# A Study of Population MCMC for estimating Bayes Factors over Nonlinear ODE Models

Thesis submitted in accordance with the requirements of
the University of Glasgow for the degree of Master of Science
by

Ben Calderhead

October 2007

# Abstract

Higher resolution biological data is now becoming available in ever greater quantities, allowing the complex behaviour of fundamental biological processes to be studied in much more detail. The area of Systems Biology is in desperate need of methods for inferring the most likely topology of the underlying genetic networks from this oftentimes noisy and poorly sampled data, to support the construction and testing of new model hypotheses. Towards that end, Bayesian methodology provides an ideal framework for tackling such challenges, and in particular offers a means of objectively comparing competing plausible models through the estimation of Bayes factors.

There are, however, formidable obstacles which must be overcome to allow model inference using Bayes factors to be of practical use. Many important biological processes may be most accurately represented using nonlinear models based on systems of ordinary differential equations (ODEs), however parameter inference over these models often produces correspondingly nonlinear posterior distributions, which are very challenging to sample from, often resulting in biased marginal likelihood estimates with large variances. Such problems are commonly encountered when modelling circardian rhythms, which exhibit highly nonlinear oscillatory dynamics and play a central role in the overall functioning of most organisms. In this thesis I investigate tools for calculating Bayes factors to distinguish between ODE-based Goodwin oscillator models of varying complexity, which form the basic building blocks for describing this ubiquitous circadian behaviour.

The main result in Chapter 3 of this thesis demonstrates how Population Markov Chain Monte Carlo may be employed in conjunction with thermodynamic integration methods to estimate Bayes factors which may accurately distinguish between two nonlinear oscillator models of varying complexity, given noisy experimental data generated from each of the models. In addition, it is shown how alternative methods may fail drastically in this setting, in particular harmonic mean based estimates. Suggestions are given regarding the optimal temperature schedule which should be employed for Population MCMC, and several ideas for future research extending this work are also discussed.

# Contents

# List of Figures

# Acknowledgements

I would like to thank my supervisor Mark Girolami for his very useful advice and support throughout this project, and also Niall Friel for useful discussions regarding thermodynamic integration methods. I am also grateful to the Computing Science department for providing me with funding for my studies.

Finally, I would like to thank everyone in the Inference Research Group in the Computing Science department at Glasgow University for their part in proofreading this thesis and generally helping to create a fun environment in which to work.

# Chapter 1

# Introduction

A recent trend in the field of biology is the change of emphasis from studying the individual components of a biological system, to studying the system as a whole and examining how the *interactions* between individual components bring about an observed phenomenon ([74]). When looked at from this holistic point of view, determining the underlying network of interactions of a system becomes a crucial task, for which new tools have to be developed. Such tools must be able to accurately evaluate and compare plausible hypotheses regarding the structure of a system, as this is essential for driving towards a more complete understanding of the core biological mechanisms at work.

Mathematical models based on systems of ODEs (ordinary differential equations) can be considered codifications of these underlying network topologies and associated dynamics, and they provide surprisingly accurate mechanistic representations of biological systems (See e.g. [27]). There are however many difficulties associated with modelling biological networks, particularly when investigating nonlinear systems such as those used to describe the very important circadian control processes (Section 1.1). Circadian rhythms play a central role in the function of most organisms and will be focussed on in this thesis. The problem of defining how well a model describes inherently stochastic and possibly incomplete observations of biological systems may be dealt with in a consistent manner by employing the Bayesian framework (Section 1.2). A particularly appealing aspect of this approach is the possibility of calculating Bayes factors ([35]), which provide an objective method of comparing model hypotheses. Monte Carlo methods (see e.g. [70]) (Section 1.3) are often used to sample from the resulting non-analytic posterior distributions, and this thesis is concerned with examining how best to calculate Bayes factors over nonlinear systems using such techniques.

## 1.1 Modelling Biological Processes

The rate at which biological genetic data can be produced is rapidly increasing due to technological advances in high-throughput experimental methods. An important task is to translate this plethora of available genetic data into knowledge regarding the structure of the underlying biochemical networks (see e.g. [78]).

One approach is to use mathematical models to shed light on the underlying design principles of biological processes, which can greatly aid our understanding of the relationship between the structure and function of complex systems ([66]). A deterministic mechanistic mathematical model is a set of ordinary differential equations[1], the outputs of which may be interpreted as corresponding directly to the levels of the various chemical species present in the biochemical system being modelled. Although current technology allows great quantities of certain types of data to be collected, measurements at the cell level are inherently stochastic and most kinetic rate constants still cannot be measured directly for the majority of biological systems under investigation ([38]). Given a mechanistic model it is therefore necessary to find a set of parameters with which the model can reproduce the observed behaviour. The increase in the amount of biochemical data becoming available is making it possible to consider the feasibility of estimating parameters for such models at a systems level using optimisation based algorithms, with large groups of parameters being estimated simultaneously. Accurately estimating parameter values for a *nonlinear* mathematical model can be greatly challenging, however, as there are often multiple parameter sets offering equally plausible solutions.

The problem of parameter estimation has been tackled in the past using various approaches, for example estimating the parameters for a model individually or using linear approximations of the observed behaviour (for an overview see [3]), however for more complex models it is known that the dynamics of individual components or linear approximations do not necessarily match the dynamics of a nonlinear system as a whole ([46]). Using a systems approach, all the parameters are estimated together in an attempt to capture all the possible types of behaviour produced by a particular system, in which the complex interactions and interdependencies produce a result which is more than just the sum of their parts. It is of vital importance that the method employed accurately identifies all of the most likely parameter sets, to be sure that any deficiencies a model has in describing the experimental data are due to the chosen structure of the model and not just a suboptimal choice of parameters. Once we are able to sample

[1]Stochastic differential equations also fall into the category of mechanistic mathematic models, however we focus here purely on ODEs since the observed behaviour being modelled is averaged over populations of cells.

from this optimal distribution of parameter values, it opens the door to being able to feasibly compare models in a more objective manner using the Bayesian framework, which shall be described later in this chapter. In the next section I describe circadian networks in more detail, as they provide the focus throughout this thesis for developing and evaluating tools to compare competing model hypotheses.

### 1.1.1 Circadian Networks

Many important biological processes are oscillatory in nature and display highly nonlinear dynamics. Oscillatory behaviour has been observed in many different contexts in a great number of living organisms and the most easily observed behaviour of this type is without a doubt the circadian rhythm ([12]). These rhythms are due to the 24 hour cycles of light and darkness on this planet and are possibly induced by organisms trying to gain a competitive advantage by anticipating periods of change in their environment. Circadian rhythms are fundamental biological processes and their oscillatory nature is genetic in origin. They have been discovered in almost all eukaryotic, and some prokaryotic, organisms and display very similar properties ([12]). Circadian rhythms impact on numerous biological processes, ranging from transcription regulation in cyanobacteria to regulating sleep-wake cycles in humans. In the model plant *Arabidopsis Thaliana* it has been estimated from oligonucleotide array experiments ([25]) that at least 6% of the genome is under the influence of output pathways leading from circadian biochemical networks. This equates to over 1000 genes being expressed rhythmically. Over the past couple of decades, evidence has emerged showing that the oscillatory behaviour of these circadian networks is based on underlying negative feedback loops ([11]), with proteins forming autoregulatory systems whereby their production rate is linked directly to their own levels of concentration. Such information has led to the knowledge driven construction of plausible models describing this rhythmic behaviour, derived from an understanding of the physical mechanisms involved (see e.g. [73, 83]).

Scientists have long speculated about the nature of oscillatory systems in living organisms. Early on there were very few clues to help the construction of hypotheses to describe this ubiquitous type of behaviour, and therefore the aim of initial theoretical work was to characterise the observed common behaviour mathematically without necessarily linking it directly to the physical biology. As a result there were many different models proposed to describe the feedback loops which might drive an internal clock, none of which was robustly backed up by experimental findings ([12]). This was particularly true in the premolecular era from around 1950 to 1970, during which time the tools and techniques available

at the molecular level were of little use for developing theory that linked particular functions with specific molecules. Now that experimental procedures at the molecular level are feasible, the great amount of theory which has built up regarding the mathematical and physical characteristics of oscillations is of great use, indeed vital, for building a true understanding of their molecular foundation (see e.g. [1], [12]).

It is known that all oscillators require three underlying features. They need a positive input, which sets a change in motion, a feedback response, which sets up an autoregulatory ability, and a time delay, which increases the range of possible output dynamics. In biological systems there are however additional features which are of great importance. A *robustness* of system response is vital for an organism to adapt to small but constantly changing environmental factors, such as temperature and light. *Resettability* is also important to enable an organism to react to larger changes in the environment. The oscillations present in most organisms roughly correspond to the length of a day and 24 hours is a very long time when compared to typical interaction times at a molecular level ([80]), thus *stability* of oscillations over such a long time period is essential. By coupling simple oscillators, mathematical systems may be constructed which accurately reproduce experimental observations with an appropriate period length, and this increased complexity results in an increased stability within the system [39].

The challenge of elucidating the underlying molecular machinery driving circadian rhythms has been tackled using various approaches. One approach has been to exploit the knowledge that light is an important and universal input pathway to the internal clock. By trying to tie changes in light to changes in particular photoreceptors, the hope is that one can discover the relevant regulatory pathways which describe clock function. Another approach has been to isolate the regulatory pathways associated with a particular rhythmic process, such as leaf movement in plants, and follow it back to the core clock network. This has been successfully applied to the model plant Arabidopsis thaliana, indeed the protein CCA1 (Circadian Clock Associated 1) was discovered in this way and is now known to form a central part of its clock network ([82]).

Perhaps the most useful approach, however, has been the attempt to perturb circadian oscillations by mutating particular genes and examining the ensuing effect (see e.g. [10]). It is only relatively recently that this approach has been possible; in the 1960s it was simply not an option. This method provides an opportunity to gain data which can then be usefully compared to the output predicted by a model. For example, a gene could be knocked out completely and the resulting protein levels measured; if the corresponding gene-less model does not correctly predict the effect, then some part of its topology must be wrongly

specified. Similar experiments may be repeatedly performed and their output compared to the newly redefined model, thus improving the model in an iterative manner, in an attempt to link the mathematical theory to the biological reality ([67]).

### 1.1.2 The Goodwin model

The complex dynamics of oscillatory networks may be modelled by highly nonlinear dynamical systems based on the Goodwin model. The Goodwin model ([22]) is based on a negative feedback loop and has become the basic building block with which to design circadian models. The main reason for the continued study of the Goodwin model is that despite being relatively simple to construct, it can make strong predictions regarding the basic relationship between the period length of the oscillating system and the degradation of the clock protein and mRNA. The basic $n$-variable Goodwin model is as follows,

$$\begin{aligned} \frac{dx_1}{dt} &= \frac{k_1}{1+x_n^\rho} - m_1 x_1 \\ \frac{dx_2}{dt} &= k_2 x_1 - m_2 x_2 \\ &\cdots \\ \frac{dx_n}{dt} &= k_n x_{n-1} - m_n x_n \end{aligned} \tag{1.1}$$

where $x_1$ and $x_2$ correspond to the levels of mRNA and protein produced from a clock gene in the system, respectively, and $x_3$ to $x_n$ correspond to other proteins involved in the system, with $x_n$ ultimately inhibiting mRNA production. The number of variables in the system, $n$, corresponds to the time delay of the negative feedback loop, which is believed to be the common underlying design responsible for oscillatory behaviour in a large number of regulatory networks ([79]). Larger values of $n$ produce longer delays in the system, enabling a greater possible range of output dynamics. $\rho$ corresponds to the Hill coefficient, which is a measure of the cooperativity or affinity of molecules to bind (see e.g. [57]). For the model described in Equation 1.1, $\rho$ must be larger than 8 for oscillatory output to be possible. A biologically realistic model, however, should have a much smaller Hill coefficient, and so one of the aims of extending this system has been to create a model capable of similar cyclic behaviour but with a smaller value of $\rho$ ([41]).

The Goodwin model has been extended in a knowledge driven manner, taking account of known feedback loops and other such interactions between chemical species, for various organisms ([12]) including the mouse, the fruit fly *Drosophila Melanogaster* and the fungus *Neurospora crassa*. Properties such as light entrainment and temperature compensation effects have also been modelled ([73, 40]),

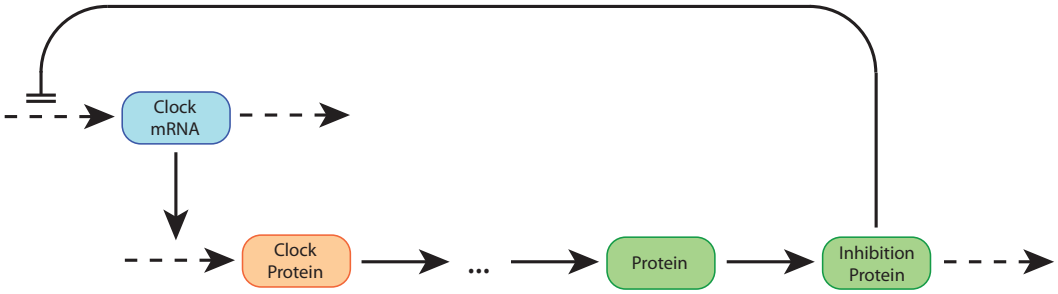


Figure 1.1: A diagram of the Goodwin model network. The arrows show which chemical species encourage production of other species, the dashed arrows show possible production routes which are not fully modelled, and the two parallel dashes at the end of a line represent an inhibition of production. The ellipsis represents other possible proteins that may play a role in the system.

however the complete set of dynamics and interactions of circadian networks are poorly understood and the circadian clock has been examined in detail only for still relatively few organisms ([12]).

## 1.2 The Bayesian Approach

Throughout the following chapters I make use of the Bayesian framework, which offers a natural way of taking uncertainty into account. It also enables us to easily incorporate prior information or beliefs about the system under study, in a principled and consistent manner, and allows us to clearly see when problems with system identifiability occur, since we calculate a posterior distribution over each parameter instead of a single point estimate of the most likely value.

Furthermore, the use of the Bayesian framework allows us to update the strength of our prior beliefs in the parameter values of the model, given the evidence of experimental observations and enables objective comparison of competing model hypotheses by calculating Bayes factors ([35]). Bayes factors compare the evidence in favour of two competing models, given a particular dataset, by considering their marginal likelihoods which, if non-analytic, may be calculated by numerically integrating out all the possible parameter values.

All the methods for estimating marginal likelihoods considered in this thesis require samples from some form of posterior distribution, which may be multimodal, as we shall see in Chapter 3. I therefore firstly investigate methods for generating samples from complex distributions, and then use these samples to compute Bayes factors, employing a variety of marginal likelihood estimation methods, the relative accuracy of which is examined in detail. Generally we shall consider a model, $H$, described by some system of differential equations, along

with an associated set of parameters $\boldsymbol{\theta}$. A series of $N$ experiments are simulated and the resulting measurements are denoted by $\mathbf{y}$.

### 1.2.1 Posterior Distribution

The posterior distribution provides us with an updated measure of our beliefs for each of the parameter values based on our prior beliefs. This distribution therefore represents the range of parameter values which most likely allows the output of a particular model to best describe the data. This can be calculated from the likelihood and prior distributions using Bayes' Theorem (see e.g. [34, 33, 15]),

$$p(\boldsymbol{\theta} \mid \mathbf{y}, H) = \frac{\overbrace{p(\mathbf{y} \mid \boldsymbol{\theta}, H)}^{Likelihood}\overbrace{p(\boldsymbol{\theta})}^{Prior}}{\underbrace{\int p(\mathbf{y} \mid \boldsymbol{\theta}, H)p(\boldsymbol{\theta})d\boldsymbol{\theta}}_{MarginalLikelihood}} \tag{1.2}$$

It can be difficult to compute the marginal likelihood in the equation above as it is usually non-analytic, other than for conjugate priors and likelihoods (see e.g. [15]). Fortunately, however, it is still possible to sample from the posterior distribution by computing only the likelihood and prior distributions, since the marginal likelihood is simply a normalising constant, which need not be explicitly calculated.

### 1.2.2 Likelihood

The likelihood is a probability distribution which accounts for the many different types of error, such as experimental variability, measurement error and the inherent stochasticity of the system under consideration. In this work, the likelihood of the experimental data given a set of parameter values is

$$p(\mathbf{y} \mid \boldsymbol{\theta}, H) = N_{\mathbf{y}}(\varphi(\boldsymbol{\theta}, H), \boldsymbol{\Lambda}) \tag{1.3}$$

where $\varphi(\boldsymbol{\theta}, H)$ is the solution of a particular system of ODEs, $N_{\mathbf{y}}$ is a normal distribution centred on $\varphi(\boldsymbol{\theta}, H)$, and the covariance $\boldsymbol{\Lambda}$ represents the covariance of the stochastic component of the system. If, for example, we assume independent and identically distributed (i.i.d.) errors[2] across all experiments, with variance $\sigma^2$, the likelihood reduces to the product over all experimental data points. In order to avoid numerical problems when dealing with the products of small probabilities, we work in log space when calculating likelihoods. The log likelihood, in this case,

[2]The errors suggested would include measurement error and model error/inadequacy.

therefore reduces to the sum of the logs of the likelihoods over all $N$ experimental data points

$$\log\left(p(\mathbf{y} \mid \boldsymbol{\theta}, H)\right) = \sum_{n=1}^{N} \log\left(N_{y_n}(\varphi_n(\boldsymbol{\theta}, H), \sigma^2)\right) \tag{1.4}$$

where $\varphi_n(\boldsymbol{\theta}, H)$ gives the model output corresponding to the $n$th data point, $y_n$. A discussion of alternative methods of characterising the stochastic components of the model is given in Chapter 4.

### 1.2.3 Prior Probabilities

A prior probability distribution is defined for each parameter and encapsulates the prior beliefs held about their most likely values. The fact that a prior must be defined for every parameter is a strength of the Bayesian method, since all previous information (or lack of information) about the parameters can be taken into account. In a systems biology context, this is important, since there is a significant amount of uncertainty regarding the hypothesised models, the actual experimental observations, and the associated parameters. In the absence of relevant experimental data, it may be possible to use information from published research to help define prior distributions, $\pi(\boldsymbol{\theta})$, over parameter values.

Generally it is useful to use gamma priors for the kinetic rate parameters of biological models, since they have positive support and may cover a wide range of possible values (See Figure 1.2). The gamma probability density function, for a single parameter $\theta$ with shape parameter $a$ and scale parameter $b$, is defined as

$$\gamma = f(\theta \mid a, b) = \frac{1}{b^a \Gamma(a)} \theta^{a-1} e^{\frac{-\theta}{b}} \tag{1.5}$$

where the mean is $ab$, the variance is $ab^2$ and $\Gamma$ is the gamma function. Priors may also be set over the systems of equations being compared to reflect the prior preference (or otherwise), $\pi(H)$, for a particular model hypothesis, $H$.

### 1.2.4 Calculating Bayes Factors

Bayes factors can be used to compute the posterior probabilities of two models, given the prior probability of each model. Given a set of data $\mathbf{y}$ and two competing model hypotheses $H_1$ and $H_2$, we wish to calculate the probability of each model hypothesis given the data. Using Bayes' theorem we obtain the following expression (for $n = 1, 2$)

$$p(H_n \mid \mathbf{y}) = \frac{p(\mathbf{y} \mid H_n)p(H_n)}{p(\mathbf{y} \mid H_1)p(H_1) + p(\mathbf{y} \mid H_2)p(H_2)} \tag{1.6}$$

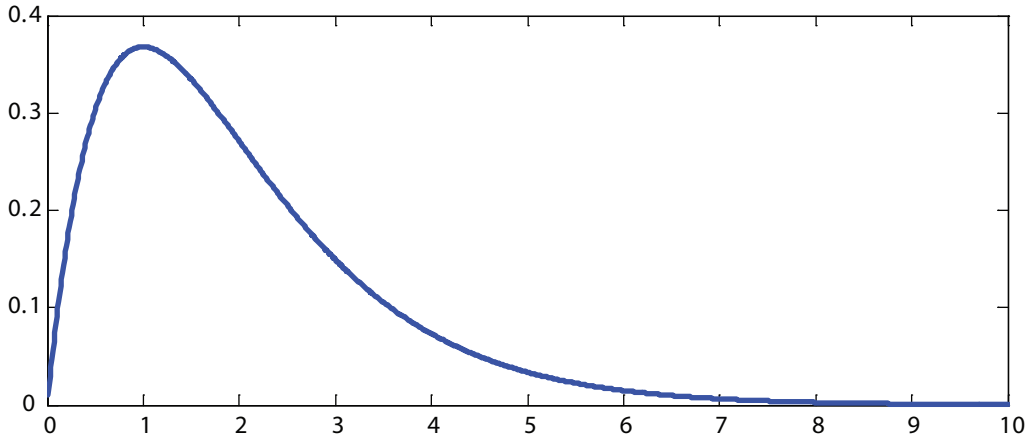


Figure 1.2: An example gamma distribution with $a = 2$ and $b = 1$. This distribution has mean 2 and variance 2. The positive support and long tail are useful for modelling biological systems with unknown kinetic rate parameters.

Given the prior probabilities $p(H_1)$ and $p(H_2) = 1 - p(H_1)$, the posterior probabilities $p(H_1 \mid \mathbf{y})$ and $p(H_2 \mid \mathbf{y}) = 1 - p(H_1 \mid \mathbf{y})$ may be calculated via some likelihood function $p(\mathbf{y} \mid H_n)$. Therefore the denominators in Equation 1.6 cancel, giving

$$\underbrace{\frac{p(H_1 \mid \mathbf{y})}{p(H_2 \mid \mathbf{y})}}_{\text{Posterior Odds}} = \underbrace{\frac{p(\mathbf{y} \mid H_1)}{p(\mathbf{y} \mid H_2)}}_{\text{Bayes Factor}} \underbrace{\frac{p(H_1)}{p(H_2)}}_{\text{Prior Odds}} \tag{1.7}$$

Often there is no preference a priori for a particular model, and so the prior probabilities of the models are usually set to be equal, which shall be the case for the experiments presented in the following chapters. Thus for $P(H_1) = P(H_2)$, the Bayes factor, denoted $B_{12}$, is equal to the ratio of the posterior probabilities of the two models.

The likelihood of the data given a model, known as the marginal likelihood, is obtained by integrating over the parameter space

$$p(\mathbf{y} \mid H_n) = \int p(\mathbf{y} \mid \boldsymbol{\theta}_n, H_n)\pi(\boldsymbol{\theta}_n \mid H_n)d\boldsymbol{\theta}_n \tag{1.8}$$

where $\boldsymbol{\theta}_n$ is a vector describing the parameters for the model $H_n$, $\pi(\boldsymbol{\theta}_n \mid H_n)$ is the prior density of the parameters, and $p(\mathbf{y} \mid \boldsymbol{\theta}_n, H_n)$ is the likelihood function. The marginal likelihood is usually intractable in all but the simplest of scenarios, in which case one must resort to numerical methods. Difficulties may arise when integrating over a high-dimensional parameter space, since the integrand may be highly peaked around its maximum, causing problems for certain types of

approximation. Quadrature methods, for example, may have difficulty finding the region of greatest mass, resulting in a poor approximation. For this reason, the use of Monte Carlo methods is often most appropriate (see e.g. [70]).

### 1.2.5 Interpreting Bayes Factors

Bayes factors have often been referred to as the "weight of evidence", since they give an indication of the relative success two models may have at predicting the data. The following table shows a standard interpretation of the Bayes factor $B_{12}$ as first introduced by Jeffreys ([34]), which compares the model $H_1$ with the model $H_2$. This is usually given in terms of evidence in favour of the first labeled model over the second.

Table 1.1: Interpretation of Bayes Factors

| $B_{12}$ | Evidence against $H_2$ |
|---|---|
| 1 to 3 | Not worth more than a bare mention |
| 3 to 10 | Substantial |
| 10 to 100 | Strong |
| $> 100$ | Decisive |

## 1.3 Monte Carlo Methods

For the purpose of computing Bayes factors to compare competing model hypotheses given a set of experimental data, it is generally necessary to calculate the marginal likelihood. In other words we wish to evaluate, for a particular model, the expectation

$$E_{\pi(\boldsymbol{\theta})}\left[p(\mathbf{y} \mid \boldsymbol{\theta})\right] = \int p(\mathbf{y} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta} \tag{1.9}$$

From now on, conditioning on a particular model $H$ will be omitted to improve readability. As mentioned previously, this integral is usually intractable, although there is an analytic solution when $p(\mathbf{y} \mid \boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$ form a conjugate pair (see e.g. [15]). Intractable integrals may be estimated using Monte Carlo integration methods (see e.g. [70]). Drawing independent samples

$$\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \ldots, \boldsymbol{\theta}^{(N)} \sim \pi(\boldsymbol{\theta}) \tag{1.10}$$

it is possible to estimate the expectation as follows

$$E_{\pi(\boldsymbol{\theta})}\left[p(\mathbf{y} \mid \boldsymbol{\theta})\right] \approx \frac{1}{N}\sum_{t=1}^{N} p(\mathbf{y} \mid \boldsymbol{\theta}^{(t)}) \tag{1.11}$$

By the Law of Large numbers, this estimator converges to the true expectation as the number of independent samples, $N$, tends to infinity

$$\frac{1}{N}\sum_{t=1}^{N} p(\mathbf{y} \mid \boldsymbol{\theta}^{(t)}) \rightarrow E_{\pi(\boldsymbol{\theta})}\left[p(\mathbf{y} \mid \boldsymbol{\theta})\right] \quad \text{as} \quad N \rightarrow \infty \tag{1.12}$$

This estimator, however, is often very unstable and inefficient for a finite number of samples (see e.g. [14]) as many samples will fall outside regions of high likelihood. An alternative is to use the harmonic mean estimator ([63]) which requires independent samples from the posterior distribution $p(\boldsymbol{\theta} \mid \mathbf{y})$. The identity used in this method states that the reciprocal of the marginal likelihood is equal to the harmonic mean of the likelihood, using samples taken from the posterior distribution. There are, however, also problems associated with this method which will be discussed in Chapter 2 where marginal likelihood estimators are examined in greater detail. An overview of some basic sampling methods is now provided, which form the foundations for more advanced methods, also described in Chapter 2.

