p(\sigma^2) \\ &= \frac{p(\mathbf{Y}|\mathbf{X}, \sigma^2) \zeta(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\gamma}) p(\mathbf{X}|\boldsymbol{\eta}) p(\boldsymbol{\theta}) p(\boldsymbol{\eta}) p(\boldsymbol{\gamma}) p(\sigma^2)}{C}, \end{aligned} \tag{27}$$

which therefore implies that the tempered posterior distribution of the latent variables and parameters is given by

$$p(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{\gamma}, \sigma^2|\mathbf{Y}, \alpha) = \frac{1}{\mathbb{Z}(\mathbf{Y}|\alpha)} \left[p(\mathbf{Y}|\mathbf{X}, \sigma^2) \zeta(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\gamma})\right]^\alpha p(\mathbf{X}|\boldsymbol{\eta}) p(\boldsymbol{\theta}) p(\boldsymbol{\eta}) p(\boldsymbol{\gamma}) p(\sigma^2), \tag{28}$$

and $\mathbb{Z}(\mathbf{Y}|\alpha)$ as

$$\mathbb{Z}(\mathbf{Y}|\alpha) = \int_{\mathbf{X}} \int_{\boldsymbol{\theta}} \int_{\boldsymbol{\eta}} \int_{\boldsymbol{\gamma}} \int_{\sigma^2} \left[p(\mathbf{Y}|\mathbf{X}, \sigma^2) \zeta(\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\gamma})\right]^\alpha p(\mathbf{X}|\boldsymbol{\eta}) p(\boldsymbol{\theta}) p(\boldsymbol{\eta}) p(\boldsymbol{\gamma}) p(\sigma^2) d\mathbf{X} d\boldsymbol{\theta} d\boldsymbol{\eta} d\boldsymbol{\gamma} d\sigma^2. \tag{29}$$

Taking the derivative of $\log \mathbb{Z}(\mathbf{Y}|\alpha)$ will yield

$$\begin{aligned}
&\frac{d}{d\alpha}\log \mathbb{Z}(\mathbf{Y}|\alpha)\\
&= \frac{1}{\mathbb{Z}(\mathbf{Y}|\alpha)}\frac{d}{d\alpha}\mathbb{Z}(\mathbf{Y}|\alpha)\\
&= \frac{1}{\mathbb{Z}(\mathbf{Y}|\alpha)}\frac{d}{d\alpha}\int_{\mathbf{X}}\int_{\boldsymbol{\theta}}\int_{\boldsymbol{\eta}}\int_{\boldsymbol{\gamma}}\int_{\sigma^2}\\
&\quad\left[p(\mathbf{Y}|\mathbf{X},\boldsymbol{\sigma}^2)\zeta(\mathbf{X},\boldsymbol{\theta},\boldsymbol{\gamma})\right]^{\alpha} p(\mathbf{X}|\boldsymbol{\eta})p(\boldsymbol{\theta})p(\boldsymbol{\eta})p(\boldsymbol{\gamma})\\
&\quad p(\boldsymbol{\sigma}^2)d\mathbf{X}d\boldsymbol{\theta}d\boldsymbol{\eta}d\boldsymbol{\gamma}d\boldsymbol{\sigma}^2\\
&= \frac{1}{\mathbb{Z}(\mathbf{Y}|\alpha)}\int_{\mathbf{X}}\int_{\boldsymbol{\theta}}\int_{\boldsymbol{\eta}}\int_{\boldsymbol{\gamma}}\int_{\sigma^2}\log\left[p(\mathbf{Y}|\mathbf{X},\boldsymbol{\sigma}^2)\zeta(\mathbf{X},\boldsymbol{\theta},\boldsymbol{\gamma})\right]\\
&\quad\left[p(\mathbf{Y}|\mathbf{X},\boldsymbol{\sigma}^2)\zeta(\mathbf{X},\boldsymbol{\theta},\boldsymbol{\gamma})\right]^{\alpha} p(\mathbf{X}|\boldsymbol{\eta})p(\boldsymbol{\theta})p(\boldsymbol{\eta})p(\boldsymbol{\gamma})\\
&\quad p(\boldsymbol{\sigma}^2)d\mathbf{X}d\boldsymbol{\theta}d\boldsymbol{\eta}d\boldsymbol{\gamma}d\boldsymbol{\sigma}^2\\
&= \int_{\mathbf{X}}\int_{\boldsymbol{\theta}}\int_{\boldsymbol{\eta}}\int_{\boldsymbol{\gamma}}\int_{\sigma^2}\log\left[p(\mathbf{Y}|\mathbf{X},\boldsymbol{\sigma}^2)\zeta(\mathbf{X},\boldsymbol{\theta},\boldsymbol{\gamma})\right]\frac{1}{\mathbb{Z}(\mathbf{Y}|\alpha)}\\
&\quad\left[p(\mathbf{Y}|\mathbf{X},\boldsymbol{\sigma}^2)\zeta(\mathbf{X},\boldsymbol{\theta},\boldsymbol{\gamma})\right]^{\alpha} p(\mathbf{X}|\boldsymbol{\eta})p(\boldsymbol{\theta})p(\boldsymbol{\eta})p(\boldsymbol{\gamma})\\
&\quad p(\boldsymbol{\sigma}^2)d\mathbf{X}d\boldsymbol{\theta}d\boldsymbol{\eta}d\boldsymbol{\gamma}d\boldsymbol{\sigma}^2\\
&= \int_{\mathbf{X}}\int_{\boldsymbol{\theta}}\int_{\boldsymbol{\eta}}\int_{\boldsymbol{\gamma}}\int_{\sigma^2}\log\left[p(\mathbf{Y}|\mathbf{X},\boldsymbol{\sigma}^2)\zeta(\mathbf{X},\boldsymbol{\theta},\boldsymbol{\gamma})\right]\\
&\quad p(\mathbf{X},\boldsymbol{\theta},\boldsymbol{\eta},\boldsymbol{\gamma},\boldsymbol{\sigma}^2|\mathbf{Y},\alpha)d\mathbf{X}d\boldsymbol{\theta}d\boldsymbol{\eta}d\boldsymbol{\gamma}d\boldsymbol{\sigma}^2\\
&= \mathbb{E}_{\alpha}\left[\log p(\mathbf{Y}|\mathbf{X},\boldsymbol{\sigma}^2)\right] + \mathbb{E}_{\alpha}\left[\log\zeta(\mathbf{X},\boldsymbol{\theta},\boldsymbol{\gamma})\right]. \qquad (30)
\end{aligned}$$

where the posterior distribution in the second last step of Eq. 30 comes from Eq. 28. This in turn means that

