Using Gradient Matching to Accelerate Parameter Inference in Nonlinear Ordinary Differential Equations

Alan Lazarus
School of Mathematics and Statistics
University of Glasgow

Thesis submitted for the degree of Master of Science

September 2018

Abstract

Ordinary Differential Equations are becoming more widely used throughout all branches of science to model systems of interacting variables. Although researchers can often postulate the structure of the ODEs, there remains a desire to better infer the parameters of these systems. After all, it is these parameters that provide improved understanding of the dynamics involved. Traditionally, parameter inference was done by solving the system of ODEs and assessing fit of the estimated signal with that of the observations. However, nonlinear ODEs often do not permit closed form solutions. Using numerical methods to solve the equations results in prohibitive computational cost, particularly when one adopts a Bayesian approach in sampling parameters from a posterior distribution.

The difficulties above have led to the introduction of gradient matching to the parameter inference problem. Instead of quantifying how well the solutions of the ODEs match the data, we quantify how well the derivatives predicted by the ODEs match the derivatives obtained from an interpolant to the data. These methods aim to more efficiently infer the parameters of the equations, but inherent in these procedures is an introduction of bias to the learning problem as we no longer sample based on the exact likelihood function. It is desirable that we obtain a method for parameter inference that is both accurate and efficient, necessitating the involvement of the exact likelihood at some point in the algorithm. Combined with the problems faced in ODE parameter inference, this idea will motivate the main result of this thesis, the introduction of a multiphase scheme in parameter inference that allows us to benefit from the efficiency of the gradient matching likelihood function and the accuracy of the exact likelihood function. The performance of this proposed method is assessed on four benchmark ODE systems, comparing with some standard MCMC sampling techniques from the literature.

Acknowledgements

I thank my parents for their continued emotional support and my supervisors, Professor Dirk Husmeier and Dr Theodore Papamarkou, for their academic support throughout the year of my project. I'd also like to thank my examiners for their very helpful comments.

Contents

List of Figures

List of Tables

Symbols

$\tilde{\mathbf{x}}$	a random interpolant of the observed values
$\dot{\tilde{\mathbf{x}}}$	the gradient of the random interpolant of the observed values
$\hat{\mathbf{x}}$	fixed point estimate of the interpolant
$\dot{\hat{\mathbf{x}}}$	fixed point estimate of the gradient
$\hat{p}(\mathbf{y}\|\boldsymbol{\theta})$	a surrogate (approximate) likelihood function
$k(\cdot,\cdot)$	a kernel function
$k'(x,x')$	derivative of kernel function with respect to x
$'k(x,x')$	derivative of kernel function with respect to x'

Chapter 1

Introduction

1.1 Background

Consider a scenario where one is provided with some noisy data from an experiment and wishes to fit a specific system of equations to capture the dynamics of the observed data. For example, Figure 1.1 presents the observed lynx and hare populations obtained from Hudson Bay over the period from 1900 to 1920 [2]. If one were able to express the dynamics of this system mathematically, then we would be able to estimate the population evolution beyond the timescale for which data have been provided. Moreover, assuming this mathematical model contained some explicit parameters, we could infer the effects of different factors on the system.

Differential equations (DEs) have long been used to express variations in different measurements over time. Modelled as a function of the different variables in the system, these expressions can formulate the rate of change of different dependent variables, $\mathbf{x}(t)$, varying across some independent variable, t, as a function of the variables in the system and some set of parameters:

$$\frac{d\mathbf{x}(t)}{dt} = f(\mathbf{x}(t), \theta). \tag{1.1}$$

Ordinary differential equations occur when the system depends on only one set of independent variables as in eq. 1.1. Returning to the problem of fitting a model to the Hudson Bay data, allow the prey data to be modelled by x and the predator y. In the absence of predators, the prey population is free to thrive, growing at a rate $dx = \alpha x dt$. Interactions between the two species occur at a rate proportional to the populations of

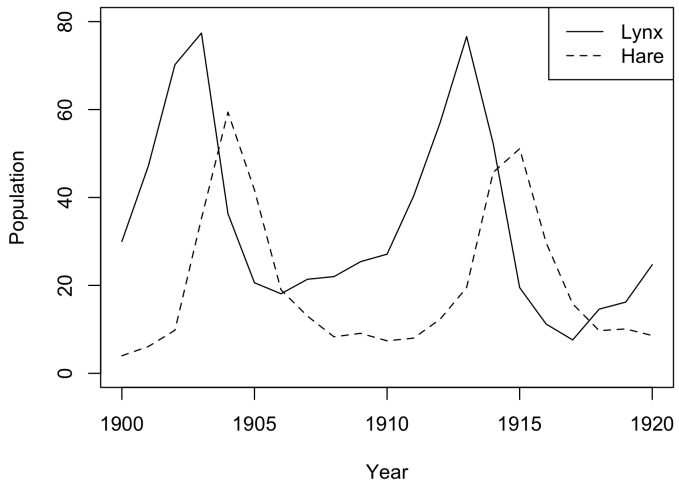

FIGURE 1.1: Evolution of lynx and hare populations over a period of 20 years in Hudson bay.

each species, expressed as $dx = -\beta xydt$. Expressions for the evolution of the predator variable are identical up to negation of the right-hand sides. The resultant ODE system, known as the Lotka-Volterra system of equations [3] may be expressed as follows:

$$\frac{dx}{dt} = \alpha x - \beta xy$$
$$\frac{dy}{dt} = -\gamma y + \delta xy$$

At this point, I stress that the correct choice of model is not always immediately obvious. In these scenarios, we may use critera such as Bayes factors and likelihood ratios to select a suitable model $\mathcal{M}$ [4, 5]. Under a Bayesian framework, we may now proceed with parameter inference, accounting for uncertainty in the model selection procedure by conditioning on the posterior distribution of the model (this is briefly discussed in Section 2.1).

Isolating the variables of interest, we may formulate the recovery of solutions of the ODEs as a recovery of the input-output pairs of a map:

$$h : \Theta \to \mathcal{Y}. \tag{1.2}$$

Under an assumption of injectivity, the capability to recover the signals by solving the forward problem (given in eq. 1.2) permits recovery of the parameters of the ODEs—the inverse problem—and vice-versa. Traditionally influenced by this mapping, the ability to

solve the inverse problem depended on an ability to recover the signal at particular parameter configurations. Minimising the misfit between solutions of the ODEs at parameter values and the noisy observations, this inference scheme relies on an ability to integrate the ODEs. In the case of linear ODEs, analytic solutions exist, and the inference problem can rely on standard model fitting methods. The inference problem becomes more interesting when one considers non-linear systems of equations for which closed form solutions are not permitted, such as the case of the Lotka-Volterra model where xy terms limit the existence of closed form solutions. This property necessitates the use of numerical procedures, making evaluation of the mapping in eq. 1.2 computationally onerous.

Recently, there has been a paradigm shift in the field of ODE parameter inference with the reemergence of Bayesian statistical methods and, in particular, the development of Markov chain Monte Carlo (MCMC) methods for sampling from probability distributions. Whereas much of the inference literature used to focus on optimisation, a large bulk of the recent work adopts an uncertainty quantification approach to the parameter learning procedure. Problematically, when one adopts a Bayesian approach, the computational cost is compounded by a requirement for MCMC sampling and the subsequent solving of the ODE system at each iteration. Induced by this computational complexity has been a shift towards more efficient metrics with which to match parameter configurations with the observed data values, perhaps replacing the space $\mathcal{Y}$ from eq. 1.2 with some surrogate space that allows us to mimic the fitting of parameters to the observations in a cheaper environment [6–8].

1.2 Bayesian Inference

Bayesian inference is the subset of statistical inference in which all beliefs and assumptions are quantified as probabilities. Although we accept that there is a unique, true solution to the inverse problem of the system, we do this while acknowledging the lack of belief we have regarding the true value of these parameters, representing this as a prior distribution $P(\theta)$. Randomness here is thus an artefact of our own personal uncertainty as opposed to being a property of the parameters themselves. In light of the data, one is in a position to update their belief about the true value of the parameters, introducing this information via a likelihood function $P(x|\theta)$. Combining the information from the prior belief and the data, we can summarise the knowledge of the parameters in a posterior distribution

via Bayes Theorem:

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)} = \frac{P(x|\theta)P(\theta)}{\int p(x|\theta')p(\theta')d\theta'}. \tag{1.3}$$

which is the product of the likelihood and prior, normalised by the marginal likelihood (evidence), $P(x) = \int p(x|\theta')p(\theta')d\theta'$. More detail on the different probabilities is given in the proceeding subsections.

1.2.1 Likelihood

Reconsidering the mapping in eq. 1.2, we may consider the sampling distribution, $f(y|\theta)$ as a function providing a measure of the probability of parameter-observation pairs. Regarded as a function of fixed θ it represents a probability distribution over the codomain, $\mathcal{Y}$, of the mapping. In the Bayesian context, we fix a point in $\mathcal{Y}$; thus, defining a distribution over the observation space would be nonsense. Instead, we reparametrize $f(y|\theta)$—the likelihood function—as a function providing the probability that the observations were produced by a particular set of parameters. The likelihood provides a mapping from the space of parameters Θ to the real numbers $\mathbb{R}$ that provides a measure of how likely it is that, given the parameter value, we observe the observed data. Ultimately, this likelihood function is the choice of the modeller, but in most cases, the nature and structure of the data will strongly influence this decision. In Bayesian inference, it is often more computationally convenient to consider the log-likelihood function as this avoids taking the product of very small values which, due to numerical float precision, can make likelihood evaluation imprecise.

1.2.2 Prior Probability Distribution

The prior probability distribution quantifies our belief about the parameters before we have seen the data. This may be capturing information from previous experiments, or simply quantifying constraints that we know the parameters are subjected to. Unless adopting an empirical Bayes approach, the prior should never rely on properties of our data. The argument here being that any structural information from the data could be accounted for by a more astute modelling process. In the case of ODE inference, the prior allows us to take into account any previous experiments carried out by the researcher,

as well as any constraints that are placed on the parameter value by their physical interpretation. When no prior information is available, one may adopt an uninformative prior which expresses vague information about the parameter in question. This is referred to as the objective Bayesian approach as we allow the data to speak for itself without being influenced by our preconceived beliefs. As a result, subjectivity is neglected as the procedure is informed purely by the nature of the likelihood function (there is, of course, the knowledge of the subset of $\mathbb{R}$ in which the parameters will lie). This line of thought may be considered as a bridge point between the frequentist and Bayesian inference approaches.

Rather fittingly, uninformative priors remain a fairly vague concept in Bayesian statistics: how exactly does one represent a lack of belief about the outcome of a problem? Laplace, motivated by the principle of insufficient reason, adopted a uniform prior to quantify a state of ignorance about the true parameter values [10]. However, under reparameterization, this becomes strictly informative (see Figure 1.2) and begins to favour points near the boundary of the uniform prior support. Viewing from a purely logical perspective, given no knowledge about $\boldsymbol{\theta}$, we certainly should also have no knowledge of $g(\boldsymbol{\theta})$. Therefore, the uniform prior is not sufficiently non-informative.

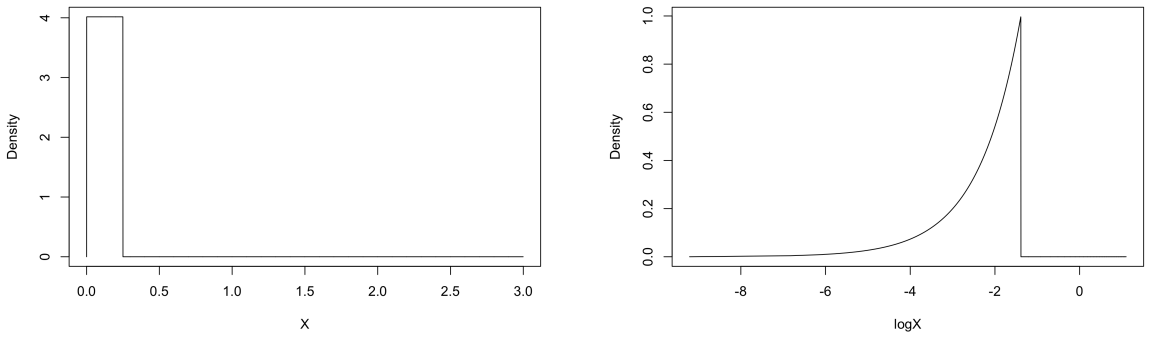

FIGURE 1.2: "Noninformative" uniform prior on original scale (left) and log scale (right). There is no ambiguity on the log scale—this is an informative prior under the log transformation.

Jeffreys [11] proposed the use of the function:

$$p(\boldsymbol{\theta}) \propto \sqrt{\det(\mathcal{I}(\boldsymbol{\theta})} \tag{1.4}$$

known as Jeffreys prior, as a solution to the non-informative prior search [10]. Based on the notion of maximising the Fisher information and placing the prior density at regions of high information, thus favouring points for which the likelihood contains high

levels of information, we effectively dilute the effect of the prior distribution. In addition, the Fisher information is closed under reparameterization of the inference problem (see Section 5.4.2 of [12]). This formulation has proven highly popular and although a lack of generality to higher dimensional parameter spaces is unappealing, Jeffreys prior provides the motivation for our noise variance prior in the case of ODE parameter inference (see section 2.1.2).

As previously discussed, intractability of the denominator of eq. 1.3 can make posterior probability calculations impossible. Although the likelihood function is fixed as a by-product of the data generating process, we have some flexibility in our prior distribution. Furthermore, different prior densities can formulate similar assumptions on the parameters in question. As such, it makes sense that, where possible, we simplify calculations through careful choice of the prior distribution. By careful selection of our prior distribution, we may form a posterior distribution belonging to the same general family of distributions as the prior distribution. These prior distributions, referred to as conjugate priors [13], can be used to reduce computational burden by leading to posteriors with general distribution equal to that of the prior (with updated hyperparameters).

1.2.3 Posterior Distribution

The posterior distribution, $p(\theta|y)$, follows from the formula in eq. 1.3. This distribution quantifies our belief about the parameters in question upon taking account of the observed data and represents an updated form of the prior (with support equal to that of the prior distribution). Whereas the prior merely formalises the learning problem in the correct region of parameter space, allowing us to use information from the likelihood for the learning process, the posterior represents our belief about the values of the parameters in light of the information provided by the data. The conclusions drawn from the posterior distribution are obviously affected by the nature of both the prior distribution and the likelihood function. Adopting a more uninformative prior leads to an inference scheme dominated by the likelihood function in an objective Bayes approach, allowing the data to speak for itself. Stronger prior information lessens the effect of the likelihood on the posterior distribution and can enable inference in cases where the mapping in eq. 1.2 violates the injectivity property.

The challenge is to provide some representation of our knowledge of the parameter values from this posterior distribution. Numerical integration methods allow evaluation of the likelihood function of nonlinear ODEs, but this property does not provide sufficient

knowledge of the distributional properties of the posterior distribution. Considering Bayes theorem, the lack of closed form solutions of the ODEs necessitates the use of sampling methods to provide an empirical estimate of the true posterior distribution. Elementary sampling procedures require closed form expressions leading to a reliance on Markov chain Monte Carlo methods that sample based on evaluation of the likelihood function.

1.3 Difficulties of Parameter Inference in ODEs

The standard ODE inference procedure assumes iid additive Gaussian noise in our observational data model, leading to a Gaussian distribution for the data and a likelihood function that, essentially, acts as a metric measuring the fit of ODE solutions at particular parameter settings to the observed data. Nonlinear systems of ordinary differential equations often do not possess closed solutions and so we must solve these systems numerically. Robinson provides a background on explicit solutions of ODEs [14]. Particularly relevant to this thesis are Runge-Kutta methods which rely on discretisation of the independent variable domain and gradients obtained at multiple points to propagate through from an initial value to the inferred state measurements [15]. Inference may be carried out by evaluating solutions at each individual parameter set and comparing with our observations via some metric. The main sticking point in this process is the onerosity involved in repeated numerical integration of the ODEs. The other downside here is the stance of ignorance taken with regards to the numerical errors involved in the repeated numerical solving of the ODEs, which, according to Xue et al. [16] is only justified in cases where the maximum step size of a p-order numerical algorithm goes to zero at a rate faster than n^{-1/p^4} as the magnitude of the observational error dwarves that of the numerical error. In this thesis we are dealing with artificial data and so the numerical error resulting from this numerical integration is neglected from the uncertainty involved.

In general, there are two routes to overcoming these inefficiencies. Firstly, we may adopt a cheaper likelihood function allowing us to take a greater number of samples in reasonable time frames, similarly to the work done with pseudolikelihoods [17]. However, this then reduces accuracy of our inferred parameter values. Alternatively, we can take the approach of population MCMC and simulated annealing, and find a smoothed version of the likelihood surface, enabling more liberated movement through the parameter space. This time, the obstacle is inefficiency as we are dependent upon a numerically expensive likelihood function in our sampling routine.

One method that attempts to circumvent this numerical solution is the use of a surrogate approach to the inference problem where we first find a smooth surrogate interpolant representation of the observed data. Dattner et al. [18] explore systems of ODEs where the parameters enter the systems linearly. Exploiting this linearity, the authors implement a two-phase scheme to avoid the numerical integration. Ranciati et al. [19] present a Bayesian smooth-and-match approach where they adopt a penalised spline framework for smoothing the data and use ridge regression to infer the parameters of ODEs in which the parameters enter linearly; the crux of the method being an assumption of two generative models for the observations. Similar to the problem faced with the approach of Wang and Barber [20], this leads to an inability to deal with partially observed systems. Like Ranciati et al, Ramsey et al. [6] implement a collocation approach, where they adopt a penalised smoothing approach to parameter inference by optimising a penalised log-likelihood function that trades off fit of a spline interpolant to the ODEs with fit to the noisy data observations. This paper was the first instance in which there was an attempt to regularise the interpolant of the noisy data, making it consistent with the nature of the ODEs.

Recently, surrogate methods based on gradient matching have been proposed. These methods allow us to circumvent the numerical approximation step by finding a smooth interpolant to the noisy data from which we may find an explicit gradient corresponding to the signals from the ODEs. This may then be matched with a functional gradient obtained by evaluating the functional relationship of the differential equations at these smoothed signals providing a surrogate likelihood which represents some approximation to the exact, complex likelihood function. These surrogate likelihoods enable more efficient inference, but one must be wary of the estimative power of these functions with respect to the exact likelihood function. Although present methods of implementing the gradient matching approach assume a Bayesian setting, there are instances of frequentist approaches to the problem.

1.4 Thesis Outline

Motivated by previously outlined parameter inference issues, the aim of this work is to develop a novel method for carrying out parameter inference in nonlinear systems of ODEs. The route to parameter estimation in these systems can adopt one of two approaches: point estimation versus estimating the posterior distribution, the latter usually

by sampling or via variational inference. Chapter 2 will consider the comparison between these two approaches and outline the necessity for us to quantify our uncertainty in the parameter values. I will introduce sampling in Bayesian inference as a method for performing parameter inference in ODEs while giving some elementary, yet problem dependent, sampling methods. The inapplicability of the elementary approaches will motivate a discussion of MCMC methodology, and some benchmark algorithms with which a proposed method may be compared. In spite of these MCMC methods, Chapter 3 considers the problems caused by properties of the ODEs in MCMC methods for parameter inference through the introduction of four benchmark ODE systems that will be used to compare the different inference methods discussed in this thesis. In an attempt to alleviate the problems encountered with traditional MCMC inference, Chapter 4 will briefly discuss Gaussian process interpolation, before introducing the proposed multi-phase MCMC scheme through which I hope to reduce computational complexity and improve accuracy of the parameter inference procedure. Chapter 5 contains a thorough comparison of the proposed scheme with some of the traditional MCMC approaches. In Chapter 6 I will conclude by discussing the advances made with this method and the potential for further improvement in ODE parameter inference, with particular attention paid to the potential extensions to the multi-phase approach.

Chapter 2

Theoretical Overview

The process of parametric model fitting necessarily involves a stage of parameter inference. We may formulate the inference problem as an exploration of the likelihood which quantifies the fit of parameter-observation pairs. In simple cases where the system is solvable in closed form, recovery of this function is trivial, and the inference procedure can rely on standard methods. The problem, however, is the lack of closed form solutions in the case of complex nonlinear ODE systems and the subsequent lack of a closed form posterior distribution for the parameters. In this case, we require more sophisticated sampling methods which, in fact, do not tend to sample directly from the distribution of interest at all but rather rely on the construction of correlated chains that, asymptotically, converge to our stationary distribution of interest—the posterior distribution of the ODE parameters. This chapter gives a general overview of the use of MCMC in parameter inference, becoming more specific to the ODE inference problem when deemed necessary. Starting with a discussion of the inference problem, I acknowledge the frameworks of both Bayesian and Frequentist inference methods through the use of both point estimation and sampling methods. Having justified the adoption of a Bayesian approach, I discuss the components of a standard Bayesian model for parameter inference, before presenting the more sophisticated Markov Chain Monte Carlo (MCMC) methods required for parameter inference in complex non-linear systems.

2.1 Bayesian and Frequentist Inference in ODEs

Consider a process defined by the system of equations:

$$\frac{d\mathbf{x}}{dt} = f(\mathbf{X}, \boldsymbol{\theta}, t) \tag{2.1}$$

and suppose we possess noisy measurements $\mathbf{y}$ corresponding to noisy realisations of latent variables governed by this system of equations corrupted by iid additive Gaussian noise ϵ.

$$y(\mathbf{t}) = x(\mathbf{t}) + \epsilon \tag{2.2}$$

Considered as a couplet, eqs. 2.1 and 2.2, along with any initial conditions, fully specify our parameter estimation problem. That is, given noisy observations, we recover the smoothed signal (and therefore the noise variance) and based on fit to the noisy observations, look to infer the parameters of our ODEs (in addition to the initial conditions when these are assumed unknown). This fit to the observations is assessed with the likelihood function, a measure of the fit of the parameters to the data that is specified by the assumed distribution of the additive noise.

We hope to recover the model structure based on the fit of the ODEs to the observed data. In the most uninformed case we have three degrees of freedom in this problem: the model, initial conditions and the model parameters where degrees of freedom refer to a component of the model that leads to variation in the obtained signal. The latter two of these are nested within the first and, given this dependence of our space of parameters and initial conditions on the parametric form of the model, we cannot begin to estimate the parameters and initial conditions before this model selection has taken place. Inference for these parameters and initial conditions may be considered as fine tuning of our selected model. Girolami [4] provides an overview of the model selection problem and the use of Bayes factors for selection of our ODE structure (frequentist approaches would rely on likelihood ratios, with the disadvantage of being unable to compare unnested ODE models). It is often the case in nonlinear ODEs that the marginal likelihood is intractable. In these cases, we must estimate the marginal likelihood using approximate methods [21]. A Bayesian formulation enables us to sample parameters from the conditional distribution $p(\boldsymbol{\theta}|\mathbf{y}, \mathcal{M}_i)$, taking account of added uncertainty in the model. In our case, we assume the given parametric form of the equations (and for the large part, known initial conditions) and so have at most two degrees of freedom in the inference problem. Nonetheless, it is

important to be aware of this additional degree of uncertainty in the general inference problem with real observed data.

2.1.1 Representing the Posterior Distribution

Parameter inference may proceed under different assumptions on the nature of the ODE parameters. One could assume the existence of a fixed parameter value and consider the distribution of some estimator, the MLE, of the parameter when applied over multiple conceptualised datasets. Uncertainty in parameter estimates may be accounted for using an approximation of the asymptotically Gaussian sampling distribution obtained using methods such as bootstrapping. It makes sense that for successful inference, this sampling distribution be centred on the true parameter value, providing an unbiased estimator for the ODE parameters. However, this property alone does not provide an attractive estimator. If variance of the sampling distribution is high, then estimates of parameters will tend to be more inaccurate than the unbiasedness would suggest. Thus, we wish to provide a minimum variance unbiased estimator for the parameters. The variance can be quantified by the curvature of the likelihood surface. With high (low) curvature at the maximum, we expect low (high) variance in the estimator. The Fisher Information matrix provides a quantification of this uncertainty forming the Cramér-Rao lower bound for a lower bound on the variance of an unbiased estimator. Any unbiased estimator that attains this lower bound on the variance provides a minimum variance unbiased estimator for the ODE parameters. Under these previous assumptions, we would be adopting a frequentist approach to parameter inference. If instead we choose to place a prior distribution over the parameters, updating the parameter distribution based on the likelihood function, we would be adopting a Bayesian approach. In ODE parameter inference, use of the prior distribution proves beneficial as the possible domain of the parameters in question can often be constrained based on physical interpretations. The probabilistic assumptions associated with Bayesian inference mean that uncertainty quantification naturally follows under the Bayesian line of thought and, while acknowledging the ability of Frequentist methods to perform successful parameter inference in ODEs [6], the work of this thesis adopts a Bayesian approach to the inference problem. Kevin Murphy provides a useful overview of the general paradigms of Frequentist and Bayesian statistics [12].

Abiding by a Bayesian framework, we may place a prior on the parameters and take the mode of the posterior distribution as the Maximum aposteriori (MAP) estimate of the parameters. Merely representing a regularised version of the maximum likelihood

estimate, this fails to fully account for the knowledge we have obtained in the posterior distribution of the parameters. Alternatively, one can take a decision theoretic approach by specifying a loss function and taking the estimate as the value minimising the mean squared error—the mean of the posterior distribution [12]. However, inherent in point estimation schemes are two assumptions that appear fairly inconsistent in our task of estimating parameter values of ODEs. Firstly, we are assuming a level of confidence in the ability to find the optimal parameter. This is a fairly optimistic assumption given the possibly complex nature of the distributions stemming from ordinary differential equations. Secondly, and a problem commonly witnessed in complex nonlinear ODEs, we must also be assuming identifiability of the parameters in question. This corresponds to injectivity of the mapping in eq. 1.2 and cannot be diagnosed using a single point estimate.

Representation of uncertainty in the parameter posterior distribution relies on sampling methods. Use of elementary methods such as the inverse-transform and accept-reject depends on the posterior distribution being tractable, a property violated by nonlinear ODEs. Conceptualised in the 1990s, Monte Carlo methods were able to overcome the problem of intractable posterior distributions by sampling based on the relationship:

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{2.3}$$

However, for most interesting ODE models an extra level of intractability in the likelihood function means that these methods become inapplicable in this context. Inference in this case relies on more sophisticated Markov Chain Monte Carlo methods which may be presented simply as extensions of the more elementary sampling scheme where we sample from some proposal distribution and, through some transformation resulting from a Markov kernel, produce a Markov chain that is invariant with respect to the desired limiting distribution. Before discussing MCMC methodology, it is worth considering a final, crucial, part of the inference process—the likelihood and prior distribution.

2.1.2 Likelihood and Prior

Vital to the inference procedure is careful selection of the likelihood function and prior distribution. Although not immediately apparent from the data, the likelihood function is fixed by the data generating process and successful inference requires be suitably selected.

Under an assumption of Gaussian noise, we have the distribution:

$$\mathbf{y} \sim \mathcal{N}(\mathbf{x}, \sigma^2 \mathbf{I}) \tag{2.4}$$

where $\mathbf{x}$ denotes the numerical solution of the ODE and σ is assumed constant across all time points. This provides a log-likelihood function of the form:

$$\log p(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) = -\frac{mn}{2}\ln(2\pi) - \frac{mn}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{m}\sum_{i=1}^{n}(y_{ji} - x_{ji})^2 \tag{2.5}$$

where m is the number of variables of the ODE and n is the number of time points. It appears that, with real life data, this assumption of a Gaussian data generating distribution may be overly simplistic. After all, the likelihood function should only portray information immediately obtained from the data. However, in these scenarios we should be able to estimate the first two moments and consideration of the principal of maximum entropy [22] dictates that a Gaussian distribution minimises the unnecessary assumptions that we place on the data through the likelihood, becoming a natural choice when a more obvious distribution is not apparent. Notice that an assumption of Gaussian noise naturally includes a measure of observational error in our likelihood function and, in cases where the noise level is known, the log-likelihood reduces to the residual sum of squares. In eq 2.5 I have assumed constant Gaussian noise variance across time and variables. This is done throughout the thesis both in the noisy observation generation and in the parameter estimation procedures. Of course, when dealing with actual data, this assumption could prove restrictive and with a more careful modelling of the correlation between the different time points, we could hope for improved parameter inference in the model [4]. This would, however, require a massive increase in the dimension of the parameter space, introducing the curse of dimensionality to the inference problem.

Selection of the prior distribution introduces subjectivity to the Bayesian process as it allows more freedom to specify any a priori knowledge of the values of our parameters, such as positivity constraints often adopted for ODE parameters. I adopt a Gamma distribution as the prior over the ODE parameters. This allows the limitation of the sampling space to $\mathbb{R}^+$ and grants sufficient coverage of different parameter values (assuming a suitable choice of prior hyperparameters). For noise variance parameters, the inverse-gamma prior provides a conjugate prior in the case of a Gaussian likelihood. As observed by Berger [23] (and displayed in Figure 2.1), the inverse-gamma prior provides a proper approximation to Jeffreys prior —$p(\sigma^2) \propto \sigma^{-2}$— for very small hyperparameters (a common choice is Inverse-Gamma$(0.001, 0.001)$). Recalling the discussion of the

uninformative property of Jeffreys prior in section 1.2.2, we may refer to the inverse-gamma as an approximation to an uninformative distribution in the realm of conjugate priors. Given the conjugacy, we know that for given mean μ, the conditional posterior distribution for the noise variance parameter, σ^2, will be:

$$\sigma^2|\mu, y \sim \text{Inv-Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{\sum_i (x_i - \mu)^2}{2}\right) \tag{2.6}$$

and so for a large number of observations from a noisy dataset with large σ^2, the posterior distribution becomes detached from the prior hyperparameters. In Figure 2.1, the inverse-gamma prior is shown on the original domain and on the log transformed space. We see that, over the untransformed space, the prior density tends to be concentrated on a fairly small region of its support. However, when one takes a log transformation, the distribution becomes approximately an improper uniform distribution over the entire domain (the posterior, of course, is a proper distribution).

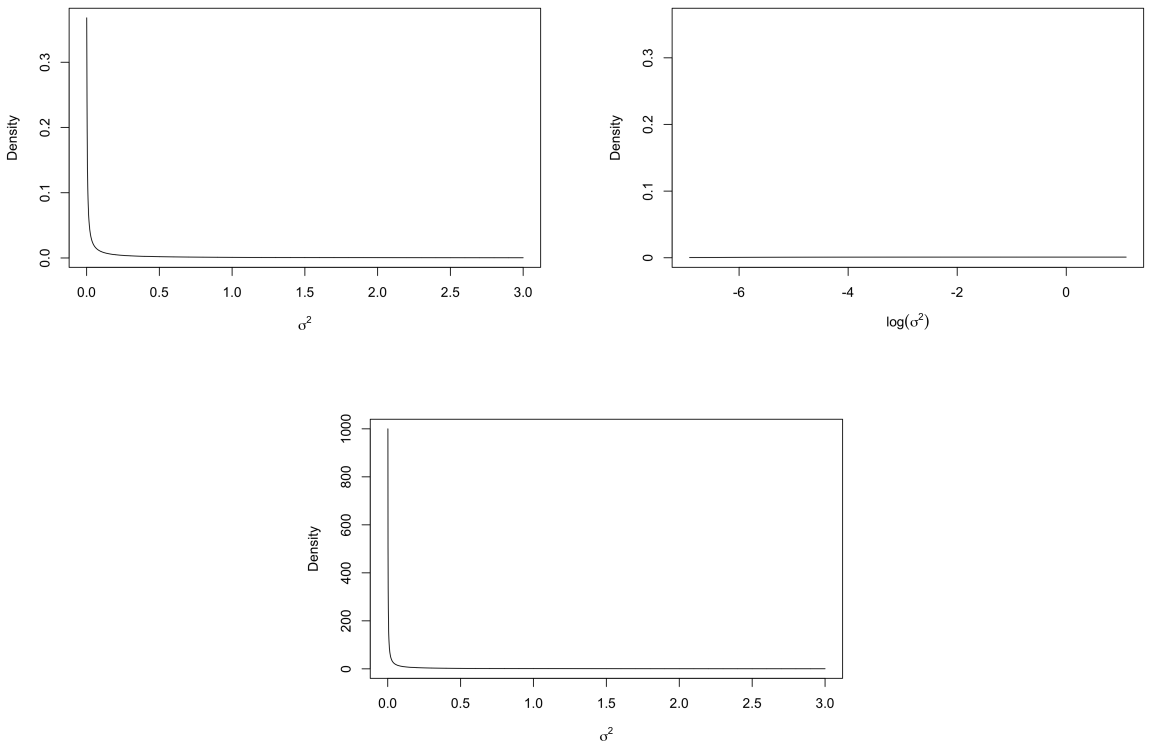

FIGURE 2.1: Top: Non-informative inverse gamma prior for the variance parameter on the original scale and on the log scale. We see that it is approximately an improper uniform on the log scale. Bottom: Jeffreys prior $p(\sigma^2) \propto \sigma^{-2}$. Comparing with the top left plot we observe the similarities between the two prior distributions.