### 1.3.1 Importance Sampling Methods

Generally it is not possible to sample directly from the particular distribution required for calculating a Monte Carlo estimate. A very naive method of sampling would be to uniformly sample from the target space, however this is usually extremely inefficient, especially for higher dimensional spaces, since the majority of the density is quite often condensed into small compact regions. Few samples will fall into these sought after regions and the resulting expectation calculated will be extremely inaccurate. (A more detailed discussion of this is given in e.g. [50])

Importance sampling may help one more accurately calculate expectations by employing some easy-to-sample-from distribution, $q$, which is in some way similar to the true distribution, $p$, where both $q$ and $p$ are distributions over some parameter space. $q$ need not be normalised and its support should cover the support of $p$, since each sample $\mathbf{x}_i$ generated from $q$ will be assigned an importance weight $w_i$ to adjust for the difference between the two distributions. This is calculated as follows

$$w_i = \frac{p(\mathbf{x}_i)}{q(\mathbf{x}_i)} \tag{1.13}$$

and an estimator takes the importance weight for each of the $i$ samples into account

$$E[\phi(\mathbf{x})] \approx \frac{\sum_i w_i \phi(\mathbf{x}_i)}{\sum_i w_i} \tag{1.14}$$

where $\phi(\mathbf{x})$ is the function over which the expectation is calculated. Clearly if $q(\mathbf{x}_i)$ is less than $p(\mathbf{x}_i)$ a large weight ($w_i > 1$) will be assigned to the sample $\mathbf{x}_i$, since it will not be sampled as often as it would have been from the correct distribution. If $q(\mathbf{x}_i)$ is greater than $q(\mathbf{x}_i)$ a small weight ($w_i < 1$) will be assigned, since $\mathbf{x}_i$ will be sampled too often compared to the true distribution.

Importance sampling is a very useful procedure, however the variance associated with such estimators tends to be very large for finite numbers of samples. Estimators based on importance sampling, although unbiased, generally tend to be unreliable because the true variance of such an estimator is difficult to ascertain due to it being based on a quotient of two distributions, and care must therefore be taken when using them. Indeed, research is still going on into how to stabilise estimators based on importance sampling methods ([64]).

This instability is particularly a problem when using nonlinear ODE models, which tend to induce multimodal posterior distributions. We shall see in the next chapter, however, that this basic idea of importance sampling is at the heart of the Sequential Monte Carlo framework ([6]), in which context it may then usefully be applied to a range of complicated problems, which involve sampling from nonlinear distributions.

### 1.3.2 Markov Chain Monte Carlo Methods

Obtaining independent samples from a nonlinear distribution may be difficult to achieve efficiently using basic sampling techniques such as uniform sampling, importance sampling or rejection sampling (see e.g. [50]). A more widely used approach is to run a Markov chain to produce samples (see e.g. [70]). A Markov chain is generated by moving a point, $\mathbf{x}$, about a target space according to some transition function $p(\mathbf{x}|\mathbf{x}^t)$, where $\mathbf{x}^t$ denotes the position of the chain at time $t$. Each move depends only on the current position of the chain, not on any of its previous positions, so that

$$\mathbf{x}^{t+1} \sim p(\mathbf{x}|\mathbf{x}^t), t = 1, 2, \ldots \tag{1.15}$$

Such a chain converges to a unique stationary distribution (assuming one exists) if it is *irreducible*, i.e. the chain may reach any set of states from any other set of states in a finite number of moves. If the chain is also *aperiodic*, i.e. the greatest common divisor of the time taken to return to any particular state is equal to 1, then an ergodic theorem holds (Equation 1.12), i.e. an estimator using samples

generated by the Markov chain converges to the required expectation over time, as $t \rightarrow \infty$.

### Standard Metropolis-Hastings Sampling

A solution to the problem of how to create such a Markov chain was produced in 1953 by Metropolis ([54]) using a symmetric proposal distribution for repositioning the chain, and this was then generalised in 1970 by Hastings ([26]) to allow the use of nonsymmetric proposal distributions. The Metropolis-Hastings algorithm generates a Markov chain whose stationary distribution is the target distribution in which we are interested.

To apply the Metropolis-Hastings algorithm, all that is needed is a proposal distribution $q$ from which samples can easily be generated. A useful proposal distribution to use when examining biological models is a lognormal distribution centred around the current parameter value, since it only allows $\Re^+$ space to be explored; it would not make sense for biological rate constants to have negative values. During each iteration, each parameter in each chain is sampled and either accepted, in which case the current position is updated, or rejected, in which case the current position is retained. Assuming we have a current parameter $\boldsymbol{\theta}^c$ we can draw a new parameter sample from the proposal distribution $q(\boldsymbol{\theta}^n \mid \boldsymbol{\theta}^c)$ which is accepted with probability

$$\alpha(\boldsymbol{\theta}^n \mid \boldsymbol{\theta}^c) = min\left[1, \frac{p(\boldsymbol{\theta}^n \mid \mathbf{y})q(\boldsymbol{\theta}^c \mid \boldsymbol{\theta}^n)}{p(\boldsymbol{\theta}^c \mid \mathbf{y})q(\boldsymbol{\theta}^n \mid \boldsymbol{\theta}^c)}\right] \quad (1.16)$$

$$= min\left[1, \frac{p(\mathbf{y} \mid \boldsymbol{\theta}^n)p(\boldsymbol{\theta}^n)q(\boldsymbol{\theta}^c \mid \boldsymbol{\theta}^n)}{p(\mathbf{y} \mid \boldsymbol{\theta}^c)p(\boldsymbol{\theta}^c)q(\boldsymbol{\theta}^n \mid \boldsymbol{\theta}^c)}\right] \quad (1.17)$$

Note that in Equation 1.16 the first term, top and bottom, on the right hand side of the *min* function is the posterior of the new parameter divided by the posterior of the current parameter. Equation 1.17 is the expanded form of the posterior in terms of the prior and likelihood functions, where the marginal likelihoods have been conveniently cancelled out. If a symmetric proposal distribution $q(\cdot|\cdot)$ is used, then the proposal distributions also cancel out leaving just the prior and likelihood functions, as in the original Metropolis algorithm.

The above calculation often takes place in log space to prevent the numerical difficulties which sometimes occur when dealing with very small probabilities. In this case the value 1 becomes $log(1) = 0$ in the left side of the *min* function in the above equations.

### Practical Implementation Issues with Metropolis

The starting positions of chains are usually randomised by sampling from the prior distribution, if it is in any way informative. At the start of a run, a chain will therefore not necessarily be sampling from the correct target distribution and it will take some time before it converges. The time taken for a chain to converge is known as the burn-in time and samples recorded during this period are discarded. When employing a Markov chain to sample from a complex distribution it is important to consider the issue of how to monitor convergence. How can one be sure that a chain is in fact sampling from the correct target distribution? Geyer [18] proposes the use of a single chain being run for a very long time, with the hope that it is more likely for such a chain to overcome any burn-in period. The problem is, however, that it only supplies one set of data points, with nothing to compare it to, so one can never be sure that all high density regions of the target distribution have indeed been visited.

An alternative method is to run multiple chains in parallel with dispersed starting positions (see [17]). For an equivalent amount of computational effort, these chains will not be as long as a single chain run in isolation, however with multiple chains one can see clearly whether they have converged to a common distribution, likely, it is hoped, to be the target. The convergence of the chains may be monitored by calculating an $\hat{R}$ value for each parameter, as described by Gelman in ([15]), which tends to 1 as the chains converge and as the number of samples $N$ tends to infinity. It is important to note that even if the $\hat{R}$ values are close to 1, the simulation may still be far from convergence if the chains have not covered all areas of the target distribution. This risk may be minimised by increasing the number of chains, or by running the simulation multiple times with different initial parameter values.

Another challenge is that of choosing efficient proposal steps. Proposal scale factors can be implemented to adjust the size of the steps made by a Markov chain. These can be adjusted during the burn-in period, based on monitoring acceptance ratios of the proposed steps, and then held constant once convergence has been judged to have occurred and posterior samples are being taken. The proposal distribution can also be adapted to the local topology of the target distribution by sampling groups of proposed parameter values from a multivariate Gaussian with an adaptive covariance matrix defining its shape. The covariance matrix can be calculated every so often during the burn-in period based on the previously accepted steps. This generally increases the probability of new proposal steps being accepted, since information regarding correlations between parameters will be discovered and exploited, which results in a more efficient algorithm with shorter burn-in time if the algorithm is able to adapt quickly to the local topology.

A proposal scale factor can be engineered manually to exactly suit many target distributions, but only with prior knowledge of the topology of the target distribution; it is however this topological knowledge that we are trying to obtain in the first place using these sampling techniques, and therein lies the difficulty. Even with engineering attempts, there are cases where the Metropolis algorithm is unable to sample from the required target distribution in an amount of time which makes its use feasible. For example, in multimodal distributions where the modes are sufficiently far apart, a proposal distribution covering these modes is likely to have a very low acceptance rate as it may frequently propose points which fall in regions of low likelihood between the modes.

The efficiency of Metropolis is measured by the acceptance rate of proposed parameters. If the proposal distribution is very wide, proposed step sizes may be large, resulting in a high rejection rate since the posterior density is likely to vary more over larger distances. On the other hand, if the proposal distribution is very concentrated around the current point, then the acceptance rate will be high, since the posterior density is not likely to vary much over short distances, and as a result the burn-in time is likely to drastically increase.

Robert and Casella ([70]) suggest that the acceptance rate for a single parameter change should be between 20% and 40%. It is therefore necessary to tune the variance of the proposal distribution in order to optimise the algorithm. The acceptance rate should be closer to 25% when updating groups of parameters in one go. Gelman ([15]) suggests the covariance of the proposal distribution for such groups should be estimated by calculating the covariance of previous accepted parameters and scaling it. The scale factor may be initialised by using the value $2.4/\sqrt{d}$, where $d$ is the number of parameters. An adaptive step size algorithm may therefore be implemented, whereby the scale factor for each chain is increased or decreased if the chain's acceptance rate is too high or low, respectively. Once the acceptance rate for each chain appears to be stable within the required range, and all the $\hat{R}$ values are close to 1, the scale factor is no longer adapted, and samples may be assumed to be coming from the required stationary distribution.

As dimensionality increases, the burn-in time generally increases as it takes longer for the chains to discover the regions of high density. Thus the efficiency of such sampling methods becomes a great concern. Many algorithms lend themselves to parallelisation, allowing for example multiple chains to be simulated on multiple computer processors. Efficiency is especially an issue when dealing with complex systems of ODEs, since in order to calculate the probability of a proposed step being accepted, the system of equations must usually be solved[3],

[3]There is a discussion in Chapter 4 regarding a possible method of inference without explic-

which can be very time consuming for large numbers of iterations.

## 1.4 Conclusions

Modelling the interactions between the multiple components that drive the behaviour of a biological system is essential for gaining a deeper understanding of both the underlying biological mechanisms at work and the organism as a whole. Many forms of uncertainty, for example in the observations or even in the proposed structure of the models, create great challenges when modelling biological processes, especially when the dynamics are highly nonlinear, as in the case of the Goodwin oscillator model. The Bayesian framework may be employed to deal with this uncertainty in a consistent and principled manner, and it offers a method of objectively comparing competing model hypotheses through the calculation of Bayes factors.

In Chapter 2, I present a review of methods which are suitable for the purpose of system identification and model comparison. I start by discussing a more naive approach to system identification involving optimisation-based methods, and then move on to methods for calculating Bayes factors using more advanced sampling methods, which extend the basic ideas introduced in this chapter. Recent extensions to the original Metropolis-Hastings algorithm include ideas involving temperature schedules and populations of interacting chains in an attempt to improve sampling efficiency, especially when dealing with higher dimensional and multimodal target distributions. These shall be examined in the next chapter and will be put into practice in Chapter 3 to tackle the problem of sampling from the posterior distributions generated by nonlinear Goodwin oscillator models.

itly solving the systems of ODEs.

# Chapter 2

# Parameter Estimation and Model Comparison Methods

Objectively distinguishing between models describing a particular biological process can be very challenging. Many systems display highly nonlinear dynamics with much stochasticity at a molecular level ([68]). Current methods of measuring mRNA and protein levels, such as the use of western blotting, microarrays and mass spectrometry, are still imprecise and although larger numbers of experimental observations may be collected for particular proteins, many kinetic rate constants do not permit themselves to be measured at all. There may be multiple plausible solutions to explain these missing components, which must therefore be inferred from the available data. As mentioned in Chapter 1, a mathematical model may be considered a codification of the hypothesised underlying biochemical network. Originally one was faced with the decision of whether to work with analytically tractable mathematical models, which were amenable to analysis but perhaps not very realistic, or to work with more complex models based on available knowledge of the underlying biology, which might be more likely to describe the phenomenon under observation, but be faced with the problem of parameter estimation and the risk that suboptimal parameter values might be chosen. Indeed, the issue of parameter inference is still one of the main challenges today when modelling biochemical networks ([75]).

How can one be sure of picking optimal, or even good, parameters for a model in order to reproduce a particular type of behaviour observed in a dataset? Many complex systems do not lend themselves to analytic mathematical examination and, moreover, when the rate constants may assume the value of any positive, albeit generally low, real number it becomes clear that the number of possible choices is quite bewildering. The advent of greater, more easily available computing power has allowed the use of parameter optimisation algorithms (see e.g. [52]). Being able to discover the optimal parameters for a particular model to

accurately describe biological data is useful for making predictions about the possible behaviour of the system under different conditions.

Certainly one can easily see from the output whether a proposed mathematical model roughly reproduces a biological dataset or, more formally, one can define some metric which quantifies the similarity, but given multiple models which can roughly reproduce the correct behaviour, how can we decide which underlying network topology *most accurately* reproduces the observed dynamics? A naive method of system identification could be based on point estimates of the "optimal" parameter values for a system. A model with a higher likelihood value using the optimal set of parameters could be considered better than another model with a lower likelihood value. However there is the obvious problem of the more complex model always being favoured, since such models are usually capable of a wider range of response dynamics and are therefore more likely to be able to reproduce the observed "noisy" dynamics. Ideally we want to be able to identify the simplest model capable of reproducing the observed behaviour, in order to gain a clearer understanding of the potential underlying mechanisms at work. The use of an overly complex model to describe data is known as overfitting, and this problem could be tackled by incorporating some kind of penalty term depending somehow on the "complexity" or number of free parameters used in the model. The problem of overfitting, in the context of linear and nonlinear regression models, is discussed in detail in e.g. ([8]). A consistent method of taking all these uncertainties into account is to use the Bayesian framework ([34], [33]), which intrinsically balances the descriptive power of a model with its complexity, since the models are marginalised over all the possible likely parameter values. This requires much more computational effort than simply finding a global set of maximal parameters, since we are now required to solve an integral over the whole parameter space.

When mechanistic ODE-based models are employed to describe a biological process, the computational efficiency of such optimisation or sampling algorithms becomes of vital importance, due to the time it takes to solve the system of differential equations at each iteration of the search procedure. It is worth noting that a similar problem crops up in the field of phylogenetics ([28]), where the search space is large and the likelihood function is computationally expensive to calculate, since all possible paths over a phylogenetic tree must be considered. Additionally, a common problem is that many optimisation methods, such as Simulated Annealing which will be described later in this chapter, often find local optima in the search space, which hinders or even stops further exploration. This occurs particularly over highly nonlinear models, as this nonlinearity often translates into correspondingly complex search spaces, as we see in Chapter 3.

Any method used to tackle these types of problems must therefore balance local steps with effective global exploration strategies.

Firstly, I describe some optimisation algorithms which search for a single point estimate of the global optimum. Parameter values may be estimated by comparing the output of a model to some experimental data and using a cost function to measure the mismatch, which must then be minimised. Equivalently, a likelihood function may be employed, which must then of course be maximised with respect to the experimental data. I then describe some more advanced sampling methods which can be used within the Bayesian framework. These produce samples from the whole target posterior distribution, which can then be used to calculate marginal likelihoods by methods which are described in the final section of this chapter. The statistical accuracy of such methods is examined in Chapter 3.

## 2.1 Optimisation-Based Methods

In this section some optimisation-based algorithms are described which have been developed over the last 20 years and have been previously applied to the area of Systems Biology, with apparent success, estimating the "optimal" parameters for a mathematical system by minimising a cost function based on some biological data (See e.g. [36]). Optimisation methods differ from Bayesian methods in that they search only for a global maximum and generally locate a degenerate distribution around the optimal mode, as opposed to sampling from the complete posterior distribution. These algorithms have been shown to produce very good results when searching for such a global maximum (see e.g. [21]). However, as they can only ever identify a single mode in a possibly complex distribution, these methods will at best only paint a partial picture of how well a model is able to reproduce a particular dataset. The algorithms can give no indication of the confidence that the chosen mode is indeed the "correct" one, nor of the robustness of the set of parameter values to small perturbations. When there are multiple parameter sets which can reproduce the available data, the issue of identifiability crops up, which manifests itself through a highly multimodal search space. Such issues may also not be picked up on using optimisation methods if only one mode can be identified. For a comparison of global optimisation methods applied to biochemical system modelling see ([56]).

### 2.1.1 Simulated Annealing

Markov Chain approaches have been introduced in attempts to overcome the many challenges associated with parameter estimation and in particular that of

trying to find the global maximum in a space containing many local maxima. Simulated Annealing ([37]) makes use of a temperature schedule to "melt out" any roughness in the parameter landscape allowing the Markov chains to escape from local maxima more easily. Note that if a cost function is employed, which measures the error associated with the parameter estimation, it is then a local minimum we wish to find, instead of a maximum.

The true target distribution is raised to a power, $t$, which can be thought of as a variable inversely proportional to temperature. This temperature variable generally starts at a very low number close to 0, i.e. at a high temperature, which has the effect of "melting" and flattening the parameter landscape allowing for easy exploration of the whole space. A Markov chain starts exploring the space at this temperature and gradually the temperature is decreased. The slower the rate of temperature decrease, the better the chances are of the Markov chain finding the global maximum ([31]). The algorithm can be stopped once the Markov chain stops accepting proposed steps, or once some other target criterion has been reached, for example once the temperature parameter reaches a certain value.

In the context of parameter estimation, assume we want to find the optimal set of parameters, $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_D]^T$ $(\theta_i \in R)$, for a model, $f$, used to describe some experimental data, $\mathbf{y} = [y_1, \ldots, y_N]^T$ $(y_n \in R)$. If we use a loss function, then we wish to find the global *minimum* of $\exp\{L(\boldsymbol{\theta}, \mathbf{y})\}$, where $L$ is some loss function which can be evaluated at each point and measures the error between the model output and the data, for example using the mean squared error between the simulated data, $f(\boldsymbol{\theta})$, and the experimental data, $\mathbf{y}$. Using a simple simulated annealing method, we could find the minimum of

$$\exp\left\{\frac{L(\boldsymbol{\theta}, \mathbf{y})}{t}\right\} \tag{2.1}$$

where $t$ is gradually increased from near 0. As $t$ becomes large, the algorithm will sample from an increasingly degenerate distribution centred on a, hopefully global, minimum. Algorithm 1 details the Simulated Annealing procedure in greater detail.

The most obvious drawback of this method is its inability to explore multiple modes simultaneously. Care must also be taken when constructing a cooling schedule. If the temperature is reduced too quickly, there is a chance the chain will get stuck in a local mode and be unable to escape. Generally, the more complex the target distribution is, the slower the cooling schedule should be, but unfortunately there are no hard and fast rules concerning the optimal size of the temperature steps, although there have been attempts to introduce adaptive temperature schedules ([81]). Usually they must be hand-picked for each search

**Algorithm 1** Simulated Annealing

1: Initialise starting position, $\boldsymbol{\theta}$, and temperature, $t$
2: **repeat**
3: Propose new position, $\boldsymbol{\theta}_{New}$, based on the current position
4: Calculate cost function for new position, $\exp\left\{\frac{L(\boldsymbol{\theta}_{New},\mathbf{y})}{t}\right\}$
5: Accept or reject proposed move according the the Metropolis probability, $\min\left[\exp\left\{\frac{L(\boldsymbol{\theta},\mathbf{y})-L(\boldsymbol{\theta}_{New},\mathbf{y})}{t}\right\},1\right]$
6: Increment temperature, $t$
7: **until** Termination criteria are met

space, and multiple runs with different temperature schedules can help confirm whether the global maximum has been found. It should be noted that combining multiple runs of a simulated annealing approach with importance sampling results in an algorithm with similarities to Annealed Importance Sampling ([61]). This can be considered under the Sequential Monte Carlo framework, which will be examined later in this chapter.

The biochemical networks which drive circadian rhythms exhibit highly nonlinear behaviour and have been examined closely in an important recent paper by Locke et al. ([48]). A great contribution of this work was to introduce the use of a general optimisation method, in the form of Simulated Annealing, to estimate parameters over mechanistic models describing the circadian networks of the model plant *Arabidopsis thaliana*. The use of a simulated annealing approach is motivated by the nonlinear distributions produced when using a cost function to optimise the parameters. The model employed in ([48]) consists of 23 free parameters, which equates to sampling from a 23 dimensional target distribution. The results using a simulated annealing algorithm are very dependent on the chosen starting position in such a high dimensional space and therefore a random search of the space was undertaken, before applying simulated annealing. A SOBOL random number generator ([77]) was used to spread out the search starting points more evenly across the space, and a cost function was employed to evaluate around 1,000,000 possible starting points, from which the top 100 solutions were then refined further using a simulated annealing routine. One weakness of this paper is the use of a hand-crafted cost function which is dependent on various measurements such as the period and amplitude of oscillations. It could be argued that such a cost function is rather arbitrary, and it would be interesting to see whether a cost function constructed slightly differently would affect the results of the optimal parameters found. I suggest a more consistent approach would be to model the data points directly, and infer the range of most likely parameters using Bayesian methods, assuming that the experimental observations are contaminated by some stochastic process. This stochastic component could

be modelled by the likelihood function, using for example a Gaussian process, as discussed in Chapter 4. More advanced sampling methodology, with global exploration capabilities, could also be employed, reducing the need for such a large initial separate random search. This is described later in this chapter.

It should be noted that the goal of applying Bayesian methods to models of this size is extremely challenging, both in terms of computational requirements as well as tuning an algorithm to successfully explore such a large parameter space and find all the regions of high likelihood.

### 2.1.2 Genetic Algorithms

Genetic Algorithms (GAs) were first created by trying to mimic ideas that were emerging from advances in our understanding of genetics and the idea was to mirror natural selection and evolution that occurs in real life. They have been applied to a wide variety of optimisation problems and are generally successful in seeking out good solutions ([55]).

The general principles of these types of algorithms is as follows. A group of individuals, called a population, explore the parameter space using two different methods. The first is mutation, whereby one of the coordinates describing the position of an individual is perturbed, effectively moving the individual to a new point in the local parameter space. The second is crossover, whereby the coordinates of two individuals combine to produce two "offspring", resulting in a more global jump through parameter space. The individuals are usually chosen proportional to their "fitness", calculated by some cost or likelihood function. For a more detailed overview of the algorithm, see ([29]).

Such algorithms are often used because they work well in practice, however exactly why they work well is difficult to analyse mathematically. GAs are not always guaranteed to find the "best" solution, although they do often find good solutions with respect to the cost function they are trying to minimise. Concrete analytical results regarding convergence and theoretical bounds on numerical estimates are available only for very specific problems and under certain restrictive conditions ([2]).

Aside from the current lack of general theoretical results, Genetic Algorithms suffer similar limitations as Simulated Annealing. They produce point estimates of variables which generally correspond to good solutions, however they fail to provide other types of useful information, such as the robustness of the system and confidence levels on the solutions found. This is ultimately the reason I believe they have limited potential for the purpose of system comparison. The methods in the next section offer solutions to the shortcomings of these optimisation methods.

## 2.2 Parameter Inference Methods

The algorithms I describe here extend the Metropolis and importance sampling methods presented in Chapter 1. They are not simply optimisation methods, but rather methods for parameter inference which generate samples from a posterior distribution of the likely solutions. The ability to sample from and examine the structure of such posteriors allows confidence levels to be calculated around optimal parameter values, and permits a more global view of how well a model describes experimental data.

In particular, the types of methods I look at involve the addition of two main ideas, firstly the idea of using an auxiliary variable, representing an inverse temperature for example, and secondly the idea of using a population of chains which explore the target space simultaneously and "communicate" in order to find regions of high density more efficiently. There are of course other methods of sampling such as slice sampling ([62]) and nested sampling ([76]), and indeed this is a very active research area which is constantly expanding. Given such a vast literature on MCMC sampling methodology it would be impossible to give a detailed review of every such method in this thesis, and therefore I focus solely on population and temperature based methods which, having been successfully employed in many areas of physics, have so far, to the best of my knowledge, had very little impact in the area of Systems Biology. This is partly because many models previously examined have either been linear in the parameters or have not exhibited very complex posterior distributions and thus simpler methods have sufficed. Recently, however, more complex nonlinear models have been examined using Simulated Annealing ([48]), and I would suggest that the following methods would provide more useful information when estimating parameters from multimodal posteriors and help highlight potential problems of identifiability which may not appear using other simpler methods.

### 2.2.1 Advanced Markov Chain Monte Carlo Methods

The basic Metropolis-based sampling method described in Chapter 1 has been developed extensively over the last couple of decades in an attempt to improve the efficiency and accuracy of generating samples from a stationary multimodal distribution. Indeed, an often used test for a newly proposed sampler is to use a stationary distribution consisting of multiple Gaussian distributions, as in ([45]). The two main ideas which have been used to advance the development of more efficient sampling methods are the idea a population and the idea of a temperature schedule.

One method of incorporating the idea of a population is Adaptive Direction

Sampling ([71]), in which chains in the population are moved according to the position of the other chains in the population. In order to implement ADS, one chain is selected as an "anchor" point, and its proposal step is sampled from a line going through both the "anchor" and another chain randomly selected from the population. The idea is that iteratively finding the highest likelihood positions on the lines joining the population of chains will result in the chains converging to the target distribution. This method was improved upon by Liu et al. ([47]), who suggested the use of local optimisation at the position of another randomly chosen chain in the population, so that the sampling would be from a line through the "anchor" pointing in the direction of a local mode. This method they named Conjugate Gradient Monte Carlo (CGMC), since they proposed conjugate gradient iterations to perform the local optimisation steps. This method also has links to the Multiple-Try Metropolis (MTM) algorithm ([47]), as MTM provides a possible method for sampling from the distribution on the line between two chosen chains, which is almost never analytic. In the MTM algorithm, instead of a single proposal step being made for a chain, a collection of possible proposal steps are carried out. The weights for these multiple proposed steps are calculated, and one of these steps is then chosen with probability relative to its weight and accepted according to a modified Metropolis-Hastings ratio. Generally the MTM algorithm allows larger step sizes to be made, since it chooses the best step from a collection of proposed steps, resulting in the algorithm being more effective than basic Metropolis algorithms.