$$\begin{aligned}
\log \mathbb{Z}(\mathbf{Y}) &= \log \mathbb{Z}(\mathbf{Y}|\alpha=1) - \log \mathbb{Z}(\mathbf{Y}|\alpha=0)\\
&= \int_0^1 \frac{d}{d\alpha}\log \mathbb{Z}(\mathbf{Y}|\alpha)d\alpha\\
&= \int_0^1 \mathbb{E}_{\alpha}\left[\log p(\mathbf{Y}|\mathbf{X},\boldsymbol{\sigma}^2)\right]d\alpha\\
&\quad + \int_0^1 \mathbb{E}_{\alpha}\left[\log\zeta(\mathbf{X},\boldsymbol{\theta},\boldsymbol{\gamma})\right]d\alpha, \qquad (31)
\end{aligned}$$

where $\mathbb{Z}(\mathbf{Y}|\alpha=0)=1$ follows from Eq. 29. Note then, that $\mathbb{Z}(\mathbf{Y})=\mathbb{Z}(\mathbf{Y}|\alpha=1)$ and by setting $\alpha=1$ in Eq. 29, it follows from Eq. 25 that $\mathbb{Z}(\mathbf{Y})=p(\mathbf{Y})C$. Hence,

$$\log p(\mathbf{Y}) = \log \mathbb{Z}(\mathbf{Y}) - \log(C). \qquad (32)$$

C can depend on the ODE model structure, but it does not depend on the data. This term can be estimated even before the data are collected, in order to speed up the whole process. We would assume that the integrand for C would be a lot smoother than for the likelihood, and so we try to approximate Eq. 26 using a simple Monte Carlo sum i.e.

$$C = \frac{1}{N_{iter}}\sum_{i=1}^{N_{iter}}\zeta(\mathbf{X}_i,\boldsymbol{\theta}_i,\boldsymbol{\gamma}_i), \qquad (33)$$

where the draws required to compute $\zeta(\mathbf{X}_i,\boldsymbol{\theta}_i,\boldsymbol{\gamma}_i)$ are sampled from the priors $p(\boldsymbol{\eta})$, $p(\boldsymbol{\gamma})$, $p(\boldsymbol{\theta})$ and $p(\mathbf{X}|\boldsymbol{\eta})$, with acceptance probability 1. In the examples looked at in Sect. 5, the simple Monte Carlo sum was quick to converge and thus Eq. 33 was used to compute C. For ODE models that cause the integrand for C to be unsuitable for approximation using a simple Monte Carlo sum, it is possible to compute $\log(C)$ using thermodynamic integration. We have included this description in Section 9.3 of the supplementary material (SM).

We finally note that there is an alternative version of the thermodynamic integration scheme that we have proposed in Sect. 3, which we describe in Section 9.2 of the SM and which, naively, appears to be more straightforward. We discuss, in the same section of the SM, the disadvantages of that scheme, and justify the choice of the present scheme.

4 Benchmark ODE systems

The ODE systems used as benchmark models throughout this paper are detailed in this section. Details of the specific parameter setting used to simulate data for a particular set-up, can be found in Sect. 5.

4.1 The Lotka–Volterra system (LV)

The Lotka–Volterra system is a simple model for prey–predator interactions in ecology (Lotka 1932), and autocatalysis in chemical kinetics (Atkins 1986). The standard model, which we refer to as LV1, is given by Eq. 34 without the quadratic decay term, $\theta_5=0$.

$$\begin{aligned}
\frac{d[x_1]}{dt} &= [\dot{x}_1] = \theta_1[x_1] - \theta_2[x_1][x_2] - \theta_5[x_1]^2\\
\frac{d[x_2]}{dt} &= [\dot{x}_2] = -\theta_3[x_2] + \theta_4[x_1][x_2] \qquad (34)
\end{aligned}$$

Here, $[x_1]$ and $[x_2]$ denote the time-dependent concentrations of two species, prey (x_1) and predator (x_2).

Equation 34 has one extra parameter than the standard form, θ_5, to account for intra-species competition; we refer to this model as LV2. The most complex version, referred to as LV3 and given by Eq. 35, is described using a saturation term (similar to a Michaelis–Menten term that can appear in biological systems described by chemical kinetics).

$$\frac{d[x_1]}{dt} = [\dot{x}_1] = \theta_1[x_1] - \frac{\theta_2[x_1][x_2]}{1+\theta_5[x_1]}$$

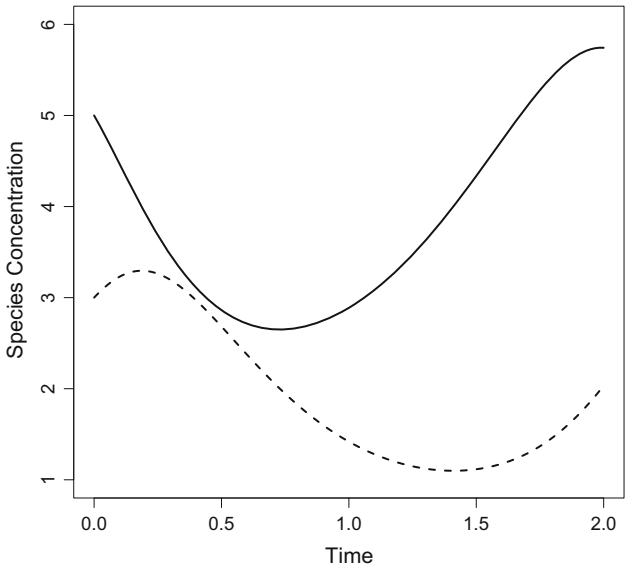

Fig. 1 An example of the signals produced from the Lotka–Volterra model (Eq. 34 with $\theta_5 = 0$). The solid line is $[x_1]$ (abundance of prey) and the dashed line is $[x_2]$ (abundance of predators)

$$\frac{d[x_2]}{dt} = [\dot{x}_2] = -\theta_3[x_2] + \frac{\theta_4[x_1][x_2]}{1 + \theta_5[x_1]} \tag{35}$$

An example of the signals produced from the standard Lotka–Volterra model (LV1, given by Eq. 34 with $\theta_5 = 0$) can be found in Fig. 1.

4.2 Protein signalling transduction pathways (PSTPs)

These equations describe protein signalling transduction pathways in a signal transduction cascade (Vyshemirsky and Girolami 2008), where the parameters indicate the kinetic rates of the reactions. There are 6 parameters ($k_1, k_2, k_3, k_4, V, K_m$) and 5 "species" ($S, S^*, R, RS, Rpp$). The system describes the phosphorylation of a protein, $R \rightarrow Rpp$, catalysed by an enzyme S, via an active protein complex (RS), where the enzyme is subject to degradation ($S \rightarrow S^*$). The chemical kinetics are described by a combination of mass action kinetics and Michaelis–Menten kinetics. A graphical representation of this system can be seen in the left panel of Fig. 2, and is referred to as PSTP1. Species in [ ] denote the time-dependent concentration for that species and a dot over a symbol is shorthand for the temporal derivative $\frac{d}{dt}$ of that symbol:

$$\begin{aligned}
[\dot{S}] &= -k_1[S] - k_2[S][R] + k_3[RS] \\
[\dot{S^*}] &= k_1[S] \\
[\dot{R}] &= -k_2[S][R] + k_3[RS] + \frac{V[Rpp]}{K_m + [Rpp]} \\
[\dot{RS}] &= k_2[S][R] - k_3[RS] - k_4[RS] \\
[\dot{Rpp}] &= k_4[RS] - \frac{V[Rpp]}{K_m + [Rpp]}
\end{aligned} \tag{36}$$

An example of the signals produced from these ODEs can be found in Fig. 3.