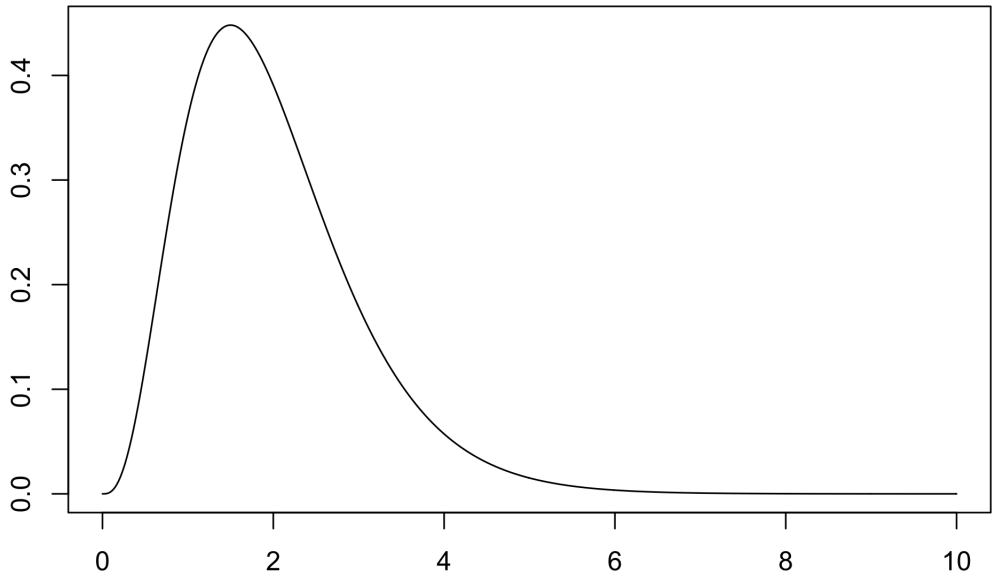

FIGURE 2.2: Gamma prior that we adopt for the parameters of our ODE models with hyperparameters $\alpha = 4$ and $\beta = 0.5$. The positive support makes it well suited to parameter inference in ODEs.

2.2 Markov Chain Monte Carlo Methods

Ideally, if we wish to provide an empirical distribution for ODE parameters, we would hope to be able to obtain independent samples from their posterior distribution. As discussed above, for more complex distributions associated with nonlinear ODEs, this sampling proves difficult. Adopting an MCMC sampling technique, we may take dependent samples of a random variable where each new proposal is proposed conditional on the previous sample:

$$\theta_t \sim q(\cdot|\theta_{t-1}) \tag{2.7}$$

Under an assumption of ergodicity, the ergodic theorem for Markov chains predicates that, despite introducing correlation to the samples from the posterior distribution, inferences in the presence of correlation are still valid when based on Markov chains that have converged to the stationary distribution.

Although the asymptotic convergence properties of MCMC methods can seem rather inaccessible compared with the large sample theory convergence properties of more elementary sampling methods, one property proves them to be far more apposite than their more elementary counterparts. Under certain regularity conditions, MCMC converges to the correct posterior distribution despite a lack of ability to obtain the marginal distribution in the denominator of Bayes theorem. More specifically, MCMC sampling provides a way of sampling from intractable posterior distributions and quantifying our uncertainty in the parameter values of the ODE with lower variance than standard sampling methods. This generality comes at a cost, introducing high levels of correlation between points, making the methods more inefficient than if we were able to sample independently as the effective sample size becomes smaller, quantifying the level of information that our full sample is able to provide. In an effort to overcome the inefficiency introduced by the correlation, adaptive MCMC methods [24] have been developed which constantly change the proposal mechanism of the sampler, improving mixing and increasing the effective sample size of the resultant chain.

2.2.1 Metropolis-Hastings

Naturally, the question of determining some probability distribution may be posed as the desire to construct a sample from the parameter space where the relative proportions of samples from each subset of the domain are representative of the value of the probability density over each subset. Intuitively, considering an iterative sampler, the way in which one should accept points would be to determine the ratio

$$R(x, y) = \frac{\pi(y)}{\pi(x)} \tag{2.8}$$

where y is the proposed point and x is the current point. However, we must also consider the sampler's action when the ratio decreases. Accepting proposals based on eq. 2.8 being greater than 1 attempts optimisation of the function in question and the chain would converge to some optimal point estimate (global or local). Instead, we must accept the new point with probability $R(x, y)$, allowing the possibility to accept proposals in the direction of decreasing likelihood and asymptotic traversal of the entire support of the distribution (as length of the chain tends to ∞). This is precisely the strategy adopted in the Metropolis algorithm [25], which was developed in 1953 to evaluate energy integrals.

The Metropolis algorithm samples by moving through the support of some prior distribution, accepting moves where the value of the posterior probability density increases and accepting with probability $A(x, y) = \min\big(R(x, y), 1\big)$ when the value of the probability density decreases (we could view this as a Markovian extension of the Accept-Reject method where a Markov property enables more efficient, representative, sampling). This is done by assessing whether $A(x, y) \geq u$ for $u \sim unif(0, 1)$. By accepting and rejecting in this way, we are able to construct a representative sample of the distribution based on the relative size of the density function in each region of the support. The Metropolis algorithm only applies when the proposal mechanism is a constant isotropic normal or symmetric distribution. In his paper in 1970, Hastings [26] generalised the concept to include non-symmetric proposal distributions. The Metropolis-Hastings routine constructs a Markov chain that, by use of a suitably defined Markov kernel, ensures that a sample of random variables converges in distribution to the true target distribution.

For initialisation of the Metropolis-Hastings algorithm, we must choose some starting value and some proposal distribution, say q. This proposal distribution is only an initial guess of the transition kernel. In order for the sampling scheme to be successful, the Markov chain produced must be invariant with respect to the target distribution:

$$\pi(y)dy = \int_{\mathbb{R}} P(x, dy)\pi(x)dx \tag{2.9}$$

where $\pi(\cdot)$ denotes the density with respect to Lebesgue measure of the invariant target distribution and $P(x, \cdot)$ is the kernel constructed by the MH sampler. Adhering to detailed balance provides the simplest way of ensuring convergence to the correct stationary distribution [27]. Given the freedom to specify the acceptance criterion, this is chosen in order to satisfy the detailed balance condition [28]:

$$A(x, y) = \min\left(\frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}, 1\right) \tag{2.10}$$

and in the presence of a symmetric proposal, this reduces to a ratio of the two densities. Billera and Diaconis [29] give a geometric interpretation of the Metropolis-Hastings algorithm where they present the MH algorithm as a projection of some arbitrary Markov chain onto the space of reversible Markov chains, characterised by the product of the Markov chains defined by $A(x, y)$ and $q(x, y)$. In spirit, this paper presents the Metropolis-Hastings algorithm as a natural (non I.I.D.) extension of elementary sampling procedures where samples are proposed from the support of the prior distribution using some proposal kernel and, through construction of a Markov kernel, the chain is projected onto

the space of reversible Markov chains such that invariance is satisfied with respect to the target density of interest.

Although theoretical properties always ensure ergodicity and asymptotic convergence with respect to the predefined target distribution, different proposal distributions in Metropolis-Hastings provide more efficient Markov chains than others. The easiest way of monitoring efficiency of the algorithm is through consideration of the acceptance rate of the procedure. The proposal mechanism is central to this as the acceptance rate varies depending on the shape of the proposal distribution. For a proposal with large scale, the algorithm experiences a decrease in acceptance rate as it begins to propose points that are further from the current position and thus tend to attempt moves between points with different densities. Therefore, despite the obvious improvement in mixing, convergence and exploration of the parameter space deteriorate due to the lack of short proposal moves. Although allowing an increased acceptance rate, a proposal distribution with too small dispersion will prove to be inefficient. As proven by Gelman et al. [30], we should target an optimal acceptance rate of around 23.4% for sufficient exploration of the target distribution. In the hypothetical case where the covariance of the target distribution is known, Roberts et al. posited the use of a Gaussian proposal with covariance given by $2.38^2\Sigma/d$ for optimal use of the MH algorithm, balancing the trade-off between optimism and pessimism in the algorithm's proposal moves. However, this method is highly hypothetical since in most cases our target distribution, including its covariance, is unknown.

2.2.2 Delayed Rejection Adaptive Metropolis

2.2.2.1 Adaptive Metropolis

MCMC methods are used when we encounter complicated distributions from which it is difficult to sample. Ironically, the nature of these complex distributions means it is unlikely that the standard MH algorithm and its associated proposal distribution will be universally applicable in these circumstances. Instead of relying on our ability to select an appropriate proposal distribution at the beginning of the algorithm, adaptive MCMC methods [24] implement on-line tuning of the proposal distribution that repeatedly updates the proposal covariance based on an empirical covariance estimate. Essentially, AM corrects overly optimistic or pessimistic proposal distributions based on the empirical covariance of the sampled points. The technique of Haario et al. [31] relies on the estimated covariance of previously sampled points in a tuning mechanism which adapts both the

magnitudinal and spatial properties of the proposal distribution:

$$\text{cov}(X_0, ...X_n) = \frac{1}{n}\left(\sum_{i=0}^{n} X_i X_i^T - (n+1)\bar{X}_i\bar{X}_i^T\right) \tag{2.11}$$

Considering the specific case of ODE parameter inference, identifiability problems in ODE systems are caused by a correlation between the parameters of the system (possibly resulting from the structure of the equations). Associated with a convergence to the target distribution will be the prevalence of a correlation between the parameter samples, leading to rank deficiency of the resultant empirical covariance matrix. In their algorithm, Haario et al. adopt a covariance of the form:

$$C_t = \begin{cases} C_0, & t \neq t_0 \\ s_d \text{cov}(X_0, ..., X_{t-1}) + s_d \epsilon I_d & \text{else,} \end{cases} \tag{2.12}$$

where addition of a small diagonal term to the covariance matrix helps to avoid any degeneracy issues. Considering the optimal proposal distribution derived by Gelman et al. [32], an attempt is made to replicate this by multiplying the proposal covariance by a dimensional-dependent constant factor s_d. Given that optimality is achieved when one adopts the covariance matrix of the target distribution, use of the empirical covariance will inevitably prove suboptimal but this suboptimality should decrease as the algorithm proceeds since the proposal distribution covariance converges to the optimal proposal distribution covariance, $C_t \to \Sigma$.

Ignoring the lack of a Markovian property, the adaptive Metropolis algorithm adopts the same acceptance criterion as the standard MH algorithm. Lacking motivation from detailed balance, this criterion must be shown to produce an ergodic Markov chain. Haario et al. [31] prove this assumption under some mild regularity conditions on the support of the distribution.

2.2.2.2 Delayed Rejection

In the standard Metropolis-Hastings algorithm, computation power is wasted on discarded proposal steps. Obviously, the rejection of these points is necessary to build a representative sample of the target distribution, but their rejection also provides information on the proposal distribution from the current point. For ODE parameter inference, the computational cost of evaluating the system of equations in order to discard a point seems

rather imprudent. An adaptation of MCMC, known as delayed rejection (DR), allows past rejected points to inform further sub proposals, making use of these previously wasted steps. As in standard Metropolis-Hastings, a stage 1 proposal $x \to y$ is accepted with probability

$$\alpha_1(x, y) = \min\left(1, \frac{\pi(y)(y, x)}{\pi(x)q(x, y)}\right) \tag{2.13}$$

If this point is accepted, then the algorithm proceeds as done in Metropolis-Hastings. Upon rejection, rather than discarding the point, we use it to inform another sub proposal, $x \to y'$ (see the left of Figure 2.3) by accepting based on the criterion

$$\alpha(x, y, y') = \min\left(1, \frac{\pi(y')q_1(y', y)q_2(y', y, x)[1 - \alpha_1(y', y)]}{\pi(x)q_1(x, y)q_2(x, y, y')[1 - \alpha_1(x, y)]}\right), \tag{2.14}$$

where q_2 is the proposal distribution for y', dependent on both x and the previously rejected proposal, y. The acceptance criterion is derived by imposing detailed balance at each stage of the delayed rejection; bearing in mind that, under detailed balance, the proposal $y' \to x$ is foreshadowed by a rejected move $y' \to y$. Essentially, in the case of a two stage delayed rejection, we wish to show probabilistic equivalence of the following diagrams:

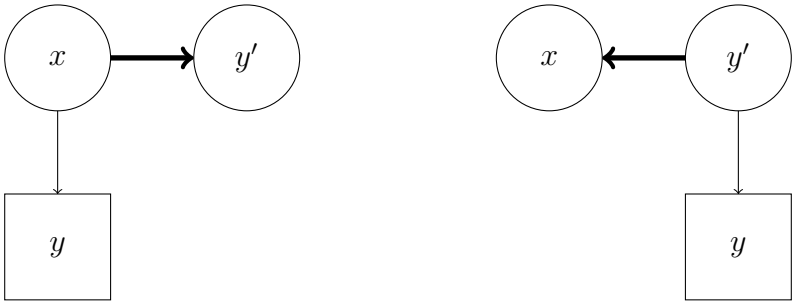

FIGURE 2.3: Demonstrating the detailed balance condition in delayed rejection. In both cases, a move to y is proposed and rejected (indicated by a square) before proposing a move to an alternative position that is informed by this rejected y move.

where the thick horizontal line is the second proposal and the thin vertical line is the initial rejection. Use of delayed rejection greatly improves the overall efficiency of MCMC, with the overall chain still abiding by a Markovian property since dependencies do not occur between retained points. A combination of delayed rejection with adaptive metropolis provides the first benchmark algorithm considered for simulations in this thesis.

2.2.2.3 DRAM—Combining Adaptive Metropolis and Delayed Rejection

Delayed Rejection Adaptive Metropolis [33] (DRAM) is an extension of the standard Metropolis-Hastings method that attempts to improve efficiency of the Markov chain by combining the concepts of adaptive Metropolis and delayed rejection. By adapting the covariance of the proposal distribution throughout the procedure, the sampler is able to learn the topologies involved and make more efficient proposals than in the standard MH algorithm. Adaptation occurs on both a global and local scale. The former is induced by the use of adaptive Metropolis and the latter results from the delayed rejection component of the algorithm. The DR is often accompanied by a scaling of the proposal distribution where the stage 1 proposal is the most optimistic, followed by shorter proposals that are scaled to enable higher probability of acceptance. I combine the AM and DR components as done by Haario et al. [33], proposing a first move that is based on the globally adapted proposal distribution and then, in the case of rejection, making n scaled proposals (proposing another stage 1 proposal move if one of these is accepted or once n sub proposals have been made). An important point to note is the violation of the Markovian property due to the dependence on past samples induced by the adaptive property of the algorithm. In standard MCMC methods, this Markovian property enables satisfaction of detailed balance, ensuring ergodicity with respect to the target distribution. Nonetheless, Haario et al provide a proof of ergodicity [31] in the absence of reversibility, making the assumption that the target density is bounded from above and has bounded support. For DRAM pseudocode, see page 14 of [34][1].

2.2.3 Population Markov Chain Monte Carlo

The MCMC methods discussed thus far have been developed extensively over the last few decades. However, these methods are not without their limitations since, despite efforts to alleviate poor convergence through adaptivity, severe multi-modality still causes convergence of the MCMC sampler to local optima. A solution is to consider an annealing procedure in the MCMC. Kirkpatrick et al. [35] present a simulated annealing approach to optimisation, making use of a temperature parameter which 'melts' the likelihood, smoothing over any local optima. Starting the algorithm at a high temperature, we repeatedly decrease the temperature as the algorithm continues, ending at a value of $T = 0$. In the initial stages, the system moves freely towards general regions of low energy (low values of the negative log-likelihood) which become narrower as the algorithm continues

[1] Our implementation of DRAM is in R using the modMCMC function from the FME R package.

and the temperature decreases. For functions with many local optima, the annealing schedule allows the system to move freely from its initial position without exploring false regions of low energy before allowing more detailed exploration in the region of interest—the global optimum. The effect of different temperatures on several MCMC samplers is shown in Figure 2.4 where, for high temperatures, the probability of making big jumps (transitions) cools down. In simulated annealing the effect of temperature would be inverted (becoming consistent with the theory of thermodynamics) and large transitions would be more likely at $t = 1$ and less likely at $t = 0$. Nonetheless, the figure still allows us to observe the general effect of temperature on movement in the parameter domain.

Suppose we wish to sample from some multimodal distribution:

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto L(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{2.15}$$

Population MCMC [36–38] draws on the idea of simulated annealing, coupled with a population approach in order to improve convergence of MCMC methods for likelihoods with multiple local optima. Defining a temperature ladder $\mathbf{t} = (t_1, ..., t_n)$ from 0 to 1 with $t_{i+1} > t_i$, we wish to set up a sequence of parallel distributions:

$$p(\boldsymbol{\theta}_{t_i}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta}_{t_i})^{t_i}\pi(\theta_{t_i}) \tag{2.16}$$

for $i \in 1, .., n$ where $p(\boldsymbol{\theta}_{t_i}|\mathbf{y})$ is the ith distribution. Whereas in simulated annealing the aim is optimisation where the temperature gradient changes in a constant direction as the algorithm continues, population MCMC introduces the change in temperature across different MCMC chains by implementing multiple parallel chains at different temperatures, with the chain at $t = 1$ always sampling from the correct posterior distribution. Each of these parallel chains may be explored using a standard Metropolis-Hastings sampler, with the sampler at temperature $t_n = 1$ providing the samples from the target distribution. Deviation from standard Metropolis-Hastings occurs with the introduction of swap moves between the parallel chains, allowing the smoothed likelihoods at temperatures approaching zero to influence the ability of the sampler to explore the multimodal posterior distribution at $t = 1$. Suppose the algorithm proposes a move from configuration $(\boldsymbol{\theta}_{t_1}, ..., \boldsymbol{\theta}_{t_k}, ..., \boldsymbol{\theta}_{t_l}, ..., \boldsymbol{\theta}_{t_n})$ to $(\boldsymbol{\theta}_{t_1}, ..., \boldsymbol{\theta}_{t_k}, ..., \boldsymbol{\theta}_{t_l}, ..., \boldsymbol{\theta}_{t_n})$ (swap between the lth and kth temperatures). By consideration of the joint probability distribution across all the parallel chains; that is, we choose whether to accept or reject a new parallel chain

configuration:

$$J(\boldsymbol{\theta}_{t_k}, \boldsymbol{\theta}_{t_l}) = \frac{p(\boldsymbol{\theta}_{t_1}|\mathbf{y}, t_1)...p(\boldsymbol{\theta}_{t_k}|\mathbf{y}, t_l)...p(\boldsymbol{\theta}_{t_l}|\mathbf{y}, t_k)...p(\boldsymbol{\theta}_{t_n}|\mathbf{y}, t_n)}{\prod_{i=1}^{n} p(\boldsymbol{\theta}_{t_i}|\mathbf{y}, t_i)} \tag{2.17}$$

$$= \frac{p(\mathbf{y}|\boldsymbol{\theta}_{t_k})^{t_l} p(\mathbf{y}|\boldsymbol{\theta}_{t_l})^{t_k} p_l(k)}{p(\mathbf{y}|\boldsymbol{\theta}_{t_k})^{t_k} p(\mathbf{y}|\boldsymbol{\theta}_{t_l})^{t_l} p_k(l)} \tag{2.18}$$

where $p_k(l)$ is given by a discrete Laplacian distribution [38] and provides the probability of proposing temperature l from temperature k:

$$p_k(l) \propto \exp(\beta \|k - l\|) \tag{2.19}$$

and $p(\mathbf{y}|\boldsymbol{\theta}_{t_i})^{t_i}$ denotes the power likelihood at the ith temperature. In eq. 2.18, prior probabilities have been suppressed as these cancel out in the numerator and denominator. Symmetry of the discrete Laplacian distribution enables further simplification of the acceptance ratio. Accepting based on comparison of $\min(1, J(\boldsymbol{\theta}_{t_k}, \boldsymbol{\theta}_{t_l}))$ with $u \sim unif(0, 1)$ allows acceptance with probability $J(\boldsymbol{\theta}_{t_k}, \boldsymbol{\theta}_{t_l})$. In Algorithm 1, I give pseudocode for the implementation of population MCMC used for all simulations in this thesis[2].

Obviously, since we are now proposing across multiple chains and also proposing swap moves, the number of numerical solutions required increases drastically. However, we hope that acceleration of convergence in terms of number of iterations is so great, compared with the asymptotic reliance of DRAM and Metropolis-Hastings, that the increased computational complexity is overcome by improved efficiency in terms of total quantity of MCMC steps required (especially when faced with multimodal likelihood functions).

My implementation of the population MCMC algorithm involves various different tuning parameters. In all versions of population MCMC, we take a product of power posterior densities:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n} p_{t_i}(\boldsymbol{\theta}|\mathbf{y}) = \prod_{i=1}^{n} p(\mathbf{y}|\boldsymbol{\theta})^{t_i} p(\boldsymbol{\theta}) \tag{2.20}$$

where the prior choice of temperature scale determines the nature of the gradient from posterior distribution to prior distribution. Taking $t_i \in (0, 1]$, the power posterior gradually shifts towards the prior as we move down the temperature ladder, the schedule of which can influence the convergence of the algorithm. For temperature distributions that are highly dense at $t = 0$, efficiency of the algorithm improves as mixing at $t = 0$ is

[2]This is an adaptation of the code found here: https://darrenjw.wordpress.com/2013/09/29/parallel-tempering-and-metropolis-coupled-mcmc/

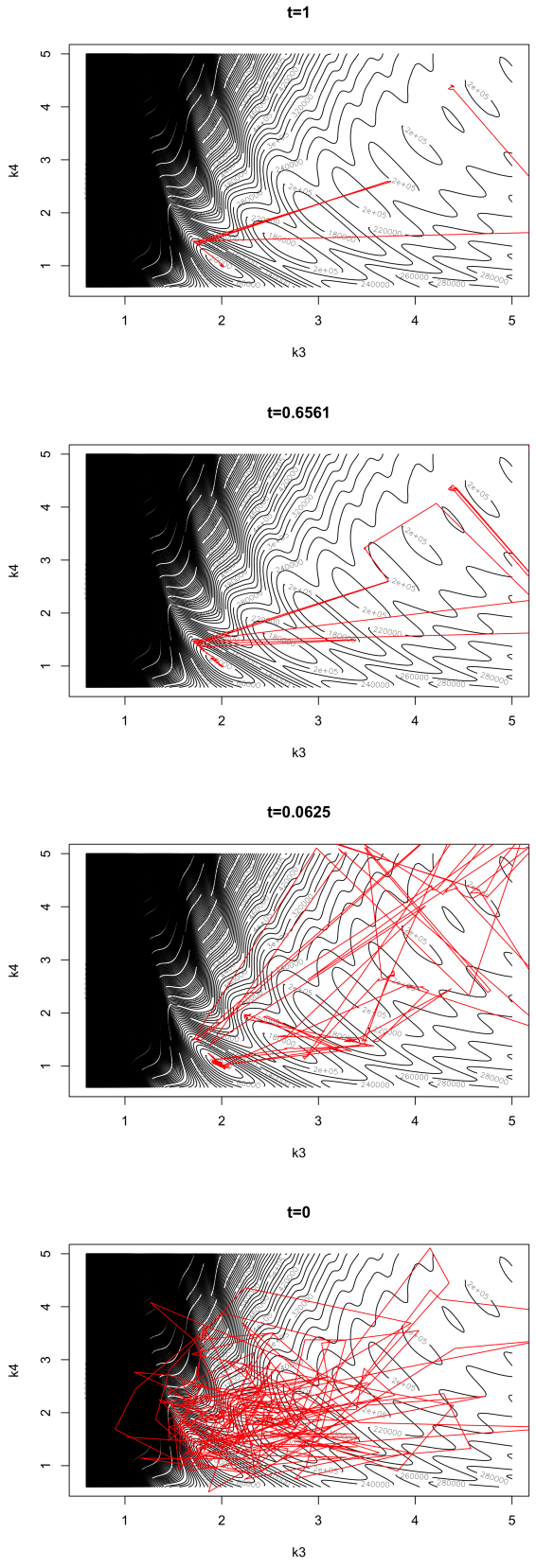

FIGURE 2.4: Displaying the effect of varying temperature on the exploration of the parameter domain in population MCMC. The effect of the likelihood increases as we move up the temperature ladder from t=0 to t=1. At the bottom of the ladder, we are sampling from the gamma prior distribution.

Algorithm 1 Population Markov Chain Monte Carlo

1: Assign starting positions
2: Decide temperature ladder (T temperatures), number of swap moves m and iterations, S, at which to perform m swap moves
3: **repeat**
4: Propose $\boldsymbol{\theta}_t^*$ for t=1,...,T and accept based on standard Metropolis-Hastings criterion.
5: **if** $i \mod S = 0$ **then**
6: **for** j in 1:m **do**
7: Randomly sample temperature, t_k and propose temperature t_l for swap via discrete Laplace distribution.
8: Propose swap between two chains and accept with probability $J(\boldsymbol{\theta}_{t_k}, \boldsymbol{\theta}_{t_l})$ from eq. 2.18.
9: **end for**
10: **end if**
11: **until** convergence

encouraged by fast exploration in the densely populated low temperature region of the ladder. In this thesis, I adopt a temperature schedule of the form $t_i = (i/10)^4$, leading to 10 parallel chains with higher frequency towards the lower end of the temperature ladder. In eq. 2.19, it was mentioned that a Laplacian distribution will be used to propose temperature swap moves. This involves the choice of a parameter β that dictates the scale of the distribution. Increasing the value of β results in an increase in the average distance of proposal moves between the different temperatures decreasing the number of swaps that occur, particularly at the end of the sampling phase. However, higher values of β lead to improved mixing in the early stages of sampling. Where applicable, the choice of tuning parameter value is inspired by the literature [38]. Otherwise, I make no claim of careful selection of the algorithm's conditions, but instead highlight the necessity for tuning as one of the key criticisms of this expensive method.

2.2.4 Delayed Acceptance Metropolis-Hastings

Delayed Acceptance Metropolis-Hastings (DAMH) [17] is a novel method to MCMC sampling that implements a surrogate pre-sifting phase where moves are tested using a surrogate likelihood function (an estimate of the expensive likelihood function) before accepting moves based on an expensive likelihood function. The idea is that, by first assessing suitability of points in surrogate space, we save computation time by filtering out proposed parameter estimates that would otherwise be rejected by the exact likelihood. In this initial filtering stage, we accept proposed points as done in the Metropolis-Hastings

algorithm with the target given by our surrogate function and subsequent acceptance probability given by:

$$\alpha_1 = \min\left(1, \frac{\hat{\pi}(\theta')q(\theta', \theta)}{\hat{\pi}(\theta)q(\theta, \theta')}\right) \tag{2.21}$$

where $\hat{\pi}(\theta) \propto \hat{p}(y|\theta)p(\theta)$ is a surrogate posterior obtained from the surrogate likelihood function $\hat{p}(y|\theta)$. In what proceeds, let $\pi(\theta) \propto p(y|\theta)p(\theta)$ be our expensive posterior where $p(y|\theta)$ is the exact likelihood. Now, we quickly give a detailed-balance-inspired justification for the form of $\alpha_2(\theta, \theta')$. We know from simple MCMC theory that for the target distribution to be the stationary distribution of the MCMC, detailed balance must be satisfied (as well as ergodicity). For DAMH, the detailed balance condition depends on two different acceptance functions, one for the surrogate filter steps and one for the expensive likelihood. I refer to these as α_1 and α_2 respectively:

$$\pi(\theta)q(\theta, \theta')\alpha_1(\theta, \theta')\alpha_2(\theta, \theta') = \pi(\theta')q(\theta', \theta)\alpha_1(\theta', \theta)\alpha_2(\theta', \theta)$$

Similarly to the derivation of the acceptance criterion in Metropolis-Hastings, assume that detailed balance is not satisfied. Suppose that the following holds:

$$\pi(\theta)q(\theta, \theta')\alpha_1(\theta, \theta')\alpha_2(\theta, \theta') \geq \pi(\theta')q(\theta', \theta)\alpha_1(\theta', \theta)\alpha_2(\theta', \theta)$$

In this scenario, we make more moves from θ to θ' than the reverse and so we make $\alpha_1(\theta', \theta)\alpha_2(\theta', \theta) = \alpha(\theta', \theta)$ equal its maximum, 1. Now, we know that $\alpha_1(\theta, \theta') = \min(1, \frac{\hat{\pi}(\theta')q(\theta', \theta)}{\hat{\pi}(\theta)q(\theta, \theta')})$ since this is just a standard Metropolis-Hastings step. But, $\alpha(\theta', \theta) = 1 \Rightarrow \alpha_1(\theta', \theta) = 1$ and so $\alpha_1(\theta, \theta') \leq 1$ since $\alpha_1(\theta, \theta') = 1/\alpha_1(\theta', \theta)$. Since $\alpha_1(\theta, \theta') = \min\left(1, \frac{\hat{\pi}(\theta')q(\theta', \theta)}{\hat{\pi}(\theta)q(\theta, \theta')}\right) = \frac{\hat{\pi}(\theta')q(\theta', \theta)}{\hat{\pi}(\theta)q(\theta, \theta')}$, we obtain:

$$\pi(\theta)q(\theta, \theta')\frac{\hat{\pi}(\theta')q(\theta', \theta)}{\hat{\pi}(\theta)q(\theta, \theta')}\alpha_2(\theta, \theta') = \pi(\theta')q(\theta', \theta)$$

Simplification and cancellations leave us with:

$$\alpha_2 = \frac{\pi(\theta')/\hat{\pi}(\theta')}{\pi(\theta)/\hat{\pi}(\theta)}$$

and since it is a probability, we take $\alpha_2 = \min\left(1, \frac{\pi(\theta')/\hat{\pi}(\theta')}{\pi(\theta)/\hat{\pi}(\theta)}\right)$. In order to introduce an adaptive mechanism to the method, I take inspiration from DRAM [33] and use the empirical covariance from past samples in order to update the covariance in a Gaussian

Algorithm 2 Delayed Acceptance Metropolis-Hastings

Assign starting position $\boldsymbol{\theta_0}$ and select cheap surrogate distribution $\hat{\pi}(\boldsymbol{\theta})$, an approximation of $\pi(\boldsymbol{\theta})$ (the exact likelihood)
repeat
  Sample proposal $\boldsymbol{\theta}_t$ from some proposal distribution $q(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$.
  Evaluate $\hat{\pi}(\boldsymbol{\theta_t})$ and therefore α_1
  if $\alpha_1 \leq u_1$ **then**
    $\boldsymbol{\theta_t} = \boldsymbol{\theta_{t-1}}$
  else
    Determine $\pi(\boldsymbol{\theta}_t)$ and evaluate α_2
    if $\alpha_2 \leq u_2$ **then**
      $\boldsymbol{\theta_t} = \boldsymbol{\theta_{t-1}}$
    else
      Accept $\boldsymbol{\theta}_t$
    end if
  end if
until convergence

proposal distribution.

For optimal use of the DAMH algorithm, we would hope for a surrogate posterior distribution that, in a topological and statistical sense, matches well with the exact posterior distribution obtained using the numerical solution likelihood function. The former requires an alignment of the ridges of the two likelihood surfaces. The latter desires a correspondence between the tails of the resultant posterior distributions. In reality, this is often not the case, leading to inaccuracies and inefficiencies in the sampling routine. Firstly, consider the case of a surrogate distribution where the tails are more weighted than the tails of the true distribution. The second acceptance criterion, α_2 should be able to filter out any steps that, despite representing the surrogate distribution, are a poor sample from the true distribution. However, heavier tails in the surrogate mean that a higher proportion of proposals will pass through the initial filter stage and so the gain in computational efficiency becomes less significant as the tails widen. The second inconsistency results when we have short tails in the surrogate distribution relative to the true distribution (or indeed a realignment as shown in Figure 4.3). Although remaining accurate, the method fails to give a sufficient representation of our uncertainty in the parameter value.

