Theoretical Computer Science 408 (2008) 4–16

Contents lists available at ScienceDirect

Theoretical Computer Science

journal homepage: www.elsevier.com/locate/tcs

Bayesian inference for differential equations

Mark Girolami *

University of Glasgow, Department of Computing Science, Sir Alwyn Williams Building Room 302, G12 8QQ Scotland, United Kingdom

ARTICLE INFO

Keywords:
Bayesian statistics
Differential equations
Biochemical pathway models

ABSTRACT

Nonlinear dynamic systems such as biochemical pathways can be represented in abstract form using a number of modelling formalisms. In particular differential equations provide a highly expressive mathematical framework with which to model dynamic systems, and a very natural way to model the dynamics of a biochemical pathway in a deterministic manner is through the use of nonlinear ordinary or time delay differential equations. However if, for example, we consider a biochemical pathway the constituent chemical species and hence the pathway structure are seldom fully characterised. In addition it is often impossible to obtain values of the rates of activation or decay which form the free parameters of the mathematical model. The system model in many cases is therefore not fully characterised either in terms of structure or the values which parameters take. This uncertainty must be accounted for in a systematic manner when the model is used in simulation or predictive mode to safeguard against reaching conclusions about system characteristics that are unwarranted, or in making predictions that are unjustifiably optimistic given the uncertainty about the model. The Bayesian inferential methodology provides a coherent framework with which to characterise and propagate uncertainty in such mechanistic models and this paper provides an introduction to Bayesian methodology as applied to system models represented as differential equations.

1. Introduction

The use of differential equations[1] to model biochemical processes and systems has a long and well established history especially within the more recent context of Systems Biology e.g. [1–3]. On the other hand the systematic characterisation of model uncertainty has been given much less consideration with the majority of effort being focused on parameter estimation from an optimisation perspective [4] . The Bayesian methodology (see e.g. [5,6]) is well suited for this particular problem in that posterior parameter inference can characterise uncertainty in the estimates of unknown parameter values and indeed the uncertainty over a set of candidate models [7].

This paper will present the appropriate Bayesian statistical methodology in its general form in Section 2, with Section 3 providing the concrete example of a linear regression model. Performing Bayesian inference over models based on systems of ordinary differential equations is discussed and Section 7 an experimental illustration is then provided.

2. Bayesian model inference

Consider observed data $\mathcal{D} = \{\mathbf{y}, \mathbf{t}\}$ such that $\mathbf{y} \in \mathbb{R}^N$ and $\mathbf{t} \in \mathbb{R}^N$. This can be considered as some input stimulus $\mathbf{t}$ and the corresponding model response $\mathbf{y}$, and the aim is to obtain a model of the functional relationship $y \leftarrow \varphi(t)$ in order that the

* Tel.: +44 0 141 330 1623; fax: +44 0 141 330 2673.
E-mail address: girolami@dcs.gla.ac.uk.
URL: http://www.dcs.gla.ac.uk/inference.

[1] For clarity this paper will simply focus on Ordinary Differential Equations however the inferential methodology presented is sufficiently general that other forms such as delay and partial differential equations can be given the same treatment.

0304-3975/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2008.07.005

inferred functional relationship can be employed in a predictive manner. A model class $\mathcal{M} = \{\mathcal{M}_1 \cdots \mathcal{M}_K\}$ is defined which enumerates the models which will be considered. Each model $\mathcal{M}_k$ may have a parametric form in which case an associated set of parameters , $\boldsymbol{\theta}_k$, will be linked with each model. When making a prediction of possible values y_* given a particular value of the dependent variable t_* the uncertainty in both model parameters and in the models themselves should be taken into account. We therefore wish to obtain the predictive probability distribution of the unobserved random variable y_* which is denoted as $p(y_*|t_*, \mathcal{D})$ and provides the probability distribution of the values which y_* may take given, or conditioned upon, the known value of t_* and the previously observed data, $\mathcal{D}$.

This probability distribution is obtained by averaging the predictive distributions obtained from each model, $p(y_*|t_*, \mathcal{M}_k, \mathcal{D})$, with respect to the discrete probability measure $P(\mathcal{M}_k|\mathcal{D})$ which defines the *posterior* probability of model $\mathcal{M}_k$ given the observed data.

$$p(y_*|t_*, \mathcal{D}) = \sum_{k\in\mathcal{M}} p(y_*|t_*, \mathcal{M}_k, \mathcal{D})P(\mathcal{M}_k|\mathcal{D}). \tag{1}$$

All predictions are therefore based on the averaged model-based predictive distributions where the averaging is carried out with respect to the posterior distribution over models within the model class thereby taking into account the posterior uncertainty of all models.

The posterior probability distribution over models $\mathcal{M}_k$ is given in the standard form as

$$P(\mathcal{M}_k|\mathcal{D}) = \frac{p(\mathcal{D}|\mathcal{M}_k)\pi(\mathcal{M}_k)}{\sum_{l\in\mathcal{M}} p(\mathcal{D}|\mathcal{M}_l)\pi(\mathcal{M}_l)} \tag{2}$$

where the *prior* probability distribution over models is denoted by $\pi(\mathcal{M}_k)$ and the *integrated likelihood* of the observed data given the model is denoted as $p(\mathcal{D}|\mathcal{M}_k)$. Note that the integrated likelihood can also be denoted as $p(\mathbf{y}|\mathbf{t}, \mathcal{M}_k)$ which makes explicit that the observations $\mathbf{y}$ are conditional on the values of $\mathbf{t}$ and the model employed $\mathcal{M}_k$.

Now each model has an associated set of possible parameters in which case the *likelihood* of model $\mathcal{M}_k$ with parameter values $\boldsymbol{\theta}_k \in \Theta$ giving rise to observations $\mathbf{y}$ is the probability density $p(\mathcal{D}|\boldsymbol{\theta}_k, \mathcal{M}_k)$. As with the individual models a prior probability density will be placed on the model parameters $p(\boldsymbol{\theta}_k|\mathcal{M}_k)$ and the integrated likelihood is obtained by taking the expectation of the data likelihood with respect to the parameter prior.

$$p(\mathcal{D}|\mathcal{M}_k) = \int_{\boldsymbol{\theta}_k\in\Theta} p(\mathcal{D}|\boldsymbol{\theta}_k, \mathcal{M}_k)p(\boldsymbol{\theta}_k|\mathcal{M}_k)\mathrm{d}\boldsymbol{\theta}_k. \tag{3}$$

The final component of the averaged predictive likelihood (1) that needs to be considered is the model specific predictive likelihood. This distribution will be obtained by further averaging such that

$$p(y_*|t_*, \mathcal{M}_k, \mathcal{D}) = \int_{\boldsymbol{\theta}_k\in\Theta} p(y_*|t_*, \boldsymbol{\theta}, \mathcal{M}_k)p(\boldsymbol{\theta}_k|\mathcal{D}, \mathcal{M}_k)\mathrm{d}\boldsymbol{\theta}_k. \tag{4}$$

where the expectation is taken with respect to the parameter posterior density obtained as

$$p(\boldsymbol{\theta}_k|\mathcal{D}, \mathcal{M}_k) = \frac{p(\mathcal{D}|\boldsymbol{\theta}_k, \mathcal{M}_k)p(\boldsymbol{\theta}_k|\mathcal{M}_k)}{p(\mathcal{D}|\mathcal{M}_k)} \tag{5}$$

It can be seen that the integrated likelihood appears in the parameter posterior distribution as a normalising term. From the predictive distribution defined in (1) it is straightforward to obtain summary statistics such as the predictive mean and variance which follow as

$$E\{y_*|t_*, \mathcal{D}\} = \sum_{k\in\mathcal{M}} E\{y_*|t_*, \mathcal{M}_k\}P(\mathcal{M}_k|\mathcal{D}) \equiv \mu_*^2$$

$$Var\{y_*|t_*, \mathcal{D}\} = \sum_{k\in\mathcal{M}} \left(Var\{y_*|t_*, \mathcal{M}_k\} + E\{y_*|t_*, \mathcal{M}_k\}^2\right) P(\mathcal{M}_k|\mathcal{D}) - \mu_*^2.$$

With this small number of identities a powerful inferential framework is provided in devising prediction from models which takes into account the sources of uncertainty based on model structure and parameter values. To give a concrete example of the Bayesian framework in action the next section provides a Bayesian treatment of a standard linear functional response model.

3. Linear response model

As an illustrative example of this inferential framework a linear model is considered, further examples and details for applications in regression and classification can be found in [8]. Given a single random variable t and a corresponding observed response variable a linear combination of basis expansions of the dependent variable is employed such that $f = \sum_{m=1}^{M} \beta_m \varphi_m(t)$ which forms the deterministic component of the overall statistical model. In addition a noise or error term is included such that the observation $y = f + \epsilon$ where ϵ is a stochastic term which if Normal errors are assumed

then this will have a Gaussian density with a mean of zero and an unknown variance or standard deviation denoted as σ. The notation $\epsilon \sim \mathcal{N}(0, \sigma)$ indicates that the random variable ϵ is Normally distributed with mean zero and standard deviation σ.