The following are different alternative candidate models of the protein signalling transduction pathway, all with varying degrees of complexity.

Equation 37 is a simplified version of Eq. 36, where now a less detailed description of the activation process is considered. It uses Michaelis–Menten kinetics to describe the phosphorylation of protein R and no longer has an intermediate complex RS. A graphical representation of this system can be seen in the top centre panel of Fig. 2, and we refer to it as PSTP2.

$$\begin{aligned}
[\dot{S}] &= -k_1[S], \quad [\dot{S^*}] = k_1[S] \\
[\dot{R}] &= \frac{-V_1[R][S]}{k_2 + [R]} + \frac{V_2[Rpp]}{k_3 + [Rpp]} \\
[\dot{Rpp}] &= \frac{V_1[R][S]}{k_2 + [R]} - \frac{V_2[Rpp]}{k_3 + [Rpp]}
\end{aligned} \tag{37}$$

The least complex form of the pathway is given when there is no degradation of protein S to S^*, and hence the concentration of S is constant. Mathematically, this corresponds to setting $k_1 = 0$ in Eq. 37. A graphical representation of this system, which we refer to as PSTP3, can be seen in the bottom centre panel of Fig. 2.

Equation 38 is the most complex of the candidate models, it describes how the phosphatase PhA deactivates the protein Rpp. All reactions are defined by mass action kinetics. A graphical representation of this system can be seen in the right panel of Fig. 2, and we refer to it as PSTP4.

$$\begin{aligned}
[\dot{S}] &= -k_1[S] - k_2[S][R] + k_3[RS] \\
[\dot{S^*}] &= k_1[S] \\
[\dot{R}] &= -k_2[S][R] + k_3[RS] + k_7[RppPhA] \\
[\dot{RS}] &= k_2[S][R] - k_3[RS] - k_4[RS] \\
[\dot{Rpp}] &= k_4[RS] - k_5[Rpp][PhA] + k_6[RppPhA] \\
[\dot{PhA}] &= -k_5[Rpp][PhA] + k_6[RppPhA] \\
&\quad + k_7[RppPhA] \\
[\dot{RppPhA}] &= k_5[Rpp][PhA] - k_6[RppPhA] \\
&\quad - k_7[RppPhA]
\end{aligned} \tag{38}$$

5 Simulation

The proposed method was tested on data generated from each of the LV models and from the PSTP1 model. Ten datasets

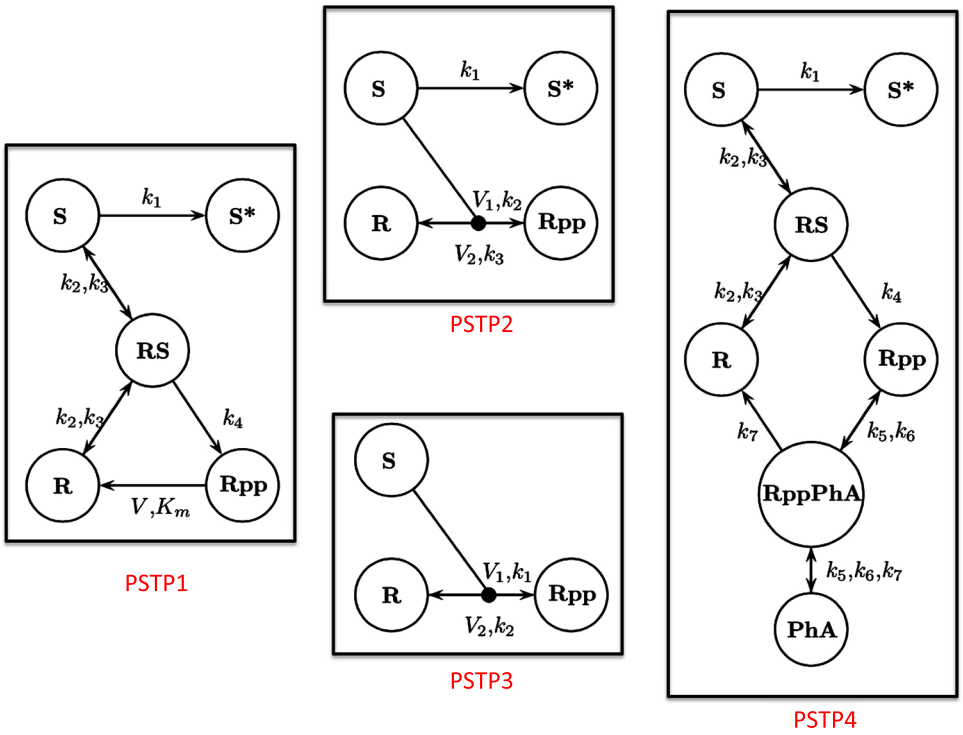

Fig. 2 **PSTP1** Graphical representation of the protein signalling transduction pathway in Eq. 36. There are 5 "species" (S, S^*, R, RS, Rpp) and 6 parameters ($k_1, k_2, k_3, k_4, V, K_m$). The system describes the phosphorylation of a protein, $R \rightarrow Rpp$, catalysed by an enzyme S, via an active protein complex (RS), where the enzyme is subject to degradation ($S \rightarrow S^*$). **PSTP2** Graphical representation of the protein signalling transduction pathway in Eq. 37. This is a simplified version of Eq. 36, where now a more general description of the activation process is considered. There are 4 "species" (S, S^*, R, Rpp) and 5 parameters (k_1, k_2, k_3, V_1, V_2). **PSTP3** Graphical representation of the protein signalling transduction pathway in Eq. 37 when setting $k_1 = 0$. It is the least complex of the candidate models and does not describe the degradation of protein S to S^*. There are 3 "species" (S, R, Rpp) and 4 parameters (k_1, k_2, V_1, V_2). **PSTP4** Graphical representation of the protein signalling transduction pathway in Eq. 38. The most complex of the candidate models, it describes how the phosphatase PhA deactivates the protein Rpp. There are 7 "species" (S, S^*, R, RS, Rpp, $RppPhA$, PhA) and 7 parameters ($k_1, k_2, k_3, k_4, k_5, k_6, k_7$). Figures adapted from Vyshemirsky and Girolami (2008)

were generated from each model in turn and iid Gaussian noise was added with a standard deviation (SD) chosen such that the average signal to noise ratio (SNR) was 10 (which we call medium SNR throughout). All simulations were then repeated with two additional SNRs, a lower value, and a higher value. The choice of these values is discussed in Section 9.4 of the supplementary material.