2.2.5 MCMC Prerequisites

2.2.5.1 Convergence Diagnostics

Adoption of MCMC methods for parameter inference necessitates the use of convergence diagnostics. The concept of convergence in MCMC refers to convergence to the target distribution of the distribution from which the samples are obtained. Ideally, we would determine a length of Markov chain that guarantees this property. As observed by Cowles and Carlin [39], this approach is rarely feasible, leaving a choice between two different approaches. A more theoretical approach would consider convergence properties of the sampling mechanism. More pragmatically, we can perform some form of output analysis [40] which relies on a measure of similarity between the empirical distributions of multiple parallel chains, all initiated from different starting points. This method is best implemented with an overly disperse starting distribution and in cases where knowledge of the posterior distribution is insufficient to allow this property, quasirandom sequences [41] can offer a space filling sequence which provides good coverage of the parameter domain[3].

Introduced by Gelman and Rubin [42], the potential scale reduction factor (PSRF) is a common tool for assessing lack of convergence in MCMC samples. Suppose we possess multiple parallel chains initiated from some disperse initial distribution. Assuming convergence to the stationary distribution, variance within each of these chains should equal the variance between the chains, estimators for which are given in eq. 2.22:

$$B = \frac{n}{m-1}\sum_{j=1}^{m}\left(\bar{\theta}_{\cdot j} - \bar{\theta}_{\cdot\cdot}\right)^2 \qquad W = \frac{1}{m(n-1)}\sum_{j=1}^{m}\sum_{i=1}^{n}\left(\theta_{ij} - \bar{\theta}_{\cdot j}\right)^2 \tag{2.22}$$

where θ_{ij} is the ith sample from the jth chain. If these variances are significantly different, we can assume that convergence has not occurred (PSRF only diagnoses non-convergence, there does not exist a method of diagnosing definite convergence of Markov chains). Eq. 2.23 provides a biased estimate of the variance in the posterior distribution:

$$\hat{\sigma}^2 = \frac{n-1}{n}W + \frac{1}{n}B \tag{2.23}$$

which tends to overestimate the posterior variance $\mathrm{var}(\boldsymbol{\theta}|\mathbf{y})$. Contrastingly, W should underestimate the variance in our posterior distribution due to correlation between samples and the fact that, in a finite number of steps, the sampler is not able to traverse the

[3]In particular, I adopt a sobol sequence using the R package randtoolbox

entire support of the distribution. Therefore, if we consider the square root of the ratio of $\hat{\sigma}^2$ and W:

$$\hat{R} = \sqrt{\frac{\hat{\sigma}^2}{W}} \tag{2.24}$$

then we can expect that, in the case of convergence of the Markov chain, $\hat{R}$ would be close to or equal to 1. Otherwise, we can conclude longer chains would either increase W or decrease B as the chains continue to better explore the posterior space. Importantly, the theory of the Gelman Rubin statistic does not necessitate a Markovian property be prevalent in the chains, meaning that this diagnostic may be applied to all methods considered in this thesis. There are countless other numerical MCMC convergence diagnostics that could be used to assess convergence of samples but, similar to PSRF, these all diagnose lack of convergence instead of convergence and PSRF represents a more established method among the different possibilities [13]

2.2.5.2 Sampling on Bounded Support

In ODE parameter inference, we often have to place some constraint on the value of the parameters. For the positivity constraint often placed on parameter in dynamical systems, we have to determine the most effective way of adopting this constraint in our samples. Immediately, one would consider a proposal that rejects any points proposed outwith our constrained space. However, this can introduce bias to the sampling procedure as the sampler can only move in one direction at the extreme boundaries of the constrained space, giving false evidence of local optima in the likelihood space. More effectively, one can fit a constrained prior distribution in the original domain and then transform the random variables (and therefore the prior distribution) to an unconstrained space for MCMC sampling. Thus, we can sample on an unbounded space while still satisfying the relevant constraints. As an example, consider the sampling of ODE parameters subjected to a positivity constraint. Applying a Gamma prior in the untransformed space before sampling in the log transformed parameter domain ensures unconstrained sampling while still being subjected to this constraint. This transformation must be accompanied by a suitable transformation of the prior distribution. For transformation $g(\cdot)$, the Metropolis acceptance criterion in this case would be, for moving from point $\tilde{\theta}_{t-1} = g(\theta_{t-1})$ to

$\tilde{\theta}_t = g(\theta_t)$

$$A(\theta_{t-1}, \theta_t) = \frac{\tilde{\pi}(\tilde{\theta}_t)p(y|\theta_t)}{\tilde{\pi}(\tilde{\theta}_{t-1})p(y|\theta_{t-1})} \tag{2.25}$$

where $\tilde{\pi}(\cdot)$ denotes the Jacobian transformation of the original prior distribution:

$$\tilde{\pi}(\tilde{\theta}_t) = \pi(g^{-1}(\tilde{\theta}))\left|\frac{dg^{-1}(\tilde{\theta})}{d\theta}\right| \tag{2.26}$$

This second suggestion is the choice I adopt for the work in this thesis as I use gamma or inverse-gamma priors and sample the parameters on the log domain.

Chapter 3

Complications in Parameter Inference for Ordinary Differential Equations

Performance of parameter inference procedures depends on the behaviour of the methods in ODE solution space. Solutions of nonlinear ODEs rely on computationally onerous numerical integration procedures, making the classic Gaussian observational log-likelihood expensive to evaluate. Nonetheless, in the case of well-behaved solutions and sufficiently smooth likelihood functions, MCMC procedures can converge in realistic timeframes especially in the case of simple toy problems. Computational complexity becomes more problematic when one considers more complex systems where properties of the signals directly influence our ability to implement the expensive likelihood function and achieve accurate parameter inference. This chapter will present the benchmark ODE systems that will be used to compare different parameter inference methods, specifying the parameter configurations implemented in each case. The presentation of these models will unravel some of the complications introduced by different ODE properties to the parameter learning problem. These complications are by no means specific to the nonlinear ODE parameter inference problem, but their combination with high computational cost of numerical integrations makes them even more troublesome in this scenario. Essentially, this chapter provides a dichotomy of ODE complications that inspire the use of more efficient parameter inference techniques considered in Chapter 4.

3.1 Lotka-Volterra

The Lotka-Volterra system of equations [3] was seen in the introductory section. Restated here, the resultant signals from this system of ODEs are given in Figure 3.1.

$$\frac{dx}{dt} = \alpha x - \beta xy \tag{3.1}$$

$$\frac{dy}{dt} = -\gamma y + \delta xy \tag{3.2}$$

The signals in Figure 3.1 were obtained with parameter values $\alpha = 0.76, \beta = 0.5$,

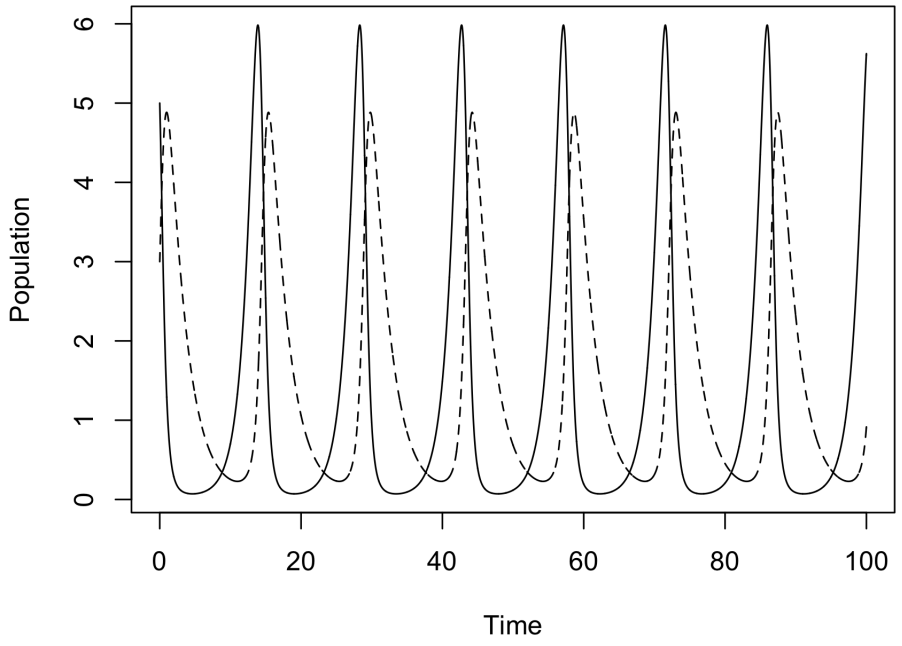

FIGURE 3.1: Evolution of prey (solid) and predator (dashed) populations produced from the Lotka-Volterra system.

$\gamma = 0.4$ and $\delta = 0.3$. These will be maintained throughout the work of this thesis and provide sufficient periodicity to allow exploration of the effect this property has on the parameter learning process.

3.1.1 Periodicity

Consider the phase planes shown in Figure 3.2 corresponding to the signals produced by the Lotka-Volterra equations. Both plots appear identical and indeed, both have been produced by simulation of the Lotka-Volterra equations with the same parameter configurations and initial conditions. The difference, however, is the period of integration. The left plot corresponds to the signal over a timeframe between 0 and 100, the right provides the signals approximated between times 0 and 15. Similarity of the two plots results from the overlapping of the different periods of data when considered in phase space. Given the deterministic nature of the system, multiple periods of data provide no better indication of the level of fit to the observed data than the single period data case. Obviously, there is an improvement in resolution of the likelihood function with a greater number of data points leading to greater discrepancy between levels of the likelihood surface; but these periods of data also act to hinder the inference process as they introduce a possibility of signal aliasing to an inference problem that has already been made more computationally onerous.

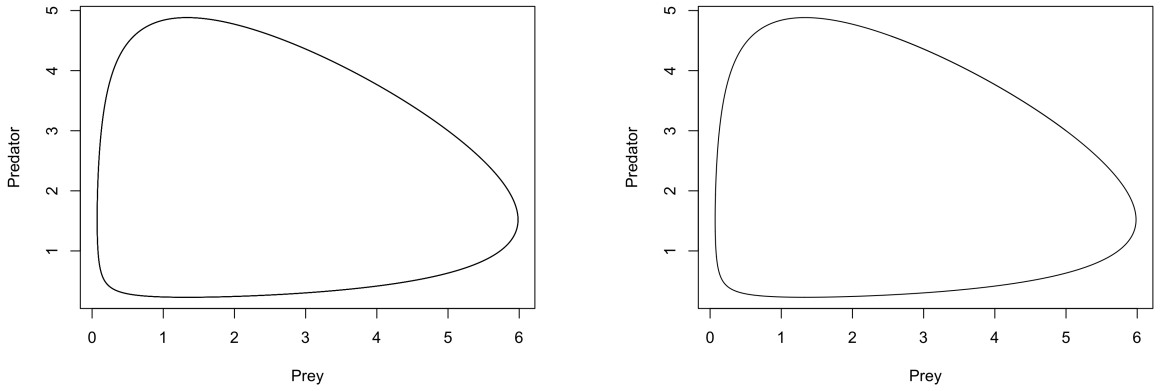

FIGURE 3.2: Left: phase space of the Lotka-Volterra model over time 0 to 100. Multiple periods of data overlap, showing the lack of information that is provided by this larger dataset. Right: phase plane for only one period of data. The level of information provided is identical. This plot presents a problem with inference in deterministic ODEs as added periods of data provide no more information about the fit of the model since we see that successive periods overlap one another on the state space plot

If one wishes to adopt an MCMC approach to parameter inference, then performance of the method depends heavily on the nature of the likelihood surface. In particular, we must be able to explore the entire parameter domain and we would hope that any attraction into local optima corresponds to an accurate fitting to the observed values.

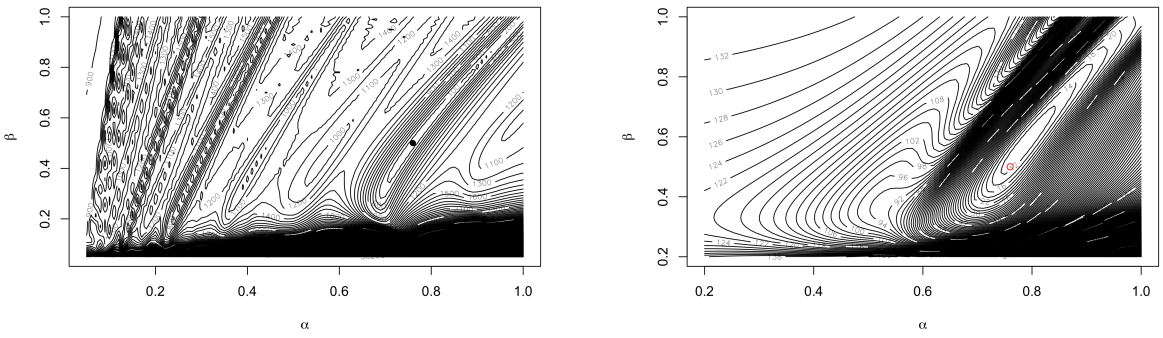

FIGURE 3.3: The effect of periodicity on the Lotka-Volterra log-likelihood surface. Multimodality is introduced as a result of signal aliasing. These local optima make the inference problem more intractable.

With multiple modes, the chain can be attracted into local optima where the fit of the signal to observations is not actually accurate. Increasing the time required for convergence, the periodicity makes a difficult problem more intractable. The negative log-likelihood surface for multiple periods of data from the Lotka-Volterra ODEs is given on the left of Figure 3.3. Comparing with the one period data case (right of Figure 3.3), increased multimodality alludes to the effect that periodicity can have on the likelihood function of the system. More careful investigation of the nature of these modes is displayed in Figure 3.5 where the signal at a local optimum is plotted with the signal at the true parameter value for both the multi-period and single-period data cases. In the case of the multi-period example, the matching at alternate peaks of the signal leads to strong correspondence in the phase space of the two signals, manifesting itself with a decrease in the value of the likelihood function. The decrease in timeframe of the single period negates the possibility of signal aliasing and so no local optima appear on the likelihood surface on the right of Figure 3.3. To consider the effect this has on the performance of MCMC, Figure 3.6 shows MCMC chains obtained using DRAM in both of the two datasets. In the multi-period scenario, attraction into the local optima induces a failure to sample from the correct posterior distribution, unlike in the single period case where the chains move freely towards the correct distribution.

Attempts to alleviate this multimodality problem likely rest on our ability to manipulate the posterior distribution. One possibility is to formulate a likelihood function that requires less computational expense, allowing an increase in the number of permissible MCMC steps and a greater opportunity for MCMC chains to evade the various local optima. At this stage, however, it remains to be seen how this could be done without introducing bias to the parameter samples. Alternatively, we could take inspiration from

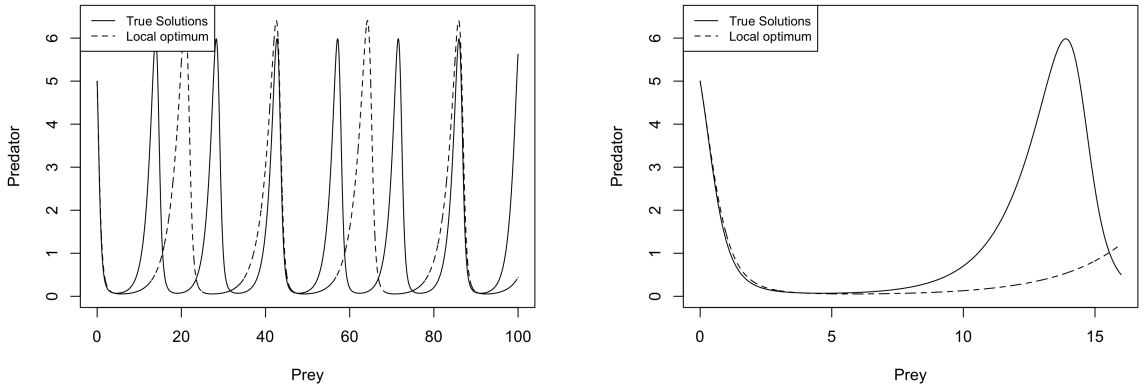

FIGURE 3.4: Simulated signal of the Lotka-Volterra model for time 0 to 100 (left) and only one period of data (right). The plot on the left displays the periodicity present in the system while the plot on the right displayed the lack of fit when considered over only one period.

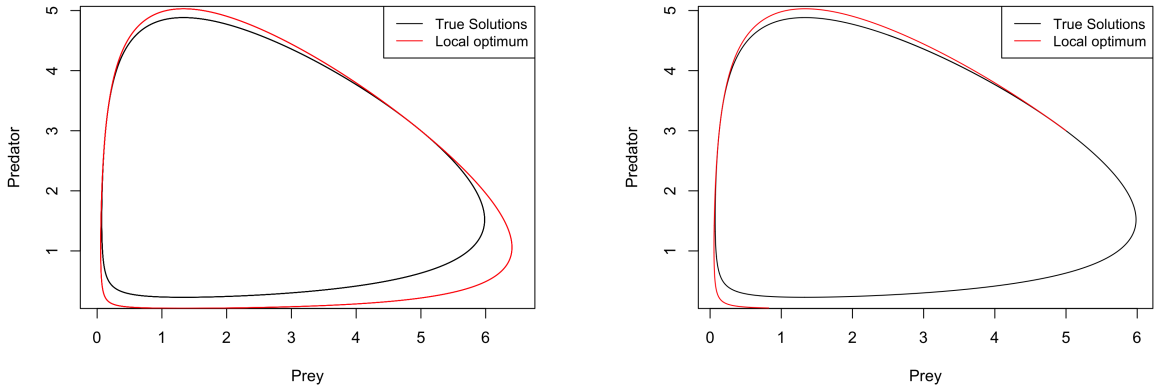

FIGURE 3.5: Presenting the phase plane of a local optimum from the Lotka-Volterra likelihood surface with multiple periods (left) and one single period (right). The level of fit to the solution can be assessed by the level of overlap of the two signals.

simulated annealing (Section 2.2.3) which optimises functions by smoothing the likelihood surface, smoothing over local optima. Optimisation, however, is not the aim of our inference since this provides little guarantee of the validity of the samples when one considers the problems of local optima and identifiability. Rather than optimization, we hope to sample from the posterior distribution of the ODE parameters which at this point has only been considered in a population approach where the vast computational expense is precisely that which we wish to avoid.

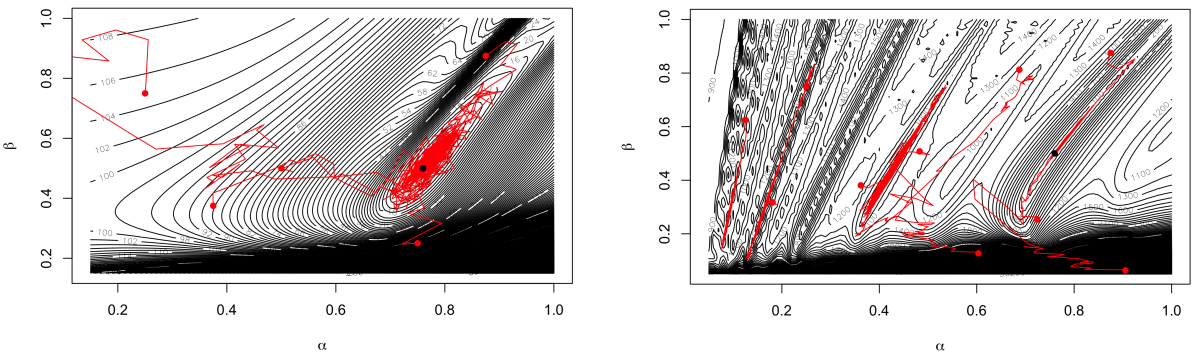

FIGURE 3.6: Left: MCMC with the single period likelihood function. The chains manage to converge to the global optimum of the likelihood function as a result of the reduction in local optima. Right: MCMC with the multi-period likelihood function. The presence of local optima causes a deterioration in the effectiveness of DRAM as the chains converge to these local optima.

3.1.2 Stiffness

For systems of nonlinear ODEs, we often rely on numerical integration techniques to approximate the signals involved. These adopt a discretisation of the time domain, propagating through from the initial conditions to the signal approximation. For instance, the Runge-Kutta methods [15] use a combination of midpoint and trapezoidal rules to build up an approximation of the underlying signal. For well-behaved solutions, this procedure already requires substantial computational cost, particularly emphasised when one adopts an MCMC approach based on the expensive log-likelihood function. If we measure the time required to obtain the samples from the posterior distribution of the parameters in the Lotka-Volterra model as displayed in Figure 3.6, we experience a vast difference in the computational efficiency for chains exploring different regions of the parameter space. The largest of these times is 300% of the smallest run time, suggesting that a parameter-dependent problem exists in construction of estimated ODE solutions.

Consider a space of solutions of the ODEs for a fixed set of initial conditions. By using numerical integration, we effectively move through this space, moving closer to the full solution of the ODEs with each numerical integration step that is performed. A problem occurs, however, when this full, correct, solution of the system is slow varying and lies close to a fast-varying solution. The ability to reject the fast-varying solution in favour of the slow one relies on a smaller step size and inefficient approximation of the slow-varying solution. By taking a larger step size, we would be removing solutions from the solution space of the ODEs, introducing a bias to the inference problem. This problem with fast-slow solution proximity—referred to as stiffness—is especially problematic in

ODEs with periodic solutions since there can be a natural correspondence between stiff solutions and local optima of the likelihood surface. An ad hoc solution to this problem can be to limit the maximum number of steps in the numerical solver since we know that these stiff solutions will not provide a sufficient solution to the ODE. However, this introduces a bias to the problem similar to the case where we adopt a larger step size. Taking a more systematic approach, we could use implicit Runge-Kutta methods in which the estimate of the signal at each point depends on evaluation of the ODE function at that point. The implicit nature of the equations now relies on solving a set of normal equations at each step and deals with rapidly deteriorating solutions associated with stiffness more efficiently than the explicit methods. However, the problem in our case is that we are numerically solving over the entire region of the parameter space that contains regions with stiff solutions and regions with non-stiff solutions. These implicit methods, with their necessary matrix inversions, are less efficient in the case of non-stiff ODEs and so if one is seeking more efficient methods of inference, it does not seem feasible to implement these universally. Rather, we seek methods that can bypass these stiff ODE configurations, relying on numerical solutions of the ODEs in regions of highest posterior probability density.

3.2 Signal Transduction Cascade and Parameter Non-identifiability

This system, outlined by Vyshemirsky and Girolami [21], models the protein signalling transduction cascade with input enzymatic protein S. This binds to protein R, forming protein complex RS, inducing phosphorylation of protein R into R_{pp} from which state the protein may be deactivated. Additionally, the system below also describes the degradation of the input S into Sd. Here, variables inside [] correspond to concentrations of different species and the rest of the components are parameters controlling the rate of

transformation of the different variables of the system.

$$\begin{aligned}
\frac{d[S]}{dt} &= -k_1[S] - k_2[S][R] + k_3[RS] \\
\frac{d[S_d]}{dt} &= k_1[S] \\
\frac{d[R]}{dt} &= -k_2[S][R] + k_3[RS] + \frac{V[R_{pp}]}{K_m + [R_{pp}]} \\
\frac{d[RS]}{dt} &= k_2[S][R] - k_3[RS] - k_4[RS] \\
\frac{d[R_{pp}]}{dt} &= k_4[RS] - \frac{V[R_{pp}]}{K_m + [R_{pp}]}
\end{aligned} \tag{3.3}$$

Figure 3.7 shows the signals produced by simulation of this system of equations. The parameter configuration used takes inspiration from the literature [7], setting the parameters as follows: $k_1 = 0.07, k_2 = 0.6, k_3 = 0.05, k_4 = 0.3, V = 0.017$ and $K_m = 0.3$.

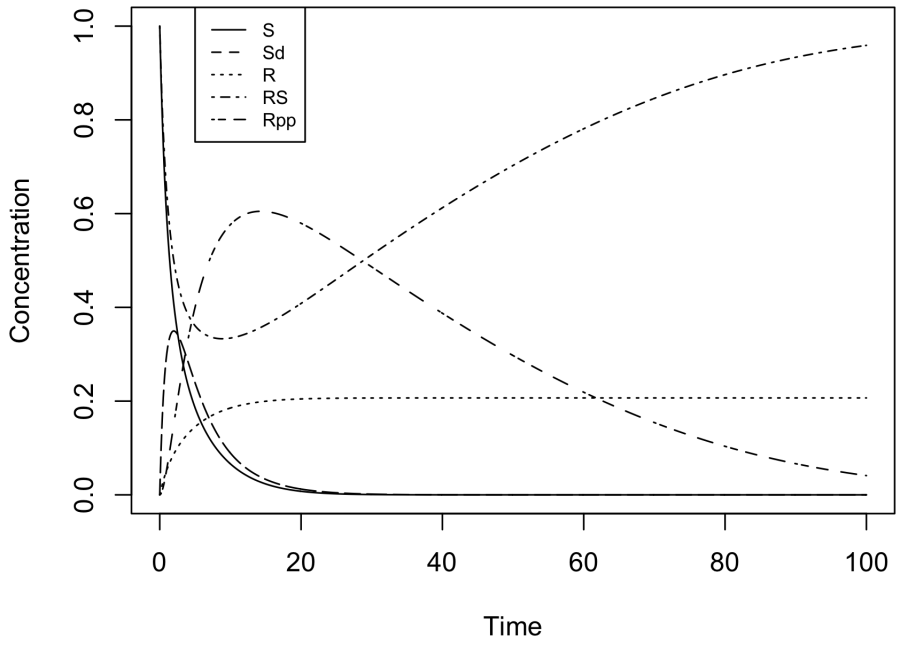

FIGURE 3.7: Evolution of the different dependent variables produced from the signal transduction cascade system.

Recall the mapping introduced in eq. 1.2:

$$h : \Theta \to \mathcal{Y}. \tag{3.4}$$

Injectivity is crucial to the parameter inference problem, allowing the ODEs to exhibit a property of parameter identifiability. Depending on the structure of the ODE system, this identifiability assumption may be violated and indeed, this is the case with the STC system of equations. Consider the ratio $V[R_{pp}]/(K_m + [R_{pp}])$, present in eq. 3.3, if V is increased, then K_m may be increased (by greater magnitude) to counteract the multiplicative change induced by the changing of parameter V. This property is termed structural non-identifiability [43] and relates to the structure of our ODEs as the functional relationship dampens the observed effect that the different parameters have on the resultant signal. Immediately, one would consider solving the problem by fixing parameter K_m. However, this does not resolve the issue completely. Although we have been able to negate the identifiability issues induced by the structure of our ODEs, a subtler violation of the identifiability property persists.

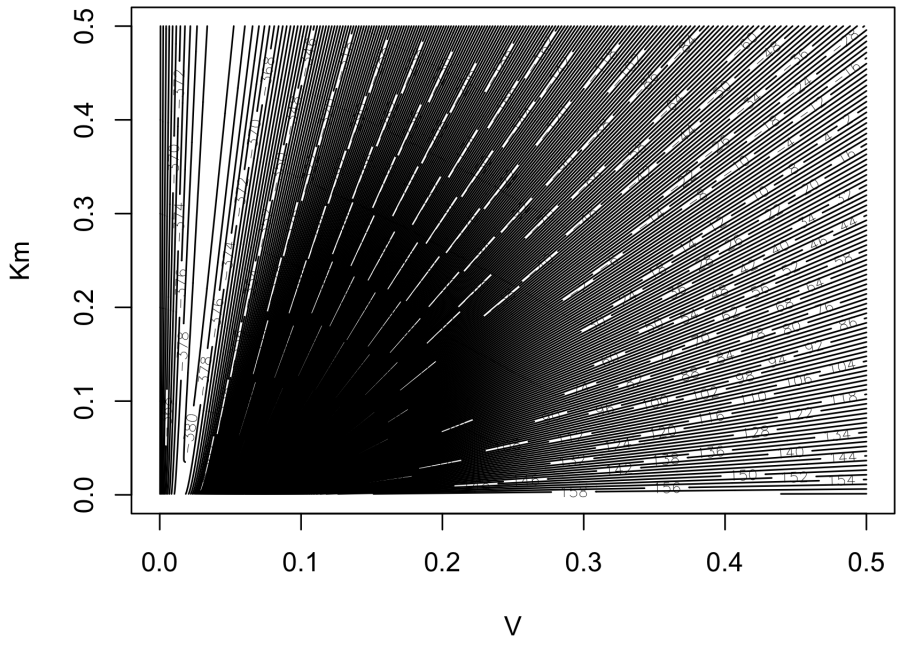

FIGURE 3.8: Contour plot showing the effect of structural non-identifiability in the signal transduction cascade system. The functional relationship between parameters K_m and V results in an obvious valley along the functional relationship in parameter space.

Consider the plots on the left of Figure 3.10 where the red line in the bottom plot corresponds to an MCMC chain attempting to sample from the posterior distribution of the STC parameters when K_m is held fixed. A lack of variation in the likelihood function results in an MCMC chain that can move freely through a valley in the posterior

parameter space, estimating large levels of uncertainty in the posterior parameter distribution. The cause of this likelihood profiling becomes clear if one considers the signals from this system plotted in Figure 3.7. The system has entered equilibrium, meaning that, over the entire duration of the evolution, only a small portion of the observed data is varying and contributing any information to the model fitting problem. Given that no changes can be made to the experimental setup, a solution is to consider inference using only the data obtained prior to equilibrium as shown in Figure 3.9. The MCMC trajectories are plotted on the right of Figure 3.10 where, rather interestingly, a reduction in number of observations leads to improved resolution in the likelihood surface. When adopting a Bayesian approach, one can also alleviate this identifiability problem by use of a sufficiently informative prior—a solution that does not generalise to the structural case.

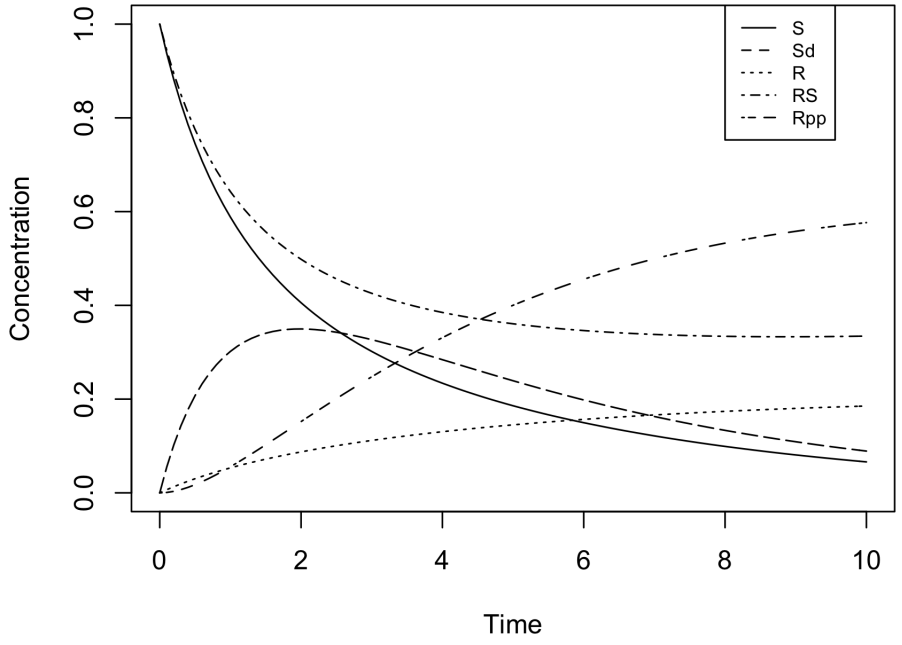

FIGURE 3.9: Signal transduction cascade data prior to reaching equilibrium. The data vary throughout the timeframe considered, providing information on the fit of the signal.