The overall model is then defined by the class of basis function employed, φ_m as well as the associated model parameters $\boldsymbol{\beta} \in \mathbb{R}^M$ and σ. Let us consider the model class of polynomial basis functions such that each $\mathcal{M}_k \equiv \varphi_k(t) = \{t^i\}_{i=1\cdots k}$ where each model is defined as $\{\mathcal{M}_k, \boldsymbol{\beta}_k, \sigma\}$ and is indexed from $k = 1 \cdots K$. In other words the model class considered is polynomials of order-k and each model-order is defined by a set of expansion coefficients $\boldsymbol{\beta}_k \in \mathbb{R}^k$ and σ.

Given data observed such that $\mathcal{D} = \{(y_1, t_1), \ldots, (y_N, t_N)\} = (\mathbf{y} \in \mathbb{R}^N, \mathbf{t} \in \mathbb{R}^N)$ an instantiation of the Bayesian inferential methodology is developed for this particular class of model and, for illustrative purposes, we will consider the class of polynomial functions up to $K = 10$.

For each model a prior over the parameters is required i.e. $p(\boldsymbol{\theta}_k|\mathcal{M}_k)$ which in this case will be $p(\boldsymbol{\beta}_k, \sigma|\mathcal{M}_k)$. In addition a likelihood requires to be defined, so given that each $\epsilon_n \sim \mathcal{N}(0, \sigma)$ and assuming that they are independent and identically distributed then it follows that $p(\mathcal{D}|\boldsymbol{\theta}_k, \mathcal{M}_k) \equiv p(\mathbf{y}|\mathbf{t}, \boldsymbol{\beta}_k, \sigma) = \mathcal{N}_{\mathbf{y}}(\Phi\boldsymbol{\beta}, \mathbf{I}_N\sigma^2)$ which denotes the N-dimensional multivariate Gaussian density at the point $\mathbf{y}$ with a mean of $\Phi\boldsymbol{\beta}$ and covariance of $\mathbf{I}_N\sigma^2$. Where $\mathbf{I}_N$ denotes an $N \times N$ identity matrix and Φ denotes the $N \times k$ dimensional matrix of polynomial responses for each t_n and $\boldsymbol{\beta}_k$ is the $k \times 1$ dimensional vector of basis coefficients. The multivariate Gaussian is defined below.

$$\mathcal{N}_{\mathbf{y}}(\Phi\boldsymbol{\beta}, \mathbf{I}_N\sigma^2) = \frac{1}{(2\pi\sigma^2)^{\frac{N}{2}}} \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{y} - \Phi\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{y} - \Phi\boldsymbol{\beta})\right\}. \tag{6}$$

Now that the likelihood has been defined a functional form for the prior $p(\boldsymbol{\beta}_k, \sigma^2|\mathcal{M}_k)$ is required. Noting that $p(\boldsymbol{\beta}_k, \sigma|\mathcal{M}_k) = p(\boldsymbol{\beta}_k|\sigma^2)p(\sigma^2|\mathcal{M}_k)$ a product of independent Gaussians can be placed on each β_m with mean zero and variance equal to σ^2 and so $p(\boldsymbol{\beta}_k|\sigma^2) = \mathcal{N}_{\boldsymbol{\beta}}(\mathbf{0}, \mathbf{I}_N\sigma^2)$. This prior distribution on the basis weights reflects the notion that the values of the coefficients should be small so as not to produce an unnecessarily complex combination of polynomials. It also has the advantage that both the likelihood and the prior having the same functional form ensures a degree of analytical tractability. The prior density for the variance σ^2 has to have support restricted to the positive half of the real line and in this case an inverse-Gamma density is a reasonable choice and has the added advantage of enabling analytic solution for all of the integrals defined in the previous section. The inverse Gamma density is defined by two parameters α and γ where $\mathcal{IG}_{\sigma^2}(\alpha, \gamma) = \gamma^{\alpha}(\sigma^2)^{-(\alpha+1)} \exp(-\gamma\sigma^{-2})/\Gamma(\alpha)$. For the case where $\alpha = \gamma = 1$ the overall prior density over the model parameters is

$$p(\boldsymbol{\beta}_k, \sigma^2|\mathcal{M}_k) = \frac{\left(\sigma^2\right)^{-(\frac{4+k}{2})}}{(2\pi)^{\frac{k}{2}}} \exp\left\{-\frac{\boldsymbol{\beta}_k^{\mathrm{T}}\boldsymbol{\beta}_k + 2}{2\sigma^2}\right\}. \tag{7}$$

It follows that an analytic form for the integrated likelihood is obtained after some straightforward but tedious manipulation. For the prior density defined above the integrated likelihood takes the following analytic form

$$\begin{aligned} p(\mathbf{y}|\mathbf{t}, \mathcal{M}_k) &= \int\int \mathcal{N}_{\mathbf{y}}(\Phi\boldsymbol{\beta}_k, \mathbf{I}_N\sigma^2)\mathcal{N}_{\boldsymbol{\beta}_k}(\mathbf{0}, \mathbf{I}_N\sigma^2)\mathcal{IG}_{\sigma^2}(1, 1)\mathrm{d}\boldsymbol{\beta}_k\mathrm{d}\sigma^2 \\ &= \frac{\Gamma(1 + \frac{N}{2})}{\sqrt{\pi^N|\Sigma_k|}} \left\{1 + \frac{1}{2}\mathbf{y}^{\mathrm{T}}(\mathbf{I}_N - \Phi\Sigma_k^{-1}\Phi^{\mathrm{T}})\mathbf{y}\right\}^{-(1+\frac{N}{2})} \end{aligned} \tag{8}$$

where $\Sigma_k = \mathbf{I}_k + \Phi^{\mathrm{T}}\Phi$, $\Gamma(\cdot)$ denotes a Gamma function and $|\cdot|$ is the matrix determinant operator. Now that the integrated likelihood is available conditional on each model the posterior distribution over models follows straightforwardly as

$$P(\mathcal{M}_k|\mathbf{y}, \mathbf{t}) = \frac{p(\mathbf{y}|\mathbf{t}, \mathcal{M}_k)\pi(\mathcal{M}_k)}{\sum_{l\in\mathcal{M}} p(\mathbf{y}|\mathbf{t}, \mathcal{M}_l)\pi(\mathcal{M}_l)}. \tag{9}$$

This distribution over models provides the means to rank models and to perform the model-based averaging of predictions.

Conveniently, due to the exponential conjugacy of the prior density and the likelihood, the parameter posterior density $p(\boldsymbol{\beta}_k, \sigma^2|\mathcal{M}_k, \mathbf{y}, \mathbf{t})$ takes the form of a product of a Normal and Inverse-Gamma density given below as

$$\frac{\delta_k^{\frac{N+2}{2}}\left(\sigma^2\right)^{-(\frac{4+k}{2})}}{(2\pi)^{\frac{k}{2}}|\Sigma_k|^{\frac{1}{2}}\Gamma(\frac{N+2}{2})} \exp\left\{-\frac{(\boldsymbol{\beta}_k - \boldsymbol{\mu}_k)^{\mathrm{T}}\Sigma_k^{-1}(\boldsymbol{\beta}_k - \boldsymbol{\mu}_k) + 2\delta_k}{2\sigma^2}\right\} \tag{10}$$

where $\boldsymbol{\mu}_k = \Sigma_k^{-1}\Phi^{\mathrm{T}}\mathbf{y}$ and $\delta_k = 1 + \frac{1}{2}(\mathbf{y}^{\mathrm{T}}\mathbf{y} - \boldsymbol{\mu}_k^{\mathrm{T}}\Sigma_k^{-1}\boldsymbol{\mu}_k)$. Finally an explicit expression for the model conditional predictive distribution can also be obtained as an analytic form for the required integral below is also available.

$$p(y_*|t_*, \mathcal{M}_k, \mathbf{y}, \mathbf{t}) = \int\int p(y_*|t_*, \boldsymbol{\beta}_k, \sigma^2)p(\boldsymbol{\beta}_k, \sigma^2|\mathcal{M}_k, \mathbf{y}, \mathbf{t})\mathrm{d}\boldsymbol{\beta}_k\mathrm{d}\sigma^2 \tag{11}$$