5.1 Lotka–Volterra original model (LV1)

Data were generated with the following parameters: $\theta_1 = 2$, $\theta_2 = 1$, $\theta_3 = 4$ and $\theta_4 = 1$. Starting from initial values of $\mathbf{x}(t = 0) = (5, 3)$ for the two "species", 11 timepoints were generated over the time course [0, 2], producing one period. Gaussian noise was added with SD $= 0.5$. The priors over the parameters were $\Gamma(4, 0.5)$ distributions, reflecting our prior knowledge that the parameters are positive. These settings were chosen to correspond with the set-up in Dondelinger et al. (2013).

5.2 Lotka–Volterra intra-species competition model (LV2)

Data were generated with the following parameters: $\theta_1 = 4$, $\theta_2 = 1$, $\theta_3 = 4$, $\theta_4 = 2$ and $\theta_5 = 5$. Starting from initial values of $\mathbf{x}(t = 0) = (5, 3)$ for the two "species", 11 timepoints were generated over the time course [0,2], producing one period. Gaussian noise was added with SD $= 0.2$. The priors over the parameters were $\Gamma(4, 0.5)$ distributions for θ_1, θ_2, θ_3 and θ_4 and a $U(0, 9)$ distribution for θ_5 as there was no indication from previous work what a suitable prior distribution would be for the parameter governing the intra-species term. These distributions reflect our prior knowledge that the parameters are positive.

5.3 Lotka–Volterra saturation term model (LV3)

Data were generated with the following parameters: $\theta_1 = 2.8$, $\theta_2 = 3.5$, $\theta_3 = 1$, $\theta_4 = 2.5$ and $\theta_5 = 1$. Starting from

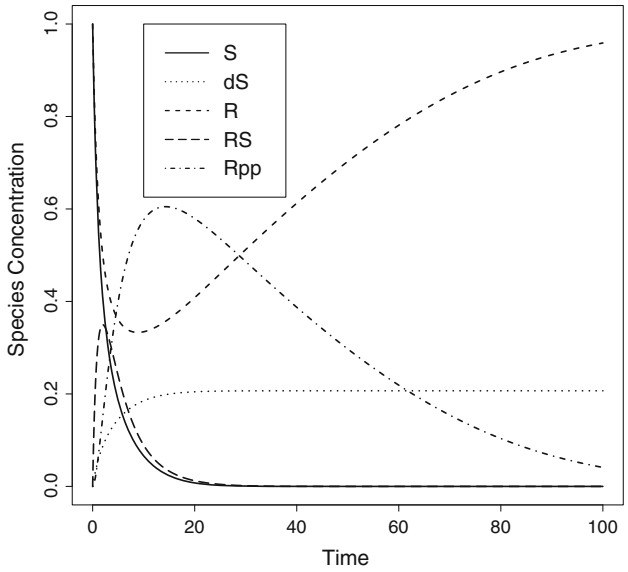

Fig. 3 An example of the signals produced from the protein signalling transduction pathway in Eq. 36. The solid line is $[S]$, the light dotted line is $[S^*]$, the dashed line near the top of the figure is $[R]$, the longer dashed line near the bottom of the figure is $[RS]$ and the dot-dashed line is $[Rpp]$

initial values of $\mathbf{x}(t = 0) = (5, 3)$ for the two "species", 11 timepoints were generated over the time course [0,2], producing one period. Gaussian noise was added with SD = 0.5. The saturation term included in these ODEs should mean that the less complex models are unable to produce signals that match the shape of the signals produced by the LV3 model. Hence, if the model selection method is working properly, this model should be clearly favoured over the other two. The priors over the parameters were $\Gamma(4, 0.5)$ distributions for $\theta_1, \theta_2, \theta_3$ and θ_4 and a $U(0, 9)$ distribution for θ_5 (reflecting the extra uncertainty surrounding the 5^{th} parameter). These distributions reflect our prior knowledge that the parameters are positive.

5.4 Protein signalling transduction pathway

For ease of reading, Eqs. 36–38 will be referred to as PSTP1, PSTP2, PSTP3 and PSTP4, respectively. Graphical representations of these models can be found in Fig. 2. Data were generated from the PSTP1 model as it provided a reasonable degree of complexity and was neither the least complex model nor the most complex model out of the four (to rule out a scenario where a method may be biased towards the least/most complex model, selecting the true model by fluke). 10 datasets were generated and iid Gaussian noise (SD = 0.0635, average SNR for each "species" = 10) was added. Since the PSTP1 model has two fewer "species" than the most complex model (PSTP4), generating data from the PSTP1 system will not produce observations for the components PhA and RppPhA (the two "species" not present in the PSTP1 model). Hence, for the less complex models (PSTP1–3), any "species" not present that are present in the PSTP4 model, had a zero rate of change included for that "species", corresponding to components that are disconnected from the rest of the system. After generating data from the PSTP1 model and adding Gaussian noise, the concentrations for PhA and RppPhA can be thought of as fluctuating slightly around their initial values, since they have a zero rate of change.

Data were generated with the following parameters: $k_1 = 0.07$, $k_2 = 0.6$, $k_3 = 0.05$, $k_4 = 0.3$, $V = 0.017$ and $K_m = 0.3$. Starting from initial values of $\mathbf{x}(t = 0) = (1, 0, 1, 0, 0, 1, 0)$ for the seven "species", 15 timepoints were generated, one at each of the following points $\{0, 1, 2, 4, 5, 7, 10, 15, 20, 30, 40, 50, 60, 80, 100\}$. The priors over the parameters were $\Gamma(4, 0.5)$ distributions, reflecting our prior knowledge that the parameters are positive. These settings were chosen to correspond with the set-up in Dondelinger et al. (2013).

5.5 Other settings

Two kernels were considered in this study (to correspond with simulation experiments that have been set up in the current literature e.g. see Dondelinger et al. 2013), the radial basis function (RBF) kernel

$$k(t_i, t_j) = \sigma^2_{\text{RBF}} \exp\left(-\frac{(t_i - t_j)^2}{2l^2}\right) \tag{39}$$

with hyperparameters σ^2_{RBF} and l^2, and the sigmoid variance kernel

$$k(t_i, t_j) = \sigma^2_{\text{sig}} \arcsin \frac{a + (bt_i t_j)}{\sqrt{(a + (bt_i t_i) + 1)(a + (bt_j t_j) + 1)}} \tag{40}$$

with hyperparameters σ^2_{sig}, a and b (see Rasmussen and Williams 2006).

The RBF kernel, Eq. 39, was used to fit the Gaussian process for the Lotka–Volterra models, and the sigmoid variance kernel, Eq. 40, was used to fit the Gaussian process for the protein signalling transduction pathway models. Examples of the signals produced from the PSTP1 model can be found in Fig. 3. Since the signals are non-stationary, we use the sigmoid variance kernel when fitting all the protein signalling transduction pathway models, since it is a non-stationary kernel (Rasmussen and Williams 2006). The initial fits from the GPs using the specified kernels were plotted against the data and showed good agreement.