Another approach is to use a population in which the Markov chains try to avoid each other, instead of being attracted to each other. So-called Pinball Sampling was suggested by Robert and Mengersen ([53]). They base their idea on a population in which chains perform random walks with corrections so that they avoid the vicinity of other chains. This approach therefore emphasises not only finding the regions of highest density, but also covering them as widely as possible. The convergence of a population using Pinball Sampling, or indeed any of the other methods mentioned above, is justified by ergodicity properties of the Markov chains. They therefore all have similar advantages and drawbacks, including the usual *curse of dimensionality*; the inital points must be assumed to provide a fair coverage of the support of the target distribution, which requires the population size to increase dramatically as the dimensionality increases. For a more detailed examination of the Pinball Sampling algorithm and the corresponding stationarity results see ([53]).

The idea of temperature has also been incorporated into many sampling schemes with great effect. Simulated Tempering is similar to Simulated Annealing ([37]) in that it makes use of intermediate distributions, associated with

a temperature index, in order to gain more accurate samples from the target distribution. A Markov chain may jump between temperatures allowing it to escape from local modes and explore the target space more widely. This method was developed independently by Parisi and Marinari ([51]), and also by Lyubartsev et al. ([49]) under the name of Expanded Ensembles. Samples may be obtained by running the Markov chain to equilibrium and recording samples only when $t = 1$, corresponding to the target distribution. It is hoped that the additional computational power spent moving between temperatures, redeems itself by exploring the true distribution more globally and producing less correlated samples. Parallel Tempering, also often known as Exchange Monte Carlo, ([19]) and Tempered Transitions ([59]) are further methods involving temperature schedules which may be employed to successfully explore multimodel distributions.

The idea of implementing a population in such a manner stems from Genetic Algorithms, where each iteration produces a new population of particles, which interact in ways mimicking natural selection, such as through mutation of the position vectors. The main difference between Genetic Algorithms and Population Markov Chain Monte Carlo (MCMC) methods however, is that while GAs are used for optimisation problems, Population MCMC methods are concerned not only with finding the global maximum of a target space, but also providing samples from all other high density regions of the space. The use of temperature ladders as a means of discovering high likelihood regions can also been seen as a process of natural selection when compared to Genetic Algorithms; the fittest samples move to a lower temperature, while the least fittest move to a higher temperature, allowing them to traverse more easily into a different region within the target space. Liang and Wong ([44]) give a good summary of the components of a Population MCMC based algorithm and the influence Genetic Algorithms have had on its development. Laskey and Myers ([43]) also provide interesting insight into how biological language and metaphors have been incorporated into stochastic search literature, as well as giving a comparison of GAs and population based Monte Carlo methods.

### 2.2.2 Population Markov Chain Monte Carlo

In this section we look at an extension of standard Markov Chain Monte Carlo algorithms, involving the ideas both of a population as well as a temperature schedule. In such population methods, chains proceed on a random walk through a product distribution space, and their movement is influenced by the position of other chains at different temperatures. The idea is that chains in low density regions will move towards chains in high density regions, so that the population will converge more quickly on the target distribution, which is at the lowest

temperature. Taking advantage of the information contained in current chains must however be done in a careful manner in order to preserve the Markovian structure of the chains.

Population Markov Chain Monte Carlo (also known as Evolutionary Monte Carlo, see [45]) consists of a population of Markov chains which explore a target distribution by means of a series of intermediate distributions with varying temperatures. A separate chain is run at each temperature and they are able to interact by jumping between temperatures and exchanging positions, thus exploiting easier exploration at higher temperatures (See Algorithm 2). This is the same as the parallel tempering method mentioned previously ([51, 60]). In addition, the chains may perform crossover steps which allow them to move to new positions within their current temperature level based on the locations of other chains, similar to the Adaptive Direction Sampling algorithm also mentioned previously.

As with other MCMC methods, a burn-in period is necessary to allow the chains to converge to the appropriate target distribution. Once the chains have converged, the chain at the lowest temperature provides samples from the true target distribution. There are quite severe restrictions on the kernels used to reposition chains, since their Markovian structure must be preserved for the algorithm to be valid. A current research area is the development of more efficient transition kernels for use with this method ([23]).

**Algorithm 2** Population Markov Chain Monte Carlo

1: Assign starting positions to each chain in a population, $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N)$
2: Define a temperature ladder attached to the population, $(\boldsymbol{\Theta}, \mathbf{t}) = (\boldsymbol{\theta}_1, t_1, \ldots, \boldsymbol{\theta}_N, t_N)$
3: **repeat**
4: Apply local move or crossover operator (as described below) to each chain in the population with probability $p_m, (1 - p_m)$ (where $p_m$ is sometimes known as the mutation rate)
5: Try to exchange $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ for $N$ pairs $(i, j)$, with $i$ sampled uniformly on $(1, ..., N)$ and $j = i \pm 1$ with probability $p_e(\boldsymbol{\theta}_j, \boldsymbol{\theta}_i)$, where $p_e(\boldsymbol{\theta}_{i+1}, \boldsymbol{\theta}_i) = p_e(\boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_i) = 0.5$ and $p_e(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = p_e(\boldsymbol{\theta}_N, \boldsymbol{\theta}_{N-1}) = 1$
6: **until** Chains converge

A standard method of implementing Population Markov Chain Monte Carlo is as follows. We assume we want to sample from a posterior distribution defined on the real space,

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto L(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \tag{2.2}$$

where $L(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood of the experimental data, $\mathbf{y}$, conditioned on the pa-

rameters, $\boldsymbol{\theta}$, and $\pi(\boldsymbol{\theta})$ is the prior distribution over the parameters. We first define an $N$-step temperature schedule, $\mathbf{t} = (t_1, \ldots, t_N)$, with $0 = t_1 < \ldots < t_N = 1$. Note that for the metaphor of temperature to make sense, the parameter schedule $\mathbf{t}$ is actually inversely proportional to temperature, with $t_1$ considered a high temperature and $t_N = 1$ considered a low temperature. A sequence of distributions[1], corresponding to each step $i = 1, ..., N$ on the temperature schedule, is then constructed

$$p(\boldsymbol{\theta}_i|\mathbf{y}) = \frac{L(\mathbf{y}|\boldsymbol{\theta}_i)^{t_i}\pi(\boldsymbol{\theta}_i)}{Z_{t_i}} \tag{2.3}$$

where $\boldsymbol{\theta}_i$ will be considered the position of the Markov chain running at temperature, $t_i$, and $Z_{t_i}$ is some, usually intractable, normalising constant

$$Z_{t_i} = \int L(\mathbf{y}|\boldsymbol{\theta}_i)^{t_i}\pi(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i \tag{2.4}$$

One can therefore picture a multimodal target distribution at $t_N = 1$, which melts at higher temperatures so that the distributions at $t_n < 1$ are easier to explore. The resulting distribution at each temperature is explored using an individual Markov chain, so that the total number of Markov chains running simultaneously is $N$. In Population Markov Chain Monte Carlo a product distribution is considered when moving individual chains, thus taking the entire population of chains throughout the temperature schedule into account. We therefore sample from

$$p(\boldsymbol{\Theta}|\mathbf{y}) = \frac{1}{Z_{\mathbf{t}}}\prod_{i=1}^{N} L(\mathbf{y}|\boldsymbol{\theta}_i)^{t_i}\pi(\boldsymbol{\theta}_i) \tag{2.5}$$

where $\boldsymbol{\Theta}$ is the population of Markov chains, $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N$, at the temperatures, $t_1, \ldots, t_N$ respectively. The (intractable) normalising constant is now

$$Z_{\mathbf{t}} = \prod_{i=1}^{N} Z_{t_i} \tag{2.6}$$

Markov chains explore the distributions according to the temperature schedule and they may also interact with one another and swap positions across temperatures. During each iteration, the algorithm updates the population by carrying out one of the following moves:

### Local Metropolis Move

A random Markov chain, $\boldsymbol{\theta}_i$, is selected from the population $\boldsymbol{\Theta}$, and a random vector is added to it to create a new proposed position, $\boldsymbol{\theta}_i'$. Thus a new population

[1]Other sequences are possible, see e.g. [16], but here we fix a geometric path between the prior and the posterior.

is defined as $\boldsymbol{\Theta}' = \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}'_i, \ldots, \boldsymbol{\theta}_N$, which is then accepted with probability $min(1, r_m)$ according to the Metropolis-Hastings rule,

$$\begin{aligned} r_m &= \frac{p(\boldsymbol{\Theta}'|\mathbf{y})}{p(\boldsymbol{\Theta}|\mathbf{y})} \frac{T(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}')}{T(\boldsymbol{\Theta}' \mid \boldsymbol{\Theta})} \\ &= \frac{\frac{1}{Z_{\mathbf{t}}} \left[ L(\mathbf{y}|\boldsymbol{\theta}_1)^{t_1} \pi(\boldsymbol{\theta}_1) \times \ldots \times L(\mathbf{y}|\boldsymbol{\theta}'_i)^{t_i} \pi(\boldsymbol{\theta}'_i) \times \ldots \times L(\mathbf{y}|\boldsymbol{\theta}_N)^{t_N} \pi(\boldsymbol{\theta}_N) \right]}{\frac{1}{Z_{\mathbf{t}}} \left[ L(\mathbf{y}|\boldsymbol{\theta}_1)^{t_1} \pi(\boldsymbol{\theta}_1) \times \ldots \times L(\mathbf{y}|\boldsymbol{\theta}_i)^{t_i} \pi(\boldsymbol{\theta}_i) \times \ldots \times L(\mathbf{y}|\boldsymbol{\theta}_N)^{t_N} \pi(\boldsymbol{\theta}_N) \right]} \\ &\quad \times \frac{T(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}')}{T(\boldsymbol{\Theta}' \mid \boldsymbol{\Theta})} \\ &= \frac{L(\mathbf{y}|\boldsymbol{\theta}'_i)^{t_i} \pi(\boldsymbol{\theta}'_i)}{L(\mathbf{y}|\boldsymbol{\theta}_i)^{t_i} \pi(\boldsymbol{\theta}_i)} \times \frac{T(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}')}{T(\boldsymbol{\Theta}' \mid \boldsymbol{\Theta})} \end{aligned} \tag{2.7}$$

where $T(\cdot \mid \cdot)$ denotes the probability of transition from one population to another. A common choice for the transition density $T$ is a Gaussian centred around the current position of the chain, which is symmetric and thus allows the transition densities in the above equation to cancel.

**Exchange**

This is similar to a standard exchange move in temperature based Monte Carlo methods. A new population $\boldsymbol{\Theta}'$ is created by swapping the positions of two chains, $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$, on the temperature ladder so that,

$$(\boldsymbol{\Theta}', \mathbf{t}) = (\boldsymbol{\theta}_1, t_1, ..., \boldsymbol{\theta}_j, t_i, ..., \boldsymbol{\theta}_i, t_j, ..., \boldsymbol{\theta}_N, t_N) \tag{2.8}$$

The new population is accepted with probability $min(1, r_e)$ according to the Metropolis-Hastings rule,

$$\begin{aligned} r_e &= \frac{p(\boldsymbol{\Theta}'|\mathbf{y})}{p(\boldsymbol{\Theta}|\mathbf{y})} \frac{T(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}')}{T(\boldsymbol{\Theta}' \mid \boldsymbol{\Theta})} \\ &= \frac{[L(\mathbf{y}|\boldsymbol{\theta}_j)^{t_i} \times L(\mathbf{y}|\boldsymbol{\theta}_i)^{t_j}]}{[L(\mathbf{y}|\boldsymbol{\theta}_i)^{t_i} \times L(\mathbf{y}|\boldsymbol{\theta}_j)^{t_j}]} \times \frac{T(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}')}{T(\boldsymbol{\Theta}' \mid \boldsymbol{\Theta})} \end{aligned} \tag{2.9}$$

where many of the terms, including the normalising constants, have conveniently cancelled out as shown previously for a local Metropolis step. Usually the two selected chains are chosen to be direct neighbours in the temperature ladder to increase the likelihood of the interaction being accepted.

**Crossover**

There are a few variations on the crossover operator. The original crossover operators for Population Markov Chain Monte Carlo were defined for a finite

binary space, however they were later extended for use in the real space ([45]). A chain, $\boldsymbol{\theta}_i$, is selected uniformly from a population, $\boldsymbol{\Theta}$. A second, different chain, $\boldsymbol{\theta}_j$, is also selected, either at random or for example with a probability proportional to its current likelihood, $f_j(\boldsymbol{\theta}_j)$. Two new chain positions, $\boldsymbol{\theta}'_i$ and $\boldsymbol{\theta}'_j$, are then produced by so-called *one-point*, *k-point* or *adaptive crossover*. The positions of the new chains replace the old positions to form a new population, $\boldsymbol{\Theta}'$, which is then accepted or rejected according to a standard acceptance probability. The one-point crossover takes place by uniformly selecting a crossover point, $c$, from $(1, ..., N)$, and then swapping all the values in the vectors $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ which occur after position $c$. The k-point crossover is similar except there are multiple uniformly selected crossover points, dictating which parts of the vector should be swapped. The adaptive crossover is more complicated and the reader is referred to ([44]) for the details.

The *snooker crossover operator* ([45]) is based on Adaptive Direction Sampling (ADS) ([20]), and offers a method of moving a chain towards a region of higher likelihood by sampling from a line going though the coordinates of the current chain and some chosen second chain. For convergence of the algorithm to occur, the proposal function used must be a Markov transition kernel satisfying ergodicity requirements. A detailed examination of stationarity properties for the Adaptive Direction Sampling algorithm may be found in ([71]). In the original ADS the second chain, which sets the direction of the line from which to sample, is chosen at random. The snooker operator improves on this by basing the choice of the second particle on their current likelihoods, thus increasing the probability of choosing a second particle near a region of high density.

A common feature of these real crossover operators is that the probability of going from the current position to a proposed position is symmetric, $P(\boldsymbol{\Theta}' \mid \boldsymbol{\Theta}) = P(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}')$, which makes the operator invariant with respect to the underlying distribution. For the theorems, and corresponding proofs, that show the snooker crossover operator is in fact a proper invariant transition, see ([45, 47]).

Obviously it is not advisable to use this type of operator too often otherwise the chains will tend to cluster together, inhibiting the exploration of the wider space.

### Application

The Population Markov Chain Monte Carlo algorithm has relatively few parameters which must be set by the user. The size of the population, $N$, which is equivalent to the number of steps on the temperature ladder, $\mathbf{t}$, the spacing of the steps on the temperature ladder, the effect of which is examined in Chapter 3, and the various probabilities, $p_m$ and $p_e$, determining how often each of the dif-

ferent types of moves are applied to the chains. By increasing $N$ (and therefore $\mathbf{t}$, since each chain is associated with a temperature) we can improve the chances of the population covering more areas of the target distribution, which is preferable in order for the system to mix well, however there is of course a corresponding decrease in speed, since there are more chains which must be moved through the product space. On the other hand, a small population size results in a steeper temperature ladder with lower acceptance probabilities, which although computationally faster may often result in poor mixing. Generally, the population size should be chosen to be comparable to the dimension of the problem. For a more detailed discussion of how best to choose these parameters see ([44, 45]).

It is useful to note that setting $p_m = 1$ turns the Population MCMC algorithm into a parallel tempering algorithm without the use of any crossover operator, and setting $p_m = 1$ and $N = 1$ turns it into a single-chain Metropolis-Hastings algorithm. The effect of the crossover operator on the performance of the Population MCMC algorithm is investigated in [32]. The authors examine autocorrelations of the likelihoods of the samples produced, firstly when running the algorithm without using the crossover operator, and then using two different variants of the crossover operator. They conclude that the addition of such a crossover operator results in a decrease in the autocorrelation, which may be interpreted as an increase in the exploration ability compared to a standard simulated tempering approach, although the authors note that there is greater computational cost and an increased complexity in coding the algorithm.

As with other MCMC algorithms, it is important to determine whether the chains have reached their target distribution, and how well the chains are mixing. These methods of diagnosis may also be described as "stopping rules". A method often used for this purpose is the $\hat{R}$ statistic, proposed by Gelman ([15]). This method looks at both the within-chain and between-chain convergence, and gives a value, $\hat{R}$, which tends to 1 as the chains converge. Using this method, multiple runs of Population MCMC can be initiated simultaneously, and convergence monitored using the multiple chains at each temperature to calculate $\hat{R}$ values for each step in the temperature schedule. Other types of convergence indicators which could be employed include autocorrelation time, which may be used on single chains, and visual aids such as histogram plots.

A slightly different approach is taken by Guihenneuc-Jouyaux et al. ([24]). They suggest that the process of determining convergence be split up into 3 stages. Firstly, one wishes to ascertain that the chain is in fact sampling from the stationary distribution. Secondly, it must be confirmed that the chain is adequately exploring all regions of the distribution containing sufficient density. Thirdly, the accuracy of any parameter estimations made must be quantified.

Their paper takes a detailed look at this approach to convergence diagnostics, and supplies some examples of their methods in action.

Laskey and Myers ([43]) compare a Metropolis-Hastings algorithm, an evolutionary algorithm and a Population MCMC algorithm. They show that the evolutionary algorithm and Population MCMC algorithm find the global maximum more efficiently than Metropolis-Hastings, with the Population MCMC method having the added advantage that it samples from a *distribution* of likely parameter values. This demonstrates the advantages of information exchange between chains. Liang and Wong ([45]) also include three illustrative examples showing the population based Evolutionary Monte Carlo (Population MCMC) method outperforming a Metropolis-based simulated tempering method. The simulation studies show how Population MCMC offers much better mixing on a 20-component two dimensional model and, most importantly, how it finds all of the modes, whereas the Metropolis based version fails to find three of the outer modes. In this thesis I will look at higher dimensional examples of multimodal distributions, and in Chapter 3 I will show that Population MCMC may be used to successfully sample from posteriors produced by nonlinear models of up to seven dimensions.

As already mentioned, it is necessary to wait for the chains to mix well before collecting samples, and this burn-in time varies in length according to the complexity and dimension of the space being explored. Population MCMC requires tuning of the temperature ladder and proposal kernels to individual problems in order to improve efficiency, however this can be time consuming. Finally, it is worth noting that the number of chains being simulated in a population (and hence the gradient of the temperature ladder) must remain constant throughout the simulation in order to achieve convergence and, in Chapter 3, I also examine how such a temperature schedule might be best chosen so as to make the algorithm most efficient.

### 2.2.3 Sequential Monte Carlo

I now describe a framework proposed in ([6, 7]) which incorporates sequential importance sampling ideas to justify the convergence of a population of samples to a target distribution. It is important to note that the validity of this framework does not rely on the ergodic properties of any Markov chains. This is a very powerful methodology and can be regarded as a general case encompassing a number of specific algorithms based on importance sampling ideas, including Annealed Importance Sampling ([61]) and Population Monte Carlo ([4]) which are described later.

The main idea is that of propogating a population of $\boldsymbol{\Theta}$ samples through a

collection of $N$ distributions in sequence $\pi_{n=1,\ldots,N}$ using some transition kernel. The first distribution, $\pi_1$, is typically easy to sample from, and the resulting samples form the starting points as the second distribution is sampled from. Intermediate distributions then help iteratively guide the samples towards the high density regions of the final distribution, $\pi_N$, which is the target we wish to investigate. Usually sequential methods have been applied to problems where the data arrives in a particular order, for example over time, however they may be equally well applied to complex static problems with large datasets ([4]).

This methodology may be used in various ways. Employing the Bayesian framework, each distribution could be taken to represent the posterior distribution of the parameters given the first $p$ datapoints. This might well have the effect of simplifying the posterior distributions early on, similar to a kind of temperature schedule effect, and since less data is used in the calculations it may also be computationally more efficient in some cases. Alternatively, the sequence of intermediate distributions could be defined artificially, for example geometrically, in a similar way to the temperature schedules encountered before. Traditional importance sampling has generally not been used for more complex problems due to the difficulty of choosing an importance distribution which approximates the target distribution well enough for the method to work efficiently. A sequence of distributions alleviates this problem by gradually moving towards the required nonlinear target distribution.

Unlike Markov Chain Monte Carlo approaches, in which the convergence monitoring of chains is vital, Sequential Monte Carlo (SMC) assigns weights to each of the samples (summing to 1) as they develop over the iterations such that their estimation of the expectation, with respect to any of the distributions, $\pi_n$, converges asymptotically

$$\sum_{i=1}^{N} W_n^{(i)} \phi(\boldsymbol{\theta}_n^{(i)}) \rightarrow E_{\pi_n}(\phi) \tag{2.10}$$

where $W_n^{(i)}$ are the weights attached to the $i$th sample in the $n$th distribution, and $\phi$ is the function over which the expectation is calculated.

SMC is similar to most temperature-based MCMC approaches in that the choice of intermediate target distributions and proposal mechanisms strongly influences how well such an algorithm performs. The main strength of SMC lies in its flexibility, and it can be shown that many well-known and often used algorithms can be considered special cases of SMC given a particular set of target distributions and proposal functions.

Annealed Importance Sampling ([61]) is an example of such an algorithm which fits into the SMC framework. It has similarities with Simulated Annealing,

described previously, in that a sample is initially taken from an easy to sample from distribution at a high temperature, and a cooling schedule is applied until one has a sample from the required target distribution. These weighted samples may then be used to estimate expectations, and so it is not necessary for the chains to have converged fully at any of the intermediate distributions. This algorithm can be considered a subset of the Sequential Monte Carlo framework by taking the proposal mechanism to be a local Metropolis style random-walk, and by choosing the intermediate distributions geometrically.

### Population Monte Carlo

Population Monte Carlo (PMC) algorithms as described by Cappé ([4]) can also be considered a subset of SMC methods ([9]). I now give an overview of the history of Population Monte Carlo and make explicit the differences between this non-Markovian method, which depends on importance sampling convergence arguments, and the Population MCMC method described previously, which depends on ergodicity properties of the Markov chains for convergence to a stationary distribution. Since this method has been influenced by particle filter methods, the terminology used to describe it sometimes differs slightly from that used to describe MCMC methods. The members of a population are thus often called particles, to distinguish them from Markovian "chains".

The idea of incorporating a population of particles into Monte Carlo methods has been around for a while. The first cross-disciplinary survey on population Monte Carlo methods was first given in ([30]), in which the basic structure of such algorithms is detailed. These algorithms have been developed in many different fields simultaneously, and can be found applied to areas such as Lattice Spin Systems, Quantum Many-Body problems and polymer science (see [30]).

**Algorithm 3** Population Monte Carlo

1: Assign starting positions to each chain in a population, $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$
2: **repeat**
3: Move each chain in the population according to some kernel and compute its end weight
4: Resample according to the weights
5: **until** Chains converge

Pseudocode for a general non-Markovian Population Monte Carlo algorithm is given by Algorithm 3. This provides the framework for developing Monte Carlo algorithms based on iterated importance sampling. The flexible choice of kernel allows for potentially easier exploration of the target distribution at both a local and global level.

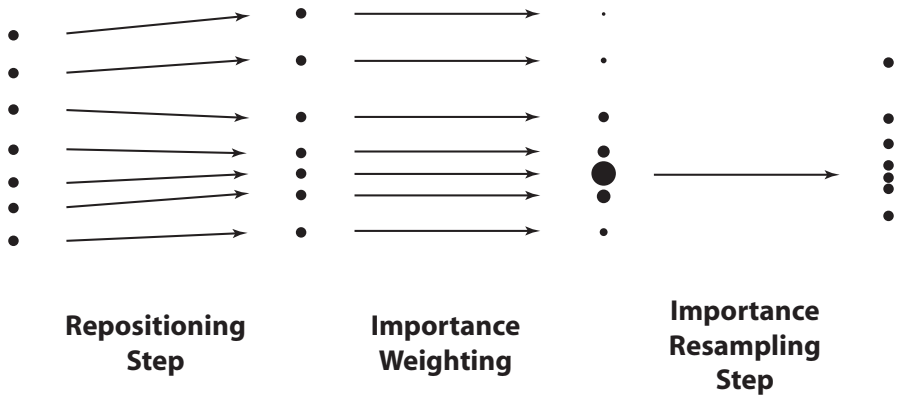


Figure 2.1: Illustration of the general resampling procedure based on the importance weighting of chains. There is no restriction on the repositioning step, in that it need not be Markovian. Such freedom however does not guarantee efficiency, and so the repositioning kernel must be carefully chosen. See ([4]) for further information on the choice of kernels.

The driving force behind Cappé's Population Monte Carlo method is the idea of iterated generations of importance sampling. This approach has many advantages over standard MCMC techniques. The algorithm consists of a population of particles which explore the parameter landscape according to some repositioning kernel. Importance sampling is then employed for resampling the population according to their relative likelihoods (see Figure 2.1). A great advantage of this importance sampling step is that it produces samples approximately simulated from the target distribution and removes dependency on a Markovian requirement of the repositioning step, making it easier to incorparate more varied repositioning kernels that may, for example, take more global steps in the parameter space. These relaxed conditions on the repositioning kernel may increase the chances that multiple modes will be found. These samples can then be used to obtain approximately unbiased estimates of expectations over the target distribution.

Adaptive proposal functions which depend on samples from past iterations may also be employed, and the method is still valid without any alterations to the rest of the algorithm, as demonstrated in ([4]). In addition, the number of particles in a population need not be kept constant over the iterations; the population size may be allowed to grow or shrink. Valid samples are still produced after a resizing of the population due to the normalisation which occurs during the importance resampling step.

The development of non-Markovian PMC methods has stemmed from several ideas. The construction of proposal functions has been strongly influenced by existing MCMC methodology. Sample equalisation and rejuvenation procedures have come from the sampling importance resampling (SIR) literature (see e.g.

[72]), while sample improvement has its roots in iterated particle systems ([5]). Population Monte Carlo is clearly similar to the Sequential Monte Carlo framework described previously, but in this case the emphasis is placed on the flexible choice of proposal mechanism. SMC adds the idea of sampling from a flexible sequence of distributions which converges to the target distribution.

## 2.3 Estimating Marginal Likelihoods

In this section I look at various methods of calculating marginal likelihoods. Being able to calculate these accurately is of vital importance for computing meaningful Bayes factors for model identification and, as will become evident, accurately calculating marginal likelihoods over nonlinear posteriors is not straightforward. The first methods I describe is based on the idea of importance sampling introduced in Chapter 1. The second is based on sampling across a path connecting the prior to the posterior.

### 2.3.1 Importance Sampling Methods

The simplest method of estimating the marginal likelihood of the data given a particular model is a Monte Carlo estimate based on importance sampling, as considered in Section 1.3.1. It has been documented that using the prior as the importance sampling function is an inefficient estimator, especially if the posterior distribution differs greatly from the prior from which the samples are being generated.