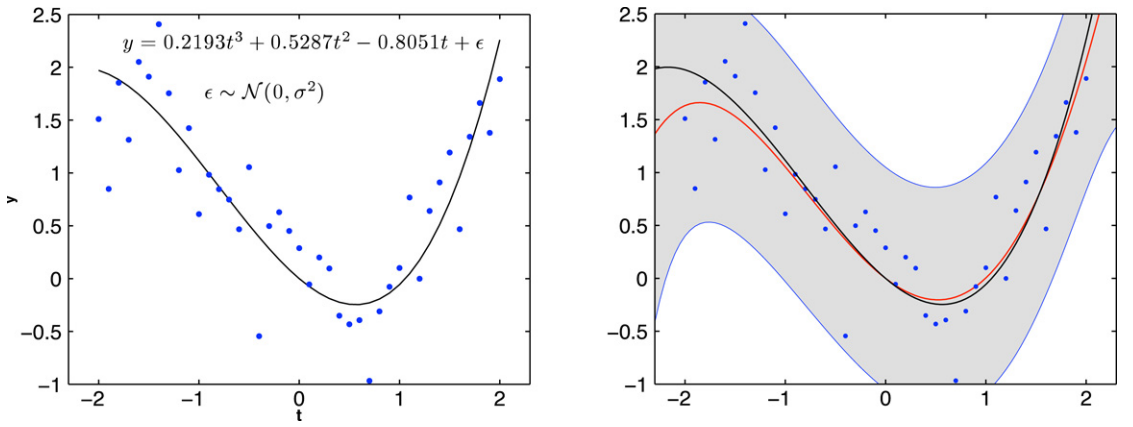


Fig. 1. The left hand figure shows the observed data points and the solid line is the noise free function which generated the finite number of noisy observations. The right hand figure shows the observed data and the noise free function. In addition the Bayesian model averaged predictive mean value (Eq. (13)), is shown in solid red and the shaded region corresponds to ± one standard deviation of the model averaged predictive distribution (Eq. (14)).

and as the predictive likelihood $p(y_*|t_*, \boldsymbol{\beta}_k, \sigma^2)$ is a univariate Gaussian then the above integral takes the form of a Student-T density where

$$p(y_*|t_*, \mathcal{M}_k, \mathbf{y}, \mathbf{t}) = \frac{\Gamma(1+\frac{N}{4})}{\Gamma(\frac{3N}{4})\sqrt{(\pi\delta_k\xi_k)}}\left\{1+\frac{(y_*-\boldsymbol{\phi}_k^{\mathrm{T}}\boldsymbol{\mu}_k)}{\delta_k\xi_k}\right\}^{-(1+\frac{N}{4})} \tag{12}$$

where $\boldsymbol{\phi}_k$ is a $k\times 1$ vector of the basis (polynomial in this example) responses at t_* the query point and $\xi_k = 1+\boldsymbol{\phi}_k^{\mathrm{T}}\Sigma_k^{-1}\boldsymbol{\phi}_k$. This predictive density has a mean value of $\boldsymbol{\phi}_k^{\mathrm{T}}\boldsymbol{\mu}_k$ and a variance $\frac{N+2}{N-2}\delta_k\xi_k$ and so the Bayesian model averaged mean prediction and associated variance follows as

$$E\{y_*|t_*, \mathbf{y}, \mathbf{t}\} = \sum_{k\in\mathcal{M}} \boldsymbol{\phi}_k^{\mathrm{T}}\boldsymbol{\mu}_k P(\mathcal{M}_k|\mathbf{y}, \mathbf{t}) \equiv \mu_* \tag{13}$$

$$Var\{y_*|t_*, \mathbf{y}, \mathbf{t}\} = \sum_{k\in\mathcal{M}}\left(\frac{N+2}{N-2}\delta_k\xi_k + \boldsymbol{\phi}_k^{\mathrm{T}}\boldsymbol{\mu}_k\boldsymbol{\mu}_k^{\mathrm{T}}\boldsymbol{\phi}_k\right)P(\mathcal{M}_k|\mathbf{y}, \mathbf{t}) - \mu_*^2. \tag{14}$$

3.1. Summary

For the case of the linear model with the choice of prior distribution over parameters chosen it is possible to obtain fully analytic expressions for the posterior distribution, the predictive distribution and the integrated likelihood which then provides the posterior distribution over all models considered. The following section gives an experimental illustration of the Bayesian methodology in practice for this type of linear model.

4. Inference for linear data models an illustrative example

Consider that measurements are made at forty one distinct points, these could be for example time points. The objective is then to devise a model of the functional relationship between, t and the measurements y, and use this to make subsequent predictions at points for which no measurement has been taken. By way of illustration we will generate data synthetically by specifying such a function which in this case is a cubic expansion $f(t) = 0.2193t^3 + 0.5287t^2 - 0.8051t$. In addition a simple Gaussian noise process with a mean of zero and $\sigma = 0.5$ is defined so that each observation is defined by $y(t) = f(t) + \epsilon(t)$ where each $\epsilon \sim \mathcal{N}(0, 0.5^2)$.

The data and the deterministic underlying function is shown in Fig. 1, the spread of the observed data points indicates the levels of uncertainty in the measurements and to provide consistent predictions this uncertainty requires to be propagated forward to all levels of inference. The observed data would suggest that there is possibly a polynomial relationship between t and y possibly of second or higher order. Given this we can select the class of models to be considered as the set of polynomial basis expansions up to a maximal order of say ten. From the plot of the data it is unlikely that tenth order effects are of any significance but we can include this and allow the Bayesian inferential mechanism to appropriately weight the models under consideration.

The most popular method for model fitting is the use of least squares estimators which are nothing more than maximum likelihood estimators. The value of the joint likelihood of the observed data when the *maximum posterior* parameter estimates are used, denoted as $p(\mathbf{y}|\mathbf{t}, \hat{\beta}_k, \hat{\sigma}_k, \mathcal{M}_k)$, is plotted as a bar chart in Figure (2) left hand diagram. What we see is that

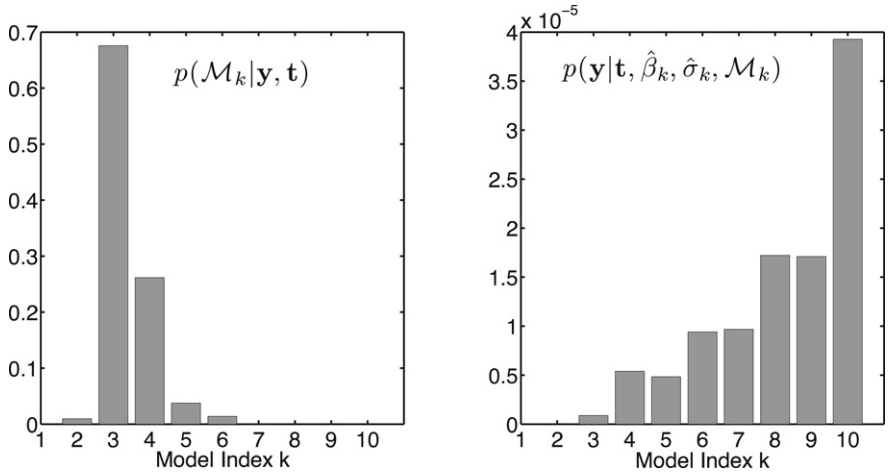


Fig. 2. The right hand bar chart displays the data likelihood for each model from $\mathcal{M}_1$ to $\mathcal{M}_{10}$. The data likelihood is obtained when the *maximum a posteriori* point estimates of the model parameters $\hat{\beta}_k, \hat{\sigma}$ are employed. The data likelihood is a raw measure of how well each model *fits* the observed data and it is clear that as the model complexity, that is the order of the polynomial expansion, increases then the misfit to the data decreases and so the data likelihood increases. The left hand bar chart shows the discrete probability distribution over the data models and here we observe a quite different pattern.

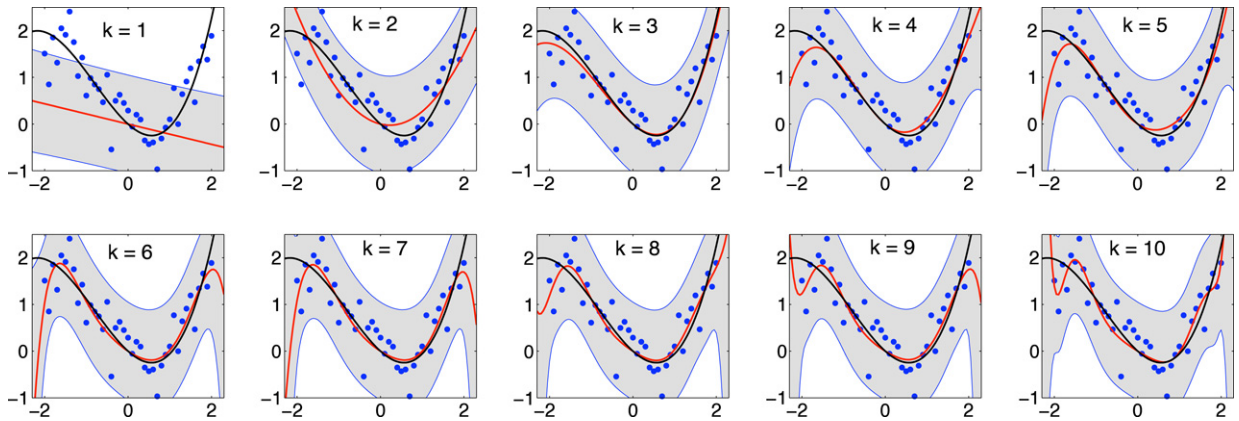
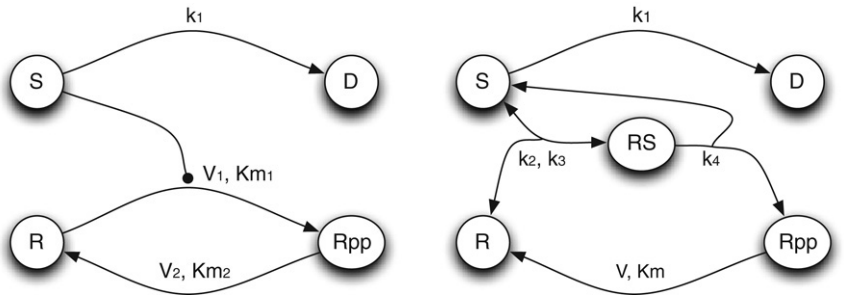