The discussion of identifiability highlights a key advantage of a sampling approach to parameter inference. If one cannot assume identifiability of the model parameters, then conclusions built on the premise of a best fitting parameter set will always be open to skepticism. If we instead construct a well-defined sampling routine then the analysis

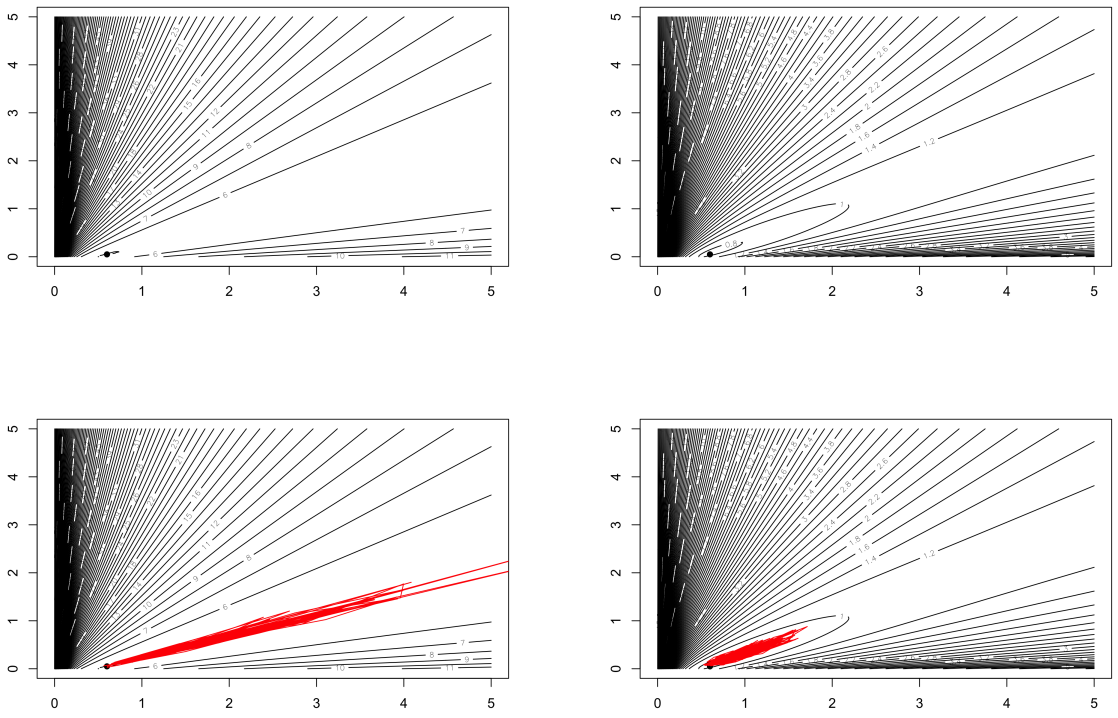

FIGURE 3.10: Top row: Likelihood surface over the k2-k3 plane in parameter space. On the left is the negative log-likelihood surface from 100 observations between 0 and 100, on the right is the negative log-likelihood surface from 20 equally spaced observations between 0 and 10. Bottom row: DRAM samples from the posterior distribution of the different datasets superimposed on the corresponding negative log-likelihood surface.

of uncertainty in the parameter space and function space can allow the diagnosis of identifiability issues in the model parameters—a property that must be accounted for in any model fitting process, especially if one wishes for reliable predictions from their fitted model.

3.3 FitzHugh-Nagumo

The FitzHugh-Nagumo equations [44, 45] model the movement of signal along excited cells via a 2-variable simplification of the more complex Hodgkin-Huxley model, enabling phase plane analysis of the system. Conceptualised to model nerve membrane excitation behaviour, Variable V denotes the voltage of the signal which accounts for self-excitation of the membrane and R a recovery variable that acts as a negative feedback mechanism. This system can be used to model signals in excitable media, providing a mathematical description of cardiac dynamics [46] and neurodegenerative diseases [47]. Nonlinearity is

introduced to the system through the equation for $\dot{V}$:

$$\frac{dV}{dt} = \gamma\left(V - \frac{V^3}{3} + R\right)$$
$$\frac{dR}{dt} = -\frac{V - \alpha + \beta R}{\gamma}$$

I take inspiration from the work of Ramsey et al. [6] and assume parameter values $\alpha = \beta = 0.2, \gamma = 3$ and noise variance equal to 0.25 (this variance is the larger of the two observation noise variances adopted by Campbell and Steele [36]). Following the same example, initial values were set to (V=-1, R=1). The signals plotted in Figure 3.11 present a periodic solution where the oscillations are not symmetric.

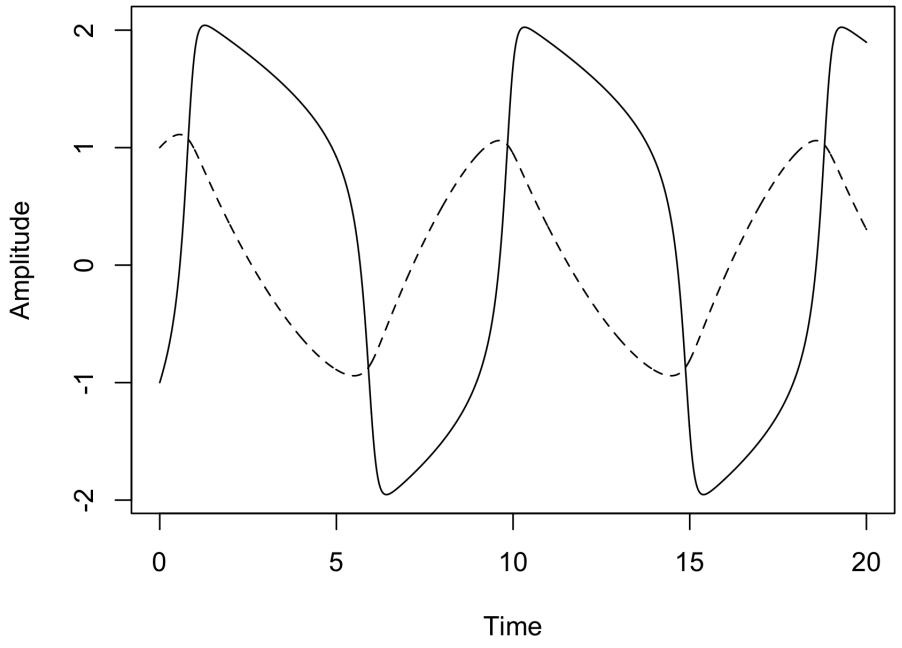

FIGURE 3.11: Evolution of Recovery (solid) and Voltage (dashed) variables produced from the FitzHugh-Nagumo system.

3.4 Goodwin Oscillator

The Goodwin Oscillator (also referred to as Circadian Oscillator and I interchange the two throughout this thesis), first introduced by Goodwin [48], models the concentration of an

enzymatic protein and messenger ribonucleic acid (mRNA) of some species. Production of mRNA p_1 is met by a binding of mRNA with ribosomes, forming a enzymatic protein p_2 which feedback, inhibiting mRNA transcription.

$$\frac{dp_1}{dt} = \frac{k_1}{36 + k_2 p_2} - k_3$$
$$\frac{dp_2}{dt} = k_4 p_1 - k_5$$

Constant decay terms in the ODEs are biologically inconsistent since we may now have negative protein concentration values. However, this encourages periodicity in the signals of the species; a trait that is more difficult to induce under the system configuration provided by Woller et al. [49] (we may show this improved periodicity by consideration of the eigendecomposition of the Jacobian matrix of the ODE system). Similar to the case of the FitzHugh-Nagumo, nonlinearity is induced by one of the equations where the protein variable, p_2, appears in the denominator. Following the work of Girolami et al [50], I adopt parameter values $k_1 = 72, k_2 = 1, k_3 = 2, k_4 = 1$ and $k_5 = 1$ and the signals produced (Figure 3.12) exhibit periodicity which leads to multimodality in the likelihood function (this was the likelihood surface shown in Figure 2.4).

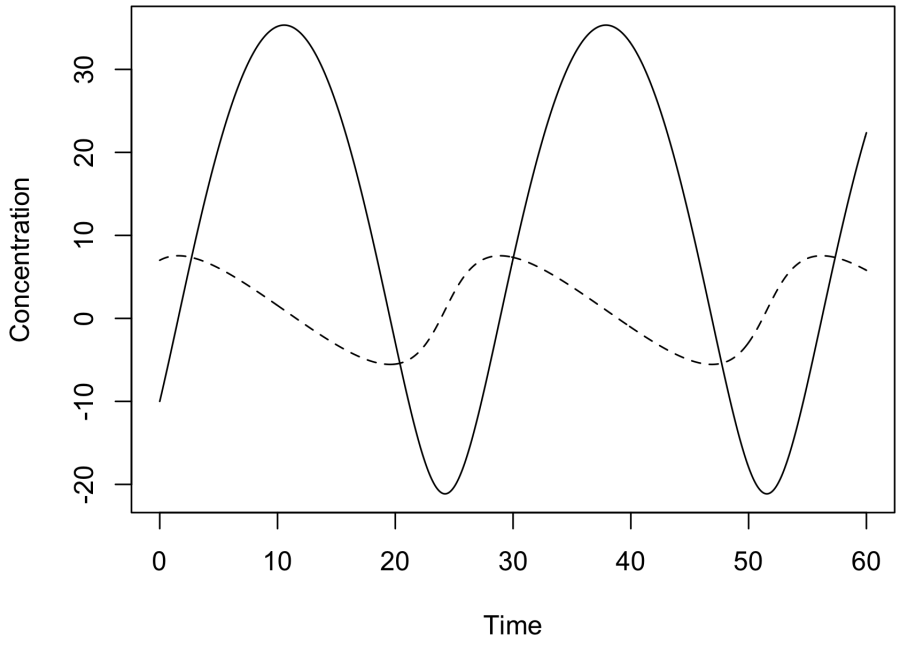

FIGURE 3.12: Evolution of p_1 (solid) and p_2 (dashed) concentrations produced from the Goodwin Oscillator system.

Chapter 4

Fixed Interpolant Gradient Matching

Having discussed the problems that ODE parameter inference presents to the MCMC inference paradigm, it now seems justified to propose a new method for parameter inference that incorporates various techniques from the literature. If one wishes to provide a computationally efficient method, then it seems inevitable that we must consider a proxy for the exact likelihood function—also termed a surrogate likelihood function. The posterior produced in this scenario is termed a surrogate posterior distribution and represents a cheap estimate to the numerical integration likelihood function that has been considered up until this point. The paradigm of gradient matching will be discussed in this chapter as a method of providing this cheap surrogate function. I will first outline some relevant Gaussian Process (GP) theory before considering some of the gradient matching literature in greater detail. The drawbacks of these schemes provide motivation for a novel implementation of the gradient matching paradigm where I neglect the need for computationally inefficient surrogate sampling, instead assuming a fixed interpolant for the noisy observations. This will involve consideration of two alternative distance metrics, one involving a simple Euclidean norm and one in which the distance measure resembles the Mahalanobis distance as obtained by Calderhead et al. [1]. This gradient matching surrogate likelihood will allow the implementation of a multi-phase scheme where a cheap burn-in phase drives the sampler towards an initialisation in the region of the global optimum, reducing the number of exact likelihood evaluations required to achieve stationarity in a post burn-in phase. As well as considering the setup with standard DRAM sampling, I will also consider the use of delayed acceptance Metropolis-Hastings

as a sampling mechanism in a two-phase scheme where the less expensive filter step is provided by fixed interpolant gradient matching.

4.1 Gaussian Process Smoothing

Consider the standard nonlinear regression problem where, given some noisy observations,

$$\mathbf{y}(t) = \mathbf{x}(t) + \boldsymbol{\epsilon}(t), \tag{4.1}$$

we wish to uncover the underlying latent state $\mathbf{x}(t)$ while making the assumption that the data distribution can be considered as $p(\mathbf{y}|\mathbf{x}(t), \sigma^2\mathbf{I}) = \mathcal{N}(\mathbf{y}|\mathbf{x}(t), \sigma^2)$. Taking a parametric approach to the problem, we could assume $\mathbf{x}(t, \boldsymbol{\theta})$ is a function of some parameter set and attempt to infer these parameters using standard Bayesian methods. This, however, seems unnecessarily restrictive as it requires us to specify a standard form for the function $\mathbf{x}$ that, being latent, can be difficult to categorize.

A Gaussian process defines a distribution over the space of functions from continuous input space $\mathcal{T}$ to the uncountably infinite set of real numbers,

$$x : \mathcal{T} \to \mathbb{R}. \tag{4.2}$$

Placing a Gaussian process prior on the codomain of x:

$$x(t) \sim \mathcal{GP}(\mathbf{m}(t), \mathcal{K}(t, t')) \tag{4.3}$$

$$m(\mathbf{t}) = \mathbb{E}[x(\mathbf{t})] \tag{4.4}$$

$$\mathcal{K}(\mathbf{t}, \mathbf{t}) = \mathbb{E}[(x(\mathbf{t}) - m(\mathbf{t}))(x(\mathbf{t}) - m(\mathbf{t}))], \tag{4.5}$$

defines it as a stochastic process such that any finite subset $(\mathbf{x}(t_1), ..., \mathbf{x}(t_n))$ is multivariate Gaussian distributed. The kernel function, $\mathcal{K}(\cdot, \cdot)$ provides much of the explanation of the relationship induced by the GP and is itself governed by a set of hyperparameters, $\boldsymbol{\gamma}$. Often, the Gaussian process is assumed to be of mean zero, providing the following distribution where explicit representation of $\boldsymbol{\gamma}$ presents the latent variable, $\mathbf{x}$, as being parametrised by these hyperparameters:

$$p(\mathbf{x}|\boldsymbol{\gamma}) = \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathcal{K}) \tag{4.6}$$

By placing a distribution over the latent variable, we sanction the marginalisation over $\mathbf{x}$ to produce a marginal likelihood:

$$p(\mathbf{y}|\boldsymbol{\gamma}, \sigma^2) = \int p(\mathbf{y}, \mathbf{x}|\boldsymbol{\gamma}, \sigma^2) d\mathbf{x} = \int p(\mathbf{y}|\mathbf{x}, \sigma^2\mathbf{I}) p(\mathbf{x}|\boldsymbol{\gamma}) d\mathbf{x}, \tag{4.7}$$

where $p(\mathbf{x}|\boldsymbol{\gamma}) = \mathcal{N}(\mathbf{0}, \mathcal{K})$ is specified by the GP prior and the Gaussian assumption on the observations permits a closed form for the marginal likelihood. This marginalisation over the latent variable produces a marginal likelihood function parametrised by the GP hyperparameters, enabling hyperparameter estimation through a type II maximum likelihood approach where the marginal likelihood function, given in eq. 4.8, contains a natural trade-off between model fit and model complexity [51], preventing overfitting or underfitting of the smoothed signal.

$$\log p(\mathbf{y}|\mathcal{T}, \boldsymbol{\gamma}) = \underbrace{-\frac{1}{2}(\mathbf{y} - \mathbf{m})^T (\mathbf{K} + \sigma^2\mathbf{I})^{-1}(\mathbf{y} - \mathbf{m})}_{\text{Model fit}} \underbrace{- \frac{1}{2}\log|\mathbf{K} + \sigma^2\mathbf{I}|}_{\text{Model complexity}} - \frac{n}{2}\log(2\pi) \tag{4.8}$$

4.1.1 Gaussian Process Posterior Distribution

In standard Bayesian regression, we obtain a posterior distribution over the parameters proportional to the product of the likelihood and the prior distribution. With a Gaussian process, we treat the function itself as a parameter of the model, seeking a posterior distribution over the latent variable of the form:

$$p(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi}, \sigma) \propto p(\mathbf{y}|\mathbf{x}, \sigma) p(\mathbf{x}|\boldsymbol{\phi}) \tag{4.9}$$

$$= \mathcal{N}(\mathbf{y}|\mathbf{x}, \sigma^2\mathbf{I}) \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{K}) \tag{4.10}$$

where the $ijth$ entry of $\mathbf{K}$ is given by the kernel function evaluated at the ith and jth timepoints. The Gaussianity of the likelihood and prior permit a closed form posterior distribution obtained using the Gaussian identity from the appendix:

$$\mathcal{N}(\mathbf{y}|\mathbf{x}, \sigma^2\mathbf{I}) \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{K}) \propto \mathcal{N}(\mathbf{x}|\mathbf{m}, \boldsymbol{\Sigma}) \tag{4.11}$$

where

$$\mathbf{m} = \boldsymbol{\Sigma}(\mathbf{K}^{-1}\mathbf{0} + \sigma^{-2}\mathbf{y}) \quad \boldsymbol{\Sigma} = (\mathbf{K}^{-1} + \sigma^{-2}\mathbf{I})^{-1} \tag{4.12}$$

from which we may make use of the Woodbury matrix identity (see Appendix) to perform the inversion in the expression for $\mathbf{\Sigma}$:

$$\begin{aligned}\mathbf{\Sigma} = (\mathbf{K}^{-1} + \sigma^{-2}\mathbf{I})^{-1} &\overset{\text{WMI}}{=} \sigma^2\mathbf{I} - \sigma^2\mathbf{I}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\sigma^2\mathbf{I} \\ &= \sigma^2\mathbf{I}((\mathbf{K} + \sigma^2\mathbf{I})^{-1}(\mathbf{K} + \sigma^2\mathbf{I}) - \sigma^2\mathbf{I}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}) \\ &= \sigma^2\mathbf{K}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\end{aligned}$$

Therefore, the posterior distribution for the latent variable $\mathbf{x}(t)$, conditional on the observed data, is given by:

$$p(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi}, \mathcal{T}) \sim \mathcal{N}\big(\mathbf{K}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y}, \sigma^2\mathbf{K}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\big) \tag{4.13}$$

This posterior distribution is inherently infinitely dimensional since its dimension is only limited by the finiteness of the GP prior. Therefore, we see that a GP prior leads to a GP posterior distribution. The posterior mean expression is a linear transformation of the observed values (a linear smoother), averaging over the noise in the data to smooth the observed values.

4.1.2 Choice of Kernel Functions for ODE Signals

For the data interpolant to be applicable in the gradient matching setting, we require that the smoothed signal be differentiable. Since GPs are closed under linear transformation and the derivative may be thought of as a linear transformation (with transformation matrix given by the Jacobian), assuming the derivative of the GP interpolant exists (which does not always hold—see Brownian motion), the gradient of a GP is itself a Gaussian process. Indeed, choice of kernel function for a Gaussian process can be influenced by a desired order of differentiability and this is essential if one wishes to adopt a gradient matching paradigm in parameter inference.

The simplest class of kernel functions is the stationary kernels, expressed as a function of the radius between two points. A typical example is the infinitely differentiable squared exponential kernel:

$$k_{SE}(r) = \exp(-\frac{r^2}{2l^2}) \tag{4.14}$$

where l is the length-scale and $r = ||x - x'||$ is the distance between two points. For high values of l, the kernel allows increased influence by distant points, smoothing the resultant posterior sample. One problem with the squared exponential kernel is the high level of smoothing in the resulting estimated signal. For more complicated signals we may consider the Matern class kernel function:

$$k(r) = \frac{2^{1-\eta}}{\Gamma(\eta)}\left(\frac{\sqrt{2\eta}r}{l}\right)^{\eta} K_{\eta}\left(\frac{\sqrt{2\eta}r}{l}\right) \tag{4.15}$$

which tends to the squared exponential kernel in the limit as $\eta \to \infty$. Often we choose to take $\eta = z + 0.5$ for $z \in \mathbb{Z}^+$ since the kernel then reduces to the product of an exponential and an order z polynomial as shown below for z=2:

$$k_{5/2}(r) = \left(1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2}\right)\exp\left(-\frac{\sqrt{5}r}{l}\right) \tag{4.16}$$

We may then observe that the Matern class kernel is $z-1$ order differentiable and the desired level of smoothness of the function provides motivation for the selection of the hyperparameter η. For $\eta = \frac{1}{2}$ we obtain the non-differentiable kernel of the Ornstein-Uhlenbeck process used for defining the velocity of a particle under Brownian motion [51].

Consider the Lotka-Volterra data as shown in Figure 3.1. Periodicity in the data is difficult to model accurately using either of the covariance functions previously considered. Since valid covariance functions remain valid under smooth transformation of the input variables, we can transform the input variables into some alternative space in which the periodicity in the signal is automatically accounted for. Under a smooth mapping of the inputs:

$$\mathbf{u} : x \to \left(\cos\left(2\pi\frac{x}{p}\right), \sin\left(2\pi\frac{x}{p}\right)\right). \tag{4.17}$$

$k_{SE}(x, x')$ becomes a valid kernel function for periodic signals [52]:

$$k_{SE}(\mathbf{u}(x), \mathbf{u}(x')) = k_{per}(x, x') = \sigma^2 \exp\left(-2\frac{\sin^2\left(\pi\frac{x-x'}{p}\right)}{l^2}\right) \tag{4.18}$$

where l is the length-scale dictating the roughness (smoothness) of the interpolant. With highly noisy periodic data, smoothing becomes difficult as we are faced with a trade-off in the lengthscale between underaccounting of the amplitude and smoothness of the signal.

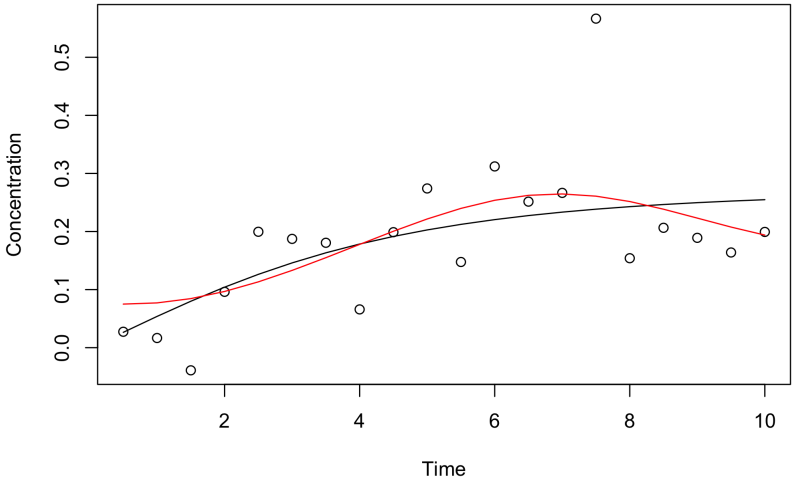

FIGURE 4.1: Comparing the interpolant produced by the squared exponential kernel (red) and the neural network kernel (black).

The signal transduction cascade is a slightly more interesting proposition due to the nonstationary signal produced by the ODEs. The kernel functions discussed so far have all been translation invariant, making them incompatible with this example. Adoption of a Neural Network Kernel [53] enables varying lengthscales for nonstationary signals. In Figure 4.1, we observe the interpolant obtained for the S_d variable using the neural network kernel and the squared exponential kernel. The nonstationarity of the Neural Network kernel leads to constant values at large positive or negative input values.

4.2 Gradient Matching Literature

4.2.1 Gradient Matching—A Frequentist Approach

Although this thesis will adopt a Bayesian approach to parameter inference, it is vital that we acknowledge the work done in gradient matching under the frequentist paradigm. Swartz and Bremermann realised the ability to bypass the numerical integration step of ODE parameter inference by instead adopting a difference metric that is applied directly in the gradient space [54]. In some sense, their method is very similar to the multi-phase method proposed in this thesis as they also acknowledge the fact that any surrogate metric may only be used as an approximation to the true function. They adopt a gradient

matching measurement function to achieve an initial estimate of the parameter value:

$$F(\boldsymbol{\theta}, \hat{\mathbf{X}}, \dot{\hat{\mathbf{X}}}) = \sum_i \sum_j w_{ij}^2 (\dot{\hat{x}}_{ij} - f(\mathbf{y}_j, \theta))^2, \tag{4.19}$$

where w_{ij} is some weight coefficient and $\dot{\hat{\mathbf{X}}}$ is a matrix with entries $\dot{\hat{x}}_{ij}$ denoting the value of the gradient of a smooth interpolant of variable i at time j obtained by fitting polynomials to segments of the data and $\mathbf{y}_j$ denotes the vector of species measurements at time j. The authors then exploit linear dependence of the variables on the parameters to find the partial derivative with respect to the parameters, which can be set equal to zero to obtain initial parameter estimates. Beginning from the estimates obtained using the criterion in eq. 4.19, Swartz and Bremermann then propose using a global optimisation technique to iteratively optimise the following function:

$$F(\theta, x) = \sum_i \sum_j w_{ij}^2 (y_{ij} - x_{ij})^2 \tag{4.20}$$

where y_{ij} is the measurement of variable i at time j and x_{ij} is the numerical solution of the ODE system for species i at time j. Varah takes a step closer to the current gradient matching paradigm by fitting a cubic spline to the noisy observations (cubic splines may be thought of as Gaussian Processes for particular kernel functions) and minimising the deviation of the gradient of the smoothed signal from the functional gradient obtained by substituting the smoothed interpolant into the ODEs [55].

Clearly, these two approaches to ODE parameter inference provide a frequentist utilisation of gradient matching. Whereas the authors chose to minimise the deviations between the two gradients, we choose to sample from the posterior whilst adopting this metric as our objective function. The problem with simple minimisation of the least squares function is the possibility of local optima in the functional space, and the difficulty in diagnosing a lack of convergence to the true global optimum. Whereas when one adopts a Bayesian approach, we are able to track convergence of the sampler to the stationary distribution induced by the gradient matching metric. Additionally, uncertainties are naturally accounted for by our posterior distribution derived from the gradient matching likelihood function.

4.2.2 Gradient Matching—A Bayesian Approach

In the recent literature a Bayesian approach to gradient matching has been introduced that attempts to bypass the need to solve the system of ODEs [1, 7, 56]. Similar to the above, the idea is to find a smoothed form of the observed signal and match the gradient obtained from this smoothed curve with that obtained from the ODEs. However, this time we obtain a distribution of sampled values from the surrogate space instead of point parameter estimates. In this reformulation of the parameter inference problem, we are no longer concerned with the sub manifold of solutions of the ODE [6], eradicating the need for constraints as we profile over the initial conditions [7].

4.2.2.1 Calderhead et al.

On a fundamental level, gradient matching is based on the notion of having two different distributions (experts) which formulate different dependencies of the gradient—a structural dependency through eq. 2.1 and a more implicit dependency resulting from the Gaussian Process (GP) interpolant (if unsure of the notion of Gaussian processes then please refer to Section 4.1) [1]. Consider the joint distribution of the GP interpolant and its gradient where $p(\mathbf{x}|\boldsymbol{\phi}) = \mathcal{N}(\mathbf{x}|\mathbf{0}, \mathbf{K})$ and $p(\dot{\mathbf{x}}|\boldsymbol{\phi}) = \mathcal{N}(\dot{\mathbf{x}}|\mathbf{0}, \mathbf{K}'')$. The joint distribution will have mean vector $\mathbf{0}$ and in the equation below I outline the components of the covariance matrix. Consider an arbitrary GP kernel $k(t_i, t_j)$ such that $cov(x(t_i), x(t_j)) = k(t_i, t_j)$ where k is assumed differentiable. We may obtain derivatives of this function as follows:

$$cov(\dot{x}(t_i), x(t_j)) = \frac{\partial k(t_i, t_j)}{\partial t_i} = k'(t_i, t_j) \tag{4.21}$$

$$cov(x(t_i), \dot{x}(t_j)) = \frac{\partial k(t_i, t_j)}{\partial t_j} = {}'k(t_i, t_j) \tag{4.22}$$

$$cov(\dot{x}(t_i), \dot{x}(t_j)) = \frac{\partial^2 k(t_i, t_j)}{\partial t_j \partial t_i} = k''(t_i, t_j) \tag{4.23}$$

which leads to a joint distribution:

$$\begin{bmatrix} \dot{\mathbf{x}} \\ \mathbf{x} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{K}'' & \mathbf{K}' \\ {}'\mathbf{K} & \mathbf{K} \end{bmatrix} \right) \tag{4.24}$$

where elements of the different submatrices are given by evaluation of the kernel derivatives from eqs. 4.21-4.23. We obtain, via an algebraic derivation (see page 87 of Bishop, 2006), a conditional distribution for the GP gradient that is dependent on a sampled

interpolant and sampled hyperparameters:

$$p(\dot{\mathbf{x}}_n|\mathbf{x}_n, \boldsymbol{\phi}_n, \boldsymbol{\sigma}) = \mathcal{N}(\mathbf{K}'_n\mathbf{K}_n^{-1}\mathbf{x}_n, \mathbf{K}''_n - \mathbf{K}'_n\mathbf{K}_n^{-1\prime}\mathbf{K}_n) \tag{4.25}$$

where $\mathbf{x}_n$ corresponds to the GP interpolant for the *nth* state of the ODE and the matrices $\mathbf{K}''_n, \mathbf{K}_n, {}'\mathbf{K}_n^{-1}$ and ${}'\mathbf{K}_n$ correspond to the matrices of the respective functions (or derivative) evaluated with hyperparameters obtained by fitting the GP interpolant. From this point, let $\mathbf{C}_n = \mathbf{K}''_n - \mathbf{K}'_n\mathbf{K}_n^{-1\prime}\mathbf{K}_n$ and $\mathbf{m}_n = \mathbf{K}'_n\mathbf{K}_n^{-1}\mathbf{x}_n$.

In order to formalisea joint posterior distribution, dependent on both the GP hyperparameters and the ODE parameters, we must introduce a probabilistic dependence between the gradient from the GP and the gradient from the ODE. Assuming a Gaussian distribution, $\dot{\mathbf{x}}_{GP} \sim \mathcal{N}(f(\mathbf{x}, \boldsymbol{\theta}), \gamma\mathbf{I})$ (where the dotted line in Figure 4.2 indicates that this is a non-generative model) we are able to find a correspondence between the ODE and GP dependencies. Considering the distribution $p(\boldsymbol{\theta}, \gamma, \dot{\mathbf{X}}|\mathbf{X}, \boldsymbol{\phi})$ where γ represents the

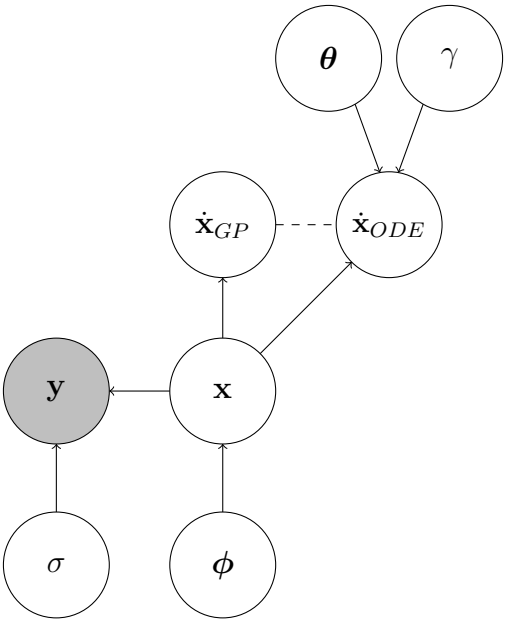

FIGURE 4.2: Graphical model representation of the method introduced by Calderhead et al. [1] where $\dot{\mathbf{x}}_{ODE}$ corresponds to the gradient obtained using the ODE system and $\dot{\mathbf{x}}_{GP}$ is the gradient obtained by differentiation of the GP interpolant.

mismatch between our two gradients (given that $\mathbf{x}$ is an estimate of the true signal, it would be naive to assume this relationship is exact), marginalisation over $\dot{\mathbf{x}}$ becomes analytically tractable when one considers a product of experts (POE) approach:

$$p(\dot{\mathbf{X}}, \mathbf{X}, \boldsymbol{\phi}|\boldsymbol{\theta}, \gamma) \propto \prod_i p(\dot{\mathbf{x}}_i|\mathbf{m}_i, \mathbf{C}_i)p(\dot{\mathbf{x}}_i|f_i(\mathbf{X}, \boldsymbol{\theta}), \gamma_i\mathbf{I}), \tag{4.26}$$

giving an objective function for the surrogate likelihood sampler that is dependent on sampled interpolants and hyperparameters. The motivation for this marginalisation has been questioned by Wang and Barber [20] and indeed, it appears that the marginalisation will encourage the distribution of the ODE gradient to be forced towards the limiting delta distribution on the mean of some t-distribution resulting from the various Gaussian interpolant samples. Irrespective of this, we may use the Woodbury matrix identity and some Gaussian distribution identities (see Appendix) to obtain:

$$p(\boldsymbol{\theta}, \gamma|\mathbf{X}, \boldsymbol{\phi}, \sigma) \propto \frac{\pi(\boldsymbol{\theta})\pi(\boldsymbol{\gamma})}{\prod_i F(\gamma_i)} \exp\Big\{ -\frac{1}{2}\sum_i (\mathbf{f}_i - \mathbf{m}_i)^T(\mathbf{K}_i + \mathbf{I}\gamma_i)^{-1}(\mathbf{f}_i - \mathbf{m}_i)\Big\} \tag{4.27}$$

where $\mathbf{f}_i$ is evaluation of the ith equation of the ODEs $F(\gamma_i) = |2\pi(\mathbf{C}_i + \gamma_i\mathbf{I})|$ and $\pi(\boldsymbol{\theta})$ and $\pi(\boldsymbol{\gamma})$ denote the prior distributions over the parameters and the mismatch parameter of the gradients. By using this surrogate likelihood (a cheap approximation to the expensive likelihood function), we may adopt a far cheaper sampling routine compared with the traditional numerical integration approach that still obtains parameter estimates with the fit of the latent variable to the ODE as the fitting criterion. Since we are assuming random interpolants and hyperparameters, these must also be sampled at each stage of the algorithm. Due to the high dimensionality of the problem, as well as the dependence between the different variables, Calderhead et al propose a Gibbs sampling routine:

$$\boldsymbol{\phi}, \sigma \sim p(\boldsymbol{\phi}, \sigma|\mathbf{Y}) \tag{4.28}$$

$$\mathbf{X} \sim p(\mathbf{X}|\mathbf{Y}, \sigma, \boldsymbol{\phi}) \tag{4.29}$$

$$\theta, \gamma \sim p(\boldsymbol{\theta}, \gamma|\mathbf{X}, \boldsymbol{\phi}, \sigma) \tag{4.30}$$

where they sample parameters $\boldsymbol{\theta}$ and γ via a Metropolis-Hastings sampler based on the surrogate objective function. Therein lies the main problem with gradient matching as a cheaper alternative to sampling using the expensive likelihood: we are not sampling from the correct posterior distribution. Although the surrogate approximates the true likelihood, it does not account for the correct level of uncertainty in our final parameter estimates. On top of this, the marginalisation may lead to stability issues due to the reduction of the diagonal correction in $F(\gamma_i)$ and so the method can become heavily dependent on the use of an informative prior to restrict the magnitude of this γ parameter. We do, of course, prioritise robustness with respect to varying prior distributions and so this situation is not desirable. There have been various attempts to alleviate the faults of this method, one of which was introduced by Dondelinger et al. [7].