#### Sampling from the Posterior

To get around the inefficiency of sampling from the prior, demonstrated in Chapter 3, a common approach is to employ importance sampling. The Monte Carlo estimate using posterior importance sampling is

$$ML_{IS} = \frac{\sum_{i=1}^{S} w_i p(\mathbf{y} \mid \boldsymbol{\theta}^{(i)})}{\sum_{i=1}^{S} w_i} \tag{2.11}$$

where $w_i = p(\boldsymbol{\theta})/\pi^*(\boldsymbol{\theta})$, and the density function $\pi^*(\boldsymbol{\theta})$ is the importance sampling function. (Note that $\pi^*(\boldsymbol{\theta})$ is not strictly required to be a normalised density function). Choosing the importance sampling function to be the posterior, and substituting this into the last equation gives

$$ML_{Posterior} = \left\{ \frac{1}{S} \sum_{i=1}^{S} p(\mathbf{y} \mid \boldsymbol{\theta}^{(i)})^{-1} \right\}^{-1} \tag{2.12}$$

which is the harmonic mean of the likelihood values, where the parameters are sampled from the posterior, $\boldsymbol{\theta} \sim p(\boldsymbol{\theta} \mid \mathbf{y})$. It has been shown that this converges almost surely to the correct value, however it does not always satisfy a central limit theorem, which sometimes manifests itself in the form of unstable results ([64]).

### 2.3.2 Thermodynamic Integration

Thermodynamic integration is also known as path sampling and is based on a more elaborate MCMC sampling scheme ([16, 13]). It is much more computationally expensive than the importance sampling estimators previously described since it requires sampling from intermediate probability distributions at various steps of a temperature ladder. Statistically however it behaves in a much more consistent manner compared to methods involving prior or posterior sampling ([13, 42]), as will be demonstrated in Chapter 3.

Given an unnormalised density, $q(\boldsymbol{\theta})$, the normalised probability density is given by,

$$p(\boldsymbol{\theta}) = \frac{1}{Z} q(\boldsymbol{\theta}) \tag{2.13}$$

where

$$Z = \int_{\boldsymbol{\theta}} q(\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{2.14}$$

is the normalisation constant. Normally in the Bayesian framework we take $q(\boldsymbol{\theta}) = p(\mathbf{y} \mid \boldsymbol{\theta}, H) p(\boldsymbol{\theta} \mid H)$, $Z = p(\mathbf{y} \mid H)$ and $p(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid \mathbf{y}, H)$ for a particular model $H$. In order to calculate the marginal likelihood using thermodynamic integration, however, we define the so-called power posterior,

$$p_t(\boldsymbol{\theta}) = \frac{\{p(\mathbf{y} \mid \boldsymbol{\theta}, H)\}^t \, p(\boldsymbol{\theta} \mid H)}{Z_t} \tag{2.15}$$

so that,

$$Z_t = \int_{\boldsymbol{\theta}} \{p(\mathbf{y} \mid \boldsymbol{\theta}, H)\}^t p(\boldsymbol{\theta} \mid H) d\boldsymbol{\theta} \tag{2.16}$$

We note that when $t = 0$, $Z_t$ is the prior marginalised over $\boldsymbol{\theta}$ which is simply equal to 1, and that when $t = 1$, $Z_t$ is the marginal likelihood. If we therefore consider the log ratio of $Z_1$ and $Z_0$ we see that

$$\begin{aligned}
\log\left(\frac{Z_1}{Z_0}\right) &= \log(Z_1) - \log(Z_0) \\
&= \log\left[\int_{\boldsymbol{\theta}} p(\mathbf{y} \mid \boldsymbol{\theta}, H) p(\boldsymbol{\theta} \mid H) d\boldsymbol{\theta}\right] - \log\left[\int_{\boldsymbol{\theta}} p(\boldsymbol{\theta} \mid H) d\boldsymbol{\theta}\right] \\
&= \log\left[\int_{\boldsymbol{\theta}} p(\mathbf{y} \mid \boldsymbol{\theta}, H) p(\boldsymbol{\theta} \mid H) d\boldsymbol{\theta}\right] \\
&= \log\{p(\mathbf{y})\}
\end{aligned}$$

The following identity is then used to calculate the marginal likelihood, where the expectation is calculated with respect to the power posteriors,

$$\log\{p(\mathbf{y})\} = \log\left(\frac{Z_1}{Z_0}\right) = \int_0^1 E_{\boldsymbol{\theta}|\mathbf{y},t} \log\{p(\mathbf{y} \mid \boldsymbol{\theta})\} dt \tag{2.17}$$

which may be derived as follows,

$$\begin{aligned}
\frac{d}{dt}\log(Z_t) &= \frac{1}{Z_t}\frac{d}{dt}Z_t \\
&= \frac{1}{Z_t}\frac{d}{dt}\int_{\boldsymbol{\theta}} \{p(\mathbf{y} \mid \boldsymbol{\theta})\}^t p(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \frac{1}{Z_t}\int_{\boldsymbol{\theta}} \{p(\mathbf{y} \mid \boldsymbol{\theta})\}^t \log\{p(\mathbf{y} \mid \boldsymbol{\theta})\} p(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \int_{\boldsymbol{\theta}} \frac{\{p(\mathbf{y} \mid \boldsymbol{\theta})\}^t p(\boldsymbol{\theta})}{Z_t} \log\{p(\mathbf{y} \mid \boldsymbol{\theta})\} d\boldsymbol{\theta} \\
&= E_{\boldsymbol{\theta}|\mathbf{y},t} \log\{p(\mathbf{y} \mid \boldsymbol{\theta})\}
\end{aligned}$$

Equation (2.17) follows by integrating with respect to t.

Work in [13] demonstrates just how good an estimator thermodynamic integration is compared to other importance sampling based estimators. This is also shown in ([42]), where thermodynamic integration is used in a phylogenetic context.

## 2.4 Conclusions

In this chapter I have presented a range of methods which may be useful for the purpose of system identification and model inference. I started by describing the difficulties associated with the process of parameter estimation when modelling biochemical networks, and also the challenges of comparing competing model hypotheses. A naive optimisation-based approach to model comparison, using Simulated Annealing and Genetic Algorithms, was presented along with a discussion of its limitations. More advanced methods for parameter inference were then

introduced, in particular Population Markov Chain Monte Carlo, which extends the original Metropolis algorithm using ideas of population and temperature to allow sampling from nonlinear multimodal distributions. An overview was given of importance sampling based methods of exploring a target distribution, and it was noted that many existing algorithms may be viewed as special cases of the more general Sequential Monte Carlo framework. Finally, three methods of calculating marginal likelihoods were given, which are extremely useful for calculating Bayes factors for a more objective form of model comparison, since all the possible parameter values are marginalised. In the next chapter I shall provide a numerical comparison of some of these methods, with a particular focus on how the Population Markov Chain Monte Carlo algorithm may be combined with thermodynamic integration to estimate marginal likelihoods accurately over both linear and nonlinear models.

# Chapter 3

# Population Markov Chain Monte Carlo in Action

This chapter considers the feasibility of employing Bayes factors to discriminate between models with different topologies using various sampling and marginal likelihood estimation methods. Firstly, linear regression models with conjugate priors are investigated in Section 3.1, since they allow for an analytic expression of the marginal likelihood to be calculated. This analytic expression then acts as a benchmark against which we can make an accurate numerical comparison of the various approaches considered in the previous chapter. Poor performance of a sampling method on such a simple statistical model would then cast serious doubt on the suitability of that method for the more demanding application of ODE models. I examine how the number of samples used in the Monte Carlo estimates affects the mean and variance of the end result. In addition, one of the approaches I look at is thermodynamic integration, for which I investigate possible choices of temperature schedules which may be employed, and suggest an optimal scheme in terms of minimising the variance of the estimates produced. In Section 3.2 an example is given of how drastically Metropolis methods of sampling can fail when exploring a multimodal posterior induced by a nonlinear Goodwin oscillator model, the canonical model for describing circadian rhythms, which further motivates the use of more advanced sampling methodology. Finally I show how Population Markov Chain Monte Carlo may be successfully employed to gain estimates of marginal likelihoods, and demonstrate how its ability to sample from multiple modes simultaneously results in the calculation of Bayes factors accurate enough to discriminate between competing model hypotheses described using nonlinear ODEs.

## 3.1 Linear Regression Models

Linear regression models were used to determine the relationship between some response variable $y$ and a set of predictor variables or covariates $\mathbf{x} = (x_1, \ldots, x_d)$, where $d$ is the dimension of the model. General models of the following form were used,

$$g(\mathbf{x}) = \sum_{i=1}^{k} \beta_i B_i(\mathbf{x}) \tag{3.1}$$

so that the function $g$ comprised of a linear combination of basis functions $B_i(\mathbf{x})$ with coefficients $\beta_i$. In particular the responses were assumed to be related to the variables through the relationship

$$y = g(\mathbf{x}) + \epsilon \tag{3.2}$$

where $\epsilon$ is a Gaussian distribution with zero-mean and known variance $\sigma^2$. This can also be written in matrix form,

$$\mathbf{y} = \mathbf{B}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{3.3}$$

where $\mathbf{y} = (y_1, \ldots, y_m)^T$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_m)^T$, and the so-called design matrix

$$\mathbf{B} = \begin{pmatrix} B_1(\mathbf{x}_1) & \ldots & B_k(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ B_1(\mathbf{x}_m) & \ldots & B_k(\mathbf{x}_m) \end{pmatrix} \tag{3.4}$$

For each pair of models, $H_1, H_2$, an "experimental" dataset of $m$ points, $\mathbf{D} = \{y_i, \mathbf{x}_i\}_{i=1}^m$, was produced by one of the linear models by calculating $g(\mathbf{x}_i)$ at some randomly selected positions and adding some noise, $\epsilon$. The two models were then compared by using this "observed"dataset to calculate $P(\mathbf{y} \mid \mathbf{X}, H_n)$, where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]$, from which the Bayes factors could be obtained.

### 3.1.1 Analytic Expressions

#### Priors

A conjugate prior distribution was used so that an analytic expression for the marginal likelihood could be calculated. This was vital so that a benchmark was available for assessing the accuracy of the approximate methods. Independent Gaussian priors centred at zero with variance $\zeta^2$ were placed on each of the unknown parameters $(\beta_1, \ldots, \beta_n)$.

$$\pi(\beta_i) = N(0, \zeta^2) \tag{3.5}$$

and so,

$$\pi(\boldsymbol{\beta}) = \prod_{i=1}^{n} N_{\beta_i}(0, \zeta^2) \tag{3.6}$$

**Likelihood**

The likelihood for a model with a fixed design matrix $\mathbf{B}$ may be written as $p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma)$. Since the errors are normally distributed so that $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, where $\mathbf{I}$ is the identity matrix of dimension $m$, the likelihood function is given by

$$p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-m/2} \exp\left\{ \frac{-(\mathbf{y} - \mathbf{B}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{B}\boldsymbol{\beta})}{2\sigma^2} \right\} \tag{3.7}$$

**Posterior**

Since both the priors and the likelihood function are Gaussian distributions, the posterior is therefore also a Gaussian distribution for which there exists an analytic form. This Gaussian posterior is given by

$$p(\boldsymbol{\beta} \mid \mathbf{X}, \mathbf{y}, \sigma^2, \zeta^2) = N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{3.8}$$

where

$$\begin{aligned} \boldsymbol{\mu} &= \left( \mathbf{B}^T\mathbf{B} + \frac{\sigma^2}{\zeta^2}\mathbf{I} \right)^{-1} \mathbf{B}^T\mathbf{y} \\ \boldsymbol{\Sigma} &= \sigma^2 \left( \mathbf{B}^T\mathbf{B} + \frac{\sigma^2}{\zeta^2}\mathbf{I} \right)^{-1} \end{aligned}$$

From now on, we do not condition explicitly on the covariates $\mathbf{X}$ in every equation for reasons of clarity.

**Marginal Likelihood**

Similarly there is an analytic form for the marginal likelihood, which is also a Gaussian distribution. The marginal likelihood of the experimental data given a particular model, $H$, is given by

$$\begin{aligned} p(\mathbf{y} \mid \sigma^2, \zeta^2, H) &= \int p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2)\pi(\boldsymbol{\beta} \mid \zeta^2) d\boldsymbol{\beta} \qquad (3.9) \\ &= (2\pi)^{-m/2} \left| \sigma^2\mathbf{I} + \zeta^2\mathbf{B}\mathbf{B}^T \right|^{-1/2} \exp\left\{ -\frac{1}{2}\mathbf{y}^T(\sigma^2\mathbf{I} + \zeta^2\mathbf{B}\mathbf{B}^T)^{-1}\mathbf{y} \right\} \end{aligned}$$

Therefore a Bayes factor can be obtained analytically by using the above equation to calculate the marginal likelihood for two competing linear regression models. This analytical Bayes factor can be used as a benchmark against which other methods of estimating marginal likelihoods may be compared.

### Power Posteriors

One of the methods I look at for estimating marginal likelihoods and Bayes factors is thermodynamic integration, which makes use of so-called power posteriors, introduced in Chapter 2. These are the posteriors obtained at each level in the temperature schedule used. The linear regression models that we use also admit an analytic expression for power posteriors. Noting that the log of the likelihood (Equation 3.7) is given by

$$\log\{p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2)\} \quad = \quad -\frac{m}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{B}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{B}\boldsymbol{\beta})$$

The power posteriors, for a particular inverse temperature $t \in [0, 1]$, are simply given by Gaussian distributions

$$p(\boldsymbol{\beta} \mid \mathbf{y}, t, \sigma^2, \zeta^2) = N_{\boldsymbol{\beta}}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t) \tag{3.10}$$

where the mean and covariance matrices are given by

$$\boldsymbol{\mu}_t \quad = \quad \left(\mathbf{B}^T\mathbf{B} + \frac{\sigma^2}{t\zeta^2}\mathbf{I}\right)^{-1}\mathbf{B}^T\mathbf{y} \tag{3.11}$$

$$\boldsymbol{\Sigma}_t \quad = \quad \frac{\sigma^2}{t}\left(\mathbf{B}^T\mathbf{B} + \frac{\sigma^2}{t\zeta^2}\mathbf{I}\right)^{-1} \tag{3.12}$$

The log of the marginal likelihood may be calculated by integrating the expectation of the log of the likelihood with respect to a power posterior over time, $t \in [0, 1]$. This expectation may be calculated analytically making use of the analytic expression for the log likelihood and the power posterior

$$\begin{aligned} & E_{\boldsymbol{\beta}|\mathbf{y},t,\sigma^2,\zeta^2}\left[\log\{p(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2)\}\right] \\ &= \int_{\boldsymbol{\beta}} N_{\boldsymbol{\beta}}(\boldsymbol{\mu}_t, \boldsymbol{\Sigma}_t)\left[-\frac{m}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{B}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{B}\boldsymbol{\beta})\right] d\boldsymbol{\beta} \\ &= -\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{B}\boldsymbol{\mu}_t)^T(\mathbf{y} - \mathbf{B}\boldsymbol{\mu}_t) - \frac{1}{2}Tr(\mathbf{B}^T\mathbf{B}\boldsymbol{\Sigma}_t) - \frac{m}{2}\log(2\pi\sigma^2) \end{aligned} \tag{3.13}$$

Alternatively, we can estimate the above expectation by sampling $\boldsymbol{\beta}^{(j)}$s from each of the power posteriors (using the analytic expression 3.10) and using the following Monte Carlo estimate

$$E_{\boldsymbol{\beta}|\mathbf{y},t,\sigma^2,\zeta^2}\left[\log\{p(\mathbf{y}\mid\boldsymbol{\beta},\sigma^2)\}\right] \approx \frac{1}{N}\sum_{j=1}^{N}\log\left\{p(\mathbf{y}\mid\boldsymbol{\beta}^{(j)},\sigma^2)\right\} \tag{3.14}$$

The integral in equation (2.17) may then be calculated numerically by discretising over the temperature, $t \in [0,1]$, and using the trapezoidal rule with $n$ partitions. So for a discretisation $0 = t_0 < t_1 < \ldots < t_{n-1} < t_n = 1$, an approximation for the log of the marginal likelihood, where we have dropped explicit dependence on $\sigma^2$, $\zeta^2$ and $H$, is given by

$$\log\{p(\mathbf{y})\} \approx \sum_{i=0}^{n-1}(t_{i+1}-t_i)\frac{E_{\boldsymbol{\beta}|\mathbf{y},t_{i+1}}\left[\log\{p(\mathbf{y}\mid\boldsymbol{\beta})\}\right] + E_{\boldsymbol{\beta}|\mathbf{y},t_i}\left[\log\{p(\mathbf{y}\mid\boldsymbol{\beta})\}\right]}{2} \tag{3.15}$$

There are therefore two sources of error in this estimation of the expectations with respect to the marginal likelihood. Firstly there is the Monte Carlo error when estimating the power posteriors themselves, which depends on the number of samples used and the sampler accurately converging to the required stationary distribution. Secondly there is the error in estimating the integral of the power posteriors over $t$, which depends on the number and spacing of the partitions used to discretise the integral. The effects and magnitude of both of these possible errors are investigated in detail.

This discretisation of the unit line need not be uniform and so there are many ways in which the $t_i$s may be chosen, which may affect the error associated with the estimate. By defining a density $p(t)$ over the temperature values we can obtain a density over $t$ which will minimise the Monte-Carlo variance ([16]). Introducing $p(t)$ obtains,

$$\log p(\mathbf{y}) = \int_0^1 \frac{E_{\boldsymbol{\beta}|\mathbf{y},t,\sigma^2,\zeta^2}\left[\log\{p(\mathbf{y}\mid\boldsymbol{\beta},\sigma^2)\}\right]p(t)}{p(t)}dt \tag{3.16}$$

$$= E_{\boldsymbol{\beta},t|\mathbf{y},\sigma^2,\zeta^2}\left[\frac{\log\{p(\mathbf{y}\mid\boldsymbol{\beta},\sigma^2)\}}{p(t)}\right] \tag{3.17}$$

The variance associated with the Monte Carlo estimate of $\log p(\mathbf{y})$ can be minimised by finding the function $p(t)$ which minimises

$$E_{\boldsymbol{\beta},t|\mathbf{y},\sigma^2,\zeta^2}\left[\frac{\log\{p(\mathbf{y}\mid\boldsymbol{\beta},\sigma^2)\}^2}{p(t)^2}\right] = \int_0^1 E_{\boldsymbol{\beta}|\mathbf{y},t,\sigma^2,\zeta^2}\left[\frac{\log\{p(\mathbf{y}\mid\boldsymbol{\beta},\sigma^2)\}^2}{p(t)}\right]dt \tag{3.18}$$

Taking functional derivatives of the following Lagrangian

$$\int_0^1 E_{\boldsymbol{\beta}|\mathbf{y},t,\sigma^2,\zeta^2}\left[\frac{\log\{p(\mathbf{y}\mid\boldsymbol{\beta},\sigma^2)\}^2}{p(t)}\right]dt+\lambda\int_0^1 p(t)dt \tag{3.19}$$

gives

$$p(t)=\frac{p^*(t)}{\int_0^1 p^*(t')dt'} \tag{3.20}$$

where

$$p^*(t)=\sqrt{E_{\boldsymbol{\beta}|\mathbf{y},t,\sigma^2,\zeta^2}\left[\log\{p(\mathbf{y}\mid\boldsymbol{\beta},\sigma^2)\}^2\right]} \tag{3.21}$$

For the linear regression model we may in fact compute $p^*(t)$ analytically, which is proportional to the normalised density function (Equation 3.20). The derivation of this analytic form is rather long and so it is relegated to Appendix A. Thus we may compute $p^*(t)$ for $t\in[0,1]$ and use the results to guide our choice of temperature schedule to minimise the variance of estimates. In particular, the width of temperature partitions should be inversely proportional to the density, $p^*(t)$, so that the regions of greatest mass are most accurately estimated.

### 3.1.2 Experimental Results: Calculating Marginal Likelihoods

In these experiments I used a standard linear model shown in Equation 3.3, where $\mathbf{B}=\mathbf{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_m]^T$, to generate 30 experimental data points. I chose a set of parameters, $\boldsymbol{\beta}$, sampled from the prior distributions of mean 0 and variance $\zeta^2=1$, and added Gaussian noise of variance $\sigma^2=1$, for a variety of models with dimension $d=2,4,6,8,10,15,20$. For each model, $H$, the marginal likelihood, $p(\mathbf{y}\mid H)$, was calculated analytically (see Equation 3.9) using the experimental data points and then estimated using the prior and posterior sampling methods and also thermodynamic integration, as described in Chapter 2. The marginal likelihood was estimated 100 times using each method so that the means and variances could be evaluated and compared. The sample sizes used during the Monte Carlo estimations were also varied from 100 through to 100,000, increasing by a factor of 10 each time, to see how this affected the accuracy of the estimate, although with thermodynamic integration I used only up to 10,000 samples due to computational time limitations. As previously mentioned, another error which appears when using thermodynamic integration is that associated with the temperature schedule. For the purposes of comparison with other methods, the effect of various temperature schedules was examined and the optimal, in terms of smallest variance, was used for all subsequent experiments.

The mean, variance and relative error in the following sections were calculated as follows,

$$\text{Mean} = \frac{1}{s}\sum_{i=1}^{s}\widehat{M}_i \tag{3.22}$$

$$\text{Variance} = \frac{1}{(s-1)}\sum_{i=1}^{s}(\widehat{M}_i - \text{Mean})^2 \tag{3.23}$$

$$\text{Rel. Err.} = \frac{1}{M_{true}}\sqrt{\frac{1}{s}\sum_{i=1}^{s}(\widehat{M}_i - M_{true})^2} \tag{3.24}$$

where $\widehat{M}_i$ is the $i^{th}$ estimate of the marginal likelihood, and $M_{true}$ is the true analytical marginal likelihood. I show that sampling from the posterior generally produces better estimates of the marginal likelihood than sampling from the prior. However the use of thermodynamic integration offers a great improvement in accuracy over both of the importance sampling based methods in terms of lower variance and less bias, as can be seen in Figure 3.1, which provides an overview of the results for a 6 dimensional linear regression model.

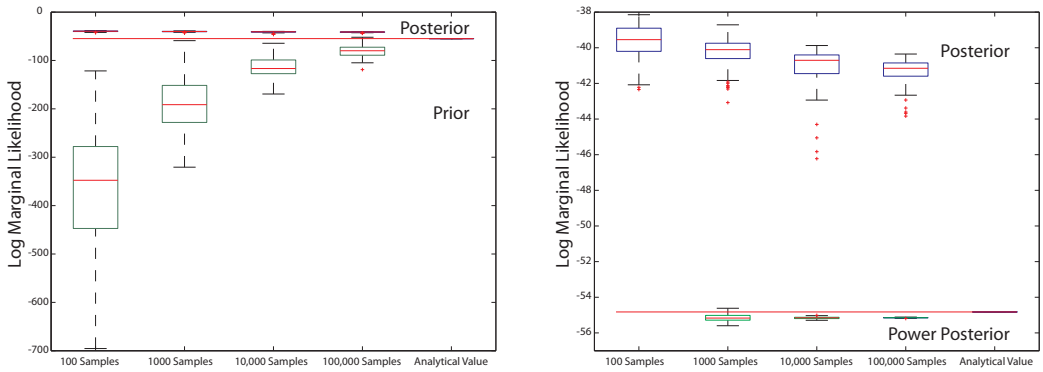


Figure 3.1: Results summary of marginal log likelihood estimation methods for 6 dimensional linear regression model, where the red line indicates the analytic value. In the left hand plot it can be seen that posterior-based estimates of the marginal log likelihood, shown above the red line, have less bias and much tighter variance than those estimated by sampling from the prior, shown below the red line. In the right hand plot, the same posterior-based estimates are displayed using a smaller scale. It is evident that the power posterior-based estimates of the marginal log likelihood are even closer to the analytic value and exhibit less variance than either of the other methods.

### Sampling from the Prior

Table 3.1 shows a comparison of the means and variances of the results generated by sampling from the prior using various sample sizes.

Table 3.1: Marginal log likelihood estimates for linear regression model sampling from prior.

| | 100 Samples | 1000 Samples | 10000 Samples | 100000 Samples | Analytic |
|---|---|---|---|---|---|
| 2D | -75.06 ± 1460 | -49.68 ± 6.39 | -47.97 ± 0.18 | -47.87 ± 0.015 | -47.87 |
| 4D | -171.59 ± 5394 | -81.77 ± 363 | -60.78 ± 36.8 | -53.74 ± 3.42 | -52.49 |
| 6D | -366 ± 15863 | -192 ± 3378 | -114 ± 423 | -80.77 ± 155 | -54.82 |
| 8D | -564 ± 18960 | -336 ± 7742 | -201 ± 2349 | -137 ± 625 | -62.69 |
| 10D | -640 ± 7707 | -417 ± 12088 | -271 ± 3480 | -188 ± 1208 | -67.20 |
| 15D | -692 ± 1477 | -694 ± 815 | -664 ± 6081 | -519 ± 6666 | -79.97 |
| 20D | -695 ± (-) | -698 ± (-) | -698 ± 125 | -672 ± 3340 | -94.05 |

Table 3.2: Marginal log likelihood relative error for linear regression model sampling from prior

| | 100 Samples | 1000 Samples | 10000 Samples | 100000 Samples |
|---|---|---|---|---|
| 2D | 97.6% | 6.47% | 0.90% | 0.26% |
| 4D | 266% | 66.4% | 19.5% | 4.23% |
| 6D | 612% | 272% | 114% | 52.4% |
| 8D | 828% | 457% | 234% | 126% |
| 10D | 862% | 545% | 315% | 186% |
| 15D | 766% | 768% | 737% | 558% |
| 20D | 639% | 641% | 642% | 618% |

The results show that as soon as the number of dimensions increases above four, in terms of the bias and variance the accuracy of the marginal likelihood estimate drastically decreases, even using a large number of samples. It was not feasible to compute the estimate using more than 100,000 samples due to the extremely long running times. From Table 3.1 we see that the variance is not computable for twenty dimensions when there is a small number of samples. This is due to computational limitations, as the calculated probabilities are extremely small. Similarly the relative error, as shown in Table 3.2, greatly increases above four dimensions, even for a large number of samples. An overview is given in Figure 3.2, where it is clear to see that the marginal likelihood estimates become extremely inaccurate as the dimension increases.