We have followed Wang and Barber (2014) and a priori restricted the range of the latent species concentration profiles $\mathbf{x}_s$ to a three-standard deviation width around an initial GP interpolant fitted to the data. The motivation is

two-fold. Firstly, it is a priori unlikely that the true concentration profiles are in drastic mismatch to the data. Secondly, this prior constraint avoids entrapment in a suboptimal attractor state corresponding to a flat latent profile of $\mathbf{x}_s = \mathbf{0}$ for all species s, leading to Eq. 21 becoming the dominating term in the calculation of the joint distribution. This phenomenon is discussed in Macdonald et al. (2016), and we also briefly discuss it in Section 9.5 of the supplementary material. In addition, the standard deviation of the noise was held at the true value and subsequent draws of the latent variables were constrained to lie within a region of 3 standard deviations around the initial GP interpolants. This was enforced for every scenario of every candidate model used to compute Eq. 31 in this paper. Following Wang and Barber (2014) and Macdonald et al. (2015) sampling from the posterior distribution was conducted via Gibbs sampling with a discretisation of the parameters and latent variables. Further details, including pseudo code, are provided in Section 9.6 of the supplementary material. When sampling from the posterior using the method that explicitly solves the ODEs (see Sect. 6 for details), the Metropolis–Hastings algorithm was implemented due to the availability of existing software. The difference in sampling schemes does not affect the conclusion of the results, since the method that explicitly solves the ODEs was used as benchmark and was not a competing method. Since we implemented Gibbs sampling by discretising the parameters and latent variables (in order to calculate the normalisation constant of the posterior by brute force marginalisation), we would expect that switching to the Metropolis–Hastings scheme (which does not require any discretisation) would only improve our results. The Gibbs sampling strategy was employed to save on computational resources.

In order to ensure the integrals in Eq. 31 are calculated to a high level of accuracy, the number of "temperature" chains was set to 20. Following Calderhead (2008), each α_i was configured as

$$\alpha_i = \left(\frac{i}{M}\right)^5, \tag{41}$$

where $i = 1, \ldots, M$ and M is the total number of "temperature" chains (20 in this case).

6 Results

In order to assess the performance of the new scheme outlined in Sect. 3, the method will be tested on two ODE systems and various candidate models of each. For comparison purposes, the results of BIC and WAIC (Watanabe 2010) will also be provided. For details on how we computed these two information criteria, see Section 9.1 of the supplementary material. There are two possible ways of defining successful model selection. (1) How well the results match the marginal likelihood scores computed using a method that explicitly solves the ODEs (since parameter inference with gradient matching additionally approximates the solution of the ODEs using an interpolant). We refer to this as the 'exact' method. (2) How often a method selects the model the data were simulated from, i.e. the network reconstruction accuracy. Both of these benchmarks will be considered and discussed.

Assuming an iid sample from a binomial distribution of sample size n, we get a standard error of $\sqrt{p(1-p)/n} = \sqrt{(1/3)(2/3)/10} = 0.15$ for the LV model (3 alternative models) and $\sqrt{p(1-p)/n} = \sqrt{(1/4)(3/4)/10} = 0.14$ for the PSTP model (4 alternative models), we can get a simple estimate of the confidence level by approximating the upper tail of the binomial distribution with a normal distribution, and calculating the standard 95% confidence interval (4-standard deviation width). This gives a critical upper significance threshold of $0.33 + 2(0.15) = 0.63$ for the LV model, and $0.25 + 2(0.14) = 0.53$ for the PSTP model. The probabilities of selecting the model the data were generated from are shown, for medium SNR values in Table 1, for low SNR values, in Table 2, and high SNR values, in Table 3. Due to space restrictions we relegate all tables showing the probability of selecting any proposed model and all figures to the supplementary material.

6.1 Lotka–Volterra original model (LV1)

By examining Tables 1, 2 and 3, we can see that the 'exact' method consistently favours the model the data were generated from for all noise levels. The proposed method also consistently favours the model the data were generated from for all noise levels. Both BIC and WAIC significantly favour the model the data were generated from, but less often than the proposed method. The percentages of the time any of the models were favoured by a particular model selection method, as well as graphical representations of the results, can be found in Tables 5–7 and Figs. 4–15 in the supplementary material.

6.2 Lotka–Volterra intra-species competition model (LV2)

The second row of Table 1 contains the probabilities for how often a model selection method favoured the data generating model, for medium SNR and when that model was LV2.

At first glance, it would appear that model selection using the method proposed in this paper, BIC and WAIC are not able to select the correct model for this scenario, since none of the methods significantly favour the LV2 model. However, the settings to simulate data from this system were arbitrarily chosen, since there was no indication from previous work

Table 1 The probability of selecting the model the data were generated from, for medium SNR values

Model	'Exact' method	Proposed method	BIC	WAIC
LV, LV1 true	**1.0**	**1.0**	**0.7**	**1.0**
LV, LV2 true	**1.0**	0.1	0.0	0.6
LV, LV2 true, shorter time domain	–	**0.8**	0.6	**0.8**
LV, LV2 true, stronger intra-species effect	**1.0**	**0.7**	0.0	0.3
LV, LV3 true	**1.0**	**1.0**	0.5	**0.8**
PSTP, PSTP1 true	**0.7**	**1.0**	0.0	0.1

Significant results are shown in bold-face fonts

Table 2 The probability of selecting the model the data were generated from, for low SNR values

Model	'Exact' method	Proposed method	BIC	WAIC
LV, LV1 true	**1.0**	**1.0**	**0.8**	**0.7**
LV, LV2 true, shorter time domain	–	0.0	0.1	0.5
LV, LV2 true, stronger intra-species effect	**1.0**	**0.8**	0.0	0.6
LV, LV3 true	**1.0**	**0.8**	**0.8**	0.6
PSTP, PSTP1 true	**0.6**	**1.0**	0.0	0.1

Significant results are shown in bold-face fonts

Table 3 The probability of selecting the model the data were generated from, for high SNR values

Data	'Exact' method	Proposed method	BIC	WAIC
LV, LV1 true	**1.0**	**1.0**	0.1	**1.0**
LV, LV2 true, shorter time domain	–	**1.0**	**0.8**	**1.0**
LV, LV2 true, stronger intra-species effect	**1.0**	0.0	0.0	0.0
LV, LV3 true	**1.0**	**1.0**	**1.0**	**1.0**
PSTP, PSTP1 true	**1.0**	**1.0**	**0.7**	**0.6**

Significant results are shown in bold-face fonts

what suitable values should be. Upon further consideration of the set-up, two issues were noted. First, the time domain over which the signals are observed was set to [0, 2], as this was consistent with simulation studies that generated data from the LV1 model. For the LV2 model, the particular parameters chosen produce signals that rapidly decrease over a short time domain ([0, 0.3]) and then plateau for the remaining time domain. Since only a small number of observations are available for these systems, only 1–2 observations were generated in the domain where the signal decreases i.e. where most of the information about the signals lies. For the 'exact' method, this does not seem to be much of an issue, but for gradient matching, having the vast majority of observations lie in the plateaued region can lead to too much information loss (gradients that are effectively zero across most of the time domain) and the results deteriorate. In order to test and demonstrate this, gradient matching was conducted using the same parameters, number of observations and a smaller time domain—only considering the time domain [0, 0.3]. This is something that could be applied in practice, should it be known that the time domain is too large to produce signals that are sufficiently informative. For this scenario, the sigmoid variance kernel was used, since the RBF kernel was unable to properly model the rapid change in signal concentration. The 'exact' method was not repeated for this scenario, as it did not struggle with the original time domain and we expect it to perform as well for LV2 with the shorter time domain.