Fig. 3. The figures show the original data, the true underlying function, as well as the mean and variance of the predictive distribution for each of the ten models considered.

the likelihood increases as the complexity of the model increases beyond that which actually produced the observations. In other words observing the data under complex models is more likely than under less complex ones. This is a fundamental flaw in model assessment within the *model fitting* paradigm.

On the other hand if we obtain the posterior distribution over models (Eq. (9)), which in the case of this particular model can be obtained analytically, we observe a quite different distribution in Fig. 2 from that obtained by measures of *data fit*. We can now see that the posterior distribution over the ten models considered gives almost no support to a linear model, this is obvious from looking at the data. It is not so obvious, visually at least, whether a third or fourth order model underlies the observations and we see from the distribution a peak at $k = 3$ with some decaying residual support from $k = 4$ onwards. It is clear then that this posterior distribution can be used as a means of ranking models in terms of their evidential support and we shall return to this in the following sections. In addition the model posterior distribution can be used to weight or average the predictions made over all models. In Fig. 3 the predictive mean and variance (from the Student-t density of Eq. (12)) for each of the ten models considered is shown graphically. For the linear model with $k = 1$ we can see the expected bias in the predictions and this decrease as the model complexity increases, however it is interesting to observe that the variance of the predictions starts to increase rapidly after $k = 3$ the *true* model order. This is most apparent for predictions made around $t = \pm 2.3$ where there was no observed data. So whilst the bias is decreasing (or put another way the *data fit* is improving) as the model complexity increases, the predictive variance is also increasing rendering the predictions of poorer quality due to the large levels of predictive uncertainty. Returning to Fig. 1 we observe the Bayesian model averaged predictions the uncertainty is at a similar level to the single $k = 3$ model however the variance is a little higher reflecting as it does the overall uncertainty in the modeling process.



(a) **Model1:** the simpler model, which has 10 parameters. This model relies on Michaelis–Menten law to define activation of protein *R*.

(b) **Model2:** the more complex model, which has 12 parameters. This model relies on mass action law to define activation of protein *R*.

Fig. 4. Models used for this case study.

4.1. Conclusions

This section has provided an example of the power of Bayesian inference within the data modeling process where a linear expansion of basis functions has been employed as the functional class. The choice of prior distributions and the essentially linear nature of the models has allowed us to obtain the required posterior and marginal densities in analytic form. The following section now considers inference over models based on nonlinear ODEs and here we will see that there is another layer of technical difficulty which has to be overcome for inference to proceed.

5. Nonlinear ODE models of biochemical pathways

The inferential framework presented and illustrated is completely general and we now demonstrate it on dynamic models defined by systems of nonlinear ordinary Differential Equations (ODE). The example we now consider is one which is directly related to systems biology, the modeling of biochemical pathways [1]. In this section two ODE-based models have been developed to demonstrate the application of Bayesian inference to realistic problems in Systems Biology. Both of the models describe a process of enzymatic activation of protein *R* into its active, or phosphorolated, form *Rpp* by an enzyme *S*. At the same time enzyme degrades into its inactive form *D*.

The first model (see Fig. 4(a)) utilises the Michaelis–Menten kinetic law to define activation of protein *R*. The system of differential equations (15) that defines Model 1 has 5 kinetic parameters: k_1, V_1, Km_1, V_2, and Km_2. In addition to the kinetic parameters, the statistical model also contains the initial values for all of the variables: $S|_{t=0}$, $D|_{t=0}$, $R|_{t=0}$, $Rpp|_{t=0}$.

The systematic component of the overall model is defined by a system of coupled ODE's denoted as $\mathcal{M}_1$, (Eq. (15)). Now the stochastic component of the model will characterise the variability in the observed data, which if the model is fully observed at each of N distinct time points can be represented by the $N \times K$, where $K = 4$ dimensional matrix $\mathbf{Y}$. The dynamic system which the ODE's define will generate four time series each of which are correlated through time and with each other thus inducing a complex covariance structure across time points and across observed species. This structure can be represented by a multivariate Gaussian Process by defining the $NK \times 1$ vector $\mathbf{y} = \text{vec}(\mathbf{Y})$ such that $\mathbf{y}$ is distributed as a Gaussian Process with mean $\boldsymbol{\mu}$ and covariance $\Sigma \otimes \mathbf{C}$ and $\otimes$ denotes the Kronecker product. Where $\boldsymbol{\mu}$ is an $NK \times 1$ vector of of the computed values for S, D, R, Rpp at each of the N time points, Σ is an $N \times N$ dimensional symmetric positive semi-definite matrix of the covariance function values at each time point and the $K \times K$ dimensional covariance matrix models the inter-species covariance.

$$\mathcal{M}_1 = \begin{cases} \dfrac{\mathrm{d}S}{\mathrm{d}t} = -k_1 \cdot S \\ \dfrac{\mathrm{d}D}{\mathrm{d}t} = k_1 \cdot S \\ \dfrac{\mathrm{d}R}{\mathrm{d}t} = -\dfrac{V_1 \cdot R \cdot S}{Km_1 + R} + \dfrac{V_2 \cdot Rpp}{Km_2 + Rpp} \\ \dfrac{\mathrm{d}Rpp}{\mathrm{d}t} = \dfrac{V_1 \cdot R \cdot S}{Km_1 + R} - \dfrac{V_2 \cdot Rpp}{Km_2 + Rpp}. \end{cases} \tag{15}$$

For the purposes of this illustrative paper we will assume a much more simplified error structure where statistical independence is assumed across time and across chemical species with a single error term describing all data variance in which case the multivariate Gaussian Process collapses to a spherical Gaussian such that $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I}\sigma^2)$.

The overall set of parameters for the ODE system and the simple error model can be represented by the vector $\boldsymbol{\theta} = (k_1, V_1, Km_1, V_2, Km_2, S_0, D_0, R_0, Rpp_0, \sigma^2)$. In defining the likelihood we will use $\boldsymbol{\mu}(\boldsymbol{\theta})$ to make explicit that the mean response is a nonlinear function of the model parameters. Note that in this case there is no analytic solution for the system

of ODE's and so numerical integration of the ODE's will be required to obtain $\boldsymbol{\mu}(\boldsymbol{\theta})$ and therefore to obtain the value of the likelihood $p(\mathbf{y}|\mathbf{t}, \boldsymbol{\theta}, \mathcal{M}_1)$.

In defining a prior for the model parameters $p(\boldsymbol{\theta}|\mathcal{M}_1)$ we note that all parameters have to be restricted to the positive part of the real line, however, irrespective of the choice of prior the integrated likelihood $p(\mathbf{y}|\mathbf{t}, \mathcal{M}_1) = \int p(\mathbf{y}|\mathbf{t}, \boldsymbol{\theta}, \mathcal{M}_1)p(\boldsymbol{\theta}|\mathcal{M}_1)\mathrm{d}\boldsymbol{\theta}$ will no longer admit an analytic form. As the integrated likelihood is the normalising constant for the posterior then we will lose the analytic expression for $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{t}, \mathcal{M}_1)$ the posterior. This means that the posterior averaging to obtain the predictive distribution $p(y_*|t_*, \mathcal{M}_1)$ cannot be achieved analytically and in addition the posterior over possible model structures will not have an analytic form. We shall return to these issues in the following section but for now we will introduce the second biochemical model that shall be considered.