### Sampling from the Posterior

It has been previously observed that sampling from the posterior results in an overestimation of the marginal likelihood, for example in the context of calculating

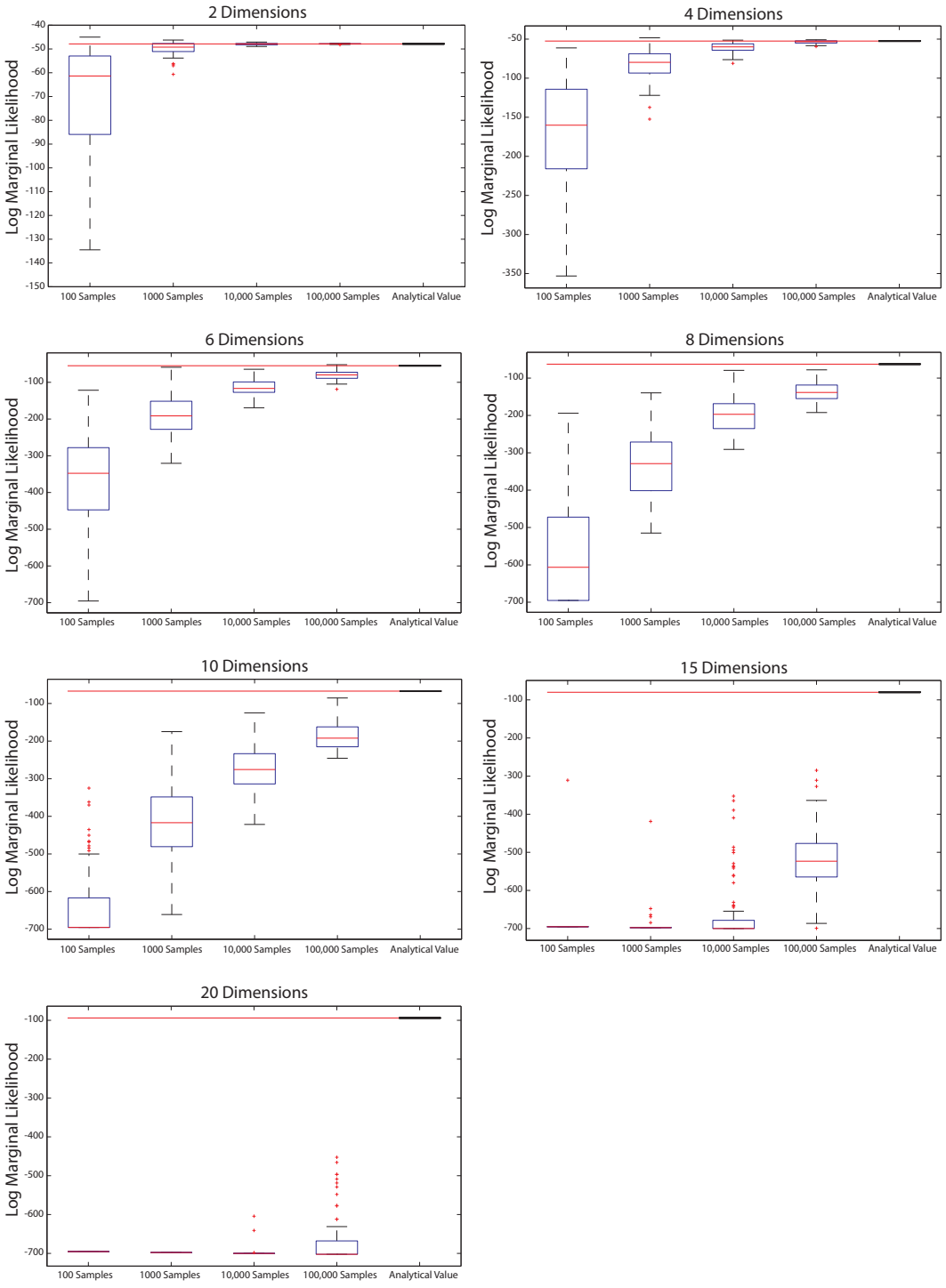


Figure 3.2: Marginal log likelihoods for linear regression model calculated from prior samples. As the number of samples increases, the estimates of the marginal log likelihoods improve as expected. Prior-based estimates provide good results for models of low dimension, however for models of greater than 6 dimensions the estimates exhibit much greater bias and variance. The results are wildly inaccurate for models of 15 and 20 dimensions.

Bayes factors over phylogenetic networks (see [42]), and this is indeed seen to be the case with these results. An important difference, however, is that in ([42]) there was no way to calculate the true marginal likelihood analytically, whereas in this work an analytic marginal likelihood may be calculated using Equation 3.9, allowing for more confident observations to be made regarding the harmonic mean estimates. Table 3.3 shows that the variances and biases are generally much smaller than those of estimates using samples from the prior.

Table 3.3: Marginal log likelihood estimates for linear regression model sampling from posterior

| | 100 Samples | 1000 Samples | 10000 Samples | 100000 Samples | Analytic |
|---|---|---|---|---|---|
| 2D | -41.95 ± 0.47 | -42.21 ± 0.38 | -42.35 ± 0.19 | -42.62 ± 0.34 | -47.87 |
| 4D | -42.69 ± 0.70 | -43.15 ± 0.55 | -43.56 ± 0.41 | -43.84 ± 0.49 | -52.50 |
| 6D | -39.68 ± 0.94 | -40.28 ± 0.69 | -41.09 ± 1.27 | -41.33 ± 0.49 | -54.82 |
| 8D | -43.15 ± 1.26 | -43.93 ± 1.10 | -44.59 ± 1.52 | -45.18 ± 0.63 | -62.69 |
| 10D | -44.12 ± 1.95 | -45.28 ± 1.62 | -46.03 ± 1.60 | -46.54 ± 0.80 | -67.20 |
| 15D | -44.77 ± 3.38 | -45.83 ± 1.90 | -47.08 ± 1.65 | -47.80 ± 1.08 | -79.97 |
| 20D | -48.94 ± 4.11 | -50.28 ± 2.86 | -51.63 ± 1.63 | -52.65 ± 1.14 | -94.05 |

Table 3.4: Marginal log likelihood relative error for linear regression model sampling from posterior

| | 100 Samples | 1000 Samples | 10000 Samples | 100000 Samples |
|---|---|---|---|---|
| 2D | 12.45% | 11.89% | 11.57% | 11.03% |
| 4D | 18.74% | 17.86% | 17.07% | 16.54% |
| 6D | 27.68% | 26.57% | 25.13% | 24.65% |
| 8D | 31.23% | 29.98% | 28.95% | 27.96% |
| 10D | 34.41% | 32.68% | 31.56% | 30.78% |
| 15D | 44.08% | 42.73% | 41.17% | 40.25% |
| 20D | 48.01% | 46.58% | 45.13% | 44.04% |

Although we see that the variance is quite small, at about 1%, the relative error, shown in Table 3.4, starts off fairly large at roughly 10% for 2 dimensions and increases as the number of dimensions increases, albeit not as drastically as in our previous results using prior samples. It is also interesting to note that there is not a great decrease in relative error as the sample size increases. As the harmonic mean estimate tends to consistently overestimate the marginal likelihood, these observations suggest that increasing the sample size has a much greater effect on reducing the variance of estimates than reducing the bias. An overview is given by Figure 3.3.

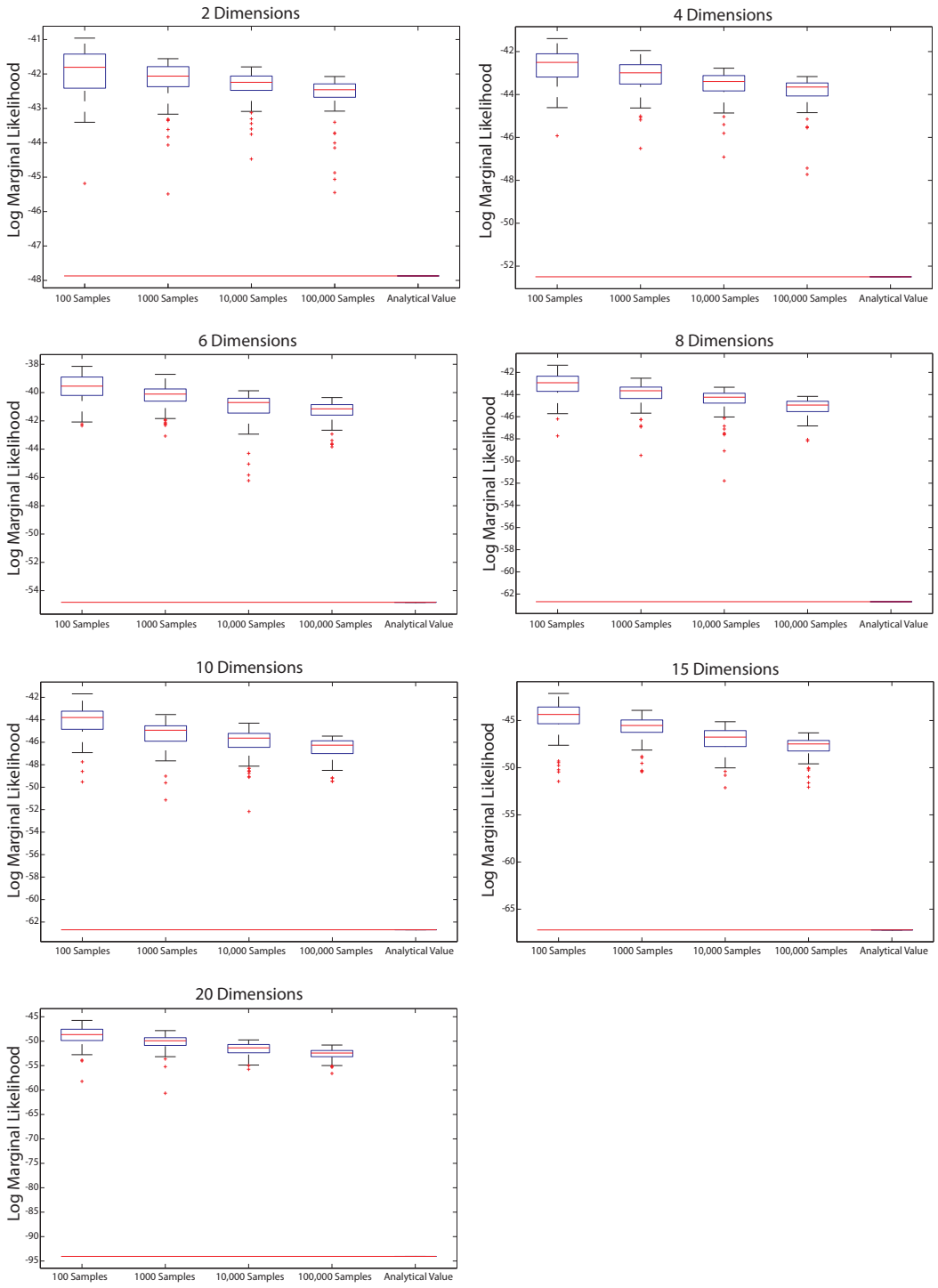


Figure 3.3: Marginal log likelihoods for linear regression model calculated from posterior samples. The bias of these estimates decreases as the number of samples increases, however the variance is not very dependent on the number of samples used, in contrast to the prior-based estimates. Again, as the dimensionality of the model increases, so does the bias in the estimates of the marginal log likelihoods.

### Thermodynamic Integration

First of all I consider which types of temperature schedules should be employed to achieve optimal results in terms of minimising the variance of Monte Carlo estimates of marginal likelihoods. This complements and extends the insights offered by Jasra et al.([32]) who examine various temperature schedules using Population MCMC to sample from mixtures of Gaussians, but only measure the accuracy induced by different spacings by considering how closely the mean parameters for each mixture component are approximated. We may use the analytic expression for the optimal density function (Equation 3.21) to visualise where the bulk of the density lies and in which regions significant changes of density occur. Since we are estimating the log of the marginal likelihood using numerical integration (Equation 3.15), it makes sense to take more estimates near regions of high density, since changes in density correspond directly to changes in the log likelihood. This means that temperature partitions should be narrow in such regions. This also makes sense when looked at from a sampling point of view, since when using Population MCMC to sample from a ladder of temperature distributions we want the transitions between densities to be as smooth as possible to allow for a reasonable acceptance rate for exchange moves, so as to encourage mixing.

Plots proportional to the optimal density functions for linear regression models of varying dimension are shown in Figure 3.4. The shape of these suggest that temperature schedules should be constructed with the intermediate temperature levels very definitely clustered towards $t = 0$, perhaps according to some kind of power law distribution, since this is where the density function most changes shape. In order to investigate whether this holds true in practice, experiments were run using a variety of temperature schedules. For these experiments the log of the marginal likelihood was calculated using numerical integration (Equation 3.15) so as to include errors introduced by the type of partitioning used. In order to exclude any other Monte Carlo errors, the expectations in this equation were calculated analytically (Equation 3.13). Table 3.5 shows the relative error of results from experiments using different partitions for estimating the power posterior integral in 2 dimensions. The number of partitions used was also varied, to see to what extent the accuracy of the estimates increases as the number of partitions used becomes larger.

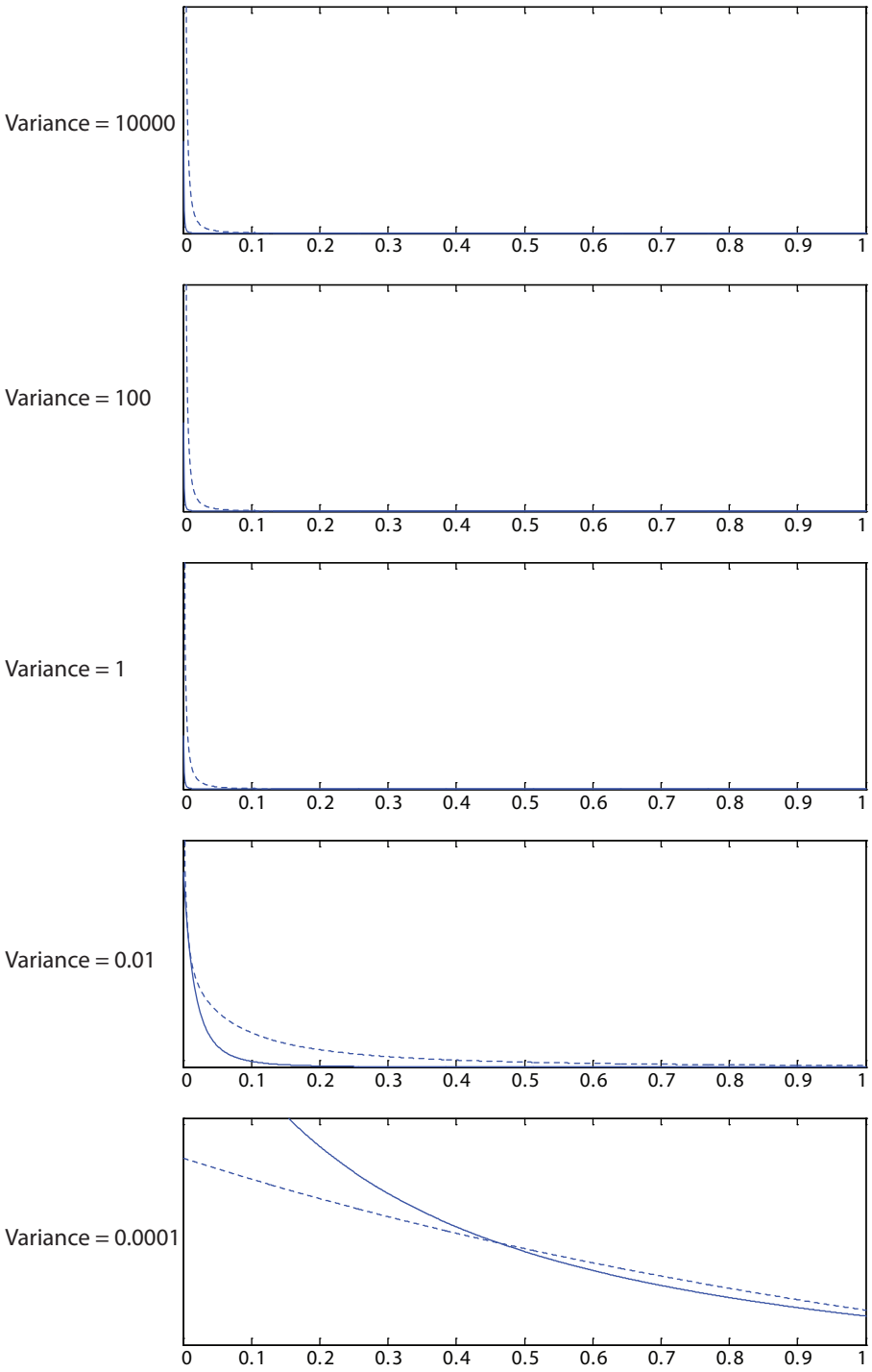


Figure 3.4: Optimal density function $p^*(t)$ plotted against temperature for linear regression model, where the continuous line represents $p^*(t)$ for a 2D model and the dotted line $p^*(t)$ for a 20D model. Notice that as the variance decreases, and the prior confidence increases, the introduction of new information (equivalent to increasing $t$) has less of an effect on the density, which defines the temperature schedule.

Table 3.5: Relative error of partitions estimating a power posterior integral for a 2 dimensional linear regression model

| Method Used | Power Raised | 10 Partitions | 20 Partitions | 30 Partitions | 40 Partitions | 50 Partitions | 60 Partitions | 70 Partitions | 80 Partitions | 90 Partitions | 100 Partitions |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Uniform | 1 | 311% | 151% | 98.9% | 72.7% | 57.1% | 46.7% | 39.3% | 33.8% | 29.5% | 26.2% |
| Centered | 2 | 568% | 296% | 198% | 148% | 118% | 97.2% | 82.7% | 71.7% | 63.2% | 56.4% |
| Centered | 3 | 774% | 425% | 291% | 220% | 176% | 147% | 125% | 109% | 96.7% | 86.6% |
| Extremes | 2 | 58.2% | 12.4% | 4.83% | 2.50% | 1.55% | 1.08% | 0.81% | 0.63% | 0.50% | 0.41% |
| Extremes | 3 | 12.8% | 3.22% | 1.51% | 0.84% | 0.54% | 0.37% | 0.27% | 0.21% | 0.17% | 0.13% |
| Extremes | 4 | 10.2% | 2.54% | 1.12% | 0.63% | 0.40% | 0.28% | 0.20% | 0.16% | 0.12% | 0.10% |
| Extremes | 5 | 11.8% | 2.50% | 1.09% | 0.61% | 0.39% | 0.27% | 0.20% | 0.15% | 0.12% | 0.10% |
| Prior | 2 | 27.3% | 5.54% | 2.19% | 1.21% | 0.79% | 0.56% | 0.42% | 0.33% | 0.26% | 0.21% |
| Prior | 3 | 4.70% | 1.35% | 0.59% | 0.33% | 0.21% | 0.15% | 0.11% | 0.08% | 0.07% | 0.05% |
| Prior | 4 | 3.74% | 0.90% | 0.40% | 0.22% | 0.14% | 0.10% | 0.07% | 0.06% | 0.04% | 0.04% |
| Prior | 5 | 3.40% | 0.82% | 0.36% | 0.20% | 0.13% | 0.09% | 0.07% | 0.05% | 0.04% | 0.03% |
| Prior | 6 | 3.47% | 0.84% | 0.37% | 0.21% | 0.13% | 0.09% | 0.07% | 0.05% | 0.04% | 0.03% |
| Posterior | 2 | 601% | 304% | 201% | 150% | 119% | 98.1% | 83.3% | 72.2% | 63.6% | 56.7% |
| Posterior | 3 | 861% | 448% | 301% | 226% | 180% | 149% | 127% | 111% | 97.8% | 87.5% |

Table 3.6: Relative error of partitions estimating a power posterior integral for a 20 dimensional linear regression model

| Method Used | Power Raised | 10 Partitions | 20 Partitions | 30 Partitions | 40 Partitions | 50 Partitions | 60 Partitions | 70 Partitions | 80 Partitions | 90 Partitions | 100 Partitions |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Uniform | 1 | 600% | 290% | 188% | 138% | 108% | 88.6% | 74.8% | 64.4% | 56.5% | 50.3% |
| Centered | 2 | 1101% | 569% | 379% | 282% | 224% | 185% | 157% | 136% | 120% | 107% |
| Centered | 3 | 1504% | 822% | 559% | 421% | 337% | 280% | 239% | 208% | 184% | 164% |
| Extremes | 2 | 113% | 26.2% | 11.1% | 5.97% | 3.66% | 2.45% | 1.74% | 1.30% | 1.01% | 0.80% |
| Extremes | 3 | 30.9% | 6.16% | 2.74% | 1.59% | 1.03% | 0.71% | 0.52% | 0.40% | 0.32% | 0.26% |
| Extremes | 4 | 21.6% | 5.49% | 2.38% | 1.33% | 0.85% | 0.59% | 0.43% | 0.33% | 0.26% | 0.21% |
| Extremes | 5 | 24.4% | 5.73% | 2.53% | 1.41% | 0.90% | 0.63% | 0.46% | 0.35% | 0.28% | 0.23% |
| Prior | 2 | 54.6% | 12.6% | 5.26% | 2.79% | 1.70% | 1.14% | 0.82% | 0.62% | 0.49% | 0.40% |
| Prior | 3 | 10.4% | 2.48% | 1.14% | 0.64% | 0.41% | 0.28% | 0.21% | 0.16% | 0.13% | 0.10% |
| Prior | 4 | 7.87% | 1.92% | 0.85% | 0.48% | 0.31% | 0.21% | 0.16% | 0.12% | 0.09% | 0.08% |
| Prior | 5 | 7.88% | 1.93% | 0.85% | 0.48% | 0.31% | 0.21% | 0.16% | 0.12% | 0.09% | 0.08% |
| Prior | 6 | 8.74% | 2.13% | 0.94% | 0.53% | 0.34% | 0.23% | 0.17% | 0.13% | 0.10% | 0.08% |
| Posterior | 2 | 1164% | 584% | 386% | 286% | 226% | 187% | 158% | 137% | 121% | 108% |
| Posterior | 3 | 1674% | 867% | 579% | 433% | 344% | 285% | 242% | 211% | 186% | 166% |

The following geometric-based temperature schedules, defining $t_{1,\ldots,N}$, were used for the comparison

Table 3.7: Equations for generating the geometric-based temperature schedules used in the experiments.

| | |
|---|---|
| Uniform: | $t_i = \frac{i}{N}$ |
| Prior: | $t_i = \left(\frac{i}{N}\right)^p$ |
| Posterior | $t_i = 1 - \left(\frac{i}{N}\right)^p$ |

In addition, *Centered* clustered the temperature points around 0.5 and *Extremes* clustered the temperature steps towards both 0 and 1 and away from the middle. Both of these schedules were generated based on scaling and combining points produced by the prior and posterior schedules shown in Table 3.7. Higher powers, $p$ correspond to a more acute clustering of points.

From Table 3.5 it can be seen that methods which cluster more partitions towards $t = 0$, corresponding to the prior, produce lower relative error in the analytic estimates than those which cluster partitions towards $t = 1$, corresponding to the posterior. This matches the prediction made using the optimal density function. Partitions skewed towards the posterior end of the scale performed very badly, indeed much worse than a uniform distribution, as would be expected. Table 3.6 shows similar results but in 20 dimensions. The results are very conclusive; even in 20 dimensions it is possible, using the right temperature schedule, to produce an estimate with a relative error of less than 1% using only 20 partitions of the unit line.

It is interesting to see that using a simple uniform distribution of points to define the temperature partitions produces relatively poor estimates of the marginal log likelihood integral, even for large numbers of partitions. This is in contrast with suggestions made by Jasra et al. [32], who advise that a uniform tempering schedule is generally a good choice when running population-based simulations. There are differences, however, in the criteria used for determining how well a temperature schedule performs, which may account for the drastic difference in conclusions. In [32] the results are drawn on the basis of the resulting estimated component means, whereas in this thesis the results are based on the estimates of the marginal likelihoods. Clearly, it may be possible to have good mean estimates, even if the samples used have quite a high variance, whereas estimates of marginal likelihoods are not as forgiving if the samples used do not accurately cover the regions of high density. In these examples the optimal results are obtained using a power law distribution of temperature points skewed towards 0 and this also makes sense when looked at from a population-based sampling viewpoint. The shape of the power posteriors changes dramatically as the temperature, $t$, moves

from 0 to 0.1, an illustration of which is given in Figure 3.8. For sampling purposes, we wish the changes between adjacent power posteriors to be as smooth as possible to encourage exchange proposals between temperatures to be accepted, as hypothesised before.

Further experiments were then undertaken to calculate the analytic density function $p(t)^*$ (Equation 3.21) for linear regression models using Gaussian priors with different variances. Plots of the results are shown in Figure 3.4.

When the variance is greater than 1, the vague prior covers a large region of the parameter space and the introduction of even a small amount of data, equivalent to a small increase in temperature, results in a large change in the density function. We observe that by setting the variance of the prior to a very small number, we are in effect stating a huge confidence that the chosen restricted region of the parameter space is the most likely. Thus it is no surprise that the introduction of data, equivalent to increasing the temperature, has only limited effects on the density function.

This can be examined from a sampling point of view. Since the density function is based on the log likelihood of the data, vague priors are likely to induce sudden changes in the power posteriors when small amounts of data are introduced. The spacing of the temperature steps should therefore be very small close to $t = 0$, to make exchanges between chains of neighbouring temperatures more likely, in order to encourage mixing. When a sharp prior is employed, adding data has a much smaller effect on the power posteriors, and so mixing between chains will be likely to occur even if the temperature steps are more uniformly distributed. These results highlight the importance the choice of prior plays when deciding on which temperature schedule should be employed. We note that when modelling most kinds of systems, we will rarely be so certain of the expected results as to be able to set such tight priors with variances of less than 0.01. Thus the majority of the time, it is likely that vague, less confident priors will be employed, and so it seems sensible to construct any temperature schedule using a power law distribution with temperature points skewed towards the prior, $t = 0$. Lartillot and Philippe ([42]) use uniform spacing and do not consider this issue at all, and Friel and Pettitt ([13]) give some preliminary discussion on the subject.

Indeed, in the experiments the partition which produced the lowest variance results was the one skewed towards the prior end of the temperature scale (towards zero) and raised to the power 5, and so this was the temperature schedule I employed for the next set of experiments, which focussed on estimating marginal likelihoods over linear regression models using thermodynamic integration and standard Metropolis MCMC.