4.2.2.2 Dondelinger et al.

Similarly to [1], Dondelinger et al [7] force an amalgamation of the two $\dot{\mathbf{x}}$ distributions via a POE approach. However, via a slightly different consideration of the inference procedure, they introduce a feedback mechanism from the ODEs to the GP interpolants, essentially making the interpolant maximally consistent with the ODEs and encouraging matching of the two gradient distributions as opposed to a delta spike at the mean. Considering the joint distribution

$$p(\dot{\mathbf{X}}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\phi}, \gamma) = p(\boldsymbol{\theta})p(\boldsymbol{\phi})p(\gamma)\prod_i p(\dot{\mathbf{x}}_i|\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\phi}, \gamma)p(\mathbf{x}_i|\boldsymbol{\phi}_i) \tag{4.31}$$

Marginalisation over $\dot{\mathbf{x}}$ becomes analytically tractable via a POE approach allowing us to introduce the distributional contribution of the GP and the ODE. As before, the basis of the learning problem is consideration of the integral

$$\int \mathcal{N}(\dot{\mathbf{x}}|\mathbf{m}, \mathbf{A})\mathcal{N}(\dot{\mathbf{x}}|f(\mathbf{X}, \boldsymbol{\theta}), \gamma\mathbf{I})d\dot{\mathbf{x}} \tag{4.32}$$

Where, in this scenario, the motivation is more theoretically consistent since instead of encouraging a matching of first moments, it encourages a matching of distributions. This is owed to the dependence of the distribution of $\mathbf{X}$ on the parameters of the ODE. Marginalisation over $\dot{\mathbf{X}}$ now encourages movement of both distributions towards the other, where fit to the noisy observations discourages movement towards matching delta distributions.

By accepting movements in an MCMC procedure based on the acceptance criterion:

$$\alpha = \min\left(1, \frac{\pi(\mathbf{Y}, \mathbf{X}', \boldsymbol{\theta}', \boldsymbol{\phi}', \gamma', \sigma')}{\pi(\mathbf{Y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\phi}, \gamma, \sigma)}\right) \tag{4.33}$$

which is a standard Metropolis-Hastings acceptance criterion (see Chapter 2.2.1), the authors ensure that acceptance of interpolant $\mathbf{X}$ is dependent on the parameters from the ODEs. Therefore, implicitly, ODE parameters inform our selection of the data interpolant, whereas Calderhead et al only introduce a dependency of the parameters on the interpolant, not the converse. Nonetheless, the problem still stands that the distributional dependencies on the GP interpolant necessitate resampling from the Gaussian process posterior at each iteration of the MCMC which requires $\mathcal{O}(n^3)$ computations.

4.3 Accelerating True Likelihood MCMC with Fixed Interpolant Gradient Matching

Considering the gradient matching procedures from the literature [1, 7], the presence of random GP hyperparameters necessitates the use of a Gibbs sampling routine to sample from the posterior distribution. If we are willing to accept a decrease in the quantification of uncertainty, fixing the kernel hyperparameters at a maximum likelihood estimate allows the adoption of a fixed estimate for the interpolant of the data, taken to be the mean of the Gaussian process posterior distribution at a single time point.

$$\hat{x}(\tau) = k(\tau, \mathcal{T})(k(\mathcal{T}, \mathcal{T}) + \sigma^2\mathbf{I})^{-1}\mathbf{y} \tag{4.34}$$

Equalling a linear combination of the observed values, this expression acts as a linear smoother (see Hastie and Tibshirani [57]), smoothing over the noise in the observed values. I assume the derivative of the interpolant $\tilde{\mathbf{x}}$ to be determined by the ODEs subject to a Gaussian-distributed mismatch error with variance γ^2:

$$\frac{d\tilde{\mathbf{x}}}{dt} \sim \mathcal{N}(f(\tilde{\mathbf{x}}, \boldsymbol{\theta}, t), \gamma^2) \tag{4.35}$$

In order to provide a fixed interpolant for the data, this distribution can then be represented by its conditional mean:

$$\dot{\hat{\mathbf{x}}} = \frac{d\hat{\mathbf{x}}}{dt} = \mathbf{k}'(t, \mathcal{T})(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y} \tag{4.36}$$

where $\mathbf{k}'(t, \mathcal{T})$ is given by eq. 4.21. A cheaper (computational efficiency wise) alternative to the standard gradient matching methods allows a cheap approximation of the numerical integration likelihood function. The paradigm of fixed interpolants results in a deviation from the formalism of Calderhead et al [1] who treat the GP and ODE gradients as random variables and relate the dependent parameters by a POE approach (see Section 4.2.2.1). The proposed gradient matching scheme fixes the GP hyperparameters and interpolant based on maximum likelihood point estimates. This is computationally cheaper than sampling from the posterior, as in [1, 7], and avoids the issue discussed by Wang and Barber [20] since my distributional derivation no longer depends on a marginalisation over the gradient of the interpolant. The resultant objective function for the MCMC is

obtained as the negative log-likelihood of a standard multivariate Gaussian:

$$-\log\hat{p}(\dot{\hat{\mathbf{x}}}|\boldsymbol{\theta}) = n\log\gamma^2 + \frac{1}{2\gamma^2}\left|\left|\frac{d\hat{\mathbf{x}}(t)}{dt} - f(\hat{\mathbf{x}}(t), \boldsymbol{\theta})\right|\right|^2 + C \tag{4.37}$$

where $\dot{\hat{\mathbf{x}}}$ corresponds to the gradient of the interpolant as in eq. 4.36 and $||\cdot||$ corresponds to the Euclidean norm. With the introduction of a prior distribution over the parameters, $\boldsymbol{\theta}$ and mismatch term, γ, will form the posterior distribution of a Metropolis-Hastings sampler, $p(\boldsymbol{\theta}, \gamma|\hat{\mathbf{X}}, \dot{\hat{\mathbf{X}}})$. For the purposes of this work, the mismatch parameter γ was assumed constant over time and across different states. However, the likelihood function is easily adjusted in order to account for varying γ across these variables.

Considering the fixed interpolant gradient matching likelihood surface as shown for Lotka-Volterra in Figure 4.3—a system for which periodicity leads to the multimodal exact likelihood surface on the right of Figure 4.3—we observe the smoother likelihood surface associated with the surrogate likelihood function. Implementation of a Gaussian process (GP) procedure acts to constrain the space in which the signals may reside, limiting the problems introduced by aliasing of periodic signals to the inference problem. Therefore, efficiency gain is twofold since the surrogate likelihood function is less computationally expensive. Equally important in Figure 4.3 is the demonstration of the realignment of the likelihood that is introduced when using the surrogate likelihood function. This suggests that, rather than treating the gradient matching method as a replacement for the brute force numerical integration approach, we could use gradient matching to provide a better initialisation for the expensive MCMC sampler, leading to a reduction in the number of required numerical integrations of the ODEs. As such, the gradient matching likelihood function is only used in an MCMC burn-in phase to provide a more informative initial distribution for the expensive MCMC sampler, without introducing increased bias to the final parameter samples.

Consideration of the mismatch between the distributions leads to the proposal of a multiphase parameter inference scheme with a burn-in phase that implements a surrogate likelihood function. Introduction of a corrective phase allows the algorithm to correct for bias that is introduced by the burn-in phase, limiting the inaccuracy in the final posterior samples. These phases are now outlined, implementing a PSRF ladder in the algorithm with the target PSRF decreasing as the algorithm progresses. This allows a sufficient number of samples to be obtained at each phase, enabling a representative sample from the distribution.

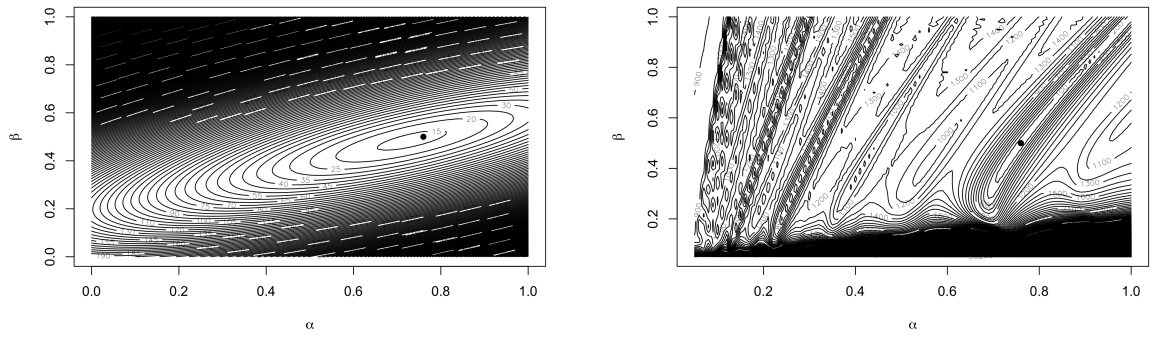

FIGURE 4.3: Considering the surrogate likelihood surface and the likelihood surface of the expensive true likelihood, we notice a realignment that introduces bias to the MCMC sampling procedure.

Surrogate burn-in phase

Begin with a surrogate burn-in phase initiated from some initial point in parameter space. This phase is continued until some target PSRF value is achieved or until some maximum number of MCMC steps have been performed. The numerical efficiency and smoothness of likelihood of the gradient matching approach enables fast convergence to a region close to the true stationary distribution. Effectively, we estimate the mean of the true posterior distribution in our burn-in phase, preventing the MCMC sampler from taking the "easier route" and just sampling high levels of noise variance such that there is extra flexibility in the fit of the signal to the observations. Although the surrogate burn-in is efficient and reduces variance in the MCMC, the approximation of the true likelihood introduces bias to the sampling scheme as displayed in Figure 4.3 and so we must correct for the introduced bias in a corrective phase of the algorithm.

Corrective phase

Beginning from the last point sampled in surrogate space, fix the parameters and sample the noise variance parameter for 200 steps, initiated at some estimate obtained during the smoothing process. This allows the ODE parameter samples to become consistent with the noise variance estimate from the GP. Beginning from the 200th parameter sample from this precorrective phase and the GP gaussian noise estimate, sample using the exact likelihood until a target PSRF value $PSRF_2$ is achieved or a preselected number of corrective steps N_{corr}, have been performed. The effect of this phase is presented in Figure 4.4, where observe the correction for the bias introduced in the samples obtained during the surrogate burn-in phase.

Sampling phase

The MCMC steps performed until now have only been used to drive us towards the correct stationary distribution. We may now initiate a sampling phase at the last sampled point from the corrective phase. The sampling phase is continued until a target PSRF value, $PSRF_3$ is reached or some number of steps, N_{samp} have been completed. The samples collected here will represent our samples from the posterior distribution for the parameters of the ODEs.

Ergodicity of this algorithm rests solely on ergodicity of the sampling phase since this is the only stage at which samples are retained Therefore, ergodicity is dependent on the choice of sampling MCMC method at this stage of the algorithm. Adopting the DRAM sampling method, I refer the reader to the literature [33] for a proof of ergodicity of the algorithm. In Algorithm 3, I outline pseudocode for the three-phase scheme while adopting the following nomenclature where $A \wedge B = \min(A, B)$:

$$A_1(\boldsymbol{\theta}, \gamma; \boldsymbol{\theta}', \gamma') = 1 \wedge \frac{\hat{p}(\dot{\hat{\mathbf{x}}}|\boldsymbol{\theta}', \gamma')\pi(\boldsymbol{\theta}')\pi(\gamma')\hat{q}(\boldsymbol{\theta}, \gamma|\boldsymbol{\theta}', \gamma')}{\hat{p}(\dot{\hat{\mathbf{x}}}|\boldsymbol{\theta}, \gamma)\pi(\boldsymbol{\theta})\pi(\gamma)\hat{q}(\boldsymbol{\theta}', \gamma'|\boldsymbol{\theta}, \gamma)} \tag{4.38}$$

$$A_2(\boldsymbol{\theta}, \sigma; \boldsymbol{\theta}', \sigma') = 1 \wedge \frac{p(\mathbf{y}|\boldsymbol{\theta}', \sigma')\pi(\boldsymbol{\theta}')\pi(\sigma^{2\prime})q(\boldsymbol{\theta}, \sigma|\boldsymbol{\theta}', \sigma')}{p(\mathbf{y}|\boldsymbol{\theta}, \sigma)\pi(\boldsymbol{\theta})\pi(\sigma^2)q(\boldsymbol{\theta}', \sigma'|\boldsymbol{\theta}, \sigma)} \tag{4.39}$$

where $p, \hat{p}$ are the exact and surrogate likelihoods, π is a prior and $q, \hat{q}$ are Gaussian proposal distributions with adaptive covariance.

4.4 Variations of the Multiphase Sampling Scheme

We have seen that various problems are posed by the problem of parameter inference in ODEs. As such, it would be naive the implement one variant of the multi-phase scheme and assume this is a best use of the methods outlined so far. For this reason, some variations of this multi-phase approach have been implemented for simulations in the thesis. In some cases, these aim to improve the standard multiphase approach. In others, the variations on implementation allow an attempt at improving the overall scheme.

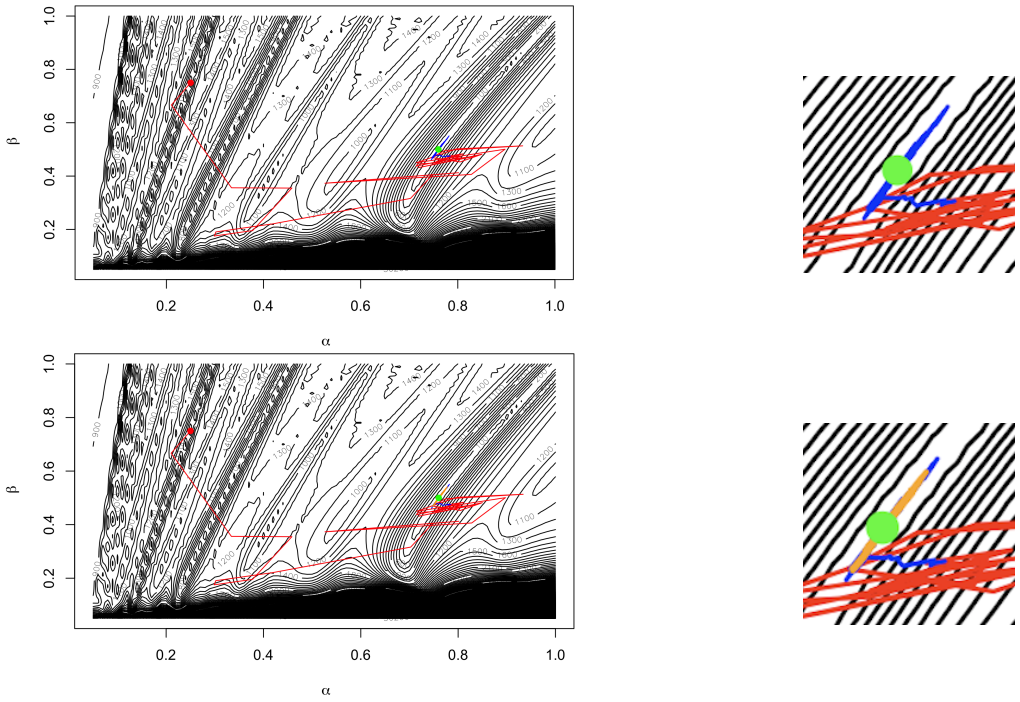

FIGURE 4.4: Considering the justification for a corrective phase in the proposed scheme. The top row shows the corrective phase on a global and zoomed scale allowing us to see the correction in distribution introduced by the corrective phase. The bottom row gives the sampling phase where, after the corrective phase, we can sample from the correct stationary distribution.

4.4.1 Multi-Phase Approach with a Selection of Interpolant

In some situations, particularly when observations have high levels of noise, the ML hyperparameters may not provide a GP posterior mean that gives a suitable approximation to the underlying signal. In these cases, I make use of a pool of interpolants from which one can sample to determine a surrogate landscape in a pooled surrogate burn-in phase. The idea is to find multiple interpolants at different lengthscale values in the GP. Then, we may find samples in each of the different surrogate landscapes and, after some burn-in period, take the mean of the sampled values to give a representation of the mode of the surrogate space. Evaluating the true likelihood at each of these point estimates provides an idea of the level of similarity between the two distributions. This stage can be parallelised, leading to a fairly insignificant computational cost. We then perform a finite state MCMC over the means, choosing between the interpolant pairs based on the true likelihood value. Alternatively, we could just choose the new interpolant based on that which gives the lowest exact likelihood value.

Algorithm 3 Multi-phase MCMC Sampling

1: Assign initial parameters $\boldsymbol{\theta}_0$ and select cheap surrogate likelihood function $\hat{p}(\dot{\hat{\mathbf{x}}}|\boldsymbol{\theta}, \gamma)$, an approximation of $p(\mathbf{y}|\boldsymbol{\theta}, \sigma)$ (our computationally expensive likelihood function).
2: **repeat**
3: Sample $\boldsymbol{\theta}_t$ and γ_t from $\hat{q}(\boldsymbol{\theta}_t, \gamma_t|\boldsymbol{\theta}_{t-1}, \gamma_{t-1})$.
4: Accept proposed point based on $A_1(\boldsymbol{\theta}, \gamma; \boldsymbol{\theta}', \gamma')$
5: **until** PSRF_1 or N_{surr} reached
6: **repeat**
7: Beginning from the last point sampled in surrogate space, keep σ^2 fixed at some estimate and sample $\boldsymbol{\theta}_t$ from $q(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})$.
8: Accept proposed point based on $A_2(\boldsymbol{\theta}, \sigma; \boldsymbol{\theta}', \sigma)$
9: **until** Until N_{pre} reached
10: **repeat**
11: Beginning from the last point sampled in the pre-corrective phase, sample $\boldsymbol{\theta}_t$ and σ_t from $q(\boldsymbol{\theta}_t, \sigma_t|\boldsymbol{\theta}_{t-1}, \sigma_{t-1})$.
12: Accept proposed point based on $A_2(\boldsymbol{\theta}, \sigma; \boldsymbol{\theta}', \sigma')$
13: **until** PSRF_2 or N_{corr} reached
14: **repeat**
15: Beginning from the last point sampled in corrective phase, sample $\boldsymbol{\theta}_t$ and σ_t from some proposal distribution $q(\boldsymbol{\theta}_t, \sigma_t|\boldsymbol{\theta}_{t-1}, \sigma_{t-1})$ (this will differ from those of the previous two phases).
16: Accept proposed point based on $A_2(\boldsymbol{\theta}, \sigma; \boldsymbol{\theta}', \sigma')$
17: **until** PSRF_3 or N_{samp} reached

4.4.2 Multi-Phase Approach with Alternative Norm

Taking inspiration from the literature [1], I propose the use of an alternative likelihood in which the distance metric is similar to the Mahalanobis distance measure. This leads to a posterior distribution of the form:

$$p(\boldsymbol{\theta}|\hat{\mathbf{X}}) \propto \frac{\pi(\boldsymbol{\theta})\pi(\gamma)}{\prod_n \mathcal{F}(\gamma_n)} \exp\left\{ -\frac{1}{2}\sum_n \left(\left(\frac{d\hat{\mathbf{x}}_n}{dt} - f_n(\hat{\mathbf{X}}, \boldsymbol{\theta}) \right)^T (\mathbf{A}_n + \gamma\mathbf{I})^{-1} \left(\frac{d\hat{\mathbf{x}}_n}{dt} - f_n(\hat{\mathbf{X}}, \boldsymbol{\theta}) \right) \right) \right\} \tag{4.40}$$

where $\hat{\mathbf{x}}$ is the smooth interpolant; $\pi(\boldsymbol{\theta})$ and $\pi(\gamma)$ correspond to prior distributions over the parameters of the ODE and the mismatch parameter; $\mathbf{A}_n = \mathbf{K}_n'' - \mathbf{K}_n'\mathbf{K}_n^{-1\prime}\mathbf{K}_n$ where $\mathbf{K}_n'', \mathbf{K}_n'$ and $'\mathbf{K}_n$ are as defined in eqs.4.21-4.23 and $\mathcal{F}(\gamma_n) = \sqrt{|2\pi(\mathbf{A}_n + \gamma\mathbf{I})|}$. Contrasting with the method outlined in Section 4.2.2.1, we no longer have a sampled hyperparameter and so all $\mathbf{K}$ and $\mathbf{A}$ matrices are constant. One would anticipate a realignment of the posterior distribution caused by this change in distance measure. However, it remains to be seen whether this will allow improved matching with the exact posterior distribution.

In additional, there will of course be a computational cost involved in the inversion of large matrices at each stage of the algorithm.

4.4.3 Delayed Acceptance Metropolis-Hastings

In Section 2.2.4, I outlined the method of Sherlock et al. [17]. For one of the comparisons presented in this thesis, this algorithm will be used for sampling in a two-phase scheme where there is a burn in phase in surrogate space that drives the sampler towards the true stationary distribution. This time, there is no need for a corrective phase as this would then negate the effect of the filter step in the DAMH method. Therefore, the algorithm begins to sample using Delayed Acceptance Metropolis-Hastings following the surrogate burn-in where the cheap distribution for the initial proposal is given by the gradient matching function from eq. 4.37.

For optimal use of the DAMH algorithm, we would hope for a surrogate distribution, $\hat{p}(\theta)$ that, in a topological sense, matches well with the true distribution. This does not only require matching of the minima, but also a fairly strong correspondence between the tails of the two distributions. In reality, this is often not the case, and so we now consider the problems encountered in two contrasting scenarios. Firstly, consider the case of a surrogate distribution where the tails are more weighted than the tails of the true distribution. The second acceptance criterion, α_2 should be able to filter out any steps that, despite representing the surrogate distribution, are a poor sample from the true distribution. However, heavier tails in the surrogate mean that a higher proportion of proposals will pass through our initial filter stage and so the gain in computational efficiency becomes less significant as the tails widen. The second inconsistency results when we have short tails in the surrogate distribution relative to the true distribution (or indeed a realignment as shown in Figure 4.3) the method becomes inefficient as the sampled points give an underrepresentation of the uncertainty in the parameter value.

Chapter 5

Empirical Method Comparison

When comparing different parameter inference schemes in ODEs, it is important that we consider the performance of the inference scheme in both parameter space and function space. As well as providing a diagnostic of non-identifiability, good performance in parameter space may not correspond to successful inference in function space since points lying close in parameter space may lie far apart in function space. Function space performance is assessed using functional RMS:

$$RMS_{func} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \|\mathbf{x} - \hat{\mathbf{x}}_i\|^2} \tag{5.1}$$

where $\mathbf{x}$ is the signal at the true parameter values and $\hat{\mathbf{x}}_i$ is an estimate of the signal at the ith parameter sample from the posterior distribution. For parameter space comparison, bias in the samples is obtained by subtracting the true parameter values from the posterior samples.

Assuming no bound on the computation time, the best method amongst those considered in this thesis for parameter inference in ODEs will inevitably be one of population MCMC or DRAM with the exact likelihood function, since these account for the true likelihood at all points of their sampling instead of just in the post burn in portion of the algorithm. However, one would expect that these methods will be far more computationally expensive than the method proposed in this thesis. As a result, we must consider the performance of different algorithms in terms of accuracy relative to the computational cost associated with their use where computational complexity can be assessed using the number of numerical integrations of the ODEs since the surrogate likelihood evaluations involve minimal computational burden.

In addition to a comparison with the standard methods, I also consider the comparison of the various implementations of the surrogate likelihood, allowing consideration of the optimal implementation of the fixed interpolant in the MCMC procedure. The methods to be compared are as described in Chapter 2. Namely, the proposed multi-phase approach with DRAM sampling (propDRAM) using the standard Euclidean norm, proposed multi-phase approach with DAMH sampling (propDAMH), population MCMC using the exact likelihood (popMCMC) and DRAM with the exact likelihood. The different abbreviations used, as well as their descriptions, are found in Table 5.1. This can be referred to for the nomenclature adopted in the final two chapters. Comparison studies are carried

Table 5.1: Abbreviations used throughout this results section.

Abbreviation	Description
PropDRAM	Three-phase proposed scheme with DRAM sampling using the expensive, exact, likelihood function.
PopMCMC	Population Markov Chain Monte Carlo with expensive likelihood function.
DRAM	Delayed Rejection Adaptive Metropolis with expensive likelihood function.
PropDAMH	Two-phase scheme with surrogate burn-in phase followed by Delayed Acceptance Metropolis Hastings sampling phase.

out for fixed parameters sets in each ODE example outlined in Chapter 3 where I demonstrated that these parameter configurations provided sufficient coverage of the difficulties encountered in parameter inference for nonlinear ODEs.

5.1 Sampling from Multiple Posteriors from Multiple Datasets

I wish to assess the performance of the method through a simulation study, providing confidence in good performance in the presence of real observed data. Given the dependence of the method on the ability to find a smooth interpolant for the data, I consider the performance of the method on ten different datasets each generated from the four Differential Equation models outlined in Chapter 3. This method of sampling parameter values and representation of results appears to be a hybrid of the Frequentist and Bayesian approaches and as such this requires more explanation.

The Bayesian approach is to condition inference on the data that have been measured or observed and to get the posterior distribution. However, since data can be generated

synthetically, we can easily get a whole ensemble of data. This seems to lend itself to a frequentist paradigm: choose an estimator, say the conditional mean, and then obtain the distribution of the estimates over all the synthetic data sets. While this is methodologically consistent (within the frequentist paradigm), it requires the mapping of a posterior distribution to a classical point estimate, which incurs an inevitable loss of information (as discussed in Section 2.1.1).

The approach proposed in this thesis uses a combination of both paradigms: For each data set, I infer the Bayesian posterior distribution, and then combine all posterior distributions over all data sets thus obtained into a super distribution. For instance, when carrying out Bayesian inference with MCMC, the super distribution is the union of all individual MCMC samples. Intuitively this approach avoids the information loss inherent in the frequentist textbook paradigm.

Consider a data set obtained by numerically solving the differential equations, D_0, corrupted with additive noise: $D = D_0 + \epsilon$. In the limit of averaging over an infinite number of such data sets we get:

$$\begin{aligned}
\frac{1}{M}\sum_{i=1}^{M} p(\theta|D_i) &\rightarrow \int p(\theta|D)p(D)dD \\
&= \int p(\theta|D_0+\epsilon)p(\epsilon)d\epsilon \text{ Since } D_0 \text{ is deterministic.} \\
&= \int p(\theta,\epsilon|D_0)d\epsilon \\
&= p(\theta|D_0)
\end{aligned}$$

So the combination of all data sets has removed the influence of the noise (conditioning on D_0 rather than D). In that way, the bias has been reduced and the results are more representative (rather than merely reflecting potential idiosyncrasies of one particular data set). Note, though, that as a consequence of the integration the variance might have increased.

For clarification, consider the example in which we observe a noise corrupted version of a quantity of interest:

$$D = \theta + \epsilon$$

where ϵ is iid noise drawn from $\mathcal{N}(0, \sigma^2)$. Given the data D and assuming a uniform prior on θ, the posterior distribution is given by

$$p(\theta|D) = \mathcal{N}(\theta|D, \sigma^2).$$

Now, assume that we have a large ensemble of data subject to the same noise corruption: $D \sim \mathcal{N}(\theta_0, \sigma^2)$. By the law of large numbers, averaging the posterior distribution converges to:

$$\int p(\theta|D)p(D)dD = \int \mathcal{N}(\theta|D, \sigma^2)\mathcal{N}(D|\theta_0, \sigma^2)dD$$

Where θ_0 is the true parameter and θ is the inferred parameter. The integral above has solution $\mathcal{N}(\theta|\theta_0, 2\sigma^2)$. To emphasise what this example shows, standard Bayesian inference on one data set, D, gives the posterior distribution $N(\theta|D, \sigma^2)$. The approach in this thesis, based on averaging the posterior distributions over all data sets, gives the posterior distribution $N(\theta|\theta_0, 2\sigma^2)$. This shows that the bias has been reduced (the distribution is now conditional on the true parameter θ_0, not the noisy observation D), whereas the uncertainty has increased (the variance is twice as large as before).

To summarise, the procedure shows the representative performance of the estimator without being misled by any potential idiosyncrasies of one particular dataset (conditioning on θ_0 rather than D) while capturing the uncertainty inherent in the data generation (increased variance). In textbook Bayesian inference, we would combine all data sets to obtain the posterior $p(\theta|D_1, , D_M)$, with reduced bias and reduced variance. However, the aim of the procedure is not to improve the estimator. The data come from a synthetic benchmark study with known ground truth; so, we know the parameters already and there is no need for estimation. The objective of this study is to quantify the typical inference accuracy and uncertainty of the estimator applied to a data set of size $|D|$. For that reason, the data sets are not combined in the textbook Bayesian sense, but the posterior distributions are combined, as described above.

5.2 Comparison with Standard Methods

Table 5.2 below gives a summary of the convergence of each of the four methods across the four different ODE systems. Considering all four methods, only the three-phase proposed scheme is able to converge in all four cases. As is shown in the proceeding sections,

lack of convergence does not necessarily correspond to poor estimative properties. Particularly in the population MCMC case, this may be a result of the default settings of the algorithm being suboptimal and could potentially be improved in extensive further explorative simulations (which were beyond the remit and time frame of my project). The remainder of this section presents more detailed comparisons of parameter inference

TABLE 5.2: Indicating whether or not the four methods were able to achieve a PSRF value of 1.01 in the four benchmark ODE systems.

ODE System	LV	FHN	GO	STC
PropDRAM	✓	✓	✓	✓
PopMCMC	X	✓	✓	X
DRAM	X	X	X	✓
PropDAMH	X	X	X	X

method performance for each of the four benchmark ODE systems.

5.2.1 Lotka-Volterra

For the purpose of this study, parameter values were chosen to allow sufficient periodicity in the resultant ODE signals. Setting $\alpha = 0.76, \beta = 0.5, \gamma = 0.4$ and $\delta = 0.3$ produced signals as shown in Figure 3.1 where observations are simulated at regular one unit intervals between 0 and 100. This parameter configuration will be used for all comparisons in the Lotka-Volterra system. In addition, for the noisy observations, I add Gaussian noise with noise variance selected to provide an average signal-to-noise ratio (SNR) equal to 10. We have seen previously, for two dimensions, (Figure 3.6) the difficulties that are present in the inference procedure for this system. Considering Figure 5.1, the inability of the exact DRAM approach to obtain accurate ODE parameter estimates is evidenced by the high levels of bias in the parameter samples. Meanwhile, the proposed method with DRAM sampling and population MCMC are able to accurately infer the parameter values. Given that my aim is to compare the performance of the proposed scheme against the other methods, the difference in absolute bias between the proposed method and the other three approaches is plotted in Figure 5.2. This time, y-axis is constrained to prevent the dominating effect of the DRAM samples, enabling better comparison of performance against population MCMC and the two-phase approach with DAMH sampling. Despite the lack of convergence of population MCMC and DAMH (indicated in Table 5.2), the performance in parameter space is similar to that of the three-phase scheme.