The second model (see Fig. 4(b)) utilises the mass action kinetic law (in its explicit form for enzymatic activation) to define activation of protein R. The system of differential equations (16) that defines model $\mathcal{M}_2$ has 6 kinetic parameters: k_1, k_2, k_3, k_4, V, and Km. In addition to the kinetic parameters, the statistical model also contains the initial values for all of the variables: $S|_{t=0}$, $D|_{t=0}$, $R|_{t=0}$, $RS|_{t=0}$, $Rpp|_{t=0}$, and the error variance σ resulting in 12 parameters overall.

$$\mathcal{M}_2 = \begin{cases} \dfrac{\mathrm{d}S}{\mathrm{d}t} = -k_1 \cdot S - k_2 \cdot R \cdot S + k_3 \cdot RS + k_4 \cdot RS \\ \dfrac{\mathrm{d}D}{\mathrm{d}t} = k_1 \cdot S \\ \dfrac{\mathrm{d}R}{\mathrm{d}t} = -k_2 \cdot R \cdot S + k_3 \cdot RS + \dfrac{V \cdot Rpp}{Km + Rpp} \\ \dfrac{\mathrm{d}RS}{\mathrm{d}t} = k_2 \cdot R \cdot S - k_3 \cdot RS - k_4 \cdot RS \\ \dfrac{\mathrm{d}Rpp}{\mathrm{d}t} = k_4 \cdot RS - \dfrac{V \cdot Rpp}{Km + Rpp}. \end{cases} \tag{16}$$

As this is a more complex mechanistic description of the biochemical process we should note that an additional chemical species RS enters this model which is not present in $\mathcal{M}_1$. If we observe only the chemical species which are in common between both models then RS at each time point is unobserved or a latent variable. However as an initial value problem has to be solved in obtaining the ODE solutions the unobserved RS only adds one additional parameter which is the unknown initial value $RS|_{t=0}$. As the overall time course of RS is not observed its effect on the likelihood is indirect via the coupling of RS to all other state-derivative equations in the system.

6. Markov chain Monte Carlo

The challenge which we now face is no longer a statistical one, it is in fact a computational problem which we have to address. The main issue in making progress here is that the multi-dimensional integrals required to obtain the integrated likelihood (Eq. (3)) and the predictive likelihood (Eq. (4)) are no longer analytic. These integrals take the generic form of

$$I(f) = \int f(\mathbf{x})\pi(\mathbf{x})\mathrm{d}\mathbf{x} \equiv E_\pi\{f(\mathbf{x})\} \tag{17}$$

where $\pi(\mathbf{x})$ is a probability density function and we can see that this is the expectation of the function $f(.)$ with respect to the density π and is written as $E_\pi\{f(\mathbf{x})\}$. For the purposes of invoking results from the Central Limit Theorem (CLT) it is assumed that the expected absolute deviation of the function is finitely bounded i.e. $E_\pi\{|f(\mathbf{x})|\} < +\infty$.

If samples can be drawn from the density π such that $\mathbf{x}^s \sim \pi(\mathbf{x})$, then from the Law of Large Numbers it follows that

$$\lim_{S\to+\infty} \frac{1}{S}\sum_{s=1}^{S} f(\mathbf{x}^s) = E_\pi\{f(\mathbf{x})\} \tag{18}$$

so the estimator $\hat{I}_S(f) = S^{-1}\sum_{s=1}^{N} f(\mathbf{x}^s)$ is an unbiased estimator of the desired integral (see e.g. [9]). In addition from the CLT the distribution of the error converges to a Normal distribution with mean zero (as the estimator is unbiased) and variance equal to $\mathrm{var}(f(\mathbf{x}))$, such that as $S \to +\infty$ then

$$\sqrt{S}\left(\hat{I}_S(f) - I(f)\right) \sim \mathcal{N}\left(0, \mathrm{var}(f(\mathbf{x}))\right) \tag{19}$$

indicating a convergence rate of $\mathcal{O}(\frac{1}{\sqrt{S}})$ irrespective of the dimensionality of the random vector $\mathbf{x}$.

So to obtain the integrated likelihood it would initially appear that the difficulty has been overcome as the following estimator can be employed.

$$\hat{p}(\mathbf{y}|\mathbf{t}, \mathcal{M}_1) = \frac{1}{S}\sum_{s=1}^{S} p(\mathbf{y}|\mathbf{t}, \boldsymbol{\theta}^s, \mathcal{M}_1) \approx \int p(\mathbf{y}|\mathbf{t}, \boldsymbol{\theta}, \mathcal{M}_1)p(\boldsymbol{\theta}|\mathcal{M}_1)\mathrm{d}\boldsymbol{\theta} \tag{20}$$

where $\boldsymbol{\theta}^s \sim p(\boldsymbol{\theta}|\mathcal{M}_1)$ which should be straightforward to sample from if the prior is from a standard family such as the Gamma. There is however a small fly in the ointment which will take some formidable computational effort to remove. Many of the samples drawn from the prior density will not necessarily lie in regions of the parameter space for which the likelihood $p(\mathbf{y}|\mathbf{t}, \boldsymbol{\theta}, \mathcal{M}_1)$ is large. This is exacerbated by the relatively high dimensionality of the parameter spaces we are now considering and the potential dynamic complexity of the systems which are being modeled thus inducing likelihood surfaces which may well vary wildly. The outcome of this is that the estimator itself will be poor having a large variance. We shall return to this issue shortly.

The other multi-dimensional integral we require is that for the predictive likelihood again we can invoke a Monte Carlo estimate of this integral such that

$$\hat{p}(y_*|t_*, \mathcal{M}_1, \mathbf{y}, \mathbf{t}) = \frac{1}{S}\sum_{s=1}^{S} p(y_*|t_*, \boldsymbol{\theta}^s) \approx \int p(y_*|t_*, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{t}, \mathcal{M}_1)\mathrm{d}\boldsymbol{\theta} \tag{21}$$

where $\boldsymbol{\theta}^s \sim p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{t}, \mathcal{M}_1)$. Now the problem we face is that samples from a density which has no analytic form are required and this presents yet another obstacle in our adoption of the Bayesian inferential framework in this setting. We consider this problem further in the following section.

6.1. Metropolis–Hastings sampler

Sampling from a target distribution $p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$ where the non-normalised form, i.e. the product of the likelihood and prior, is only available in computable form is still feasible by simulating from a Markov chain whose stationary distribution is the desired target. There is a vast literature on Markov chain Monte Carlo (MCMC) methodology and here we consider the foundational Metropolis-Hasting method [10] which employs local Markov transitions in sampling space (see for example [11]). Assume we have a transition density $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$ which defines the probability of making the transition from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}^*$. This can be something as simple as a Gaussian computed at $\boldsymbol{\theta}$ which has a mean of $\boldsymbol{\theta}^*$ with some covariance $\mathbf{C}$ i.e. $q(\boldsymbol{\theta}^*|\boldsymbol{\theta}) = \mathcal{N}_{\boldsymbol{\theta}^*}(\boldsymbol{\theta}, \mathbf{C})$. Given the current state of the Markov chain (parameter values) a transition can be proposed $\boldsymbol{\theta} \to \boldsymbol{\theta}^*$ with probability $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$. So if the transition probability is indeed a multivariate Gaussian $\mathcal{N}_{\boldsymbol{\theta}^*}(\boldsymbol{\theta}, \mathbf{C})$ then it is clear that the Markov chain will make local moves within the ellipse of equi-probability defined by the covariance matrix $\mathbf{C}$ and centered at $\boldsymbol{\theta}$. To ensure that the Markov chain has the target density as its stationary distribution the proposed transitions will then be accepted with a probability $\alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta})$ where this is defined as

$$\min\left\{1, \frac{p(\mathbf{y}|\boldsymbol{\theta}^*)p(\boldsymbol{\theta}*)q(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})q(\boldsymbol{\theta}^*|\boldsymbol{\theta})}\right\} \tag{22}$$

so proposing candidate transitions $\boldsymbol{\theta}^*$ from $q(\boldsymbol{\theta}^*|\boldsymbol{\theta})$ and accepting these with probability $\alpha(\boldsymbol{\theta}^*|\boldsymbol{\theta})$ will provide dependent samples $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \boldsymbol{\theta}^3, \ldots, \boldsymbol{\theta}^S$ which are distributed as $p(\boldsymbol{\theta}|\mathbf{y})$.

This method is particularly appealing in that it provides a means of sampling from a density for which the normalised form is unavailable as the transition acceptance probability depends only on the product of the likelihood, prior and proposal densities. So these samples obtained from the Markov Chain based Metropolis routine can be used in providing a Monte Carlo estimate of the predictive likelihood equation (21).

It would appear then that we have all the tools available to perform full Bayesian inference over this class of nonlinear models. Although it is feasible to perform Monte Carlo averaging with samples drawn from the prior in estimating the integrated likelihood (Eq. (20)) it is well known that this can be a rather poor estimator [12]. To illustrate this suppose the linear models of the previous section were set with a prior density on the basis coefficients which was Gaussian centered at zero with variance one. Furthermore if the true model had a set of basis coefficients, $\beta_i = \pm 3.5$, it is clear that more than 99.7% of values drawn from this prior will be smaller / larger than the true ± 3.5 and hence will generate values of the likelihood function that are small and contribute little to the estimate. So a large number of values, S, from the prior will need to be generated before the Monte Carlo estimate will start to converge and the variance of the finite sample estimate takes on reasonable values. This form of estimator using samples from the prior is hugely inefficient and this is studied in a detailed manner in [13].