Second, an inspection of the structure of the LV2 model tells us that the parameters were poorly chosen. When θ_5 is large, the component will decrease x_1. However, θ_5 could be set to zero and θ_2 could made large and again x_1 would decrease. Hence, the LV1 model has a term that is able to affect the signals in a way very much the same as the LV2 model, without the need for an extra parameter. This essentially makes the intra-species component weakly identifiable. Gradient matching seems to be more affected by this than an explicit solution of the ODEs, since it is an approximate method. The methods were tested again on data generated when θ_5 was more substantial. To this end, data were generated with the following parameters; $\theta_1 = 100$, $\theta_2 = 0.1$, $\theta_3 = 4$, $\theta_4 = 0.1$ and $\theta_5 = 10$. The effect this has on the system is that for x_1, this "species" concentration rises exponentially and then plateaus, since the intra-species competition term stops the population increasing without end. The LV1 model should not be able to replicate this because

concentrations for x_2 go to zero. Hence, the LV1 model should not have a way to regulate the population concentration and get good agreement with the data. For this set-up, iid Gaussian noise of SD $= 0.1414$ was added to each "species" (average SNR for each "species"$= 10$). The priors over the parameters were $\Gamma(4, 0.5)$ for θ_2, θ_3 and θ_4, $U(0, 110)$ for θ_1 and a $U(2, 11)$ for θ_5 (reflecting the extra uncertainty of these two parameters). The time domain was chosen to be $[0, 2]$. Since the dependency on θ_5 is higher, this should compensate for the information lost by the plateaued signal, allowing us to see that gradient matching can in principle deal with these types of signals, should the observed signals be informative enough.

Since we know that gradient matching cannot deal with the signals produced using the original settings we used to generate data from the LV2 model, we do not present results for low SNRs and high SNRs for the LV2 original settings scenario.

Now examining Tables 1, 2 and 3, we can see how often a model selection method favoured selecting the LV2 model when data are generated over a shorter time domain. For medium SNR, the proposed method and WAIC perform as well as one another, significantly favouring the LV2 model, and outperform BIC, which doesn't. For high SNR, all methods significantly favour the LV2 model, with the proposed method and WAIC favouring it more often. All methods fail in the low SNR case. This failure is likely a consequence of the fact that the interpolants get noisier, making the derivatives less reliable (error present in a signal is amplified when modelling the derivatives).

For the LV2 stronger intra-species effect scenario, the 'exact' method consistently favours the model the data were generated from for all SNRs. The proposed method significantly favours the LV2 model (in agreement with the 'exact' method) for low and medium SNRs, outperforming both BIC and WAIC, which do not significantly favour the LV2 model. Rather strangely, for the high SNR (lowest noise) scenario, none of the gradient matching based model selection methods (proposed method, BIC and WAIC) significantly favour the LV2 model. A closer inspection reveals that the cause of this is the prior distribution in function space, which forces functions to lie within a 3-standard deviation width around the initial interpolant (this width was discussed in Sect. 5.5 and more details can be found in Section 9.5 of the SM). Larger SNRs increase the tightness of the prior bound. For LV2 with a stronger intra-species interaction, the prior is so informative that, irrespective of the underlying model, all data fit terms are always high. Choosing the true model can thus only achieve little gain in goodness of data fit, which does not compensate the higher Bayesian penalty for model complexity. Consequently, all gradient matching methods consistently select the least complex model with the lowest complexity penalty in this case.

It is important to note that when the 'exact' method was applied to the LV2 stronger intra-species effect scenario, the initial conditions of the ODEs were fixed at their true values, since the solver encountered numerical instabilities. For this choice of parameter settings the differential equations are stiff, which drives the step size of the solver to small values below machine precision, causing the software to crash. This is discussed in the literature, for example see pp. 45–47 of Soetaert et al. (2010). It is also worth pointing out that various solvers were used (`euler`, `lsoda`, `ode23`, `ode45`, `vode`) and this issue was still present. In order to avoid this problem, the initial conditions were held fixed at the true values for the 'exact' method. Note that this information would not be available in practice and that the gradient matching approach does not require any initial conditions—a practical benefit that gradient matching has over the explicit approach.

The percentages of the time any of the models were favoured by a particular model selection method, as well as graphical representations of the results, can be found in Tables 8–14 and Figs. 16–40 in the supplementary material.

6.3 Lotka–Volterra saturation term model (LV3)

By examining Tables 1, 2 and 3, we can see that the 'exact' method consistently favours the model the data were generated from, for all SNRs. The proposed method significantly selects the LV3 model (in agreement with the 'exact' method) for all SNRs, outperforming BIC for medium SNR and WAIC for low SNR (where these methods do not significantly favour the LV3 model). BIC and WAIC significantly favour the LV3 model for the other SNRs. The percentages of the time any of the models were favoured by a particular model selection method, as well as graphical representations of the results, can be found in Tables 15–17 and Figs. 41–52 in the supplementary material.

6.4 Protein signalling transduction pathway Model 1 (PSTP1)

Tables 1, 2 and 3 show that the 'exact' method significantly favours the model the data were generated from for all SNRs. The proposed method also significantly favours the model the data were generated from for all SNRs and outperforms both BIC and WAIC for SNR low and SNR medium (where these methods do not significantly favour the PSTP1 model). All methods favour the PSTP1 model for SNR high. The percentages of the time any of the models were favoured by a particular model selection method, as well as graphical representations of the results, can be found in Tables 19–21 and Figs. 53–66 in the supplementary material.