Considering performance in function space in Figure 5.3, I provide only the difference in function space performance between the methods. It appears that the performance in function space of the the proposed scheme is similar to that of population MCMC and vulnerability of DRAM to local optima is indicated by large functional RMS values (compared with those of the three-phase scheme). An interesting observation is the existence of outlying values in function space obtained using the DAMH method. These were less apparent in the parameter space and suggest convergence to a local optimum close to the global optimum, providing an example where good parameter space performance does not lead to good performance in function space. Computational complexity involved in the four methods can be compared using the number of numerical integrations of the ODEs since the surrogate likelihood evaluations involve minimal computational burden. In Figure 5.4, we consider the difference in number of numerical integrations of the three benchmark methods compared with my proposed scheme. As expected, population MCMC and DRAM require a far greater number of evaluations of the ODE system. The number of numerical integrations required in the DAMH method is similar to that of the proposed scheme due to the lack of convergence when using the DAMH scheme.

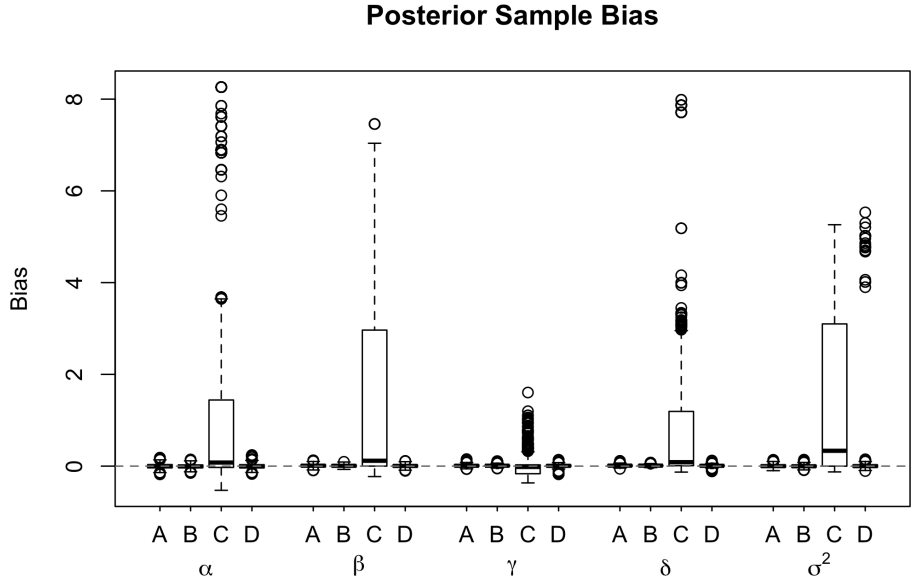

FIGURE 5.1: Bias from posterior samples using each of the four alternative methods for inference in the Lotka-Volterra model. A=propDRAM, B=popMCMC, C=DRAM and D=propDAMH. The dashed line corresponds to a bias equal to zero.

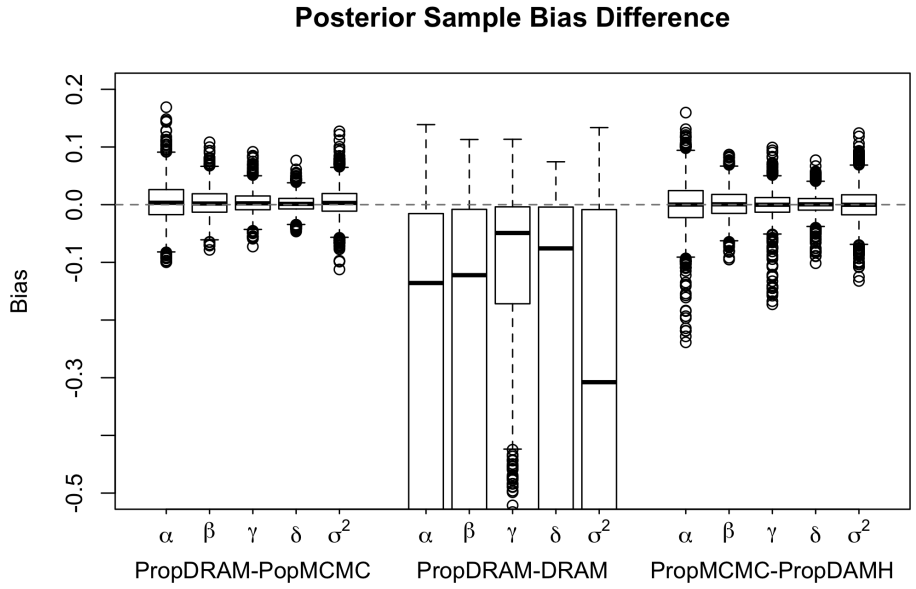

FIGURE 5.2: Difference in absolute bias using the three-phase proposed scheme compared with the three other methods for obtaining samples from the Lotka-Volterra model. Values above the dashed line at zero correspond to higher level of bias in the three-phase proposed method.

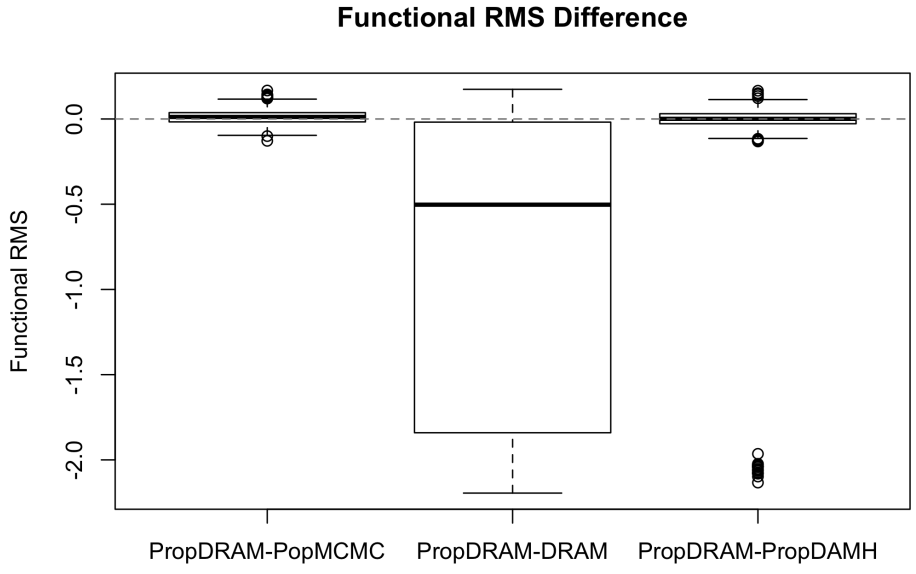

FIGURE 5.3: Difference in function space performance for each of the methods in the Lotka-Volterra model. This corresponds to determining the functional RMS for each of the posterior samples and then taking the difference between these values for the proposed scheme compared with the other three methods.

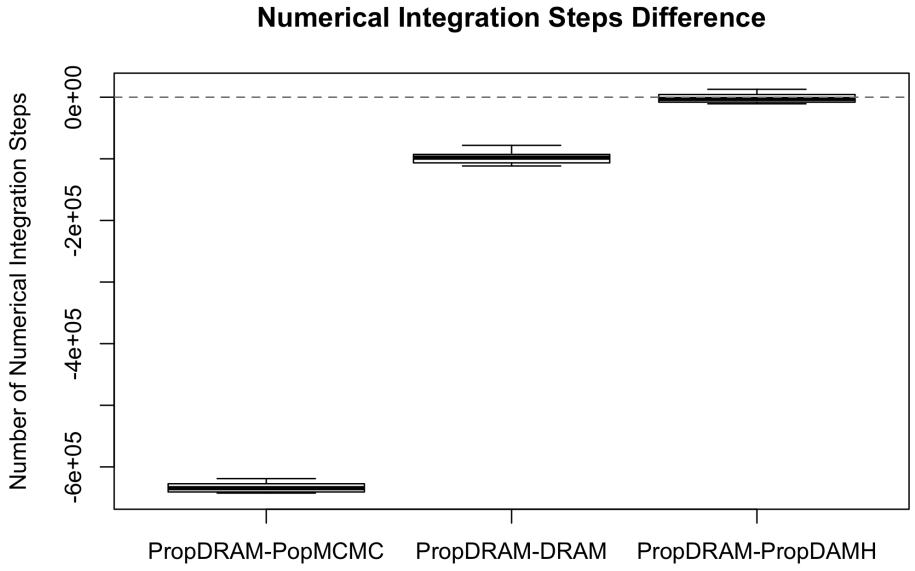

FIGURE 5.4: Difference in number of numerical integrations performed in the three-phase proposed method compared with the three other methods for inference in the Lotka-Volterra system. The proposed scheme requires the lowest number of integration steps in order to achieve the target PSRF value.

5.2.2 Goodwin Oscillator

Following the work of Girolami et al [50], I adopt parameter values $k_1 = 72, k_2 = 1, k_3 = 2, k_4 = 1$ and $k_5 = 1$. Measurements are simulated at 0.5-time intervals over the range from 0 to 60 with added Gaussian noise with variance equal to 0.5. Performance of the different methods for the Goodwin Oscillator ODEs can be considered in a similar manner to the Lotka-Volterra equations. This system proves to be less challenging than the Lotka-Volterra case. However, periodicity in the data still presents some multimodality in the likelihood surface (see Figure 5.5), leading to a reputation as a challenging learning problem. Considering the performance in parameter space via the bias in the posterior samples in Figure 5.6, we notice similar performance of population MCMC and DRAM to that observed in the Lotka-Volterra model as population MCMC converges to the stationary distribution and obtains accurate parameter samples. Meanwhile, DRAM fails to evade the local optima leading to a lack of convergence and some outlying parameter samples. Performance of the DAMH method is similar in this case to the proposed method, owing to the improved alignment of the surrogate likelihood and the true likelihood. However, Figure 5.9 shows that the number of numerical solutions required in the DAMH algorithm tends to be greater than that required to obtain convergence in

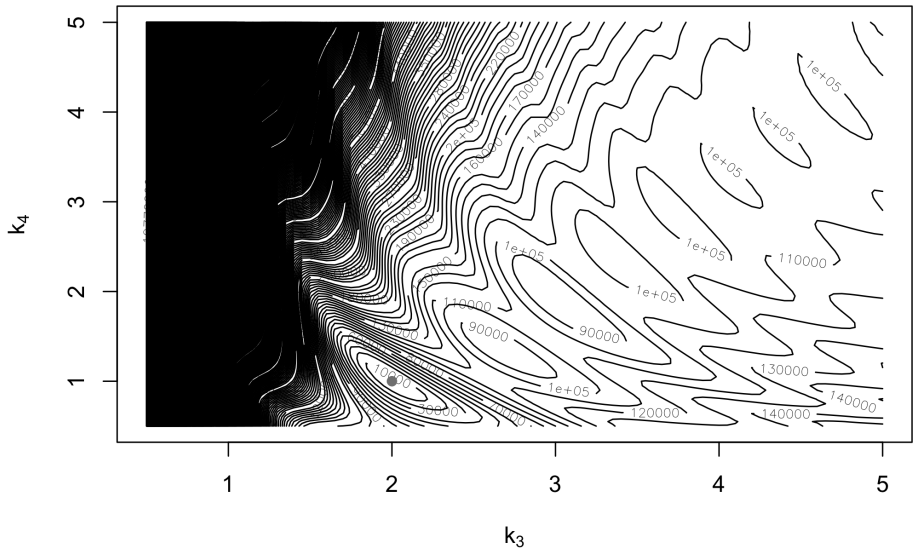

FIGURE 5.5: Negative log likelihood surface of the Goodwin Oscillator over parameters k_3 and k_4. We notice the multimodality of the surface which leads to a challenging learning problem. The red point is the true parameter value.

the three-phase scheme. In this case, this is a result of the heavier tails in the surrogate distribution which lead to a lack of discrimination against poorer fitting parameter configurations. Therefore, the implementation of the DAMH algorithm moves closer to that of the proposed method with DRAM sampling, even for the few datasets where convergence was achieved before the maximum number of steps had been performed.

5.2.3 FitzHugh-Nagumo

I take inspiration from the work of Ramsey et al. [6] and assume parameter values $\alpha = \beta = 0.2, \gamma = 3$ and noise variance equal to 0.25 (this variance is the larger of the two observation noise variances adopted by Campbell and Steele [36]). Initial values were set to (V=-1, R=1) and observations were simulated at intervals of length 0.2 between 0 and 20. Through consideration of two-dimensional contour plots (the $\alpha - \beta$ likelihood space is in Figure 5.10 and $\alpha - \gamma$ is in Figure 6.3), we can confirm that local optima are not as prevalent on the likelihood surface as in the previous two cases, but as before, DRAM is unsuccessful in converging to the stationary distribution as presented by some outlying values in the posterior bias plots in Figure 5.11. The boxplots in Figure 5.12 present the similarities between the results of population MCMC and the proposed scheme. DRAM

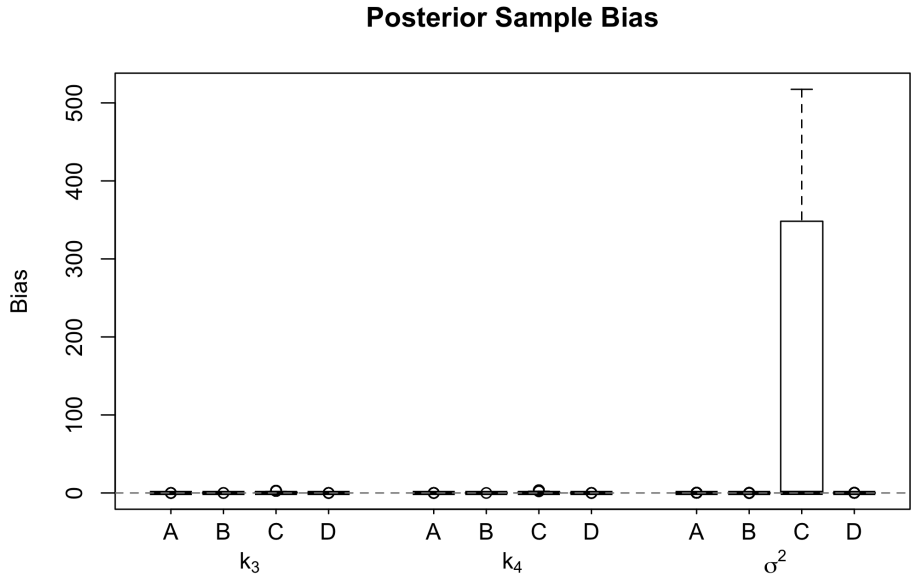

FIGURE 5.6: Bias in posterior sample for the four different methods in the Goodwin Oscillator model. A=propDRAM, B=popMCMC, C=DRAM and D=propDAMH. The proposed scheme, population MCMC and DAMH with surrogate burn-in perform similarly. DRAM is poor due to local optima and lack of mixing.

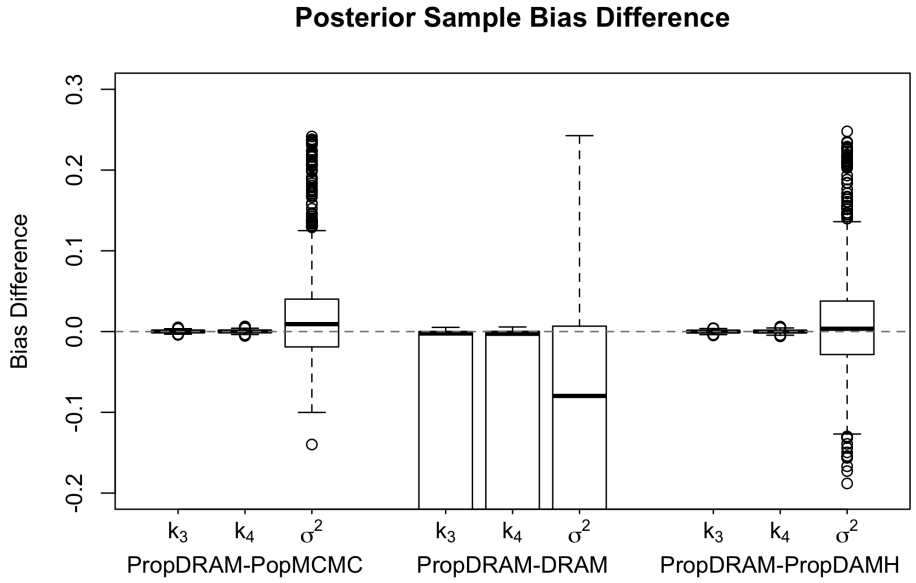

FIGURE 5.7: Difference in absolute bias between the three-phase proposed scheme and the other methods for the Goodwin Oscillator. This shows the similar performance of DAMH with surrogate burn-in and population MCMC compared with the proposed method which vastly outperforms exact likelihood DRAM sampling.

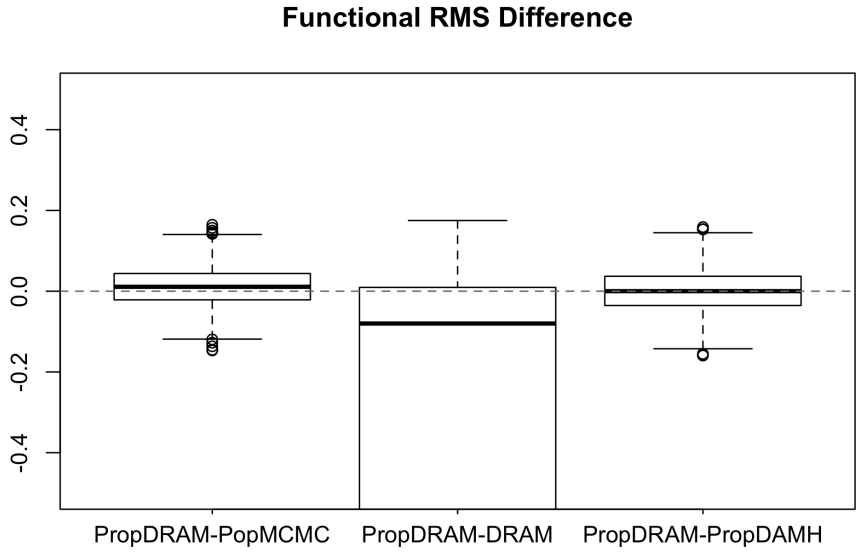

FIGURE 5.8: Difference in functional RMS of the posterior samples from the Goodwin Oscillator using the three methods compared with the proposed scheme. The performance of the proposed method is similar to that of population MCMC and DAMH with surrogate burn-in. DRAM, however, gets trapped in local optima.

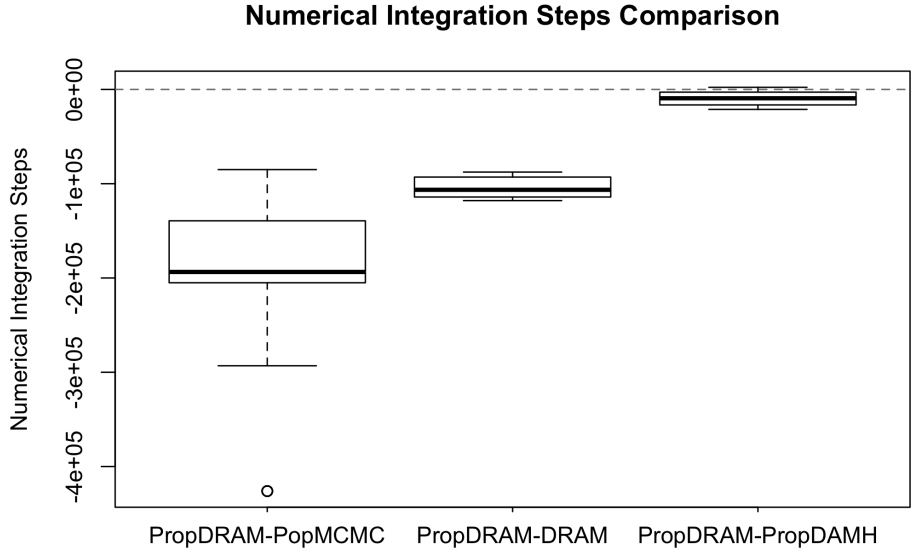

FIGURE 5.9: Difference in number of numerical integration steps required in the proposed scheme and the three alternative approaches.

is again the worst performing sampler. Despite being unable to converge, DAMH sampling with a surrogate burn-in phase performs similarly to the three-phase approach.

Again, consideration of the number of numerical integrations performed shows us that the DAMH sampler requires similar computational complexity to the three-phase scheme thus defeating its purpose as a more efficient sampling routine. Population MCMC and DRAM both require a far greater number of computational integration steps than the proposed method.

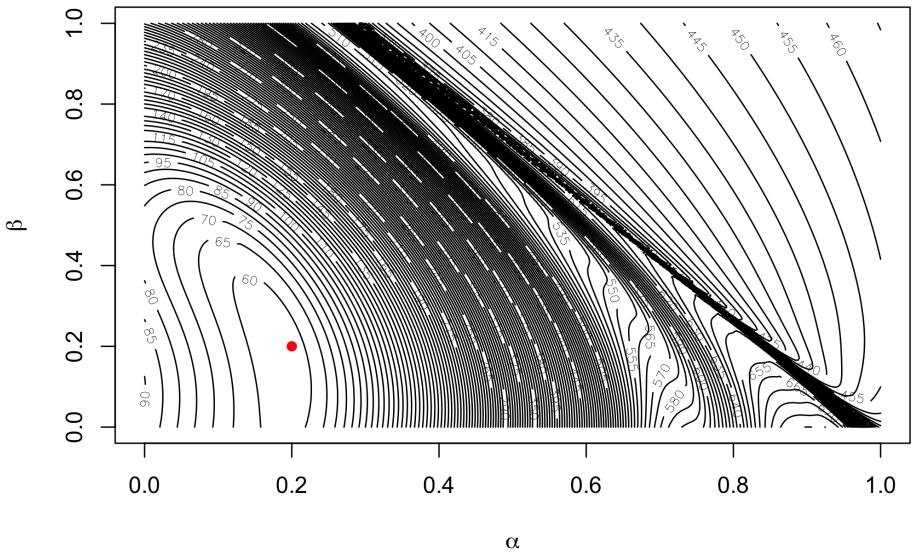

FIGURE 5.10: Negative log likelihood for the FitzHugh-Nagumo system with γ fixed. We notice that the likelihood surface is less multimodal than in previous examples.

5.2.4 Signal Transduction Cascade

The parameter configuration used takes inspiration from the literature [7], setting the parameters as follows: $k_1 = 0.07, k_2 = 0.6, k_3 = 0.05, k_4 = 0.3, V = 0.017$ and $K_m = 0.3$. For compliance with parameter identifiability, the parameter K_m is assumed known and so this parameter is fixed at its true value leaving a parameter space of dimension 6. A total of 20 equally spaced measurements were simulated between times 0 and 10. By consideration of two-dimensional likelihood surfaces (one of which was shown on the right of Figure 3.10), we observe that this model appears to have the least multimodal likelihood surface. Given this topological observation, one would postulate there being less of an advantage in adopting a parallel scheme, such as population MCMC, for this particular example. Indeed, there are other problems encountered with this system that further weaken the use of population MCMC.

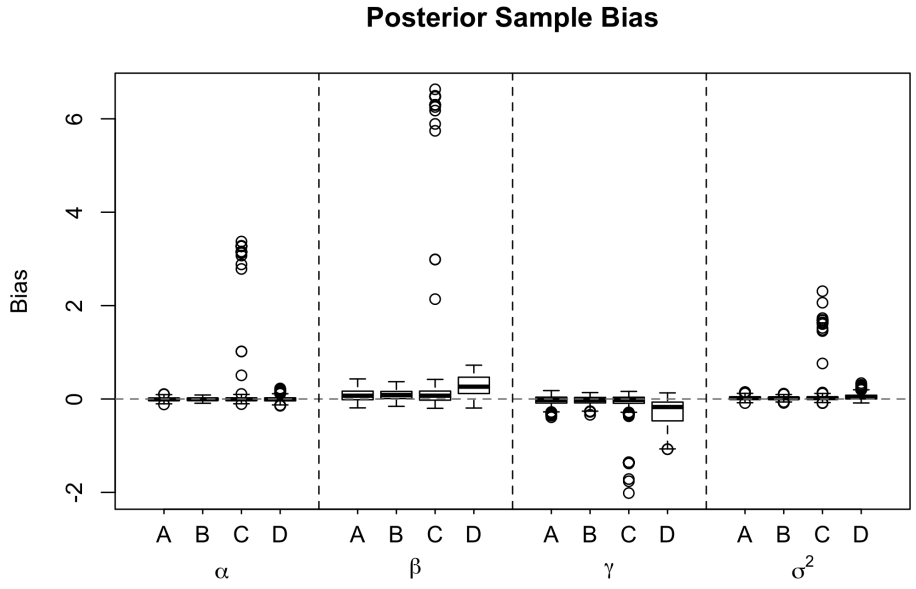

FIGURE 5.11: Boxplot giving bias in posterior samples for the four methods performing inference in the FitzHugh-Nagumo model. A=propDRAM, B=popMCMC, C=DRAM and D=propDAMH. Again we observe a tendency for DRAM to become trapped in local optima.

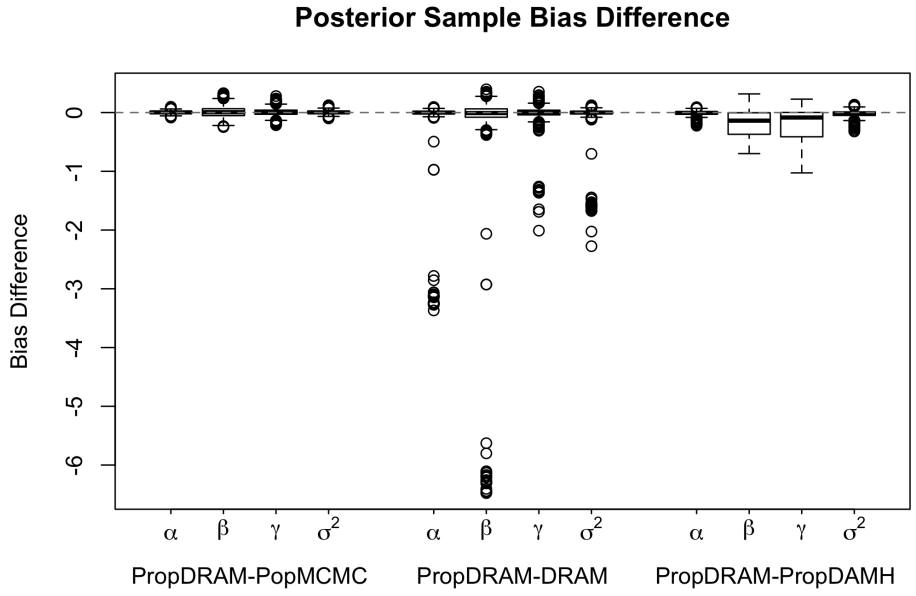

FIGURE 5.12: Difference in absolute bias of posterior samples obtained using the proposed scheme and the three other methods for the FitzHugh-Nagumo system. We observe similar performance of the proposed scheme, population MCMC and DAMH but DRAM struggles due to lack of convergence.

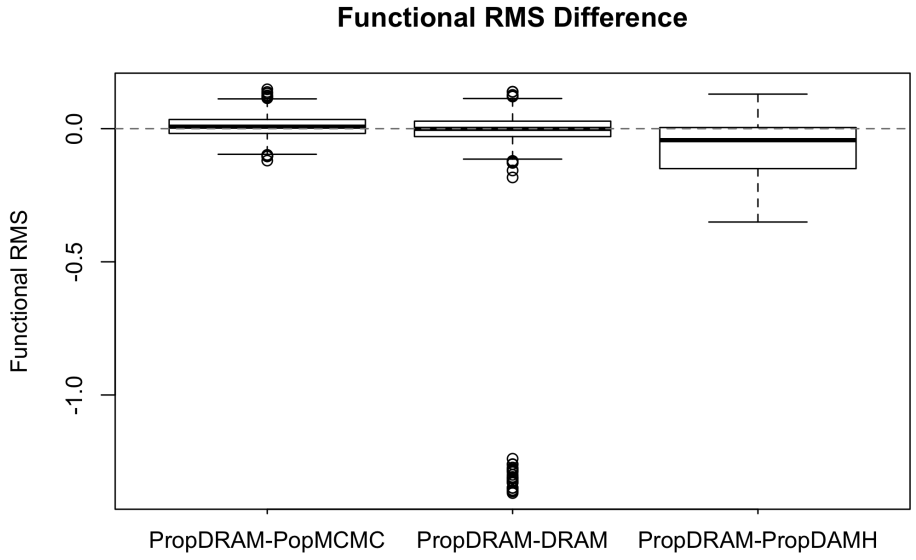

FIGURE 5.13: Difference between function space performance of the proposed scheme and the benchmark sampling methods in the FitzHugh-Nagumo model. The proposed method performs similarly to population MCMC and outperforms the other two methods.

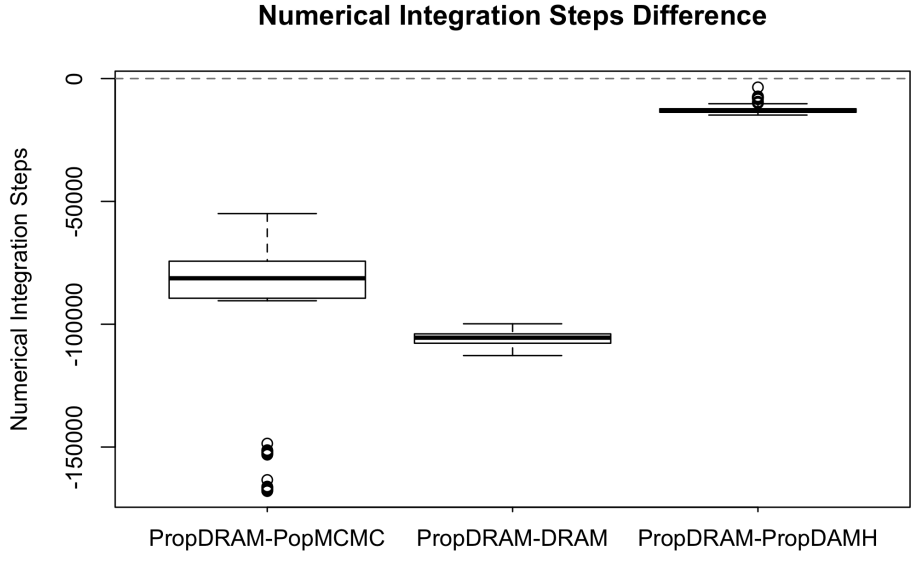

FIGURE 5.14: Difference between the number of numerical integrations required in the proposed scheme and the three comparison methods. We see that the three-phase proposed method requires the smallest number of computationally expensive numerical integration steps.

The boxplots in Figure 5.15 appear to display some identifiability issues with the $k_2 - k_3$ parameters and the plot on the left of Figure 5.19 indeed confirms some weak identifiability. Considering the bias displayed in Figure 5.19, it appears that performance of the proposed method and population MCMC are similar in terms of accuracy. However, we observe in Figure 5.18 that, for the proposed method, these similar levels of accuracy are achieved in a far smaller number of numerical integration steps. On this occasion, I choose to give both functional RMS and difference in functional RMS boxplots in Figure 5.17, allowing us to investigate the existence of identifiability issues in the inference problem. Indeed, when one considers the plot on the left of Figure 5.19, the suspicion is matched by an apparent relationship between the two parameters in question which can be confirmed by fitting a linear regression model to the samples of these parameters. Resultantly, it is important to assess the ability of the methods to infer the ratio of the parameters, k_3/k_2 and this is shown in the boxplot of Figure 5.19. The inference appears to have been more successful this time, and there is an obvious improvement in inference when we adopt an adaptive scheme as opposed to population MCMC.