Table 3.8 shows that the variances using thermodynamic integration are very

low (all less than 0.09). They increase only slightly as the number of dimensions increases and decrease as the number of Monte Carlo samples increases. The relative errors, shown in Table 3.9, start off very low (all less then 0.9%) and, as the number of Monte Carlo samples increases, these relative errors decrease further towards the base error value caused by the partition estimate of the power posterior integral. Thermodynamic integration is seen to be very stable to changes in the number of dimensions of the model. This may also be seen in Table 3.8, which shows that the mean values of the marginal likelihood estimates do not change very much as the number of samples increases; they instead stay fairly constant but with a decreasing variance. An overview is given by the boxplots in Figure 3.5, where a small but systematic bias is clear to see, due to the trapezoidal integration method employed. This issue is discussed in Chapter 4.

Table 3.8: Marginal log likelihood estimates using thermodynamic integration for linear regression model

| | 100 Samples | 1000 Samples | 10000 Samples | Analytic |
|---|---|---|---|---|
| 2D | -48.04 ± 0.0168 | -48.04 ± 0.0013 | -48.04 ± 0.0001 | -47.87 |
| 4D | -52.70 ± 0.0306 | -52.72 ± 0.0025 | -52.71 ± 0.0002 | -52.50 |
| 6D | -55.15 ± 0.0403 | -55.15 ± 0.0032 | -55.15 ± 0.0003 | -54.82 |
| 8D | -63.07 ± 0.0527 | -63.09 ± 0.0036 | -63.08 ± 0.0004 | -62.69 |
| 10D | -67.62 ± 0.0555 | -67.64 ± 0.0049 | -67.64 ± 0.0005 | -67.20 |
| 15D | -80.63 ± 0.0690 | -80.66 ± 0.0080 | -80.66 ± 0.0005 | -79.97 |
| 20D | -94.84 ± 0.0815 | -94.86 ± 0.0089 | -94.86 ± 0.0008 | -94.05 |

Table 3.9: Marginal log likelihood relative errors using thermodynamic integration for linear regression model

| | 100 Samples | 1000 Samples | 10000 Samples |
|---|---|---|---|
| 2D | 0.44% | 0.36% | 0.36% |
| 4D | 0.51% | 0.43% | 0.42% |
| 6D | 0.69% | 0.61% | 0.59% |
| 8D | 0.71% | 0.64% | 0.63% |
| 10D | 0.72% | 0.66% | 0.66% |
| 15D | 0.88% | 0.87% | 0.86% |
| 20D | 0.89% | 0.86% | 0.85% |

The experiments were also run in 50 and 100 dimensions using thermodynamic integration. The results follow the trend of stable means, the variance decreasing as the number of samples increases, and a very low relative error (less than 1.5% in both cases). Importance sampling methods failed in 50 and 100 dimensions due to computational limitations.

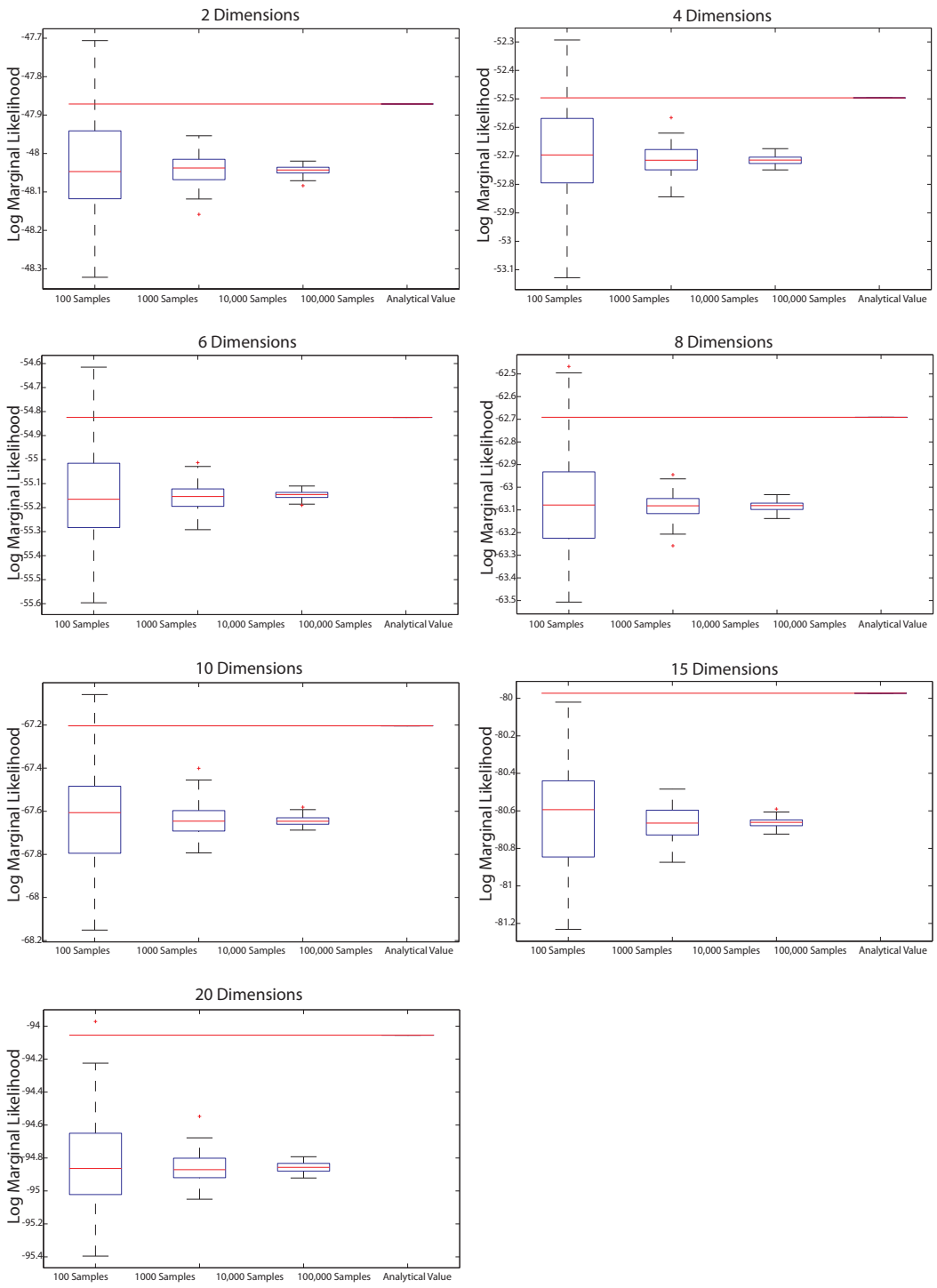


Figure 3.5: Marginal log likelihoods for linear regression model calculated from power posterior samples using 20 temperature steps. In all cases the variances associated with the estimates are less than those produced using posterior and prior-based sampling methods. Most importantly, even for models of 15 and 20 dimensions, estimates of the marginal log likelihood may be obtained with very low variance and bias. The systematic bias observed is due to the numerical integration using a finite number of temperatures.

### 3.1.3 Experimental Results: Calculating Bayes factors

In these experiments we defined two models, generated "experimental" data from one of them, and calculated the Bayes factor 100 times in order to see how accurately we could predict which model produced the data. The Bayes factors were calculated using importance sampling methods and also using thermodynamic integration. The results were then compared to the analytically calculated Bayes factors. The marginal likelihoods were calculated under the same experimental conditions as they were previously in Section 3.1.2. Note that when thermodynamic integration and sampling from the posterior were employed to calculate Bayes factors, only up to 10,000 samples were used due to computational time limitations.

#### Experiment A

Experiment A consisted of two models,

$$\text{Model 1:} \quad y = \beta_1 x_1 + \beta_2 x_2 \tag{3.25}$$

$$\text{Model 2:} \quad y = \beta_1 x_1^2 + \beta_2 x_1 + \beta_3 x_2 \tag{3.26}$$

Bayes factors were first calculated using data generated from the first model given by Equation 3.25, and then using data generated from the second model given by Equation 3.26. The parameter values used for generating the data were sampled from their prior distributions. When model 2 was used to generate data however, the experiments were run varying $\beta_1$ manually in order to simulate a strongly non-linear model (i.e. when $\beta_1 = 1$) and also a more weakly non-linear model (i.e. when $\beta_1 = 0.1$). $\beta_1$ values of 0.15 and 0.16 were also used, as these produced Bayes factors which were not classed as "decisive" and therefore represented cases where the accuracy of the estimate could most affect the interpretation of the evidence. A summary of how Bayes factors should be interpreted was given previously in Table 1.1.

#### Generating Data from Model 1:

Here we see that thermodynamic integration offers the most consistently accurate results compared to the true analytic Bayes factor value of 28.3. Sampling from the prior, see Table 3.10, results in completely uninformative results due to very high variances. When using 100,000 samples the mean Monte Carlo estimate is fairly accurate, although the variance is still high, as is the relative error at 39%.

We have already seen how sampling from the posterior results in an overestimated marginal likelihood. When we calculate Bayes factors using samples from

the posterior, we see that the Bayes factor is massively underestimated, as shown in Table 3.11. Although the variance appears to be very small, the relative error is too large for the results to be informative. Indeed, when interpreted using the standard scale described in Section 1.2.5, the Bayes factor estimates based on sampling from the posterior would suggest that the difference between the two models is "Not worth more than a bare mention", whereas the analytic Bayes factor suggests that the difference between models is in fact "Strong". This result therefore suggests that Bayes factor estimates based on posterior sampling are unable to distinguish between even simple linear models.

Table 3.12 shows the results of using thermodynamic integration with 20 temperature steps. The variance decreases rapidly as the number of Monte Carlo samples used increase, and the relative error decreases to a level which would not influence the interpretation of the Bayes factor.

Table 3.10: Experiment A, Bayes factor results, $B_{1,2}$, sampling from prior

| No. of Samples | 100 | 1000 | 10,000 | 100,000 | Analytic |
|---|---|---|---|---|---|
| Mean | 1.90E+101 | 3.16E+16 | 968 | 30.5 | 28.3 |
| Variance | 2.70E+204 | 9.88E+28 | 46887509 | 118 | - |
| Relative Error | 5.9E+102% | 1E+18% | 24300% | 39% | - |

Table 3.11: Experiment A, Bayes factor results, $B_{1,2}$, sampling from posterior

| No. of Samples | 100 | 1000 | 10,000 | Analytic |
|---|---|---|---|---|
| Mean | 2.25 | 2.39 | 2.52 | 28.3 |
| Variance | 0.07 | 0.06 | 0.04 | - |
| Relative Error | 92% | 92% | 91% | - |

Table 3.12: Experiment A, Bayes factor results, $B_{1,2}$, using thermodynamic integration

| No. of Samples | 100 | 1000 | 10,000 | Analytic |
|---|---|---|---|---|
| Mean | 33.21 | 33.46 | 33.72 | 28.3 |
| Variance | 35.72 | 3.26 | 0.42 | - |
| Relative Error | 27.4% | 19.5% | 19.5% | - |

**Generating Data from Model 2:**

Again thermodynamic integration appears to offer the most accurate results in terms of relative error. Sampling from the prior produced reasonable results, but only when using a very large number of samples. Sampling from the posterior produced very poor results, with the estimated Bayes factors having relative

errors greater than 1000% even when using a large number of samples. For $\beta = 0.1$, the difference between models based on posterior sampling are interpreted as being borderline "Substantial", when in fact it should be "Not worth more than a bare mention". For $\beta = 0.15$, the posterior-based estimates describe the difference between models as "Decisive" instead of merely "Substantial" and, for $\beta = 0.16$, as "Decisive" instead of just "Strong". This reinforces our impression that estimates based on sampling from the posterior should not be blindly trusted.

Table 3.13: Experiment A, Bayes factor results, $B_{2,1}$, for $\beta_1 = 0.1$

| **Sampling from Prior** | | | | | |
|---|---|---|---|---|---|
| No. of Samples | 100 | 1000 | 10,000 | 100,000 | Analytic |
| Mean | 1.67E+02 | 1.74E-01 | 0.166 | 0.150 | 0.156 |
| Variance | 2.39E+06 | 2.01E-01 | 0.018 | 0.002 | - |
| Relative Error | 991410% | 286% | 86.6% | 26.7% | - |
| **Sampling from Posterior** | | | | | |
| No. of Samples | 100 | 1000 | 10,000 | 100,000 | Analytic |
| Mean | 3.38 | 3.00 | 2.69 | - | 0.156 |
| Variance | 0.18 | 0.09 | 0.05 | - | - |
| Relative Error | 2083% | 1837% | 1632% | - | - |
| **Thermodynamic Integration** | | | | | |
| No. of Samples | 100 | 1000 | 10,000 | 100,000 | Analytic |
| Mean | 0.1352 | 0.1315 | 0.1302 | - | 0.156 |
| Variance | 0.00048 | 0.00004 | 0.00001 | - | - |
| Relative Error | 19.35% | 16.20% | 16.60% | - | - |

### Experiment B

Experiment B consisted of two linear models which were compared to evaluate how well this methodology could distinguish between two very similar models.

$$\text{Model 1:} \quad y = \beta_1 x_1 + \beta_2 x_2 \tag{3.27}$$

$$\text{Model 2:} \quad y = \beta_1 + \beta_2 x_1 + \beta_3 x_2 \tag{3.28}$$

Similarly, Bayes factors were first calculated using data generated from the first model given by Equation 3.27, and then using data generated from the second model given by Equation 3.28. The parameter values used for generating the data were again sampled from their prior distributions. When model 2 was used to generate data however, the experiments were run varying $\beta_1$ manually in order

Table 3.14: Experiment A, Bayes factor results, $B_{2,1}$, for $\beta_1 = 0.15$

| **Sampling from Prior** | | | | | |
|---|---|---|---|---|---|
| No. of Samples | 100 | 1000 | 10,000 | 100,000 | Analytic |
| Mean | 14.61 | 9.96 | 5.82 | 6.85 | 6.92 |
| Variance | 6702.88 | 387.75 | 14.88 | 2.07 | - |
| Relative Error | 1183% | 287% | 57.7% | 20.7% | - |
| **Sampling from Posterior** | | | | | |
| No. of Samples | 100 | 1000 | 10,000 | 100,000 | Analytic |
| Mean | 153.46 | 133.75 | 117.90 | - | 6.92 |
| Variance | 372.42 | 180.86 | 98.22 | - | - |
| Relative Error | 2137% | 1844% | 1611% | - | - |
| **Thermodynamic Integration** | | | | | |
| No. of Samples | 100 | 1000 | 10,000 | 100,000 | Analytic |
| Mean | 6.335 | 6.200 | 6.147 | - | 6.92 |
| Variance | 0.855 | 0.069 | 0.009 | - | - |
| Relative Error | 15.74% | 11.05% | 11.23% | - | - |

Table 3.15: Experiment A, Bayes factor results, $B_{2,1}$, for $\beta_1 = 0.16$

| **Sampling from Prior** | | | | | |
|---|---|---|---|---|---|
| No. of Samples | 100 | 1000 | 10,000 | 100,000 | Analytic |
| Mean | 1.50E+02 | 7.59E+01 | 48.6 | 52.4 | 52.0 |
| Variance | 1.82E+06 | 5.39E+04 | 2907 | 272 | - |
| Relative Error | 2588% | 447% | 103.4% | 31.6% | - |
| **Sampling from Posterior** | | | | | |
| No. of Samples | 100 | 1000 | 10,000 | 100,000 | Analytic |
| Mean | 1593 | 1343 | 1154 | - | 52.0 |
| Variance | 42779 | 19432 | 9822 | - | - |
| Relative Error | 2988% | 2496% | 2127% | - | - |
| **Thermodynamic Integration** | | | | | |
| No. of Samples | 100 | 1000 | 10,000 | 100,000 | Analytic |
| Mean | 45.1 | 44.1 | 43.7 | - | 52.0 |
| Variance | 51.7 | 4.09 | 0.58 | - | - |
| Relative Error | 19.10% | 15.63% | 16.00% | - | - |

Table 3.16: Experiment A, Bayes factor results, $B_{2,1}$, for $\beta_1 = 0.17$

| **Sampling from Prior** | | | | | |
|---|---|---|---|---|---|
| No. of Samples | 100 | 1000 | 10,000 | 100,000 | Analytic |
| Mean | 1.29E+08 | 7.74E+06 | 6.13E+06 | 5.84E+06 | 5.80E+06 |
| Variance | 1.38E+18 | 6.75E+14 | 3.00E+13 | 2.63E+12 | - |
| Relative Error | 20289% | 447% | 94.2% | 27.8% | - |
| **Sampling from Posterior** | | | | | |
| No. of Samples | 100 | 1000 | 10,000 | 100,000 | Analytic |
| Mean | 2.96E+08 | 2.54E+08 | 2.20E+08 | - | 5.80E+06 |
| Variance | 1.37E+15 | 6.53E+14 | 3.49E+14 | - | - |
| Relative Error | 5038% | 4300% | 3715% | - | - |
| **Thermodynamic Integration** | | | | | |
| No. of Samples | 100 | 1000 | 10,000 | 100,000 | Analytic |
| Mean | 4.94E+06 | 4.82E+06 | 4.79E+06 | - | 5.80E+06 |
| Variance | 6.08E+11 | 4.57E+10 | 6.60E+09 | - | - |
| Relative Error | 19.92% | 17.29% | 17.51% | - | - |

to simulate a more strongly differing model (i.e. when $\beta_1 = 4$) as well as a weakly differing model (i.e. when $\beta_1 = 2$).

**Generating Data from Model 1:**

Using data generated from model 1, all three methods produce good results based on 10,000 Monte Carlo samples. From Table 3.19, it is clear however that thermodynamic integration outperforms the other two, with a relative error of around just 1.4%. The true analytic Bayes factor is 1.63 which should be interpreted as meaning there is no difference between the models worth mentioning.

Table 3.17: Experiment B, factor results, $B_{1,2}$, sampling from prior

| No. of Samples | 100 | 1000 | 10,000 | 100,000 | Analytic |
|---|---|---|---|---|---|
| Mean | 7.7E+13 | 4.12 | 1.71 | 1.640 | 1.643 |
| Variance | 5.9E+29 | 140 | 0.22 | 0.026 | - |
| Relative Error | 4.7E+16% | 733% | 28.4% | 9.7% | - |

**Generating Data from Model 2:**

Some very interesting results were obtained using data generated from model 2 (Equation 3.28). $\beta_1$ was varied to show how accurate Bayes factors are for strongly and weakly differing models using these methods. Table 3.20 shows that

Table 3.18: Experiment B, Bayes factor results, $B_{1,2}$, sampling from posterior

| No. of Samples | 100 | 1000 | 10,000 | Analytic |
|---|---|---|---|---|
| Mean | 1.602 | 1.610 | 1.616 | 1.643 |
| Variance | 0.033 | 0.024 | 0.014 | - |
| Relative Error | 11.3% | 9.5% | 7.5% | - |

Table 3.19: Experiment B, Bayes factor results, $B_{1,2}$, using thermodynamic integration

| No. of Samples | 100 | 1000 | 10,000 | Analytic |
|---|---|---|---|---|
| Mean | 1.640 | 1.636 | 1.646 | 1.643 |
| Variance | 0.057 | 0.006 | 0.001 | - |
| Relative Error | 14.4% | 4.57% | 1.37% | - |

with $\beta_1$ set to 2, all three methods produced good results for 10,000 samples, with means close to the analytic value of 1.203 and variance low enough as to not affect the interpretation of the Bayes factors calculated, namely that there is no significant difference between the models. For $\beta_1$ set to 3, things begin to get interesting. 100,000 samples from the prior are now required to obtain an estimate with variance low enough as not to change the interpretation of the Bayes factor as "substantial to strong" in favour of model 2. Sampling from the posterior results in an overestimation of the Bayes factor, although it would still suggest "strong" evidence that model 2 is preferred over model 1. Finally, for $\beta_1$ set to 4, sampling from the prior produces a mean value very close to the analytic value, 83.5 compared to 82.5, however, with a variance of 606 little can be inferred from the Bayes factor with any kind of confidence. Posterior sampling once again produces an overestimate of the Bayes factor, describing the evidence as "Decisive" instead of just "Strong", however the variance is so high as to render the result meaningless. This is further evidence that harmonic mean based estimates should not be trusted for estimating Bayes factors. Thermodynamic Integration is the only one of the three methods which calculates a mean value close to the analytic value as well as having a low enough variance, 1.45, that one can confidently interpret the result as being "strong" in favour of model 2.

### 3.1.4 Discussion

When employing prior sampling, estimates of marginal likelihoods are generally poor. For a wide prior the region of high density is relatively small, resulting in fewer samples landing in this region. Therefore the estimates are smaller than analytic values. Additionally, as the dimension increases the region of high density becomes smaller relative to the size of the prior, and the estimates become

Table 3.20: Experiment B, Bayes factor results, $B_{2,1}$, for $\beta_1 = 2$

| **Sampling from Prior** | | | | | |
|---|---|---|---|---|---|
| No. of Samples | 100 | 1000 | 10,000 | 100,000 | Analytic |
| Mean | 2.20E+10 | 2.067 | 1.208 | 1.192 | 1.203 |
| Variance | 4.27E+22 | 9.355 | 0.139 | 0.010 | - |
| Relative Error | 1.72E+13% | 263% | 30.8% | 8.1% | - |
| **Sampling from Posterior** | | | | | |
| No. of Samples | 100 | 1000 | 10,000 | 100,000 | Analytic |
| Mean | 1.334 | 1.298 | 1.267 | - | 1.203 |
| Variance | 0.024 | 0.013 | 0.009 | - | - |
| Relative Error | 16.8% | 12.4% | 9.47% | - | - |
| **Thermodynamic Integration** | | | | | |
| No. of Samples | 100 | 1000 | 10,000 | 100,000 | Analytic |
| Mean | 1.221 | 1.209 | 1.201 | - | 1.203 |
| Variance | 0.0273 | 0.0029 | 0.0003 | - | - |
| Relative Error | 13.75% | 4.46% | 1.33% | - | - |

Table 3.21: Experiment B, Bayes factor results, $B_{2,1}$ for $\beta_1 = 3$

| **Sampling from Prior** | | | | | |
|---|---|---|---|---|---|
| No. of Samples | 100 | 1000 | 10,000 | 100,000 | Analytic |
| Mean | 3.25E+12 | 29.02 | 11.41 | 11.81 | 11.76 |
| Variance | 1.05E+27 | 8847 | 27.53 | 3.79 | - |
| Relative Error | 2.75E+14% | 810% | 44.5% | 16.5% | - |
| **Sampling from Posterior** | | | | | |
| No. of Samples | 100 | 1000 | 10,000 | 100,000 | Analytic |
| Mean | 21.39 | 18.21 | 16.27 | - | 11.76 |
| Variance | 7.48 | 3.46 | 1.74 | - | - |
| Relative Error | 85.1% | 57.1% | 40.0% | - | - |
| **Thermodynamic Integration** | | | | | |
| No. of Samples | 100 | 1000 | 10,000 | 100,000 | Analytic |
| Mean | 11.82 | 11.77 | 11.69 | - | 11.76 |
| Variance | 2.54 | 0.24 | 0.02 | - | - |
| Relative Error | 13.50% | 4.18% | 1.44% | - | - |

Table 3.22: Experiment B, Bayes factor results, $B_{2,1}$, for $\beta_1 = 4$

| **Sampling from Prior** | | | | | |
|---|---|---|---|---|---|
| No. of Samples | 100 | 1000 | 10,000 | 100,000 | Analytic |
| Mean | 5.65E+10 | 166 | 95 | 83.5 | 82.5 |
| Variance | 2.86E+23 | 390462 | 8333 | 606 | - |
| Relative Error | 6.49E+11% | 761% | 111% | 29.7% | - |
| **Sampling from Posterior** | | | | | |
| No. of Samples | 100 | 1000 | 10,000 | 100,000 | Analytic |
| Mean | 226 | 176 | 147 | - | 82.5 |
| Variance | 937 | 379 | 160 | - | - |
| Relative Error | 178% | 116% | 80% | - | - |
| **Thermodynamic Integration** | | | | | |
| No. of Samples | 100 | 1000 | 10,000 | 100,000 | Analytic |
| Mean | 82.46 | 82.38 | 81.80 | - | 82.5 |
| Variance | 135 | 12.22 | 1.45 | - | - |
| Relative Error | 13.99% | 4.22% | 1.67% | - | - |

correspondingly worse.

Estimates based on posterior sampling were seen to be very unstable. The analytic marginal likelihood values used to calculate the Bayes factors were of the order of $10^{-9}$ and so even small errors in the estimation of these had the effect of creating large biases in the Bayes factors, which are calculated as the ratio of two marginal likelihood values. It was observed that generally the larger the true marginal likelihood, the greater the overestimation of the estimated marginal likelihood using posterior sampling. This explains the results in Table 3.16, which show that sampling from the prior actually produces slightly better numerical results than sampling from the posterior, however both methods do produce estimates with so much variance as to render them virtually meaningless. It is known that using this harmonic mean based estimator often results in over-estimations, since every so often a sample with very small likelihood will be chosen and it will have an disproportionate effect on the overall estimate due to the calculation being based on reciprocals. Indeed a very recent paper attempts to circumvent this problem by suggesting the use of a modified harmonic mean based estimator (see [64]) and future work could investigate this method in the context of linear regression models, to see to what extent it alleviates the problems observed when calculating Bayes factors.

In conclusion, although estimators based on prior and posterior importance sampling are unbiased in the limit, the results exhibit strong biases using compu-

tationally feasible number of samples, and this effect becomes more pronounced as the number of dimensions increases.

It was demonstrated that thermodynamic integration performed the best in terms of having the lowest variance and the lowest relative error. It was also shown that the choice of prior may affect the optimal temperature schedule which should be used, in terms of minimising the variance of the resulting Monte Carlo estimates. Generally, for a wide prior a power law distribution should be employed, in which the temperature steps are smaller towards $t = 0$, in order to make a more accurate measurement of the region in which the greatest changes in likelihood take place. By investigating the estimates obtained using a variety of temperature schedules, it was seen that the lowest variance results were indeed produced using those based on the theoretical optimal density function.

In this section, methods of estimating Bayes factors were investigated using simple linear regression models. They are now applied in the following section to the problem of distinguishing between complex nonlinear models of varying dimension.