In addition to the practical inefficiency of Monte Carlo estimates of the integrated likelihood based on samples from the prior there is a further hurdle that has to be negotiated before we are ready to consider fully Bayesian inference of these nonlinear ODE based models.

Consider the following simple model of a Circadian Oscillator as originally described in [14].

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \frac{k_1}{36 + k_2 y} - k_3, \qquad \frac{\mathrm{d}y}{\mathrm{d}t} = k_4 x - k_5 \tag{23}$$

where $k_1 = 72, k_2 = 1, k_3 = 2, k_4 = 1$ and $k_5 = 1$, and the initial values are set to $x(0) = 7$ and $y(0) = -10$. The data comprised of 120 data points, $\mathbf{y}$, which were simulated using these settings, between $t = 0$ and $t = 60$ in discrete steps of 0.5, $\mathbf{t}$, to which independent Gaussian noise was added with variance $\sigma = 0.5$. The posterior was then calculated conditionally over the parameters k_3 and k_4 and plotted from 0 to 5 on each axis.

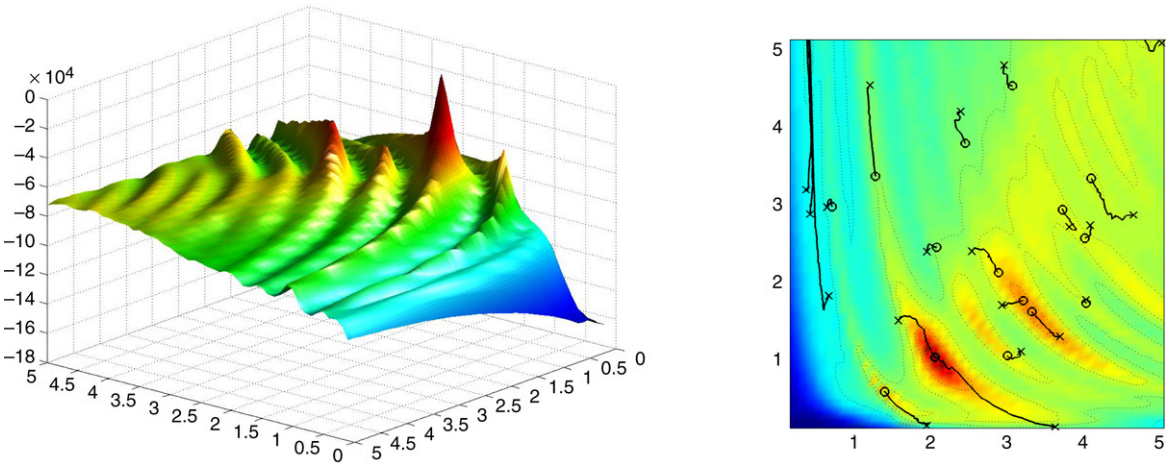


Fig. 5. The posterior $p(k_3, k_4|\mathbf{y}, \mathbf{t}, k_1 = 72, k_2 = 1, k_5 = 1, x(0) = 7, y(0) = -10, \sigma = 0.5)$. The right hand plot shows the path of twenty independent Markov chains derived from the Metropolis method, it is clear that they do not mix or cover the whole parameter space.

What can be observed is that the there is indeed a sharp peak in the posterior at the 'true' values of k_3 and k_4 however there are a series of strong ripples which form a number of ridges on the posterior surface. This has major practical implications for the ability of the Markov chain driving the Metropolis sampler to successfully explore the posterior surface fully. Imagine that the current state of the chain is located at the peak of one of the ridges of the posterior. Any proposal will result in a move into a region of lower posterior density and so the acceptance probability will be very small making it highly unlikely that the chain will make any further transitions away from the region on the ridge in finite time. This is illustrated in the contour plot of Fig. 5 where Markov chains sampled using the Metropolis method easily get caught in these local modes, even when engineering techniques, such as an adaptive step size, are employed. The plot shows the paths taken by 20 independent Markov chains generated by a Metropolis sampler. Their starting points, indicated by a $\times$, were generated randomly in the parameter space from a prior distribution, and their end positions are denoted by a $\circ$. The localised movement of the chains on the ridges is evident from the figure and this problem can be resolved by defining as the target distribution a product of densities which form a path from the smooth prior to the ridge like posterior.

6.2. Product space sampling

Consider the target density $p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\alpha}) = \prod_{i=1}^{L} p(\boldsymbol{\theta}|\mathbf{y}, \alpha_i)$ where each term in the product is defined as $p(\boldsymbol{\theta}|\mathbf{y}, \alpha_i) \propto p(\mathbf{y}|\boldsymbol{\theta})^{\alpha_i} p(\boldsymbol{\theta})$. Clearly for $\alpha_i = 0$ the prior is recovered, i.e., $p(\boldsymbol{\theta}|\mathbf{y}, \alpha_i = 0) = p(\boldsymbol{\theta})$ and likewise when $\alpha_i = 1$ the posterior is recovered i.e. i.e., $p(\boldsymbol{\theta}|\mathbf{y}, \alpha_i = 1) = p(\boldsymbol{\theta}|\mathbf{y})$. The Markov chain is now defined over the product density and hence proposals of transitions between densities $p(\boldsymbol{\theta}|\mathbf{y}, \alpha_i)$ at different values of α_i are valid in terms of the convergence of the Markov chain to the desired stationary distribution $p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\alpha})$. Intuitively densities with low values of α_i will be *smoother*[2] thus allowing freer movement of the chain in parameter space, see Fig. 6. The transitions or exchanges between different densities suggests that these freer moving samples can be propagated up this ladder of densities to the desired posterior so having the effect of enabling movement of the chain between the *ridges* of high density. More formal arguments are provided in [15,16] however the use of this population of parallel Markov chains operating at different levels of smoothness between the prior and the posterior provides a means of global sampling from the desired posterior density, Fig. 7.

If the eventual aim is to sample directly from the posterior then the intermediate distributions serve only one purpose, and that is to provide auxiliary bridging distributions to the posterior of interest. However consider the form of the intermediate densities $p(\boldsymbol{\theta}|\mathbf{y}, \alpha) = z_\alpha^{-1} p(\mathbf{y}|\boldsymbol{\theta})^\alpha p(\boldsymbol{\theta})$ where $z_\alpha = \int p(\mathbf{y}|\boldsymbol{\theta})^\alpha p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$ some schoolboy calculus shows that

$$\frac{d}{\mathrm{d}\alpha} \log z_\alpha = \int p(\boldsymbol{\theta}|\mathbf{y}, \alpha) \log p(\mathbf{y}|\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \equiv E_{\boldsymbol{\theta}|\mathbf{y},\alpha} \{\log p(\mathbf{y}|\boldsymbol{\theta})\} \tag{24}$$

and integrating with respect to α yields

$$\log z_{\alpha=1} = \log p(\mathbf{y}) = \int_0^1 E_{\boldsymbol{\theta}|\mathbf{y},\alpha} \{\log p(\mathbf{y}|\boldsymbol{\theta})\} \mathrm{d}\alpha. \tag{25}$$

This expresses the logarithm of the integrated likelihood in terms of the expectations of the log-likelihood with respect to each of the intermediate bridging distributions. Now it was previously highlighted that Monte Carlo estimates of the

[2] Consider a standarised Gaussian density raised to the power of α i.e. $\mathcal{N}(0, 1)^\alpha \propto \mathcal{N}(0, \alpha^{-1})$. Small values of α give a large variance around the mean whilst larger values concentrate the Gaussian tightly around the mean.

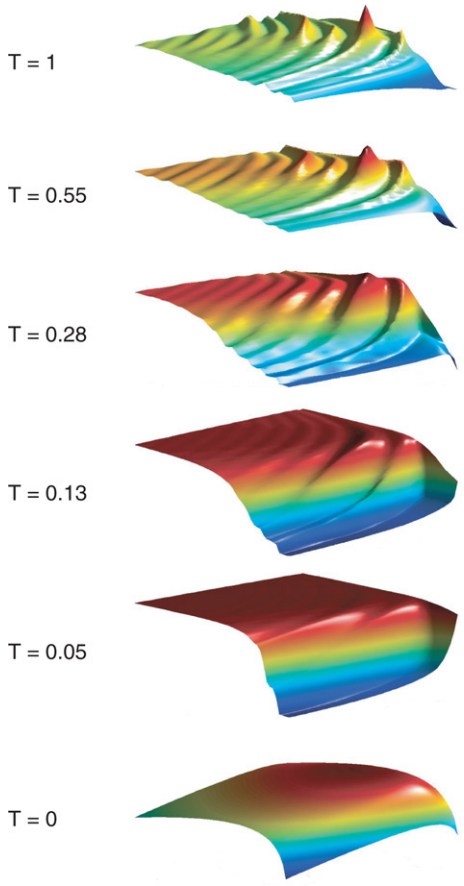


Fig. 6. Tempered posterior surfaces conditioned on two parameters of a 2-variable Goodwin oscillator model. The shapes of the power posteriors change most rapidly between between $t = 0$ and $t = 0.28$, and the overall transition from smooth prior to spiky posterior allows chains to globally explore the parameter space through exchanges between temperatures.