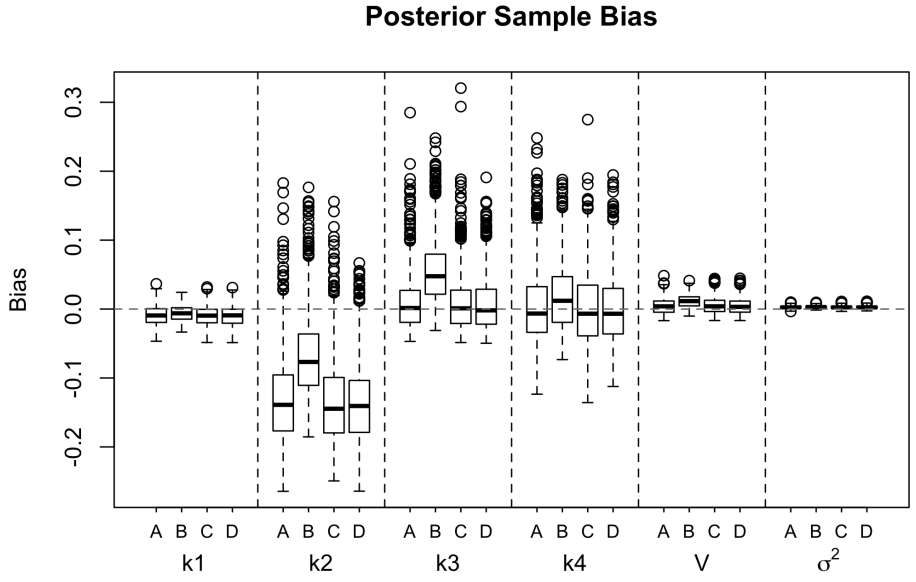

FIGURE 5.15: The bias in the posterior samples for each of the four methods performing parameter inference for the Signal Transduction Cascade where A=propDRAM, B=popMCMC, C=DRAM and D=propDAMH. It seems that the proposed method performs similarly to the others for parameter inference in this model but it looks like an identifiability issue may exist between parameters k2 and k3.

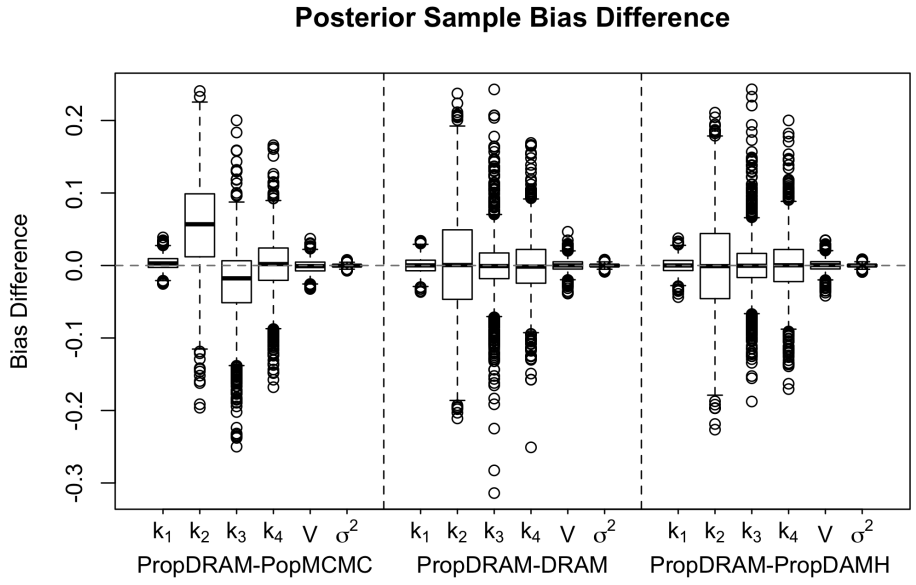

FIGURE 5.16: Difference between the absolute posterior sample bias for the three-phase method compared with the other 3 schemes sampling from the posterior parameter distribution of the signal transduction cascade. In this case, the proposed scheme seems to be more accurate than the other three methods.

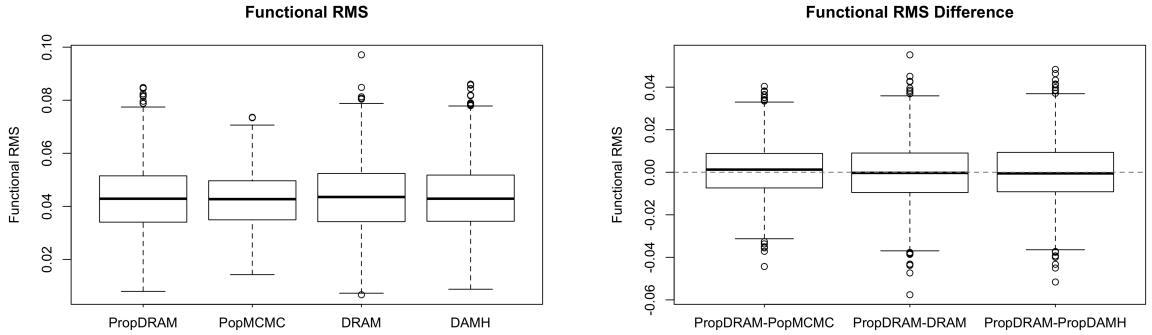

FIGURE 5.17: Left: Functional RMS corresponding to parameter samples obtained from posterior parameter space of the signal transduction cascade using each method. Despite seemingly poor performance in parameter space, all methods appear to perform well in function space (suggesting some sort of identifiability issue). Right: Difference in functional RMS value between three-phase proposed scheme and the other methods we have considered.

5.3 Other Comparisons

5.3.1 Unknown Initial Conditions

The assumption of unknown initial conditions increases the dimension of the parameter space as these are effectively treated as extra free parameters in the inference problem.

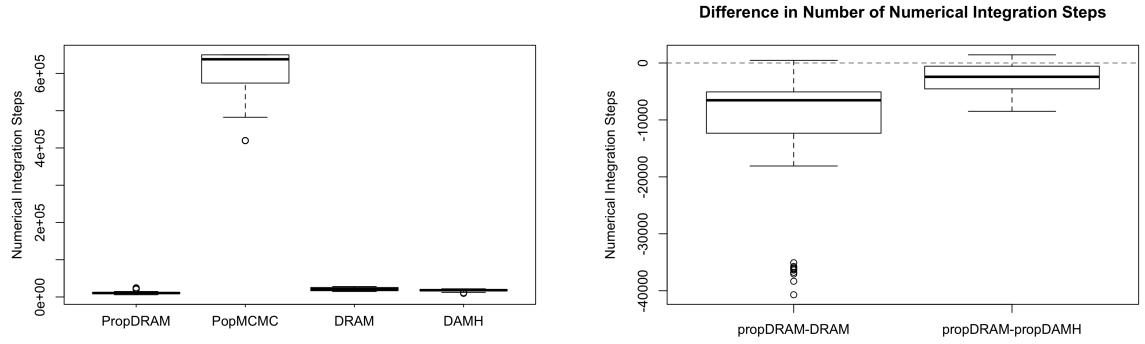

FIGURE 5.18: Left: Number of numerical integration steps attempted by the four different methods for parameter inference in the STC model. The three-phase method requires the lowest number of steps of the three considered approaches. Right: Difference between number of numerical integrations carried out in the three-phase scheme, DRAM and the two-phase scheme with DAMH sampling. Values greater than zero show that less numerical integrations have been attempted in the three-phase scheme.

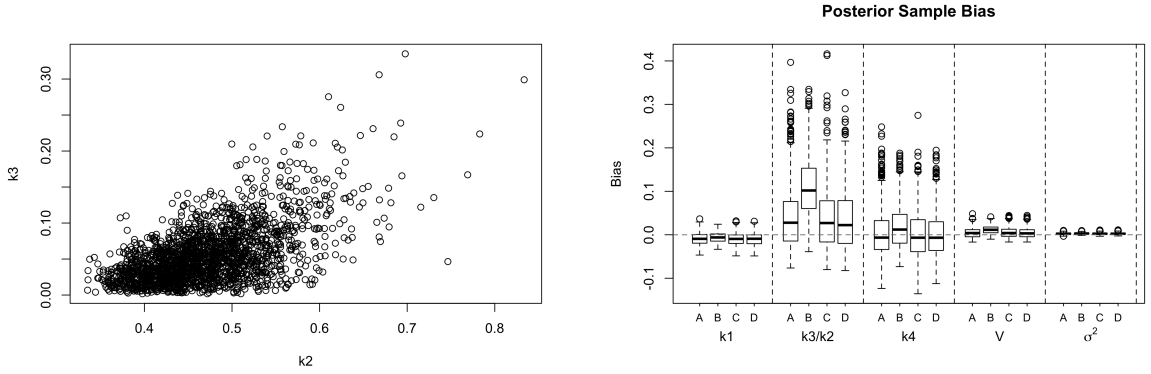

FIGURE 5.19: Left: Presenting the weak identifiability problem that exists between parameters k_2 and k_3 by plotting posterior samples of these parameters against one another. Right: Posterior bias of samples where we instead attempt to infer the ratio k_3/k_2. The inference appears more successful and we observe the improvement of adaptive algorithms, DRAM and the three-phase proposed method, over population MCMC.

These present a new challenge since in order to infer the parameters one must first infer the initial conditions which introduce added local optima to the likelihood surface. One would anticipate these initial conditions having more of an effect on the standard procedures than my proposed scheme since these are smoothed over in the initial burn in phase where the sampler is driven towards the correct stationary distribution, allowing us to better infer the initial conditions of the ODEs. Due to time restrictions, I chose to only assess the performance of the algorithms with unknown initial conditions for the Goodwin Oscillator model (of the models considered with multimodal likelihood surfaces, this had

the smallest difference in performance of the three different algorithms), but I would anticipate similar effects on the different inference schemes for the alternative models too. I will consider two priors with different levels of dispersion. The less informative given by a $unif(-50, 50)$ distribution and the more informative a $unif(-20, 20)$. In both cases, I assume no knowledge of the initial conditions and so I sample the starting value from the prior distribution.

Figure 5.20 presents the bias in the posterior samples when adopting the more informative uniform prior. As expected, there appears to be more of an influence on the traditional DRAM approach. However, it appears to fairly accurately estimate the parameters in the ODE despite its lack of ability to consistently infer the correct initial conditions. This suggests the existence of system initialisations that provide a signal with reasonable fit to the true signal at similar parameter values. Or, expressed differently, the system is robust to changes in initial conditions. Interestingly, it appears that the DAMH method underestimates the observational noise variance by an order greater than 10.

Figure 5.21 allows consideration of the performance of the four algorithms with a more uninformative uniform prior. Here we notice a fairly large difference in the performance of DRAM as it becomes trapped in an even larger number of local optima due to the increased range of support of the initial conditions. The effect this has on the three-phase proposed scheme and population MCMC is fairly minimal. The performance of DAMH is similar to before, where again the sampler seems to underestimate the observational noise variance. Given the recurrence of this mis-estimation, it will be worth discussing in the following chapter.

5.3.2 Alternative Distance Metric

As discussed in section 4.4.2, it is worth considering performance of the multiphase approach with an alternative distance measure that is motivated by the literature [1]. Performance will be compared in both the surrogate phase and sample phase. Results are given for the Goodwin Oscillator where, although we observe good accuracy, there is a lack of convergence due to the low acceptance rate observed in the surrogate burn-in phase. Viewing the plot on the left of Figure 5.22, we observe slight deterioration in accuracy of measurement of the noise variance parameter, but there seems to be little difference in accuracy of inference for the parameters of the ODE. In Figure 5.23, I consider the bias in the surrogate samples with and without the alternative distance metric. There is little difference in the parameter inference accuracy but, as previously discussed,

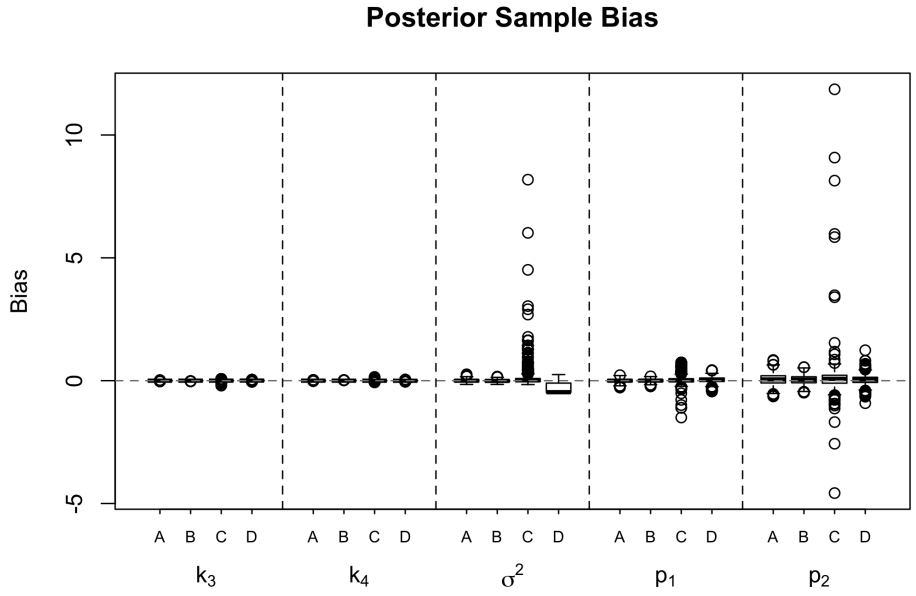

FIGURE 5.20: Sample bias for each of the four sampling methods in the Goodwin Oscillator with a more informative uniform prior on the unknown initial conditions. A, B, C and D have the same meaning as in Figure 5.15. The increased dimension in parameter space appears to be less problematic in the case of the proposed scheme and population MCMC.

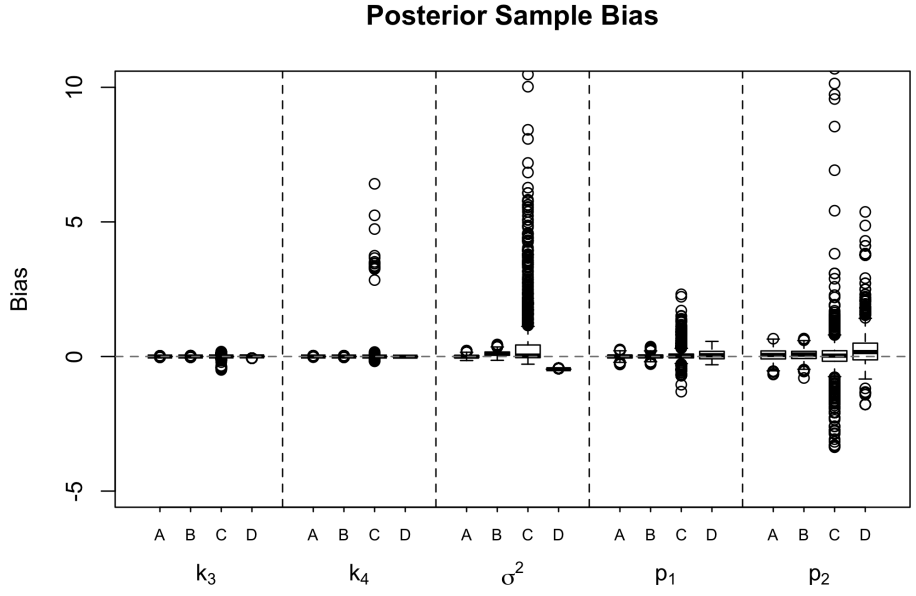

FIGURE 5.21: Sample bias for each of the four sampling methods in the Goodwin Oscillator with a more uninformative prior on the unknown initial conditions. A, B, C and D have the same meaning as in Figure 5.15. The increased dimension in parameter space appears to be less problematic in the case of the proposed scheme and population MCMC.

there is a tendency for the mismatch parameter to become as small as computationally possible (given in the informative prior) in the case of the alternative distance metric.

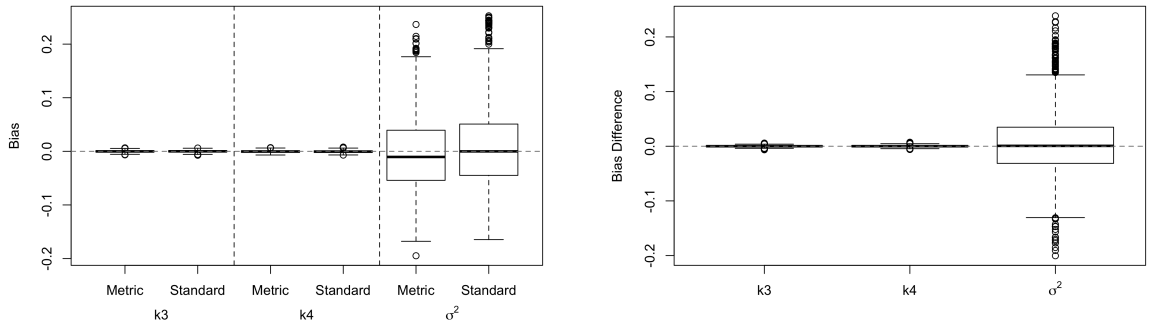

FIGURE 5.22: Left: Bias of samples using the three-phase approach with the alternative metric from [1] (Metric) and using the standard Euclidean norm (standard). Right: Comparing the bias in the posterior samples of both methods by taking the difference of absolute bias. Values lower than zero indicate more accurate inference using the standard method.

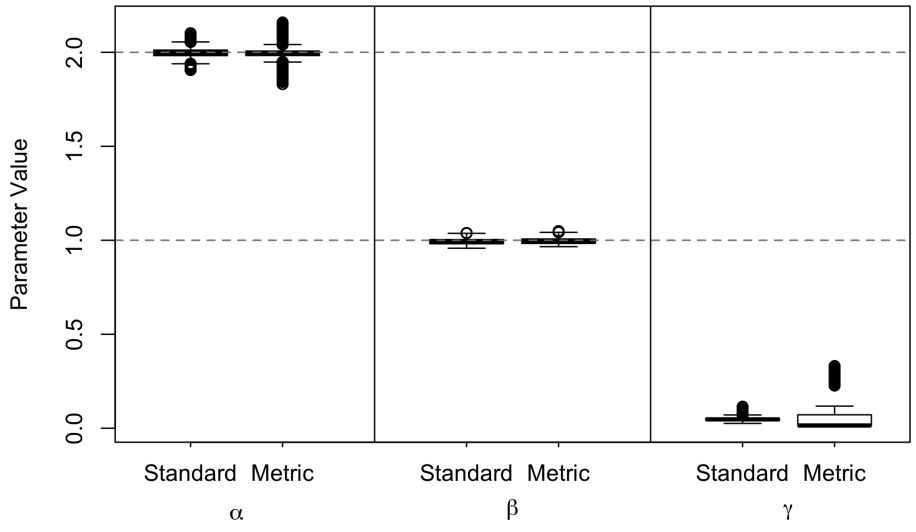

FIGURE 5.23: Surrogate samples in the burn-in phase where I adopt the surrogate likelihood with the alternative distance metric and without the alternative distance metric. Notice the tendency for the mismatch to become as small as possible in the case of the alternative distance metric.

Chapter 6

Discussion and Conclusion

6.1 Discussion

6.1.1 Multi-Phase Approach versus Traditional Methods

Traditionally, parameter inference in nonlinear ODEs relies on an ability to numerically solve the system of equations in order to sample from the posterior distribution of the parameters. However, as discussed in this thesis, there are various properties of the resultant likelihood that make this function flawed as an objective function in our MCMC schemes. Considering the argument of Korattikara et al [58], the multiphase approach is optimal (amongst the considered algorithms) in minimising the error of the MCMC samples (error here refers to bias and variance of the MCMC samples). Whereas limitations of approximation lead to increased bias in the posterior samples of gradient matching methods, the multiphase approach presented here negates this bias by in fact sampling using the exact likelihood function. The standard numerical integration approaches are limited by time constraints (population MCMC) and multimodality (DRAM) and so increased variance of our samples is introduced by an inability to take a sufficient number of samples from the posterior distribution. However, the gradient matching burn-in drives us towards the true distribution with no expense of increased bias or variance.

The proposed method is limited by a reliance on being able to find a good interpolant to the data. This is particularly problematic in the case of Lotka-Volterra for reasons that were discussed in Section 4.1.2 as choice of lengthscale hyperparameter is influenced by a trade off between accounting for the amplitude of the signal and providing a sufficiently

smooth representation at the peaks and troughs. Unfortunately, the fixed interpolant had a tendency to overfit the noisy observations in order to fully account for the amplitude of the signal. Resultantly, we had to take the ad hoc approach of sampling multiple interpolants and selecting the choice of interpolant based on the average true likelihood value of samples from the surrogate space. This has proven to work well for the case of Lotka-Volterra as we observe in Table 6.1 where we indicate convergence in the corrective phase with the initial ML interpolant in the first row (60% convergence) and the convergence after selection of interpolant in row 2 (100% convergence). It is important to note that this method may be less successful in models where we have a greater number of variables in the system. For instance, in the signal transduction cascade equations, 10 interpolants for each variable would require 10^5 different interpolant evaluations which may not be feasible (although, on a computing cluster, this would be equivalent to 1000 evaluations if we have 100 parallel processes). It would be more appealing and generalisable to find a method of sampling the interpolants at each step in the MCMC at acceptable computational cost, but it remains to be seen how we would choose to pick one surrogate space over another given that our choice is not necessarily motivated by minimisation of a surrogate likelihood but instead by alignment of the surrogate likelihood with the true likelihood. In other words, we wish to have good alignment of the minimum values. Consideration of this argument and the unimodality of the gradient matching likelihood surface, it makes sense that we select based on the mean of our samples from the surrogate space rather than wasting computational resources on estimating the true likelihood of all samples from each interpolant pair. Finally, the major problem with the method proposed here is the requirement of initial conditions. Without these, we are not able to assess the true likelihood for each individual surrogate space. Table 6.1 gives the effect of this added interpolant choosing phase on the ability to achieve the target PSRF value of 1.05. In Figure 5.1, we observed the tendency for DRAM to fail to converge when sampling parameter values for the Lotka-Volterra showing the vulnerability of the DRAM algorithm to the local optima that the surrogate burn-in allows us to avoid. This multimodality influenced the inclusion of population MCMC in the comparison where despite consistent accuracy across different models, the substantial computational cost makes this method less applicable in more complex differential equation systems.

Comparing performance of the different algorithms across the four models, we have observed the consistent good performance of the proposed scheme compared with the other methods. A summary of the accuracy and computational efficiency is provided by the plots in Figure 6.2, RMS in parameter space is plotted against the average number of numerical integrations performed. It was mentioned previously that the aim with ODE

TABLE 6.1: Indicating convergence in the corrective phase of the three phase proposed scheme on the Lotka-Volterra model. The first row corresponds to convergence after surrogate burn-in in the Maximum Likelihood surrogate space. The second row corresponds to convergence once we have performed a surrogate burn-in in the surrogate phase chosen by assessing the true likelihood value of the surrogate space.

Dataset	1	2	3	4	5	6	7	8	9	10
Achieve PSRF=1.05	X	✓	✓	X	✓	X	✓	✓	✓	X
Achieve PSRF 1.05 after interpolant selection	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

parameter inference is to accurately and efficiently infer the parameters of the systems, the simulations performed show the similar accuracy of the population MCMC and the proposed three phase scheme with a vast improvement in computational efficiency when we consider the multi-phase scheme. Additionally, use of population MCMC makes it difficult to include an adaptive component in the sampler and so it becomes more difficult to achieve convergence, especially in cases like the Signal Transduction Cascade where the optimum of the likelihood is very wide. This is even more of an issue in the signal transduction cascade, for which Figure 5.19 shows the definite improvement of the proposed scheme compared with population MCMC. The FitzHugh-Nagumo model proved to be a relatively easy learning task with our system configuration and level of noise, mainly due to a lack of local optima on the likelihood surface and the subsequent ease of convergence to the stationary distribution. Considering the notoriously difficult problem of parameter inference for the Goodwin Oscillator model (likelihood surface shown in Figure 5.5) is known to be a notoriously difficult learning problem due to the large number of local optima present on the surface of the likelihood. However, this proved to not be much of a challenge for the proposed scheme and population MCMC. Considering the performance in function space, we give a summary by way of a plot of functional RMS versus the number of numerical integrations required.

Considering the indication of convergence from Table 5.2 in tandem with the accuracy of the methods, it is interesting to observe the high level of accuracy of methods such as population MCMC and DAMH with surrogate burn-in despite the occasional inability to converge to the stationary distribution. In the case of population MCMC, this is best witnessed by performance for the Lotka-Volterra model where we saw lack of convergence (Table 5.2 and Figure 5.1). This may be due to the default settings of the algorithm parameters (as mentioned previously, requirement of extensive tuning is one of the drawbacks of the algorithm). This property of DAMH is best witnessed in the FitzHugh-Nagumo model by considering Table 5.2 and Figure 5.11. In Figure 6.3, I present the problem faced using DAMH for inference in the FitzHugh-Nagumo model. A

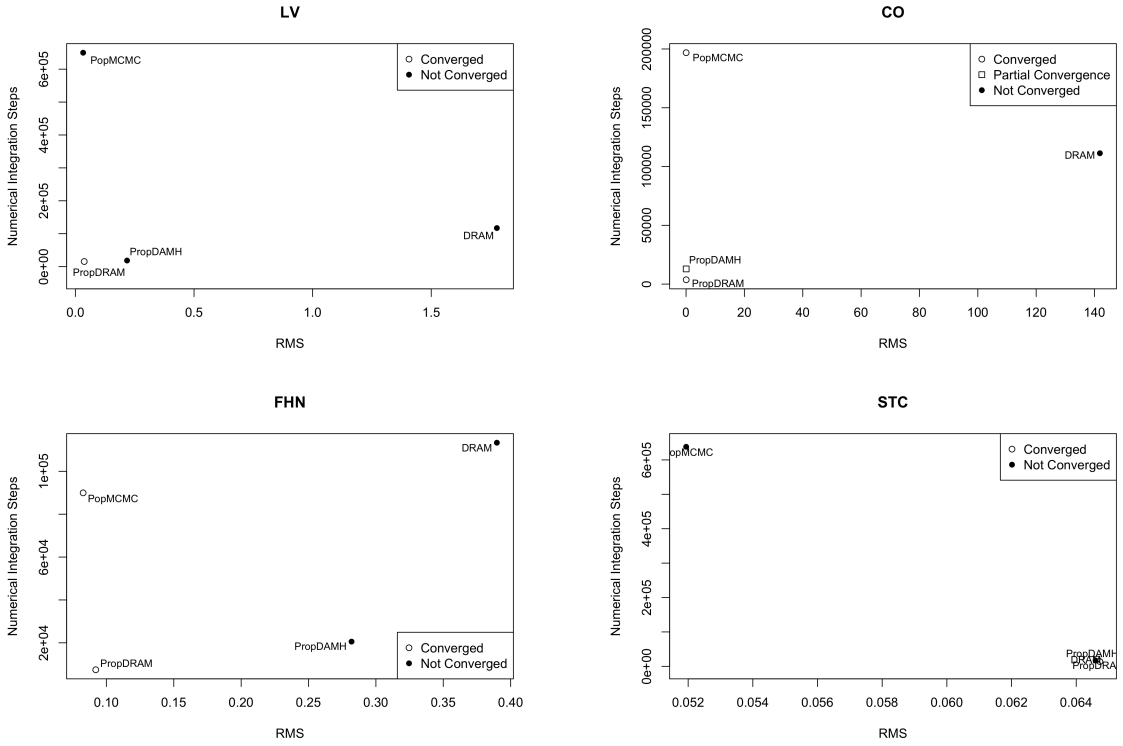

FIGURE 6.1: Plots of RMS value versus numerical integrations for each of the methods across each of the different ODE models. Good performance would be signified by a method appearing in the bottom left corner of a plot (with the exception of the STC model). The three-phase proposed scheme is the only method that appears in the bottom left hand corner of each plot.

very low acceptance rate is induced by the lack of similarity between the two likelihood surfaces (comparing the surrogate likelihood surface on the bottom of Figure 6.3 and the true likelihood surface on the top), forcing the sampler to take small steps in parameter space instead of taking the more direct route that is shown by the corrective phase of the three phase proposed scheme in the contour plot on the right of Figure 6.3. Contemplating the direct implementation of the DAMH method (without surrogate burn-in), it appears unlikely that, assuming a multimodal likelihood surface, the DAMH method would converge to the true stationary distribution in any reasonable amount of time. DAMH is used to reduce computation time by reducing numerical complexity of the MCMC steps, not by smoothing the surface. Ultimately, movement between points is still dependent on evaluation of the exact likelihood function and so local optima still pose a problem.

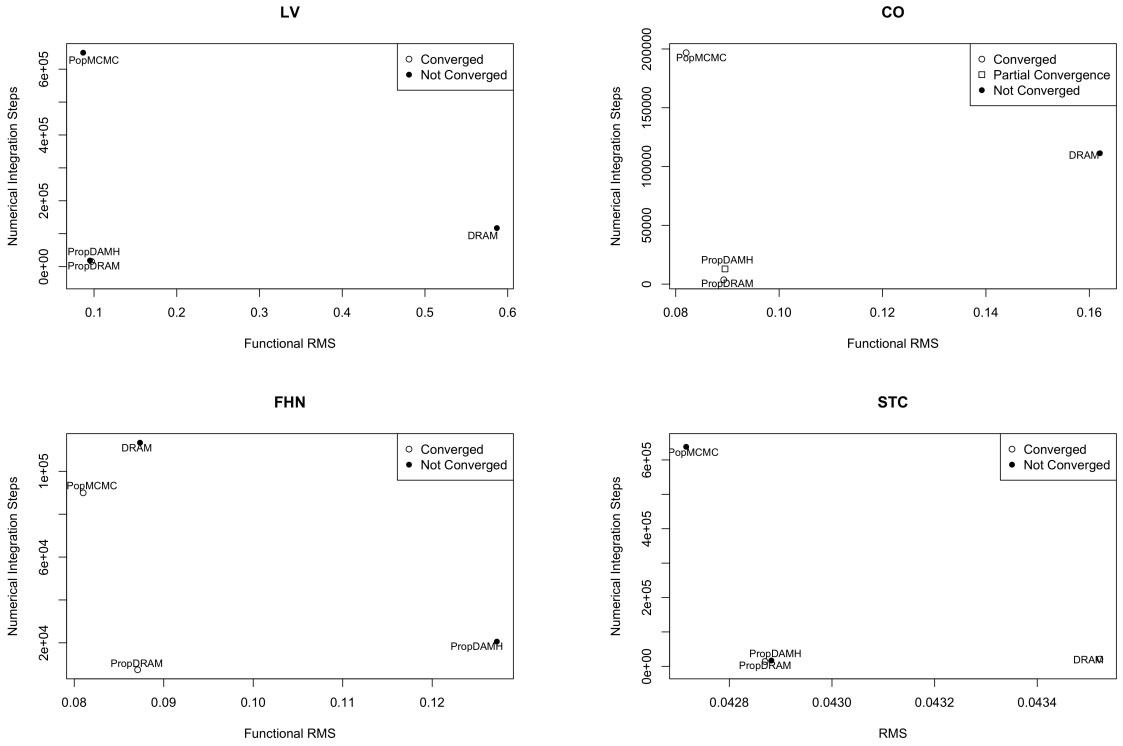

FIGURE 6.2: Plots of Functional RMS value versus numerical integrations for each of the methods across each of the different ODE models. Good performance would be signified by a method appearing in the bottom left corner of a plot (with the exception of the STC model). The three-phase proposed scheme is the only method that appears in the bottom left hand corner of each plot.