## 3.2 Nonlinear ODE Models

It is perhaps not surprising that the estimates obtained from the thermodynamic integral (Table 3.8) are so good considering the linear regression model induces relatively simple log-concave posterior densities. When each power posterior $p(\boldsymbol{\theta}|\mathbf{y}, t)$ is multimodal, however, we immediately face the danger of obtaining poor estimates for each $E_{\boldsymbol{\theta}|\mathbf{y},t}[\log p(\mathbf{y}|\boldsymbol{\theta})]$ when using a standard Metropolis method. The conditional posterior surface over two parameters of a 2 variable Goodwin circadian oscillator model is shown in Figure 3.6. This model used was introduced by Goodwin ([22]) and the exact details of its equations and parameter values are given in Appendix B. The nonlinearity of the model results in sharp ridges of high posterior values. Chains sampled using a Metropolis method easily get caught in these local modes, even when engineering techniques, such as an adaptive step size, are employed. Figure 3.7 shows the paths taken by 20 independent Markov chains generated by a Metropolis sampler. Their starting points, indicated by a $\times$, were generated randomly in the parameter space from a prior distribution, and their end positions are denoted by a $\circ$. The localisation of the chains on the ridges is evident from the figure and we will see how this adversely affects the estimation of Bayes factors for the purpose of model comparison in the following sections. As discussed in Chapter 2, recent advances in MCMC methodology suggest a possible solution to this problem in the form of the Population MCMC method, which we shall now see applied to these nonlinear

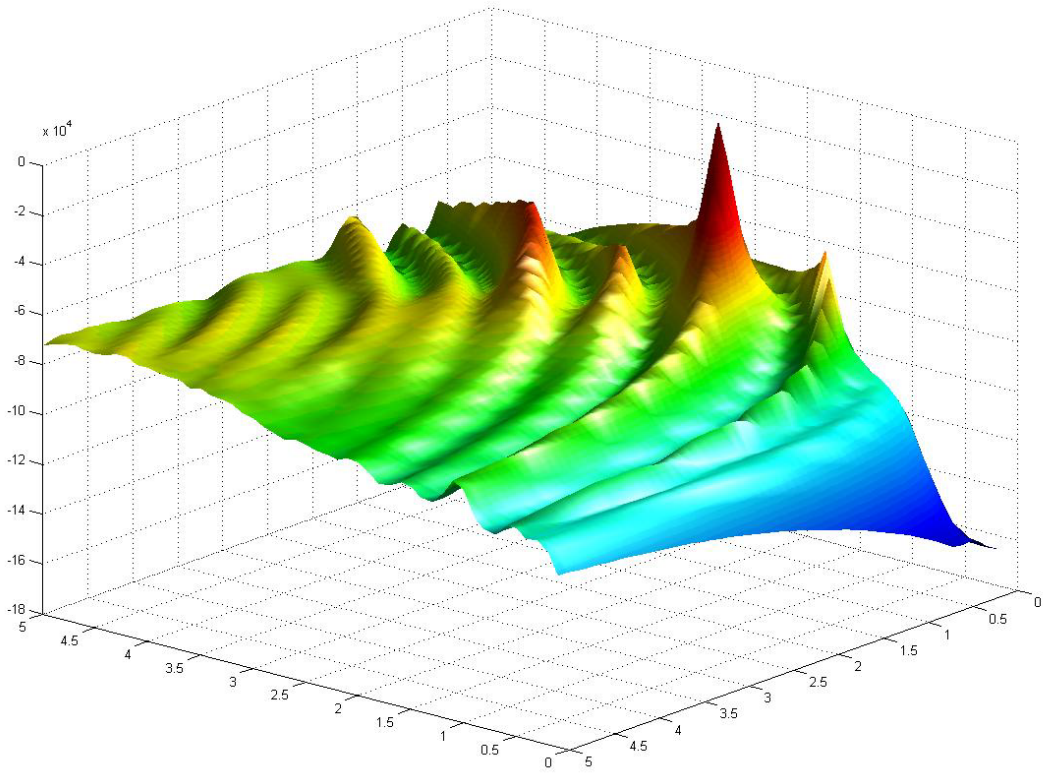


Figure 3.6: Log posterior surface conditioned on two parameters of a 2-variable Goodwin oscillator model. Details for reproducing this plot are given in Appendix B.

Goodwin models.

### Population MCMC

Population-based MCMC enables samples to be drawn from a target density $p(\boldsymbol{\theta}|\mathbf{y})$ by defining a product form of target density indexed by a temperature parameter $t$ such that

$$p(\boldsymbol{\Theta}|\mathbf{y}, \mathbf{t}) = \prod_{n=1}^{N} p(\boldsymbol{\theta}_n|\mathbf{y}, t_n) \tag{3.29}$$

and the desired target density $p(\boldsymbol{\theta}|\mathbf{y})$ is defined for one value of $t_n$. A time homogeneous Markov transition kernel which has $p(\boldsymbol{\theta}|\mathbf{y})$ as its stationary distribution can be constructed from both local proposal moves and global moves between the tempered chains of the population, thus allowing free exploration within the parameter space. Figure 3.7 shows how each of the independent chains of a Metropolis sampler get stuck at various local modes in the posterior density, as they can only make moves within the local parameter space. In contrast, Figures 3.9, 3.10 and 3.11 show three tempered chains at $t = \{0, 0.5, 1\}$ respectively, i.e.

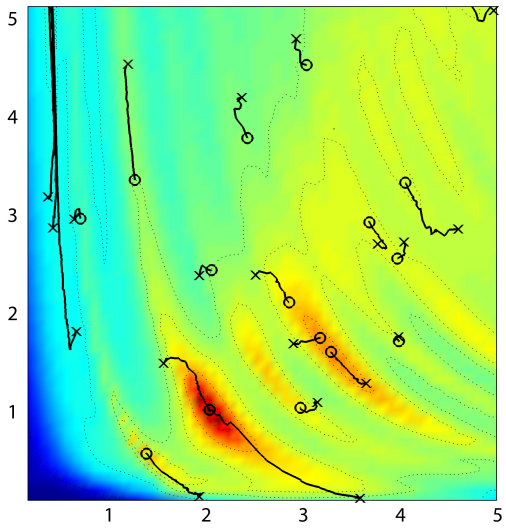


Figure 3.7: The progress of twenty independent Metropolis samplers across the posterior induced by a Goodwin model. The trapping of chains in local modes is most apparent.

ranging from the prior, to an intermediate power-posterior, to the posterior itself, at $t = 1$. At $t = 0$ the samples are drawn from a gamma prior, and thus cover a large area of the parameter space. At the intermediate temperature a free traversal of the parameter space is still possible, while the posterior shows large global mode-hopping steps at $t = 1$. Clearly the estimates of $E_{\boldsymbol{\theta}|\mathbf{y},t}\{\log p(\mathbf{y}|\boldsymbol{\theta})\}$ at each temperature will be superior to those obtained from an independent Metropolis sampler at every temperature, which will be highlighted later in this chapter. Finally, Figure 3.8 shows the conditional power posterior of the 2-variable Goodwin oscillator model at a range of temperatures. Notice that the shapes of the power posteriors change most rapidly between between $t = 0$ and $t = 0.28$, which reinforces the suggestions made in the previous section that the temperature schedule should be skewed towards the prior, $t = 0$.

### The Goodwin Model of Biochemical Oscillatory Control

As an illustrative example of a mechanistic dynamic system and the associated challenges of performing Bayesian inference of model parameters and assessing the validity of alternative model structures we employ models of oscillatory enzymatic control, specifically the Goodwin model ([22]). Note that this model differs from the model given in Appendix B, in that it has a greater number of variables, corresponding to chemical species, which more closely matches the real circadian systems being modelled. Indeed, this model has become the standard basic

mechanism for periodic protein expression, driven by a negative feedback loop which inhibits mRNA transcription. Recent experimental evidence has shown that essential elements of the circadian clock in many organisms consist of negative feedback loops, similar to those in Goodwin's original model. See Section 1.1.2 for a description of the Goodwin model.

It has been shown that this Goodwin model has unstable steady states only when $\rho > 8$, and we therefore set $\rho = 10$ so that we may be certain of oscillatory responses for a wide variety of parameter values. As $n$ increases, so does the time taken for the negative feedback to propagate through the system, enabling a more dynamic range of responses. An $n$-variable Goodwin model therefore has $n + 2$ tunable parameters.

### 3.2.1 Experimental Results

An oscillatory system response, consisting of 80 noisy observations of each of the chemical species made at equally spaced time intervals, was obtained from an $n$-variable Goodwin Model, for $n = \{3, 5\}$ with $x_{1,\ldots,n} = 0$ at time $t = 0$. The observations were made from $t = 40$ to allow the system to settle into a possible steady state from the initial conditions. It is noted that instead of allowing the system time to settle, the initial conditions could alternatively be inferred as additional parameters, although this could potentially increase the complexity of sampling. The specific values of the parameters for both models were drawn from gamma prior distributions with mean 2 and variance 1, and Gaussian noise with variance $\sigma = 0.2$ was added to the observations.

For a particular set of parameters, the error between the model output and the data set was measured using a Normal distribution with variance $\sigma = 0.2$. When using real experimental data, however, the noise variance $\sigma$ would be unknown and could also be inferred as an additional parameter. The overall likelihood was therefore the product of these errors over all data points.

#### Parameter Identification via Posterior Inference

Consider first the problem of model identification by posterior sampling. In the first case, a Metropolis sampler with an adaptive proposal distribution was employed to obtain samples from the posterior. In the second case, a population of ten Metropolis samplers, set along a quintic temperature ladder were used. In addition to standard Metropolis moves, exchange and crossover moves between temperatures were proposed, and these were tuned to ensure an acceptance rate in the range of 30% to 40%. Figure 3.12 shows the estimated marginal posteriors for the $n = 3$ oscillator model obtained using the Population MCMC scheme and

it is clear the regions of highest density are positioned around the actual parameter values. On the other hand the posteriors obtained from standard Metropolis sampling have biased estimates of the posteriors, as can be seen from Figure 3.13.

### Model Comparison using Bayes Factors

Perhaps the most important tool which the Bayesian methodology can offer to computational systems biology is the objective assessment of competing models. Bayes factors were calculated for both Goodwin models, firstly using data generated from the 3 variable model, and then using data generated from the 5 variable model. This allows us to test the discriminating capability of Bayes factors in this setting. The required marginal likelihoods were estimated using power posteriors, with a temperature ladder consisting of 10 discrete steps using a quintic power law spacing. Monte Carlo estimates of the required expectations were obtained using both an adaptive Metropolis sampler and a population MCMC method. Marginal likelihoods were calculated 5 times using each method for each combination of model and data used. Averages and variances were then calculated.

Table 3.23: Bayes Factors & Marginal Log-Likelihoods for Goodwin Models Using Metropolis

| | Simple Data | Complex Data |
|---|---|---|
| Simple Model | $-586 \pm 22,715$ | $-1623 \pm 40,710$ |
| Complex Model | $-782 \pm 116,869$ | $-600 \pm 891,103$ |
| $\log B_{S,C}$ | $195 \pm 205,745$ | - |
| $\log B_{C,S}$ | - | $1022 \pm 802,184$ |

Table 3.24: Bayes Factors & Marginal Log-Likelihoods for Goodwin Models Using Population MCMC

| | Simple Data | Complex Data |
|---|---|---|
| Simple Model | $-426 \pm 31$ | $-1432 \pm 37$ |
| Complex Model | $-536 \pm 67$ | $-190 \pm 47$ |
| $\log B_{S,C}$ | $110 \pm 93$ | - |
| $\log B_{C,S}$ | - | $1242 \pm 117$ |

Convergence of the Markov chains to a stationary distribution was carefully assessed for each sampling method using the Gelman $\hat{R}$ statistic. Normally this statistic is calculated with samples from parallel running chains, however we may also use this on single chains by comparing each 1000 iterations with the previous 1000 iterations to evaluate when the chain has reached an equilibrium. 1000 samples were stored once $\hat{R} < 1.10$ for each parameter at each temperature.

The burn-in time was found to be around 10,000 iterations for the Metropolis method, and 40,000 to 50,000 iterations for the population MCMC method.

Figures 3.14 and 3.15 show the traces obtained from the model using the parameters at the maximum of the inferred posteriors. The original noisy experimental data is also shown in these plots in red.

In Tables 3.23 and 3.24, the 3 variable Goodwin model is referred to as the Simple model, and the 5 variable Goodwin model as the Complex model. From the estimated Bayes factors we observe that the 'true' models can be discriminated, however, the variances of the estimates obtained using only Metropolis sampling at each temperature are enormous (Table 3.23) making these estimates of little practical value when using these for evidential based reasoning. These huge variances resulted from the calculated Bayes factor sometimes favouring the 'true' model and sometimes the 'wrong' model.

The variance of the estimates obtained when inter-chain moves are introduced through the population MCMC procedure are at a hugely reduced level making these low variance estimates such that they can be employed with high confidence when assessing the evidential support in favour of a particular model.

It is also interesting to note that when using the complex data the mean Bayes factor is much higher than when using the simple data. This may explained by examining the predicted model outputs, shown in Figures 3.14 and 3.15. Notice how both models are able to roughly reproduce the simple data, and so the Bayes factor in favour of the simpler model is the result of the complexity of the complex model being penalised. In contrast, the simple model is simply unable to reproduce the complex data, and the much larger Bayes factor in favour of the complex model reflects this.

### 3.2.2 Discussion

In this section I have demonstrated the problems which can occur when trying to sample from a complex posterior distribution using a standard Metropolis sampler. It was seen how multiple independent chains would not converge and got stuck in different areas of the parameter space. This resulted in marginal likelihood and Bayes factor estimates with variances so large that the results were meaningless. In stark contrast, the Population Markov Chain Monte Carlo method produces well mixed samples from each of the required power posteriors and produces Bayes factors which correctly identify the model which the experimental data came from, as well as having low enough variance for the results to be credible. One criticism often made of sampling methods such as Population MCMC, which employ a temperature schedule, is that there are a lot of wasted samples drawn from intermediate distributions. The use of thermodynamic in-

tegration counters this argument by utilising all of the samples from every intermediate distribution to obtain stable estimates of marginal likelihoods, thus minimising computational wastage, and samples from the posterior are automatically obtained at the same time giving estimates of the most likely parameters.

One drawback of using the Population MCMC method is the amount of time that is required for it to run to convergence on such nonlinear models. This is due to the time spent solving many systems of ODEs at every iteration of the algorithm. This will become a greater problem when larger models are considered and motivates further work on both improving the efficiency of the method so that the time to convergence decreases, as well as looking into ways of parallelising the algorithm to take advantages of computer clusters as a way of gaining an increase in speed. In the next chapter I discuss, as an alternative approach, the possible use of Sequential Monte Carlo, which offers a very flexible framework for sampling from complex distributions, and it will be interesting to see what kind of an impact this flexibility will have on its efficiency in such a application.

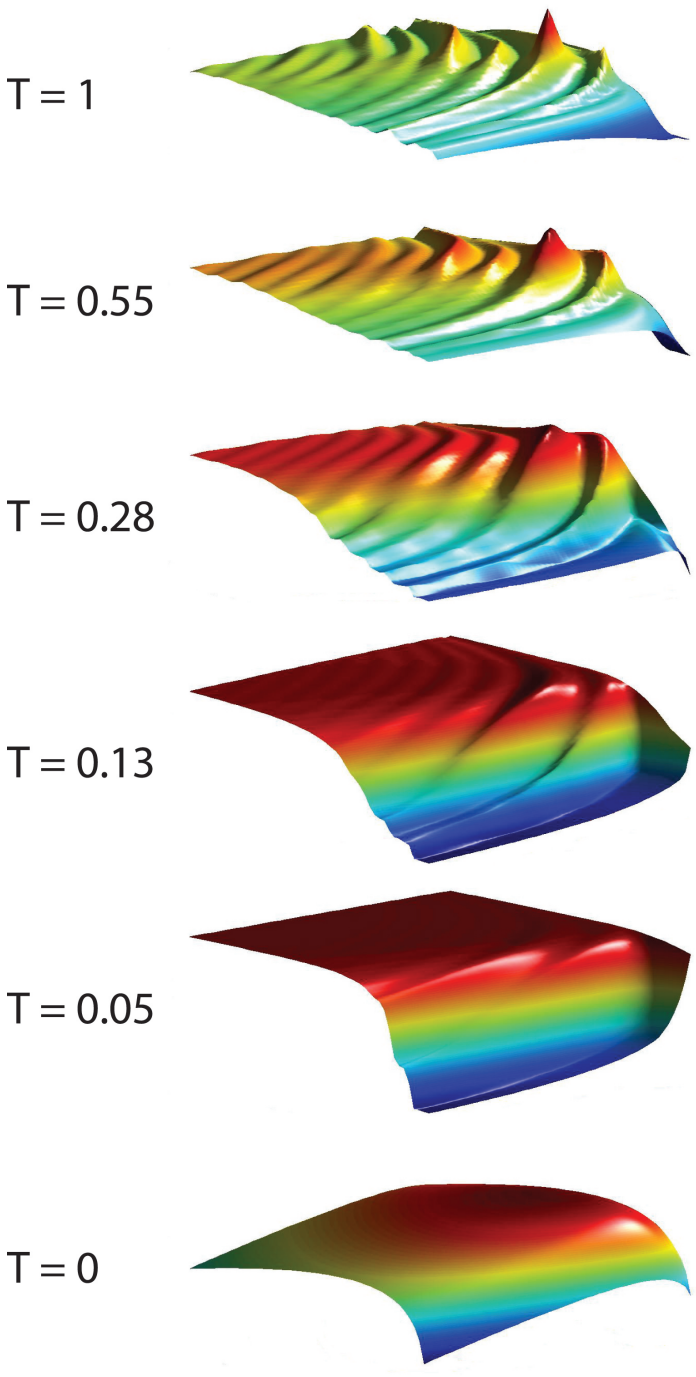


Figure 3.8: Power posterior surfaces conditioned on two parameters of a 2-variable Goodwin oscillator model, details of which are given in Appendix B. The shapes of the power posteriors change most rapidly between between $t = 0$ and $t = 0.28$, and the overall transition from smooth prior to spiky posterior allows chains to globally explore the parameter space through exchanges between temperatures.

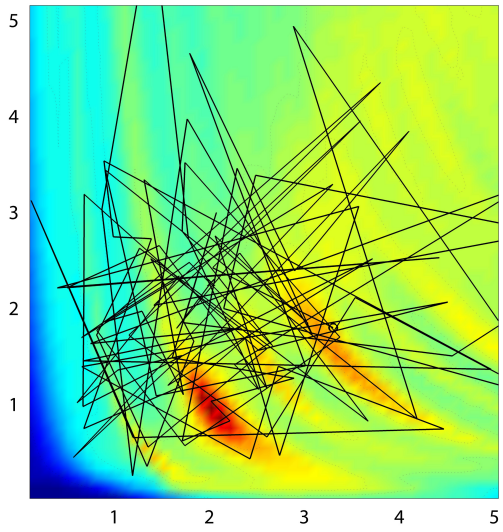


Figure 3.9: Samples obtained from a chain at $t = 0$, which is effectively sampling from the prior. The free movement within the parameter space is clear to see. The iso-contours of the posterior are also plotted in this case.

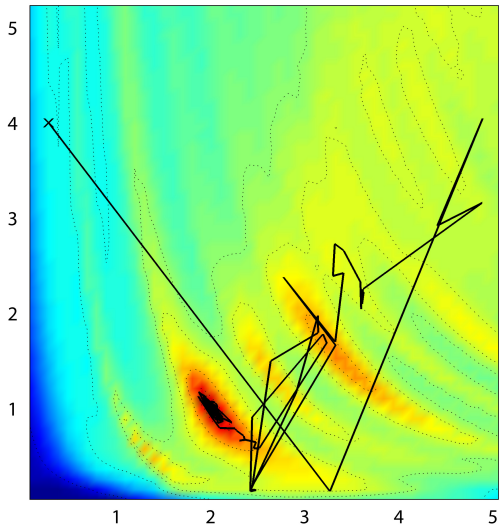


Figure 3.10: Progress of samples drawn from a chain at temperature $t = 0.5$ are shown against the iso-contours of the full posterior. The free movement across modes is most apparent and this is mainly due to the exchange proposals between temperatures.

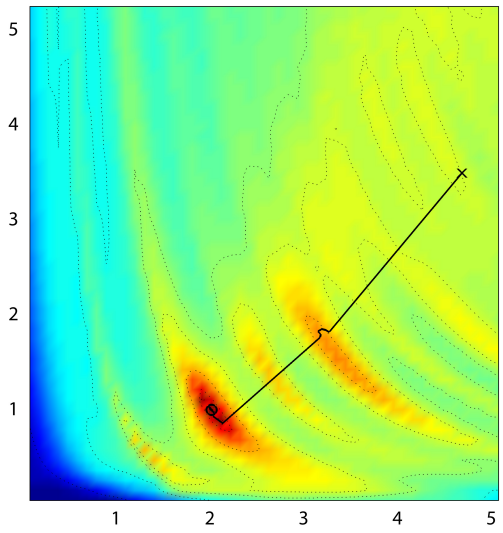


Figure 3.11: Samples drawn from the posterior, when $t = 1$. There are great differences between this and the highly localised *sticky* exploration in Figure 3.7. The Population MCMC algorithm clearly has a much greater ability to move between modes in order to find the most likely one.

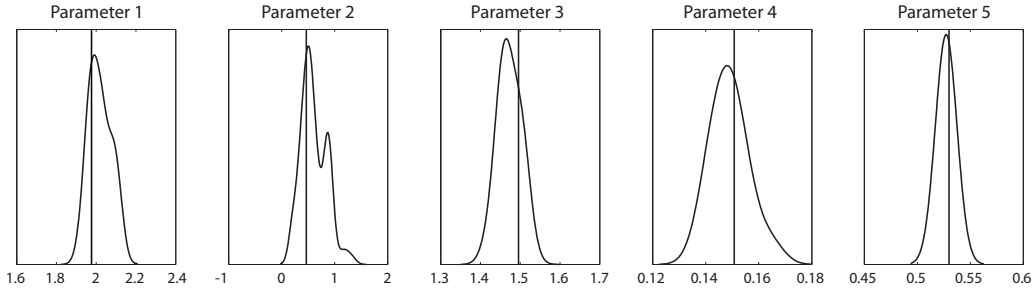


Figure 3.12: The marginal posteriors obtained from population MCMC for each of the parameters of a Goodwin oscillator model. The values of the true parameter values are indicated by a black vertical line which coincides very well with the highest density regions of the posteriors.

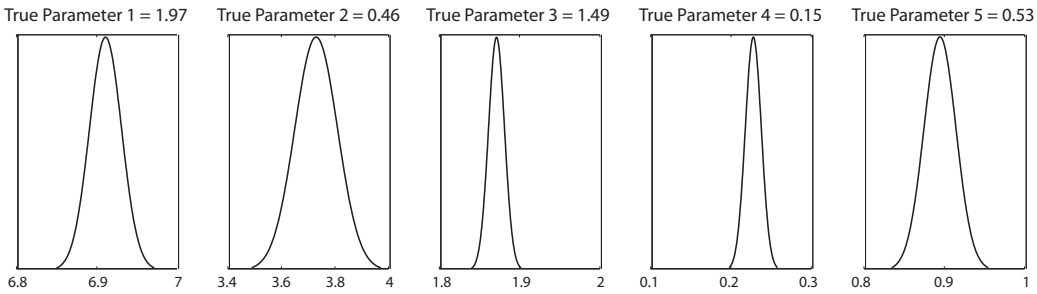


Figure 3.13: The posteriors obtained from a Metropolis sampler with adaptive proposal distributions. The woeful bias in the estimates of the posteriors is most apparent.

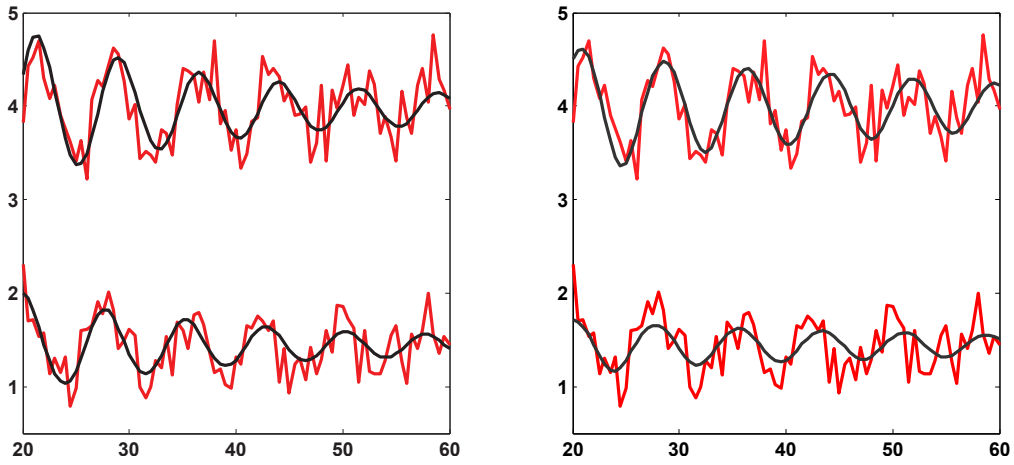


Figure 3.14: Traces obtained using data generated from the 3 variable Goodwin model. The left-hand plot shows the traces using the most likely parameters inferred from the 3 variable Goodwin model. The right-hand plot shows the traces using the most likely parameters inferred from the 5 variable Goodwin model. Experimental data is shown in red and the predicted data in black.

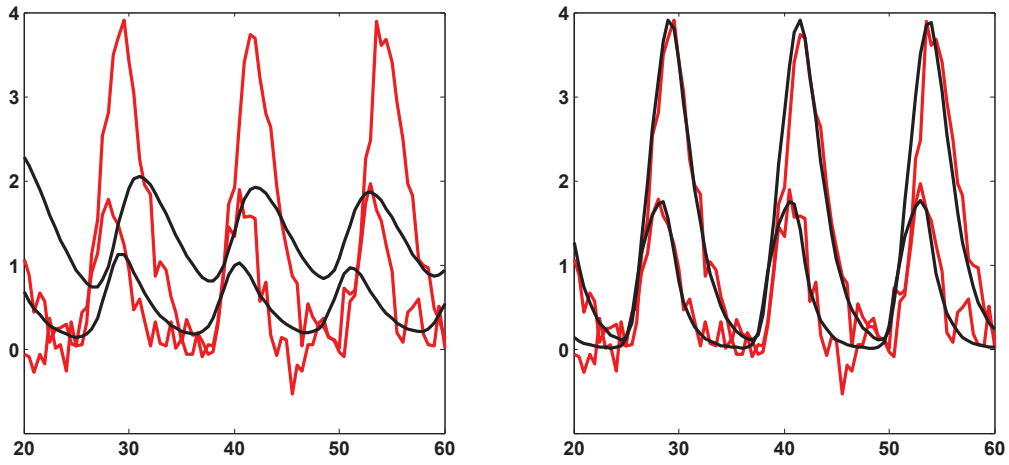


Figure 3.15: Traces obtained using data generated from the 5 variable Goodwin model. The left-hand plot shows the traces using the most likely parameters inferred from the 3 variable Goodwin model. The right-hand plot shows the traces using the most likely parameters inferred from the 5 variable Goodwin model. Experimental data is shown in red and the predicted data in black.