Table 4 Model selection scores (to 1 decimal place), for the real data observations of lynx and hare populations. Bold values indicate the model most favoured by a method

Method	LV1	LV2	LV3
Log marginal likelihood, 'exact' method (higher is better)	− 5955.1	− **4675.1**	− 17,486.6
Log marginal likelihood, proposed method (higher is better)	69.9	**78.2**	71.4
BIC (lower is better)	− 154.2	− 153.6	− **156.0**
WAIC (lower is better)	− **126.9**	− 125.1	− 126.0

7 Application to real data

We have applied the Lotka–Volterra model and its two variants, models LV1, LV2 and LV3, to the hare–lynx time series recorded by the Hudson Bay Company. These time series show the annual abundance of the snowshoe hare *Lepus americanus* and the Canadian lynx *Lynx canadensis* in the boreal forest of North America, as measured by pelts collected by the Hudson Bay Company between 1900 and 1920. We took the data from Table 2.6 in Howard (http://www.math.tamu.edu/~phoward/m442/modbasics.pdf), which are described in more detail on Carpenter (http://mc-stan.org/users/documentation/case-studies/lotka-volterra-predator-prey.html) and Mahaffy (https://jmahaffy.sdsu.edu/courses/f09/math636/lectures/lotka/qualde2.html). The Hudson Bay time series have been analysed in various articles before (Zhang et al. 2007, Nedorezov 2016), though not with the objective of model selection, as in our work. Our results are shown in Table 4.

As opposed to the simulated data, we do not have a gold-standard here, i.e. we do not know what the true model is. However, based on the previous results, one can assume that the 'exact' method gives the most accurate results. It is therefore reassuring that our proposed method concurs with the 'exact' model selection, whereas the two information criteria select different models.

8 Discussion

Distinguishing between competing hypotheses as to the structure of systems described by ODEs is a challenging problem. In practice, one must first run a sampling algorithm (such as MCMC) or optimiser given a proposed ODE model and then either compute the marginal likelihood of the data given the model or rely on predictive performance measures such as information criteria. Calculating the marginal likelihood can be difficult since the integral over the parameters is usually not available in closed form. Information criteria are easier to compute, but rely on asymptotic assumptions which can be difficult to satisfy in practice.

Since solving the differential equations in order to perform statistical inference is usually too computationally onerous to be viable in practice, this work considers the method of gradient matching instead. The additional advantage of gradient matching is that, as opposed to approaches based on numerically integrating the ODEs, it does not require knowledge of the initial conditions. Previous work on gradient matching, reported in the existing literature, has focused on parameter estimation. To the best of our knowledge, this is the first paper to focus on model selection with gradient matching, by adapting the method of thermodynamic integration.

We have compared these results with two information criteria—BIC, which is an asymptotically correct approximation of the log marginal likelihood, and WAIC, which is asymptotically equivalent to Bayesian leave-one-out cross-validation (see Watanabe 2010). Since gradient matching is an approximate inference method, the resulting marginal likelihood computed by thermodynamic integration will also be approximate. In order to have a gold standard to compare the results to, the marginal likelihood was also calculated by explicitly solving the ODEs (numerically) to obtain the necessary components to then carry out thermodynamic integration.

We have evaluated the performance of the proposed method on three different forms of the Lotka–Volterra model (Lotka 1932) and four different forms of a protein signalling transduction pathway model (Vyshemirsky and Girolami 2008) for three signal-to-noise ratios. The results of our simulation studies can be summarised as follows:

(1) Model selection with gradient matching is more challenging than parameter estimation. In particular, the model selection performance with gradient matching is not as good as with the 'exact' method (by which we mean the method that numerically integrates the ODEs). We need to point out, though, that we have run all MCMC simulations for the 'exact' method until convergence (according to established convergence diagnostics using potential scale reduction factors). For more complex models, this would not be possible due to the high computational complexity of the numerical integration step, and this could bring the advantages of gradient matching to the fore.
(2) The performance of gradient matching is close to that of the 'exact' method for medium noise levels. The performance of gradient matching sometimes deteriorates for lower and higher SNRs. The deterioration for lower

SNRs is a consequence of the fact that the interpolants get noisier, making the derivatives less reliable. The deterioration for higher SNRs is a consequence of the chosen prior in function space, whose tightness distorts the trade-off between data fit and model complexity.

(3) The proposed thermodynamic integration method for computing the marginal likelihood achieves better model selection results than BIC. This is not surprising, as BIC is an approximation of the marginal likelihood that only becomes exact in the asymptotic limit. The proposed method also tends to be better than WAIC. This is an encouraging finding, given that WAIC is a competitive method that has been shown to come close in performance to model selection with the marginal likelihood (Aderhold et al. 2017).

Reliable model selection with gradient matching is not something that one could have taken for granted. Our simulations show that the posterior means of the parameters are usually very similar between the gradient matching approach and the 'exact' method. However, there are clear discrepancies in the shapes of the posterior distributions and, consequently, the posterior credible intervals for the parameters. The widths of these intervals are sometimes larger and sometimes narrower than the 'exact' credible intervals. This deviation is intrinsic to the methodology of gradient matching per se. A good illustration is available in Lazarus et al. (2018). Figure 2 of this paper shows that, for a model similar to those investigated in our paper, the log likelihood landscapes for the exact method and gradient matching are very different, despite the fact that the maximum likelihood configurations match very well. The upshot is that gradient matching tends to be a reliable method for parameter estimation, but not necessarily for uncertainty quantification. This could, in principle, affect model selection. The marginal likelihood is an integral over the unnormalised posterior distribution, and the question is how it will be affected by reshaping the log likelihood landscape. The novelty of our paper is that it has investigated this question empirically, on a range of benchmark problems, and assesses the accuracy of model selection in comparison with the 'exact' method.

Future work could focus on improving the numerical integration in Eq. 31. Presently, this is calculated using the trapezoidal rule, which introduces bias into the estimation of the marginal likelihood; see Friel et al. (2014). The estimation could be improved with the higher-order numerical integration scheme proposed by Friel et al. (2014). However, Aderhold et al. (2017) have discussed potential limitations and occasional deteriorations of this scheme, which suggests further investigations are advisable in the context of the present work. Future work could also focus on implementing the alternative integration path proposed by Grzegorczyk et al. (2017). Rather than tempering between the prior and posterior for each model, calculating the log marginal likelihood, then comparing the values between competing models, the alternative method by Grzegorczyk et al. (2017) tempers between the posterior distributions of two competing models and calculates the log Bayes factor directly. The results in Grzegorczyk et al. (2017) show that this approach can lead to a substantial variance reduction, which can potentially boost the performance of the model selection scheme proposed in the present paper.

Acknowledgements This research is supported by The Biometrika Trust, Fellowship No. B0003, and the EPSRC Centre for Multiscale soft tissue mechanics with application to heart and cancer, Reference No. EP/N014642/1.