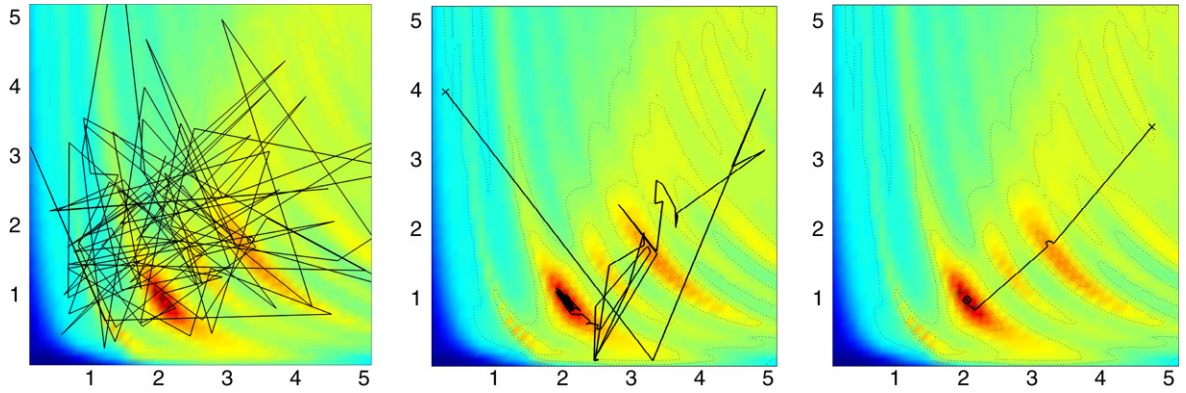


Fig. 7. Samples obtained from a chain at $t = 0$, which is effectively sampling from the prior. The free movement within the parameter space is clear to see. The iso-contours of the posterior are also plotted in this case. Progress of samples drawn from a chain at temperature $t = 0.5$ are shown against the iso-contours of the full posterior. The free movement across modes is most apparent and this is mainly due to the exchange proposals between temperatures. Samples drawn from the posterior, when $t = 1$. There are great differences between this and the highly localised *sticky* exploration in Fig. 5. The Population MCMC algorithm clearly has a much greater ability to move between modes in order to find the most likely one.

integrated likelihood employing samples from the prior will be particularly poor. Consider Eq. (25) which represents the logarithm of the integrated likelihood, it is clear that expectations of the log-likelihood with respect to the bridging densities from the prior to the posterior are integrated in obtaining this expression. So an estimator based on this representation will employ samples drawn from the prior and all intermediate densities up to the posterior thus providing a potentially much superior estimator as is indeed the case [17–19]. An approximation to this integral can be obtained by noting that it can be

represented in terms of a discrete set of α_i values and associated expectations by using the trapezoidal rule for numerical integration

$$\log p(\mathbf{y}) = \frac{1}{2}\sum_{i=1}^{L} \Delta_i \left[E_{\boldsymbol{\theta}|\mathbf{y},\alpha_{i-1}} \{\log p(\mathbf{y}|\boldsymbol{\theta})\} + E_{\boldsymbol{\theta}|\mathbf{y},\alpha_i} \{\log p(\mathbf{y}|\boldsymbol{\theta})\} \right] + \frac{1}{2}\sum_{i=1}^{L} \epsilon_i \tag{26}$$

where $\Delta_i = \alpha_i - \alpha_{i-1}$ and the discretisation error is $\epsilon_i = D(p_{i-1}||p_i) - D(p_i||p_{i-1})$. Each $D(p_i||p_{i-1})$ is the Kullback-Liebler divergence between the posteriors conditioned on α_i and α_{i-1}, defined as

$$D(p_i||p_{i-1}) = E_{\boldsymbol{\theta}|\mathbf{y},\alpha_i} \left\{ \log \frac{p(\boldsymbol{\theta}|\mathbf{y}, \alpha_i)}{p(\boldsymbol{\theta}|\mathbf{y}, \alpha_{i-1})} \right\}.$$

So as proposed in [18,19] the approximation for the integrated likelihood follows as

$$\log p(\mathbf{y}) \approx \frac{1}{2}\sum_{i=1}^{L} \Delta_i \left[E_{\boldsymbol{\theta}|\mathbf{y},\alpha_{i-1}} \{\log p(\mathbf{y}|\boldsymbol{\theta})\} + E_{\boldsymbol{\theta}|\mathbf{y},\alpha_i} \{\log p(\mathbf{y}|\boldsymbol{\theta})\} \right] \tag{27}$$

where the samples from each intermediate distribution can be used to obtain Monte Carlo estimates for each $E_{\boldsymbol{\theta}|\mathbf{y},\alpha_{i-1}} \{\log p(\mathbf{y}|\boldsymbol{\theta})\}$ appearing in the approximation above.

This also suggests a strategy for discretising the unit line in order to minimise the error which consists on varying the partition widths Δ_i to drive each $D(p_{i-1}||p_i) \to 0$ and minimise the variance of the overall estimator for the integrated likelihood. It is straightforward to obtain the minimum-variance partition as

$$\Delta_i \propto \frac{1}{E_{\boldsymbol{\theta},\Delta_i|\mathbf{y}} \left\{ \log p(\mathbf{y}|\boldsymbol{\theta})^2 \right\}} \approx \frac{2}{E_{\boldsymbol{\theta}|\mathbf{y},\alpha_i} \left\{ \log p(\mathbf{y}|\boldsymbol{\theta})^2 \right\} + E_{\boldsymbol{\theta}|\mathbf{y},\alpha_{i-1}} \left\{ \log p(\mathbf{y}|\boldsymbol{\theta})^2 \right\}} \tag{28}$$

which can be used to define an adaptive scheme to estimate the width of the partitions. Starting with a uniform partitioning $\Delta_i = L^{-1}$ the above can be used to update each Δ_i subject to $\sum_i \Delta_i = 1$. This scheme clearly will reduce the partition widths if the approximate total variance across the partition is large and increase the width as the total variance drops.

We now have all the computational tools required to obtain estimates of the measures necessary to perform posterior inference over parameters $p(\boldsymbol{\theta}_k|\mathcal{D}, \mathcal{M}_k)$ and model structures, $P(\mathcal{M}_k|\mathcal{D})$, and the following section will illustrate this with some examples of simple biochemical models described in terms of systems of non-linear ODE's.

7. Illustrative experiment

The example will consider the two ODE models described in Section 5. Given that there are two candidate models the first task is to define a set of prior distributions over the parameters.

7.1. Prior distributions

The prior densities which were used are as follows. For k_1, V_1, Km_1, V_2, Km_2 and σ a Gamma prior with shape and scale parameters set to one i.e. $\alpha = 1$ and $\gamma = 1$, denoted as $\mathcal{G}(1, 1)$ which has a mean value of $\alpha\gamma = 1$ and variance $\alpha\gamma^2 = 1$. This is a reasonable prior density to assume for these parameters as, (1) it has support on the positive part of the real line and so satisfies the positivity constraints on the kinetic parameters and the noise variance, (2) it possesses a variance which is sufficiently wide to cover the plausible range of values for the kinetic parameters so as to retain the same scaling as the variables S, D, R and Rpp of the model. The use of $\mathcal{G}(1, 1)$ as a prior for the error variance is selected based on the empirical variance of the observed data. The prior densities for the initial values S_0, D_0, R_0 and Rpp_0 were also defined using Gamma distributions. Due to no measurements being available at $t = 0$ the trend of the data at $t = 0$ was estimated and the variance at this point was reasonably assumed similar to the empirical variance of the observations. As such S_0, $R_0 \sim \mathcal{G}(5, 0.2)$ which has mean of 1 and variance of 0.2 which is large compared to the empirical variance. Likewise D_0, $Rpp_0 \sim \mathcal{G}(1, 0.1)$ having mean 0.1 and variance 0.01 again reflecting the plausible values based on the trend of the data towards $t = 0$. It may be that actual measured values for the parameters of interest are available in which case these may be used as a location parameter and an associated spread can be defined depending on how reliable the measurements are considered.