6.1.2 Further Discussion of Comparisons

Considering our choice of setup in the case of unknown initial conditions, it made sense to adopt a priori ignorance with regards to the initial conditions, allowing us to assess the methods in the most pessimistic scenario having already been assessed in the more optimistic case. In reality, we would often have a fairly good idea of the initial conditions and this level of knowledge may often be paired with a noisy observation that may be used as our starting point for the MCMC. The ignorance is portrayed by two different prior distributions with different levels of vagueness. In the case of the more informative prior, inference proved more successful than expected for the traditional DRAM sampling method. In Figure 5.20 we observed the bias in the posterior samples using each of the four methods, giving evidence of the ability of the three-phase proposed scheme to correctly infer the initial conditions and ODE parameters as a result of the surrogate burn-in. We of course had to have a short pre-corrective phase where the ODE parameters were fixed at the surrogate sample and the initial conditions were inferred using the

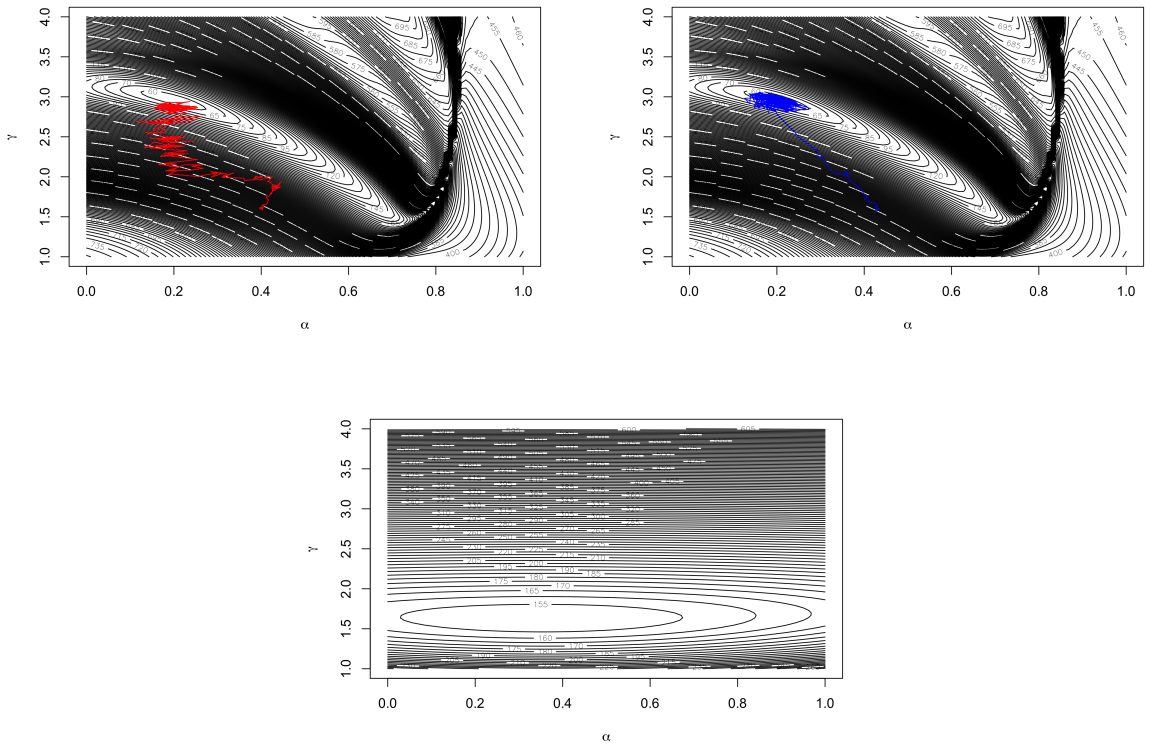

FIGURE 6.3: Top: Presenting the problem encountered in the DAMH algorithm with surrogate likelihood given by the gradient matching function for the FitzHugh-Nagumo model. This is caused by the lack of similarity between the expensive likelihood and the surrogate likelihood (as displayed by the surrogate likelihood plot in the bottom row) which, on the left, causes very slow movement to the optimum and lack of convergence whereas on the right, using the proposed scheme, we obtain improved convergence of the MCMC sampler.

true likelihood function. Without this short phase, there would be minimal benefit to a surrogate burn-in phase in this case. Figure 5.21 presents the results where we adopt the more uninformative prior for the initial conditions. Positively, the results are fairly similar to the more informative case, showing some level of robustness to the prior distribution. It was previously observed that in both Figure 5.21 and Figure 5.20 there is definite evidence of an underestimation of the noise variance from the observations which would suggest some overfitting in the inference procedure.

The use of the alternative metric proved to be successful in terms of accuracy. However, due to the difficulty in constraining the mismatch parameter to values that were able to prevent singularities or computationally negative determinants in the matrices involved. We also have too much of a dependence on the nature of the prior distribution which, of course, is not desirable in the case of MCMC sampling. Although this method was shown to be equally accurate in the case of the Goodwin Oscillator, this equal accuracy is

matched by an increase in computational complexity due to the requirement for inversion of matrices at each stage of the algorithm.

In the simulations used for this thesis, an upper bound was placed on the number of MCMC steps that were attempted by each of the different samplers. This then means that some methods are allocated extra computational overhead per MCMC step than others. If these simulations were to be repeated, it would make sense to adapt the stopping rule for the different algorithms. Instead of capping the number of MCMC steps attempted, the number of numerical integrations of the ODEs could be capped to assess performance of the algorithms when requiring equal levels of computational resources.

6.2 Assessing Robustness of Inference Methods

ODEs are very useful in the modelling of real world data. Supplied with noisy observations (often including noisy initial values) that stem from some unknown generative model, we must attempt to fit an idealistic model that does not necessarily represent a perfect fit to the underlying latent signal due to the assumptions that it relies on. In the absence of a ground truth, we cannot quantify inference performance based on a distance measure in parameter space but instead must rely on a measure of performance in function space as a (potentially suboptimal) proxy. Nonetheless, if one implements multiple parallel MCMC samplers then we would hope to be able to diagnose (lack of) convergence to the correct posterior distribution. In this section, I consider two different non-ideal model fitting paradigms, assessing performance of the proposed multiphase parameter method compared with population MCMC.

Consider the Hudson Bay dataset from the introductory section. Providing data on lynx and hare populations over a period of 20 years, the suitable model to fit is the Lotka-Volterra ODE system, outlined in Section 3.1. Using the multiphase proposed scheme, the resultant MCMC chains are able to converge within the allocated timeframe. Population MCMC failed to converge in this case, tending to sample unrealistically high noise variance parameters, flattening all parallel likelihood surfaces. Given the small number of data points, the sampler then has freedom to sample independently of the fit of the signal. Taking samples from the MCMC obtained posterior distribution provides the distribution of signals portrayed by the grey region in Figure 6.4. The multiphase scheme is able to outperform population MCMC due to the tendency for popMCMC to sample a large noise variance, flattening the likelihood surface across the different temperatures and providing

liberated movement of the MCMC sampler. The improved performance of the proposed scheme is emphasised on the right of Figure 6.7.

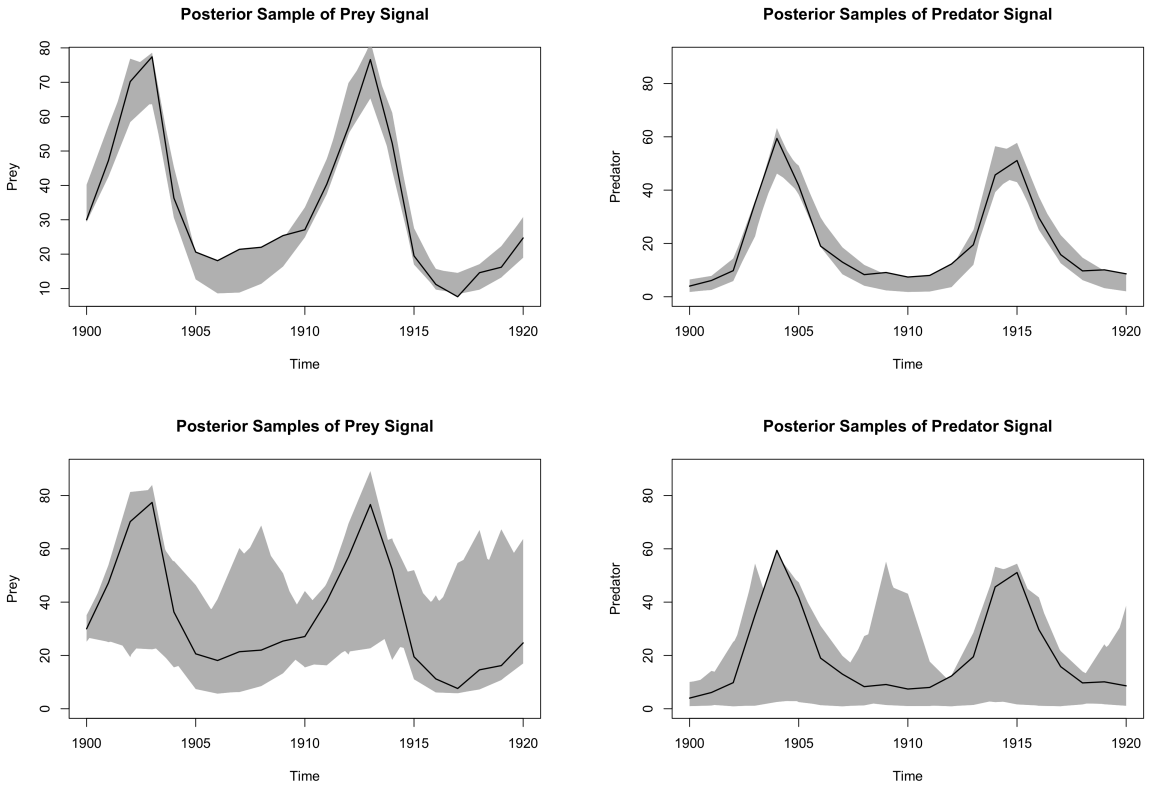

FIGURE 6.4: Displaying the distribution of signals from MCMC samples. The grey region is obtained using the signals evaluated at each of the different posterior parameter samples. The top row uses samples from the multiphase proposed scheme and the bottom results from the implementation of population MCMC.

Often in real life scenarios, we have data that do not perfectly fit the model that we are attempting to fit. Emulation of this situation by introduction of a mismatch between the model used for data generation and the model used for inference enables further assessment of the robustness of inference methods. The between-prey competition Lotka-Volterra ODE system is given in eq. 6.2 where the between prey competition parameter causes a constant decrease in the population over time and a change in the period of the data signal (Figure 6.5):

$$\frac{dx}{dt} = \alpha x - \beta xy + \epsilon x^2 \tag{6.1}$$

$$\frac{dy}{dt} = -\gamma y + \delta xy. \tag{6.2}$$

I use parameter configuration $\alpha = 0.76, \beta = 0.5, \epsilon = 0.01, \gamma = 0.4$ and $\delta = 0.3$, obtaining observations over timeframe between 0 and 50 with added Gaussian noise with variance

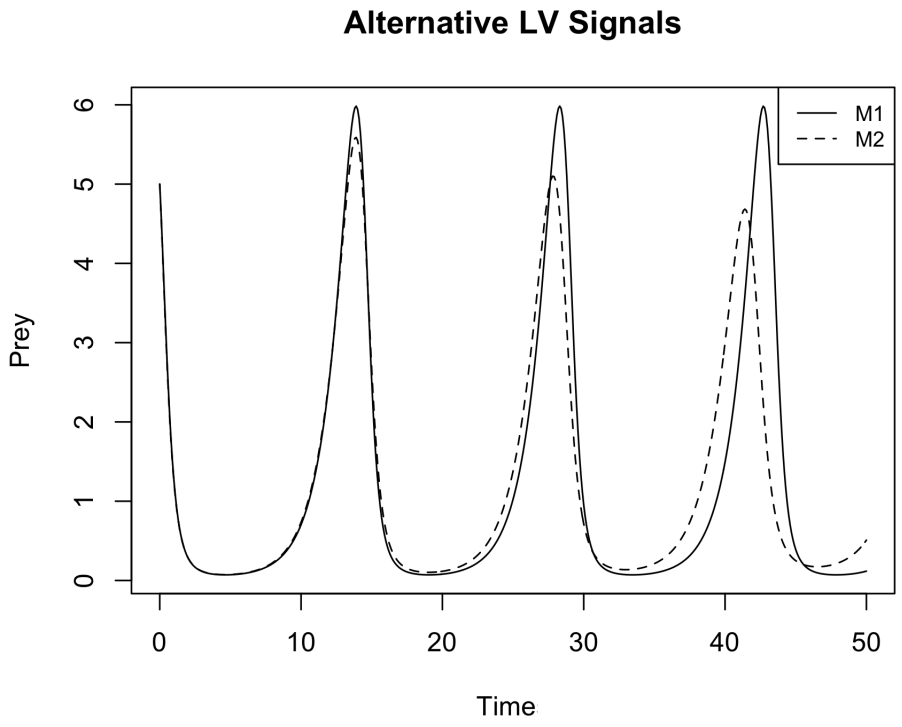

FIGURE 6.5: Prey signals produced using the LV model with between prey competition (M2) and the standard Lotka-Volterra model (M1). The addition of a prey competition term introduces constant decay to the prey signal and a gradual change in the period of the signal.

0.5. Figure 6.6 presents samples from the functional posterior distribution of the ODEs, obtained by evaluation of the ODEs at samples from the posterior distribution of the parameters. This time, both methods (PropDRAM and popMCMC) show good performance in function space as emphasised by functional RMS measurements given on the left of Figure 6.7.

6.2.1 Comments on Proposed Scheme

The multiphase approach has proven successful in estimating parameter for ODEs, but there are some implementation properties that require mentioning. Firstly, it is important to obtain an estimate of the observation noise while obtaining the smoothed interpolant. This must be selected as the starting point of the noise variance parameter in the corrective phase. Otherwise, the tendency is for the sampler to make the noise variance as high as possible, allowing flexibility in the sampling of the ODE parameters undoing all the work that has been done in the surrogate distribution to drive the sampler towards the

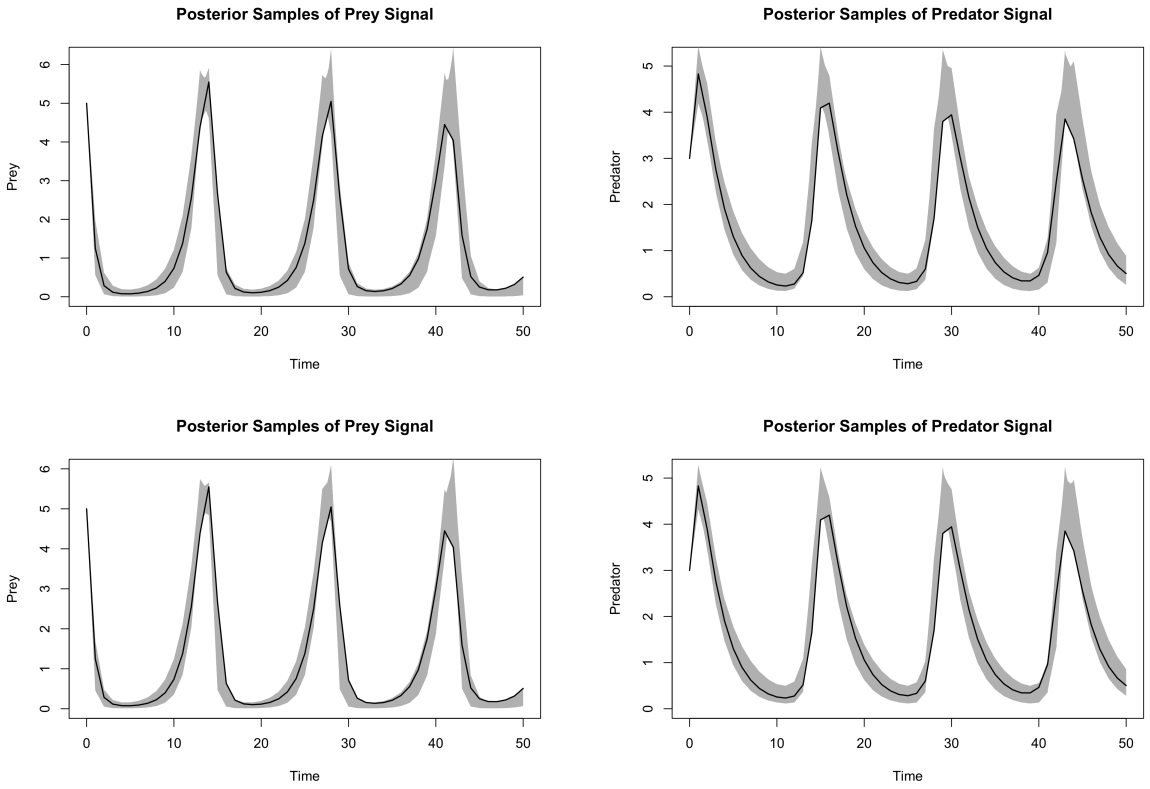

FIGURE 6.6: Posterior samples in function space for the incorrect generative model example. The top row gives the samples obtained using the proposed multi phase scheme and the bottom samples gives those corresponding to population MCMC.

correct stationary distribution. On a similar note, it can be beneficial to include a short pre-corrective phase (as outlined in Section 4.3) between the surrogate burn-in and the corrective phase where we fix the ODE parameters at the final point samples using the surrogate likelihood and allow the MCMC scheme to sample from the marginal distribution of the noise variance parameter. This allows correction for any inaccuracy in the estimated noise variance parameter which, when poorly evaluated, can lead to stability problems in the MCMC scheme.

6.2.2 Possible Extensions of the Method

It appears that multiphase sampling with gradient matching could open new avenues in the field of ODE parameter inference. The work presented has been useful in showing the ability of multiphase schemes to ameliorate convergence to the desired stationary distribution by allowing us to ignore the level of bias induced by sampling from cheap, alternative likelihoods. Provided that there is sufficient similarity between the surrogate and true likelihoods, we can greatly reduce the number of numerically expensive samples

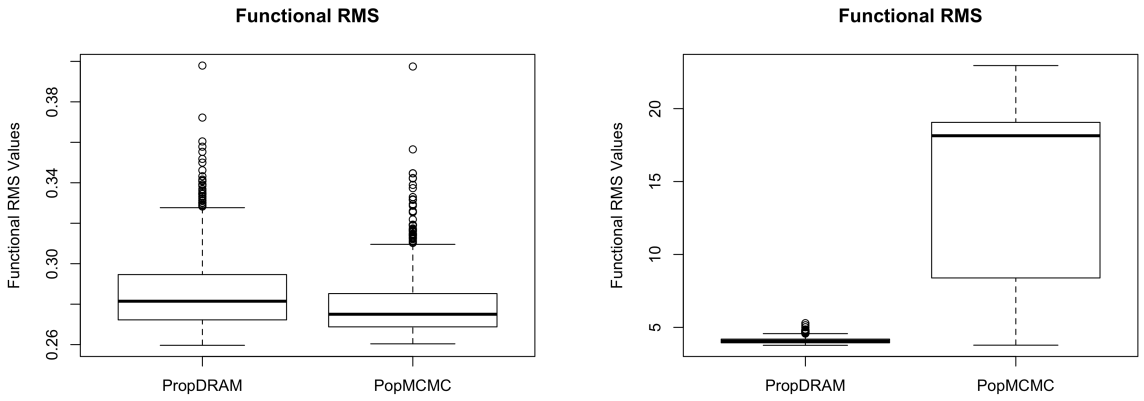

FIGURE 6.7: Comparing the functional RMS values across different samples obtained using both the proposed method and population MCMC for the incorrect model example (left) and the real data example (right). The proposed scheme achieves accurate inference in both cases whereas population MCMC struggles in the real data example.

required for convergence. There are, of course, limitations of the algorithm as presented by the necessity of a slight, situational based solution to the Lotka-Volterra inference problem—a solution that would not be applicable in cases involving noisy initial conditions. This difficulty of obtaining sufficient interpolants was specific to the case of periodic data where it becomes difficult to find suitable lengthscale hyperparameter values for the kernel function. In some sense, the fact that the method is reliant on finding better interpolants for the data is a more optimistic problem to have, given the universal use of smoothing methods across the different areas of statistics.

The most immediate extension would be to find better methods of smoothing the noisy observations. Recently, the idea of warped GPs has been considered in the literature, where we alter the time domain in order to make the observations more consistent with the Gaussian Process kernel [59] (this idea was touched on when outlining the Periodic kernel in Section 4.1.2). Another solution from the literature could be obtained through the use of a mixture of Gaussian processes [60] where different lengthscale parameters may be assumed at different regions of input space, allowing incorporation of the varying properties of the data in our smoothed curve. This may allow us to incorporate the smoothness required at peaks in the Lotka-Volterra signal, along with the inclusion of the amplitude of the signal. Taking a completely different approach to the problem, one could instead choose to adopt a partially observed system in the surrogate burn-in phase, choosing to match observations and numerical solutions at a carefully selected sample of time points. However, this may not generalise well outwith the realm of simple toy problems.

Considering improvement of the corrective and sampling phases, it is possible that the implementation of more sophisticated, state-of-the-art, proposal mechanisms such as Hamiltonian Monte Carlo which implements Hamiltonian dynamics in order to make better informed proposals (see chapter 5 of [61]) or Langevin diffusion processes [62] in the corrective and sampling phases of our multiphase scheme. These do of course come with the added necessity to carefully tune the algorithm in addition to the extra computation cost will increase as a result of the complex nature of these algorithms. It is worth emphasising that the obtainment of better interpolants would likely lead to the bigger improvement in parameter learning performance.

It would be interesting to assess the performance of the proposed scheme in the realm of partial differential equations, where numerical integration techniques require even more computational resources than the systems in this thesis, making the traditional MCMC inference schemes even less applicable than in the ordinary differential equation case.

6.3 Conclusion

Parameter inference in nonlinear ordinary differential equations is a very onerous process. Our inability to find closed form solutions of these systems necessitates computationally inefficient sampling from posterior distributions when one wishes to adopt the most immediate likelihood resulting from consideration of the data generating distribution. There do exist alternative representations of the data that one can use to coax more implicit measures of similarity from the data and systems in question, as outlined in Section 4.2. These of course come at a cost of accuracy and so we arrive at a crossroads in the search for efficient, accurate parameter inference methods. Do we abandon the wish for computational efficiency and instead take the obstinate approach of accurate, time consuming parameter inference with the numerical integration likelihood or do we take the more quantitatively inclined route and adopt cheap likelihood functions with bias introduced to the parameter samples? The multiphase approach in this thesis successfully combines the two. By taking a cheap proxy for the likelihood to sample our discarded burn-in steps, we introduce vastly improved efficiency to the learning problem by introducing bias and efficiency in the portion of the chain where the former can be deemed fairly irrelevant and the latter can prove to be at its more troublesome (owing to stiffness and local optima). This efficient sampling enables more representative computationally expensive sampling steps, for which bias becomes a prime concern. The method has been shown to perform equally to population MCMC on four benchmark ODE systems while being able

to outperform the computationally expensive DRAM method. Despite the equality in accuracy, population MCMC is prone to the problems caused by the nature of the ODE likelihood function where stiffness and computational efficiency mean that large samples require immense computation time. Resultantly, for the models considered in this thesis, method comparisons show the improvements obtained—in terms of both accuracy and computational efficiency—when one adopts a multiphase sampling routine with gradient matching burn-in for ODE parameter inference.

Appendix A

Statistical and Mathematical Identities

A.1 Gaussian Identities

Let $\mathbf{x}$ and $\mathbf{y}$ be multivariate Normal random variables such that the mean of $\mathbf{y}$ depends linearly on random variable $\mathbf{x}$, then the product of these two densities is proportional to a multivariate Gaussian density:

$$\mathcal{N}(\mathbf{x}|\mathbf{a}, \mathbf{A})\mathcal{N}(\mathbf{y}|\mathbf{C}\mathbf{x}, \mathbf{B}) \propto \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

where

$$\boldsymbol{\mu} = \boldsymbol{\Sigma}(\mathbf{A}^{-1}\mathbf{a} + \mathbf{C}^T\mathbf{B}^{-1}\mathbf{y}) \text{ and } \boldsymbol{\Sigma} = (\mathbf{A}^{-1} + \mathbf{C}^T\mathbf{B}^{-1}\mathbf{C})^{-1}$$

A.2 Matrix Identities

Woodbury matrix identity:

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1} \tag{A.1}$$

Bibliography

[1] B. Calderhead, M. A. Girolami, and N. D. Lawrence. ODE parameter inference using adaptive gradient matching with Gaussian processes. 2008.

[2] stan-dev/example-models. `https://github.com/stan-dev/example-models/tree/master/knitr/lotka-volterra`. Accessed: 2018-01-11.

[3] A. J. Lotka. The growth of mixed populations: Two species competing for a common food supply. *Journal of the Washington Academy of Sciences*, 22(16):110–120, 1932.

[4] M. Girolami. Bayesian inference for differential equations. *Theoretical Computer Science*, 408(1):4–16, November 2008.

[5] S. Hug, D. Schmidl, W. B. Li, M. B. Greiter, and F. J. Theis. Bayesian Model Selection Methods and Their Application to Biological ODE Systems. In *Uncertainty in Biology: A Computational Modeling Approach*, chapter 10, pages 243–268. Springer International Publishing, Cham, 2016.

[6] J. O. Ramsay, G. Hooker, D. Campbell, and J. Cao. Parameter estimation for differential equations: a generalized smoothing approach. *Royal Statistical Society Series B*, 69(5):247–260, October 2007.

[7] F. Dondelinger, D. Husmeier, S. Rogers, and M. Filippone. ODE parameter inference using adaptive gradient matching with Gaussian processes. In *Proceedings of Machine Learning Research*, volume 31, pages 216–228. 2013.

[8] Mu Niu, Simon Rogers, Maurizio Filippone, and Dirk Husmeier. Fast parameter inference in nonlinear dynamical systems using iterative gradient matching. In Maria Florina Balcan and Kilian Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1699–1707, New York, New York, USA, 20–22 Jun 2016. PMLR.

[9] J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. Wiley, New York, N.Y., 1994.

[10] C. P. Robert. *The Bayesian Choice*. Springer, New York, 2 edition, 2007.

[11] An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 186(1007):453–461, 1946.

[12] K. P. Murphy. *Machine Learning A Probabilistic Perspective*. The MIT Press, Cambridge MA, 2012.

[13] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Ranton, FL, 3 edition, 2013.

[14] J. C. Robinson. *An Introduction to Ordinary Differential Equations*. Cambridge University Press, Cambridge UK, 2004.

[15] J. C. Butcher. A history of Runge-Kutta methods. *Applied Numerical Mathematics*, 20(3):247–260, March 1996.

[16] H. Xue, H. Miao, and H. Wu. Sieve estimation of constant and time-varying coefficient in nonlinear ordinary differential equation models by considering both numerical error and measurement error. *The Annals of Statistics*, 38(4):2351–2387, October 2010.

[17] C. Sherlock, A. Golightly, and D. A. Henderson. Adaptive, delayed-acceptance MCMC for targets with expensive likelihoods. *Journal of Computational and Graphical Statistics*, 26(2):434–444, 2017.

[18] I. Dattner and C. A. J. Klaassen. Optimal rate of direct estimators in systems of ordinary differential equations linear in functions of the parameters. *Electronic Journal of Statistics*, 9(2):1939–1973, 2010.

[19] S. Ranciati, C. Viroli, and E. Wit. Bayesian Smooth-and-Match strategy for ordinary differential equations models that are linear in the parameters. *ArXiv e-prints*, April 2016.

[20] Y. Wang and D. Barber. Gaussian processes for Bayesian estimation in ordinary differential equations. *Journal of Machine Learning Research*, 32:1485–1493, 2014.

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

[22] E. T. Jaynes. *Probability Theory the Logic of Science*. Cambridge University Press, Cambridge UK, 2003.

[23] J. Berger. The case for objective Bayesian analysis. *Bayesian Analysis*, 1(3):385–492, 2006.

[24] G. O. Roberts and J. S. Rosenthal. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18:349–367, 2009.

[25] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, and A. H. Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21 (6):339–354, 1953.

[26] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, April 1970.

[27] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer, New York, 2004.

[28] S. Chib and E. Greenberg. Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49(4):327–335, November 1995.

[29] L. J. Billera and P. Diaconis. A geometric interpretation of the Metropolis-Hastings algorithm. *Statistical Science*, 16(4):335–339, November 2001.

[30] A. Gelman, W. R. Gilks, and G. O. Roberts. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1): 110–120, 1997.

[31] Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive Metropolis algorithm. 7, 04 2001.

[32] A. G. Gelman, G. O. Roberts, and W. R. Gilks. Efficient Metropolis jumping rules. In J. M. Bernardo, J. O. Berger, A. F. David, and A. F. M. Smith, editors, *Bayesian Statistics V*, pages 599–608. Oxford University Press, Oxford, 1996.

[33] H. Haario, M. Laine, A. Mira, and E. Saksman. Dram: Efficient adaptive MCMC. *Statistics and Computing*, 16(4):339–354, 2006.

[34] M. Laine. *Adaptive MCMC Methods With Applications in Environmental and Geophysical Models*. PhD thesis, Helsinki, Finland, 2008.

[35] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[36] D. Campbell and R. J. Steele. Smooth functional tempering for nonlinear differential equation models. *Statistics and Computing*, 22(2):429–443, November 2012.

[37] K. B. Laskey and J. W. Myers. Population Markov chain Monte Carlo. *Machine Learning*, 50:175–196, January 2003.

[38] N. Friel and A. N. Pettitt. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3): 589–607, 2008.

[39] M. K. Cowles and B. P. Carlin. Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434): 883–904, June 1996.

[40] S. P. Brooks and G. O. Roberts. Convergence assessment techniques for Markov chain Monte Carlo. *Statistics and Computing*, 8:319–335, 1998.

[41] J. Gentle. *Random Number Generation and Monte Carlo Methods*. Springer, 2005.

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

[43] A. Raue, C. Kreutz, F. J. Theis, and J. Timmer. Joining forces of bayesian and frequentist methodology: a study for inference in the presence of non-identifiability. *Philisophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984), February 2013.

[44] Richard FitzHugh. Impulses and physiological states in theoretical models of nerve membrane. *Biophysical Journal*, 1, 1961.

[45] J S. Nagumo, S Arimoto, and S Yoshizawa. An active pulse transmission line simulating a nerve axon. In *Proceedings of the IRE*, 1962.

[46] S. Gktepe and E. Kuhl. Computational modeling of cardiac electrophysiology: A novel finite element approach. 79:156 – 178, 07 2009.

[47] B. Bruggemeier, C. Schusterreiter, H. J Pavlou, and X. Cai. Improving the utility of drosophila melanogaster for neurodegenerative disease research by modelling courtship behaviour patterns. 11 2014.

[48] B. C. Goodwin. Oscillatory behavior in enzymatic control processes. *Advances in Enzyme Regulation*, 3:425–438, November 1965.

[49] A. Woller, D. Gonze, and T. Erneux. Strong feedback limit of the goodwin circadian oscillator. *Physical Review*, 87, March 2013.

[50] M. Girolami, B. Calderhead, and V. Vyshemirsky. System identification and model ranking: The bayesian perspective. In N. D. Lawrence, M. Girolami, M. Rattray, and G. Sanguinetti, editors, *Learning and Inference in Computational Systems Biology*, chapter 8, pages 201–230. MIT Press, Cambridge, MA, 2010.

[51] C. E. Rasmussen and K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA., 2006.

[52] D. J. C. MacKay. Introduction to Gaussian processes. In C.M. Bishop, editor, *Neural Networks and Machine Learning*. Springer-Verlag, 1998.

[53] C. K. I. WIlliams. Computation with infinite neural networks. *Neural Computation*, 10(5):1203–1216, July 1998.

[54] J. Swartz and H. Bremermann. Discussion of parameter estimation in biological modelling: Algorithms for estimation and evaluation of the estimates. *Journal of Mathematical Biology*, 1(3):247–260, September 1975.

[55] J. M. Varah. A spline least squares method for numerical parameter estimation in differential equations. *SIAM Journal on Scientific and Statistical Computing*, 3(1): 28–46, 1982.

[56] B. Macdonald, C. Higham, and D. Husmeier. Controversy in mechanistic modelling with Gaussian processes. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1539–1547, Lille, France, 07–09 Jul 2015. PMLR. URL `http://proceedings.mlr.press/v37/macdonald15.html`.

[57] T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman and Hall, 1990.

[58] Anoop Korattikara, Yutian Chen, and Max Welling. Austerity in MCMC land: Cutting the Metropolis-Hastings budget. 1, 04 2013.

[59] Mu Niu, Benn Macdonald, Simon Rogers, Maurizio Filippone, and Dirk Husmeier. Statistical inference in mechanistic models: time warping for improved gradient matching. *Computational Statistics*, 2017. ISSN 1613-9658. doi: 10.1007/s00180-017-0753-z. URL `https://doi.org/10.1007/s00180-017-0753-z`.

[60] V. Tresp. Mixtures of Gaussian processes. In *NIPS*, 2000.

[61] S. Brooks, A. Gelman, G. L. Jones, and X.-L. Meng. *Handbook of Markov Chain Monte Carlo*. CRC Press, Boca Raton, FL, 2011.

[62] T. Xifara, C. Sherlock, S. Livingstone, S. Byrne, and M. Girolami. Langevin diffusions and the Metropolis-adjusted langevin algorithm. *Statistics & Probability Letters*, 91:14–19, 2014.