References

Aderhold, A., Husmeier, D., Grzegorczyk, M.: Approximate Bayesian inference in semi-mechanistic models. Stat. Comput. **27**, 1003–1040 (2017)

Adon N.A., Jabbar M.H., Mahmud F. (2015) FPGA Implementation for Cardiac Excitation-Conduction Simulation Based on FitzHugh-Nagumo Model. In: Toi V., Lien Phuong T. (eds) 5th International Conference on Biomedical Engineering in Vietnam. IFMBE Proceedings, vol 46. Springer, Cham, pp 117–120

Arlot, S., Celisse, A.: A survey of cross-validation procedures for model selection. Stat. Surv. **4**, 40–79 (2010)

Atkins, P.W.: Physical Chemistry, 3rd edn. Oxford University Press, Oxford (1986)

Biktashev, V., Suckley, R., Elkin, Y., Simitev, R.: Asymptotic analysis and analytical solutions of a model of cardiac excitation. Bull. Math. Biol. **70**, 517–554 (2008)

Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Berlin (2006)

Calderhead, B.: A study of population MCMC for estimating Bayes Factors over nonlinear ODE models. Master's thesis, University of Glasgow (2008)

Calderhead, B., Girolami, M.A., Lawrence, N.D.: Accelerating Bayesian inference over nonlinear differential equations with gaussian processes. Neural Inf. Process. Syst. (NIPS) **22**, 217–224 (2008)

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

Carpenter, B.: Predator–prey population dynamics: the Lotka–Volterra model in Stan. http://mc-stan.org/users/documentation/case-studies/lotka-volterra-predator-prey.html. Accessed 27 Oct 2018

Chib, S.: Marginal likelihood from the Gibbs output. J. Am. Stat. Assoc. **90**, 1313–1321 (1995)

Dondelinger, F., Filippone, M., Rogers, S., Husmeier, D.: ODE parameter inference using adaptive gradient matching with Gaussian processes. In: The 16th International Conference on Artificial Intelligence and Statistics (AISTATS), of JMLR, vol. 31, pp. 216–228 (2013)

Fang, Y.: Asymptotic equivalence between cross-validations and akaike information criteria in mixed-effects models. J. Data Sci. **9**, 15–21 (2011)
FitzHugh, R.: Impulses and physiological states in theoretical models of nerve membrane. Biophys. J. **1**, 445–466 (1961)
Friel, N., Pettitt, A.N.: Marginal likelihood estimation via power posteriors. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **70**, 589–607 (2008)
Friel, N., Hurn, M., Wyse, J.: Improving power posterior estimation of statistical evidence. Stat. Comput. **24**, 709–723 (2014)
Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B.: Bayesian Data Analysis, vol. 3. Chapman and Hall/CRC, London (2013)
González, J., Vujačić, I., Wit, E.: Inferring latent gene regulatory network kinetics. Stat Appl Genet Mol Biol **12**(1), 109–127 (2013)
Gratie, D.E., Iancu, B., Petre, I.: ODE analysis of biological systems. In: Formal Methods for Dynamical Systems. Lecture Notes in Computer Science, pp. 29–62 (2013)
Grzegorczyk, M., Aderhold, A., Husmeier, D.: Targeting Bayes factors with direct-path non-equilibrium thermodynamic integration. Comput. Stat. **32**, 717–761 (2017)
Holsclaw, T., Sansó, B., Lee, H.K.H., Heitmann, K., Habi, S., Higdon, D., Alam, U.: Gaussian process modeling of derivative curves. Technometrics **55**, 57–67 (2013)
Howard, P.: Modeling basics. http://www.math.tamu.edu/~phoward/m442/modbasics.pdf. Accessed 27 Oct 2018
Lazarus, A., Husmeier, D., Papamarkou, T.: Multiphase MCMC sampling for parameter inference in nonlinear ordinary differential equations. Proc. Mach. Learn. Res. **84**, 1252–1260 (2018)
Lotka, A.: The growth of mixed populations: two species competing for a common food supply. J. Wash. Acad. Sci. **22**, 461–469 (1932)
Macdonald, B., Higham, C., Husmeier, D.: Controversy in mechanistic modelling with Gaussian processes. In: Proceedings of the 32nd International Conference on Machine Learning, vol. 37 (2015)
Macdonald, B., Niu, M., Rogers, S., Filippone, M., Husmeier, D.: Approximate parameter inference in systems biology using gradient matching: a comparative evaluation. Biomed. Eng. Online **15**(Suppl 1), 80 (2016)
Mahaffy, J.M.: Mathematical modeling: Lotka–Volterra models. https://jmahaffy.sdsu.edu/courses/f09/math636/lectures/lotka/qualde2.html. Accessed 27 Oct 2018
Murphy, K.P.: Machine Learning. A Probabilistic Perspective. The MIT Press, Cambridge (2012)
Neal, R.: Erroneous results in "Marginal Likelihood from the Gibbs Output". Open Letter—Department of Statistics and Department of Computer Science, University of Toronto (1998). ftp://www.cs.toronto.edu/dist/radford/chib-letter.pdf. Accessed 10 Mar 2018
Nedorezov, L.V.: The dynamics of the lynx-hare system: an application of the Lotka-Volterra model. Biophysics **61**(1), 149–154 (2016)
Newton, M., Raftery, A.: Approximate Bayesian inference with the weighted likelihood bootstrap. J. R. Stat. Soc. Ser. B (Methodol.) **56**, 3–48 (1994)
Ramsay, J.O., Hooker, G., Campbell, D., Cao, J.: Parameter estimation for differential equations: a generalized smoothing approach. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **69**, 741–796 (2007)
Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. The MIT Press, Cambridge (2006)
Robinson, J.C.: An Introduction to Ordinary Differential Equations. Cambridge University Press, Cambridge (2004)
Schlitter, J., Husmeier, D.: System relaxation and thermodynamic integration. Mol. Simul. **8**, 285–295 (1992)
Soetaert, K., Petzoldt, T., Setzer, R.: Package deSolve: solving initial value differential equations in R. R CRAN deSolve Documentation—Vignette. https://cran.r-project.org/web/packages/deSolve/vignettes/deSolve.pdf (2010). Accessed 28 Nov 2017
Solak, E., Murray-Smith, R., Leithead, W.E., Leith, D.J., Rasmussen, C.E.: Derivative observations in Gaussian process models of dynamic systems. Adv. Neural Inf. Process. Syst. **15**, 9–14 (2002)
Spiegelhalter, D., Best, N., Carlin, B., Linde, A.V.D.: Bayesian measures of model complexity and fit. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **64**, 583–639 (2002)
Vyshemirsky, V., Girolami, M.A.: Bayesian ranking of biochemical system models. Bioinformatics **24**(6), 833–839 (2008)
Wang, Y., Barber, D.: Gaussian processes for Bayesian estimation in ordinary differential equations. In: Proceedings of the 31st International Conference on Machine Learning, vol. 32 (2014)
Watanabe, S.: Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. J. Mach. Learn. Res. **11**, 3571–3594 (2010)
Wu, H., Lu, T., Xue, H., Liang, H.: Sparse additive ordinary differential equations for dynamic gene regulatory network modeling. J. Am. Stat. Assoc. **109**(506), 700–716 (2014)
Zhang, Z., Tao, Y., Li, Z.: Factors affecting hare-lynx dynamics in the classic time series of the Hudson Bay Company. Canada. Clim. Res. **34**, 83–89 (2007)