7.2. Posterior inference

The parameter and initial values detailed in Fig. 8 were used in the system of equations to run the model in simulation mode and the levels of R_{pp} were monitored at twenty time intervals. Each of the R_{pp} levels had a small amount of additive Gaussian noise with $\sigma = 0.02$ to simulate measurement and observation error. A product space sampling[3] method was employed and the marginal posterior mean and variance obtained for each of the parameters are listed in the table of Fig. 8. It is clear that the marginal posterior distributions have a main mode which are centered on the true parameter values.

[3] A software tool to perform Bayesian inference over ODE models is available at http://www.dcs.gla.ac.uk/biobayes/.

Parameters	Posterior $\mu \pm \sigma$
$k_1 = 0.05$	0.050 ± 0.001
$V_1 = 0.20$	0.168 ± 0.019
$Km_1 = 0.10$	0.094 ± 0.005
$Km_2 = 0.10$	0.107 ± 0.018
$V_2 = 0.10$	0.090 ± 0.005
$S_0 = 1.00$	0.990 ± 0.006
$D_0 = 0$	0.003 ± 0.006
$R_0 = 1.00$	0.913 ± 0.051
$Rpp_0 = 0$	0.085 ± 0.051
$\sigma = 0.020$	0.021 ± 0.001

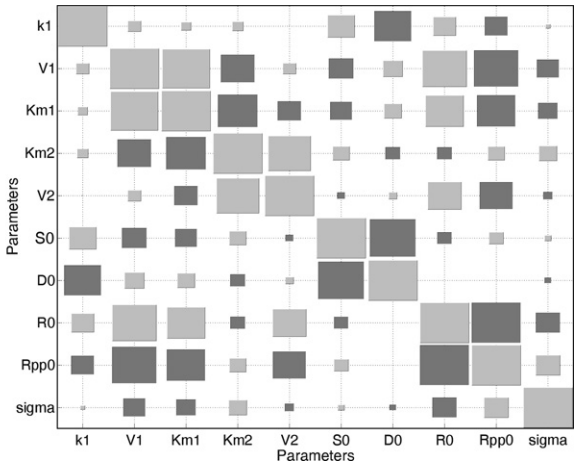


Fig. 8. The left hand table shows the posterior mean and variance of the model parameters with the right hand diagram highlighting the covariance structure of the full parameter posterior distribution.

It is illuminating to study the full posterior density obtained by the sampling $p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{t})$ where the set of parameters are $\boldsymbol{\theta} = (k_1, V_1, Km_1, V_2, Km_2, S_0, D_0, R_0, Rpp_0, \sigma^2)$. If we consider the Hinton diagram in Fig. 8 significant posterior correlations between a number of the model parameters can be observed. For example we can see that Km_1 and V_1 have a strong positive correlation as does Km_2 and V_2. If we study the system of equations (15) defining $\mathcal{M}_1$ we see that these posterior correlated parameters are employed in defining the chemical reactions using the Michaelis–Menten kinetic law. This dependency arises due to the mathematical form of the reaction equation. The maximum rate at which the enzyme S catalyses the activation reaction when the concentrations of substrate R are high is denoted as V_i and Km_i corresponds to the concentration of R that allows enzyme S to generate Rpp at half their maximum rate. Both of these parameters occur in a rational form and so there is a structural dependency in the effect they will have. Consider further the strong negative posterior correlation between the initial values of the chemical species R_0 and Rpp_0. This is due to the fact that R and Rpp are involved in a cyclic process of activation and deactivation and therefore can be converted into each other. The positive correlation between V_1 and R_0 can be explained by the requirement to sustain a sufficient rate of production of Rpp. Clearly appropriate re-parameterisation can eliminate such posterior dependencies and the information with the posterior can be employed in subsequent model simplification studies.

As it is possible now to also obtain estimates of the integrated likelihoods for each model which can then be used to rank and compare models (see [7]) given some observed data. For the two models considered we simulate data (levels of Rpp) and then compute the integrated likelihoods of the data under each candidate model. Denoting $\mathbf{y}_1$ as data generated from model 1 (the 'true' model) and likewise $\mathbf{y}_2$ as the data simulated from model 2 then we proceed to compute estimates of the integrated likelihoods $\log p(\mathbf{y}_1|\mathcal{M}_1)$ and $\log p(\mathbf{y}_1|\mathcal{M}_2)$. Obtaining these measures and assuming that *a priori* neither model is preferred i.e. $\pi(\mathcal{M}_1) = \pi(\mathcal{M}_2)$ then we can obtain the posterior odds of model 1 over model 2 which in this case is referred to as the Bayes Factor, $\log B_{12} = \log p(\mathbf{y}_1|\mathcal{M}_1) - \log p(\mathbf{y}_1|\mathcal{M}_2)$. This then provides a means of objectively assessing the evidential support for one model over another. In this case we find that $\log p(\mathbf{y}_1|\mathcal{M}_1) \approx 861.0849 \pm 0.3699$ and $\log p(\mathbf{y}_1|\mathcal{M}_2) \approx 771.7686 \pm 0.0714$ indicating very strong support for model 1, which of course was the 'true' model which gave rise to the observations. Likewise employing data generated from model 2 we obtain $\log p(\mathbf{y}_2|\mathcal{M}_1) \approx 706.8671 \pm 0.5425$ and $\log p(\mathbf{y}_2|\mathcal{M}_2) \approx 851.6974 \pm 0.5015$ indicating very strong support for the 'correct' model.

8. Conclusions

There is often a requirement to perform a statistical analysis of mechanistic models based on ordinary differential equations, in particular characterising the residual uncertainty in estimates of parameters or unobserved components and assessing the evidential support of competing model definitions. It is argued that the Bayesian inferential framework provides a consistent approach to defining and propagating uncertainty within such mechanistic models. Examples based on simple biochemical models have been provided to illustrate the power of this methodology in enabling reasoning over models.

Acknowledgments

Mark Girolami is funded by an Advanced Research Fellowship from the Engineering and Physical Sciences Research Council (EPSRC) EP/E052029/1. This work also receives funding from Microsoft Research under their Towards 2020 Science Programme.

References

[1] E.O. Voit, Computational Analysis of Biochemical Systems, Cambridge University Press, 2000.
[2] B. Schoeberl, C. Eichler-Jonsson, E.D. Gilles, G. Muller, Computational modelling of the dynamics of the MAP kinase cascade activated by surface and internalised EGF receptors, Nature Biotechnology 20 (2002) 370–375.
[3] W. Kolch, M. Calder, D. Gilbert, When kinases meet mathematics: The systems biology of MAPK signalling, FEBS Letters 579 (8) (2005) 1891–1895.
[4] C.G. Moles, P. Mendes, J.R. Banga, Parameter estimation in biochemical pathways: A comparison of global optimization methods, Genome Research 13 (11) (2003) 2467–2474.
[5] J.M. Bernardo, A.F.M. Smith, Bayesian Theory, Wiley, 1994.
[6] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, Bayesian Data Analysis, 2nd edition, Chapman & Hall/CRC, 2003, pp. 296–297.
[7] V. Vyshemirsky, M. Girolami, Bayesian ranking of biochemical system models, Bioinformatics 24 (2008) 833–839.
[8] D. Denison, C.C. Holmes, B. Mallick, A. Smith, Bayesian Methods for Nonlinear Classification and Regression, Wiley, 2002.
[9] A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, Bayesian Data Analysis, Chapman & Hall/CRC, 2003.
[10] W.K. Hastings, Monte carlo sampling methods using markov chains, and their applications, Biometrika 57 (1970) 97–109.
[11] W. Gilks, S. Richardson, D. Spiegelhalter (Eds.), Markov Chain Monte Carlo in Practice, Interdisciplinary Statistics Series, Chapman & Hall/CRC, 1995.
[12] R.E. Kass, A.E. Raftery, Bayes factors, Journal of the American Statistical Association 90 (430) (1995) 773–795.
[13] B. Calderhead, M. Girolami, Estimating bayes factors for nonlinear ode models via thermodynamic intehration and population mcmc, Computational Statistics and Data Analysis (2008).
[14] B. Goodwin, Oscillatory behavior in enzymatic control processes, Adv. Enzyme Regul. 3 (1965) 425–438.
[15] C. Geyer, Parallel tempering, in: Computing Science and Statistics Proceedings of the 23rd Symposium on the Interface, American Statistical Association, New York, 1991, p. 156.
[16] C.J. Geyer, Practical markov chain monte carlo, Statistical Science 7 (1992) 473–482.
[17] A. Gelman, X. Meng, Simulating normalizing constants: From importance sampling to bridge sampling to path sampling, Statistical Science 13 (2) (1998) 163–185.
[18] N. Friel, A.N. Pettitt, Marginal likelihoods via power posteriors, Journal of the Royal Statistical Society, Series B 70 (3) (2008) 589–608.
[19] N. Lartillot, H. Philippe, Computing bayes factors using thermodynamic integration, Syst. Biol. 55 (2) (2006) 195–207.