# Chapter 4

# Discussion

The need to perform Bayesian inference over ODE-based models is being driven by advances in systems biology. In this thesis I have investigated the challenges associated with calculating Bayes factors over nonlinear models describing important circadian control processes. We have seen how standard MCMC methodology is inappropriate for such applications, since marginal likelihood estimates based on samples generated from independent chains using a Metropolis algorithm are of such high variance as to render the Bayes factors produced from them useless. An alternative approach was suggested using a combination of Population MCMC and thermodynamic integration, which was shown to produce much lower variance results than other importance sampling based methods of estimating marginal likelihoods.

A comparison of various methods was first made using linear regression models, for which analytic marginal likelihoods could be calculated in order to gain deeper insights into the factors affecting the statistical accuracy of marginal likelihood estimates, before applying them to more complex nonlinear Goodwin style oscillator models. Several methods of calculating marginal likelihoods were compared using varying numbers of samples and temperature partitions on models of increasing dimension. It was shown that thermodynamic integration offered the most stable results, even for models of 20 dimensions using only 20 partitions in the temperature schedule. An analytic expression was then derived for the optimal density function for the temperature profile bridging the prior to the posterior, in terms of minimising the variance of marginal likelihood estimates, which was used to guide the choice of temperature partition spacing. The impact that the spacing in a temperature schedule has on the variance of marginal likelihood estimates was investigated and it was found that power law distributions, with the partitions heavily skewed towards the prior, generally offer consistent results, as predicted by the analytic optimal density function. Finally, two experiments were presented demonstrating how Bayes factors can be used to distinguish

between linear models of varying complexity, and again thermodynamic integration was seen to offer the most stable estimates resulting in meaningful Bayes factors.

I then applied the insights gained from the investigations using linear models to the problem of estimating parameters and Bayes factors over two nonlinear models of varying complexity, using data generated from first the simpler model and then the more complex model. Using oscillatory Goodwin models, commonly employed to build descriptions of circadian rhythms in a wide range of organisms, it was shown how such nonlinear models induce extremely multimodal posterior distributions, and that standard Metropolis samplers fail drastically, even using engineering techniques such as adaptive step size proposals. It was then demonstrated how Population MCMC may successfully be employed to sample from a sequence of distributions between the prior and posterior, with the inferred posterior samples closely approximating the actual parameters which had been used to generate the data. It was shown how the samples obtained at each temperature using Population MCMC could be used to estimate marginal likelihoods, and thus Bayes factors, using thermodynamic integration. Experiments were then presented comparing two nonlinear Goodwin models of varying complexity, which demonstrated how standard Metropolis sampling combined with thermodynamic integration produces estimates of Bayes factors with such high variance as to render the results meaningless. Population MCMC on the other hand produced low variance estimates of Bayes factors using thermodynamic integration, such that the true models could be successfully identified.

### Stochastic Process Models

In Section 1.2.2, the assumption that the noise across consecutive data points is i.i.d. is perhaps not very realistic in a biological setting, especially when considering oscillatory systems. A better way of modelling the noise might then be to define a likelihood function using some kind of stochastic process model, such as a Gaussian process (GP) (see e.g. [69]). A GP produces multiple instances of functions, the means of which are given by some defined underlying function, and the covariance functions model dependencies between time points. The implementation of such a GP introduces added complexity in terms of finding the correct underlying covariance function to describe data with specific characteristics, e.g. oscillatory data with a particular period and amplitude. The parameters describing the covariance function of a GP could also be inferred, along with the other parameters, so that it adapts to the data. This would however add a number of extra dimensions to the space over which the Bayesian inference takes place. For the purposes of this thesis I assumed independence between data points and left

the implementation of GP noise models in this context as future work.

## 4.1 Considerations for Population MCMC and Thermodynamic Integration

The marriage of Population MCMC and thermodynamic integration has the potential to be a very fruitful one. However there are still a number of areas which need to be investigated further, before these methods may be usefully employed to further our knowledge of the circadian system.

### 4.1.1 Scalability of Population MCMC

The two nonlinear Goodwin models considered in Chapter 3 were of 5 and 7 dimensions. Current state of the art models describing circadian networks consist of up to 50 parameters, the majority of which must be estimated without any measurable biological data. Indeed, as mentioned previously, the measurements which are available are likely to have large amounts of variance due to the stochastic effects at a molecular level and other experimental sources of uncertainty. The scalability of Population MCMC for sampling from nonlinear distributions inferred using larger models must therefore be investigated before these methods are able have an impact on the frontier of knowledge in the area of circadian research.

The length of time taken to solve the systems of differential equations which describe a biological process also becomes an important factor as the size of the models increase. Larger models result in longer running times for the algorithms, presenting new computational challenges. One approach is to code the algorithms in a low level compiled language, as a means of increasing speed, however this advantage is still limited by the processing capacity of the computer used. Parallelisation of these sampling algorithms is an attractive option, as computational requirements of simulations could then be spread out over a cluster of computers, which could drastically cut running times. Population based sampling methods appear particularly suited to parallelisation and this approach could become increasingly important in the future as larger models are considered. Another solution might be to use an alternative method of inference which avoids the need to solve the system of ODEs explicitly, and this is discussed in Section 4.3.

### 4.1.2 Thermodynamic Integral Approximation

One issue to be aware of when using thermodynamic integration is the fact that there will be a systematic, albeit small (see Section 3.1.2), bias in the results when

approximating the thermodynamic integral using a finite temperature schedule (Equation 3.15). We have already seen how the variance may be minimised by making use of the expression for the optimal density function to guide our choice of discrete spacing for the temperature schedule, however it would be interesting to investigate the feasibility of also minimising the bias by sampling jointly from the parameters and temperature, i.e. from the distribution $p(\boldsymbol{\theta}, t \mid \mathbf{y})$. This is mentioned in ([16]), and would result in an unbiased estimator of the marginal likelihood, however the feasibility of performing this in practice over high dimensional multimodal distributions remains to be seen.

## 4.2 Alternative Sampling Methods

The computational time required to perform inference could be decreased by making the sampling methods themselves more efficient. For example, new kernels could be developed for the Population MCMC approach adopted in this thesis, or other sampling methods altogether could be adopted. Indeed, there are a couple of alternative sampling methodologies currently available which promise to efficiently sample from multimodal distributions, and these are discussed in the following sections.

### 4.2.1 Sequential Monte Carlo

Sequential Monte Carlo (SMC) offers a general framework for sampling, as already described in Chapter 2, and indeed many algorithms may be seen to be special cases of this method.

The most appealing aspect of the SMC framework is its flexibility. This in itself does not result in an efficient sampling algorithm, but rather it allows efficient samplers to be *designed*. More research is needed into how an efficient sampler may be constructed with respect to choosing an appropriate sequence of target distributions and to taking advantage of the free choice of transition kernels. One interesting idea is to create the sequence of distributions based on increasing amounts of experimental data, as suggested in [6, 7]. The presumption being that the posterior distribution induced from a model given small amounts of data will be less complex and easier to sample from than when the complete dataset is used. It would also be interesting to investigate, for a particular system, the impact of reordering the data when using sequential methods. An artificial sequential reordering of the data would be possible since all the data will already have been collected beforehand, and introducing certain types of data earlier than others might have the effect of restricting the searchable parameter space. This

possible reduction in complexity may present computational advantages, allowing the algorithms to converge in a smaller number of iterations.

Another often-cited advantage of the Sequential Monte Carlo approach is the fact that, when using a Markovian transition function, there is no need to run the population of Markov chains to convergence, since the validity of this framework is based on importance sampling arguments, and therefore independent of any ergodicity properties. This would however obviously have an impact on the variance of the marginal likelihood estimates produced from the non-converging Markov chains. It would be interesting to investigate how the convergence of the chains corresponds to the variance of the resulting marginal likelihood estimates, and perhaps some computational gains would be possible by relaxing the convergence requirements.

### 4.2.2 Nested Sampling

Nested Sampling ([76]) may be used to directly calculate marginal likelihoods, and is based on sampling within a "hard constraint" on the likelihood function, so that the algorithm focusses more on the "nested" shapes of the contours as opposed to constantly changing likelihood values normally produced during a random exploration of the parameter space. Claims about its ability to sample from multimodal distributions without requiring the introduction of any auxilliary variables, such as temperature, sound very appealing. Recent results published in a PhD thesis by Murray ([58]), however, suggest that there is very little difference in performance over mixtures of Gaussian models when compared to some temperature based sampling methods, such as Annealed Importance Sampling, raising the question of whether there is indeed anything to be gained by employing such a nested sampling approach. Another question meriting investigation, is whether Nested Sampling would be able to cope with the highly nonlinear posterior distributions induced by the types of ODE models commonly used to describe complex biological processes, and so it would be interesting to examine this potentially useful method in a systems biology context.

## 4.3 Alternative Methods of Inference

The main computational cost of sampling from distributions induced by nonlinear models is incurred solving the systems of ODEs for each proposed set of parameters. As mentioned previously, one possible solution to this is the parallelisation of the sampling algorithm, allowing the computational cost to be spread accross multiple computers.

Another solution is to infer the parameters using the time derivatives described by the system of ODEs. Such *collocation methods* (see e.g. [65]) can be used to avoid the computationally expensive requirement of explicitly solving systems of ODEs in order to obtain the posterior $P(\boldsymbol{\theta} \mid \mathbf{y})$.

For example, as mentioned in Section 4, a Gaussian Process (GP) may be used as a likelihood function to model some experimental data $\mathbf{y}$ with dependent noise. These experimental observations at $T$ discrete time points are represented by $\mathbf{y}(t) = \mathbf{x}(t) + \boldsymbol{\epsilon}(t)$, where $\mathbf{x}(t) = [x_1(t), \ldots, x_N(t)]$ represents the levels of each of the $N$ chemical species present in the system at time $t$, and $\boldsymbol{\epsilon}$ is an appropriate noise process with some variance $\sigma$. By denoting the time courses for the $N$ chemical species as the $N \times T$ matrix $\mathbf{X}$, and the experimental observations for the $N$ chemical species as the $N \times T$ matrix $\mathbf{Y}$, we may place a GP prior, which has a covariance function with parameters $\varphi$, over the time course of each chemical species so that $\mathbf{X}_{n,\cdot} \sim GP(\varphi)$. The dynamics of $N$ chemical species may be modelled by a system of ODEs such that $\dot{\mathbf{X}}_{\cdot,t} = \mathbf{f}(\mathbf{X}_{\cdot,t}, \boldsymbol{\theta}, t)$. The posterior $p(\mathbf{X}_{n,\cdot} \mid \mathbf{Y}_{n,\cdot}, \sigma, \varphi)$ is therefore also a GP of the standard form, indeed samples may be obtained from the conditional posterior $p(\mathbf{X}_{n,\cdot}, \sigma, \varphi_n \mid \mathbf{Y}_{n,\cdot}, \dot{\mathbf{X}}_{n,\cdot})$ in the usual manner.

We will also obtain a posterior of the time derivatives of the levels of the chemical species, $p(\dot{\mathbf{X}}_{n,\cdot} \mid \mathbf{X}_{n,\cdot}, \boldsymbol{\theta}, \gamma)$. This then allows us to define a posterior over the parameters of the system $\boldsymbol{\theta}$ in terms of the time derivatives described by our system of ODEs, $\dot{\mathbf{X}}_{\cdot,t} = \mathbf{f}(\mathbf{X}_{\cdot,t}, \boldsymbol{\theta}, t)$. Assuming Normal errors with variance $\gamma$ and some prior over the parameters $\pi(\boldsymbol{\theta})$, the posterior over the parameters may be written as

$$p(\boldsymbol{\theta} \mid \mathbf{Y}, \dot{\mathbf{X}}, \mathbf{X}, \gamma) \propto \exp\left\{-\frac{1}{2\gamma}\sum_{t=1}^{T}\left|\dot{\mathbf{X}}_{\cdot,t} - \mathbf{f}(\mathbf{X}_{\cdot,t}, \boldsymbol{\theta}, t)\right|^2\right\}\pi(\boldsymbol{\theta})$$

Therefore samples from the joint posterior $p(\boldsymbol{\theta}, \dot{\mathbf{X}}, \mathbf{X}, \gamma, \varphi, \sigma \mid \mathbf{Y})$ can be obtained by a Metropolis within Gibbs routine, ignoring details of hyper-parameters, so that

$$\begin{aligned}
\mathbf{X}_{n,\cdot} &\sim p(\mathbf{X}_{n,\cdot} \mid \mathbf{Y}_{n,\cdot}, \dot{\mathbf{X}}_{n,\cdot}) \\
\dot{\mathbf{X}}_{\cdot,t} &\sim p(\dot{\mathbf{X}}_{\cdot,t} \mid \mathbf{X}_{\cdot,t}, \boldsymbol{\theta}, \gamma) \\
\boldsymbol{\theta} &\sim p(\boldsymbol{\theta} \mid \mathbf{Y}, \dot{\mathbf{X}}, \mathbf{X}, \gamma)
\end{aligned}$$

where $p(\mathbf{X}_{n,\cdot} \mid \mathbf{Y}_{n,\cdot}, \dot{\mathbf{X}}_{n,\cdot})$ is a conditional predictive posterior GP, and $p(\dot{\mathbf{X}}_{\cdot,t} \mid \mathbf{X}_{\cdot,t}, \boldsymbol{\theta}) = N_{\dot{\mathbf{X}}_{\cdot,t}}(f(\mathbf{X}_{\cdot,t}, \boldsymbol{\theta}, t), \gamma\mathbf{I})$. Therefore $\boldsymbol{\theta}$ may be sampled using some Markov Chain Monte Carlo method, such as Popoulation MCMC, without having to explicitly integrate the system of ODEs at each iteration. The computational

costs involved in this approach consist of sampling each $\mathbf{X}$ and $\dot{\mathbf{X}}$, as well as the hyper-parameters, but this will be dominated by scaling of order $O(NT^3)$, and when $T$ is relatively small this may prove to be substantially faster than explicitly solving a system of $N$ ODEs.

This is a very exciting approach which has the potential to drastically speed up parameter inference over large nonlinear models. There may well be challenges to overcome in a practical implementation of this method, possibly regarding a loss of information in the observations as we are inferring the parameters based on the time derivatives as opposed to just the observations themselves, although I believe the approach is a very promising one.

## 4.4 Conclusions

This thesis has focussed on investigating how Bayes factors can be accurately estimated for nonlinear ODE-based models, such as those commonly used to describe circadian control. As described in this chapter, there are many exciting possible avenues of research still to be explored, and while there is still much work to be done before these methods may be usefully applied to extending state of the art models using real experimental data, such methodology has the potential to have a great impact on the area of systems biology in the near future.

# Appendix A

# Derivation of Optimal Density for Temperature Schedule

Here I derive an analytic expression for Equation 3.21. This equation is directly proportional to the optimal density function, $p(t)$, introduced when investigating how to minimise the variance of marginal likelihood estimates for linear regression models using thermodynamic integration. This expression may therefore be used to choose the optimal distribution of points in a temperature schedule, by concentrating them around on the regions of highest mass. I make use of the following identities[1] for the expectation operator, where $\boldsymbol{\beta}$ is a stochastic vector drawn from a Gaussian distribution with mean $\boldsymbol{\mu}$, and covariance $\boldsymbol{\Sigma}$

$$E[\mathbf{A}\boldsymbol{\beta}+\mathbf{b}] = \mathbf{A}\boldsymbol{\mu}+\mathbf{b} \tag{A.1}$$

$$E\left[(\mathbf{A}\boldsymbol{\beta}+\mathbf{a})(\mathbf{B}\boldsymbol{\beta}+\mathbf{b})^T\right] = \mathbf{A}\boldsymbol{\Sigma}\mathbf{B}^T+(\mathbf{A}\boldsymbol{\mu}+\mathbf{a})(\mathbf{B}\boldsymbol{\mu}+\mathbf{b})^T \tag{A.2}$$

$$E\left[\boldsymbol{\beta}^T\mathbf{A}\boldsymbol{\beta}\right] = Tr(\mathbf{A}\boldsymbol{\Sigma})+\boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} \tag{A.3}$$

$$\begin{aligned} E\left[(\mathbf{A}\boldsymbol{\beta}+\mathbf{a})(\mathbf{A}\boldsymbol{\beta}+\mathbf{a})^T(\mathbf{A}\boldsymbol{\beta}+\mathbf{a})\right] &= \left(2\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T+(\mathbf{A}\boldsymbol{\mu}+\mathbf{a})(\mathbf{A}\boldsymbol{\mu}+\mathbf{a})^T\right)(\mathbf{A}\boldsymbol{\mu}+\mathbf{a}) \\ &\quad +Tr(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)\times(\mathbf{A}\boldsymbol{\mu}+\mathbf{a}) \end{aligned} \tag{A.4}$$

$$\begin{aligned} &E\left[(\mathbf{A}\boldsymbol{\beta}+\mathbf{a})^T(\mathbf{B}\boldsymbol{\beta}+\mathbf{b})(\mathbf{C}\boldsymbol{\beta}+\mathbf{c})^T(\mathbf{D}\boldsymbol{\beta}+\mathbf{d})\right] \\ &= Tr\left(\mathbf{A}\boldsymbol{\Sigma}(\mathbf{C}^T\mathbf{D}+\mathbf{D}^T\mathbf{C})\boldsymbol{\Sigma}\mathbf{B}^T\right) \\ &\quad +\left((\mathbf{A}\boldsymbol{\mu}+\mathbf{a})^T\mathbf{B}+(\mathbf{B}\boldsymbol{\mu}+\mathbf{b})^T\mathbf{A}\right)\boldsymbol{\Sigma}\left(\mathbf{C}^T(\mathbf{D}\boldsymbol{\mu}+\mathbf{d})+\mathbf{D}^T(\mathbf{C}\boldsymbol{\mu}+\mathbf{c})\right) \\ &\quad +\left(Tr(\mathbf{A}\boldsymbol{\Sigma}\mathbf{B}^T)+(\mathbf{A}\boldsymbol{\mu}+\mathbf{a})^T(\mathbf{B}\boldsymbol{\mu}+\mathbf{b})\right)\left(Tr(\mathbf{C}\boldsymbol{\Sigma}\mathbf{D}^T)+(\mathbf{C}\boldsymbol{\mu}+\mathbf{c})^T(\mathbf{D}\boldsymbol{\mu}+\mathbf{d})\right) \end{aligned} \tag{A.5}$$

We wish to find an analytic expression for the following expectation (A.6) with respect to a power posterior distribution for a particular temperature. For the linear regression model considered in Chapter 3, the power posterior distributions are Gaussian, with mean $\boldsymbol{\mu}$, and covariance $\boldsymbol{\Sigma}$ (see equations 3.11, 3.12). We proceed by first multiplying out the brackets and noting that the expectation operator is linear

[1] See The Matrix Reference Manual, http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/intro.html, 2005.

$$
\begin{aligned}
& E\left[\left(-\frac{m}{2}\log 2\pi\sigma^2 - \frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{B}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{B}\boldsymbol{\beta})\right)^2\right] \qquad \text{(A.6)} \\
&= \frac{m^2}{4}(\log 2\pi\sigma^2)^2 + E\left[\frac{m}{2\sigma^2}\log 2\pi\sigma^2(\mathbf{y}-\mathbf{B}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{B}\boldsymbol{\beta})\right] \\
&\quad + E\left[\frac{1}{4\sigma^4}(\mathbf{y}-\mathbf{B}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{B}\boldsymbol{\beta})(\mathbf{y}-\mathbf{B}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{B}\boldsymbol{\beta})\right]
\end{aligned}
$$

An analytic expression for the second term in A.6 may be found using identity A.2

$$
\begin{aligned}
& E\left[\frac{m}{2\sigma^2}\log 2\pi\sigma^2(\mathbf{y}-\mathbf{B}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{B}\boldsymbol{\beta})\right] \\
&\quad = \frac{m}{2\sigma^2}\log 2\pi\sigma^2\left[Tr(\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T) + (\mathbf{y}-\mathbf{B}\boldsymbol{\mu})^T(\mathbf{y}-\mathbf{B}\boldsymbol{\mu})\right]
\end{aligned}
$$

The third term also has an analytic form, however a bit more work is required to calculate it. We start by multiplying out the middle two brackets and then multiplying the result by the outer two brackets, which splits the third term down into the following three expressions

$$
\begin{aligned}
& \frac{1}{4\sigma^4}E\left[(\mathbf{y}-\mathbf{B}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{B}\boldsymbol{\beta})(\mathbf{y}-\mathbf{B}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{B}\boldsymbol{\beta})\right] \\
&= \frac{1}{4\sigma^4}E\left[(\mathbf{y}-\mathbf{B}\boldsymbol{\beta})^T[\mathbf{y}\mathbf{y}^T - 2\mathbf{B}\boldsymbol{\beta}\mathbf{y}^T + \mathbf{B}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{B}^T](\mathbf{y}-\mathbf{B}\boldsymbol{\beta})\right] \\
&= \frac{1}{4\sigma^4}E\Big[\underbrace{(\mathbf{y}-\mathbf{B}\boldsymbol{\beta})^T\mathbf{y}\mathbf{y}^T(\mathbf{y}-\mathbf{B}\boldsymbol{\beta})}_{\text{Expression 1}} \\
&\qquad - \underbrace{2(\mathbf{y}-\mathbf{B}\boldsymbol{\beta})^T\mathbf{B}\boldsymbol{\beta}\mathbf{y}^T(\mathbf{y}-\mathbf{B}\boldsymbol{\beta})}_{\text{Expression 2}} \\
&\qquad + \underbrace{(\mathbf{y}-\mathbf{B}\boldsymbol{\beta})^T\mathbf{B}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{B}^T(\mathbf{y}-\mathbf{B}\boldsymbol{\beta})}_{\text{Expression 3}}\Big]
\end{aligned}
$$

The expectation of Expression 1 may be calculated by multiplying out the brackets and using the identities A.1 and A.2

$$
\begin{aligned}
& E\left[(\mathbf{y}-\mathbf{B}\boldsymbol{\beta})^T\mathbf{y}\mathbf{y}^T(\mathbf{y}-\mathbf{B}\boldsymbol{\beta})\right] \\
&= E\left[(\mathbf{y}^T\mathbf{y}\mathbf{y}^T - \boldsymbol{\beta}^T\mathbf{B}^T\mathbf{y}\mathbf{y}^T)(\mathbf{y}-\mathbf{B}\boldsymbol{\beta})\right] \\
&= E\left[\mathbf{y}^T\mathbf{y}\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{y}\mathbf{y}^T\mathbf{B}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{B}^T\mathbf{y}\mathbf{y}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{B}^T\mathbf{y}\mathbf{y}^T\mathbf{B}\boldsymbol{\beta}\right] \\
&= (\mathbf{y}^T\mathbf{y})^2 - 2E\left[\mathbf{y}^T\mathbf{y}\mathbf{y}^T\mathbf{B}\boldsymbol{\beta}\right] + E\left[\boldsymbol{\beta}^T\mathbf{B}^T\mathbf{y}\mathbf{y}^T\mathbf{B}\boldsymbol{\beta}\right] \\
&= (\mathbf{y}^T\mathbf{y})^2 - 2\mathbf{y}^T\mathbf{y}\mathbf{y}^T\mathbf{B}\boldsymbol{\mu} + Tr(\mathbf{B}^T\mathbf{y}\mathbf{y}^T\mathbf{B}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T\mathbf{B}^T\mathbf{y}\mathbf{y}^T\mathbf{B}\boldsymbol{\mu}
\end{aligned}
$$

The expectation of Expression 2 may be broken down into four further expressions

$$
\begin{aligned}
& E\left[-2(\mathbf{y}^T - \boldsymbol{\beta}^T\mathbf{B})\mathbf{B}\boldsymbol{\beta}\mathbf{y}^T(\mathbf{y} - \mathbf{B}\boldsymbol{\beta})\right] \\
&= -2E\left[(\mathbf{y}^T\mathbf{B}\boldsymbol{\beta}\mathbf{y}^T - \boldsymbol{\beta}^T\mathbf{B}^T\mathbf{B}\boldsymbol{\beta}\mathbf{y}^T)(\mathbf{y} - \mathbf{B}\boldsymbol{\beta})\right] \\
&= -2E\left[\mathbf{y}^T\mathbf{B}\boldsymbol{\beta}\mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{B}\boldsymbol{\beta}\mathbf{y}^T\mathbf{B}\boldsymbol{\beta} - \boldsymbol{\beta}^T\mathbf{B}^T\mathbf{B}\boldsymbol{\beta}\mathbf{y}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{B}^T\mathbf{B}\boldsymbol{\beta}\mathbf{y}^T\mathbf{B}\boldsymbol{\beta}\right] \\
&= -2\underbrace{E\left[\mathbf{y}^T\mathbf{B}\boldsymbol{\beta}\mathbf{y}^T\mathbf{y}\right]}_{\text{Expression 2a}} + 2\underbrace{E\left[\mathbf{y}^T\mathbf{B}\boldsymbol{\beta}\mathbf{y}^T\mathbf{B}\boldsymbol{\beta}\right]}_{\text{Expression 2b}} \\
&\quad + 2\underbrace{E\left[\boldsymbol{\beta}^T\mathbf{B}^T\mathbf{B}\boldsymbol{\beta}\mathbf{y}^T\mathbf{y}\right]}_{\text{Expression 2c}} - 2\underbrace{E\left[\boldsymbol{\beta}^T\mathbf{B}^T\mathbf{B}\boldsymbol{\beta}\mathbf{y}^T\mathbf{B}\boldsymbol{\beta}\right]}_{\text{Expression 2d}}
\end{aligned}
$$

Expression 2a admits an analytic form trivially as follows

$$
\begin{aligned}
E\left[\mathbf{y}^T\mathbf{B}\boldsymbol{\beta}\mathbf{y}^T\mathbf{y}\right] &= \mathbf{y}^T\mathbf{y}E\left[\mathbf{y}^T\mathbf{B}\boldsymbol{\beta}\right] \\
&= \mathbf{y}^T\mathbf{y}\mathbf{y}^T\mathbf{B}\boldsymbol{\mu}
\end{aligned}
$$

Expression 2b admits an analytic form using identity A.3

$$
\begin{aligned}
E\left[\mathbf{y}^T\mathbf{B}\boldsymbol{\beta}\mathbf{y}^T\mathbf{B}\boldsymbol{\beta}\right] &= E\left[\boldsymbol{\beta}^T\mathbf{B}^T\mathbf{y}\mathbf{y}^T\mathbf{B}\boldsymbol{\beta}\right] \\
&= E\left[\boldsymbol{\beta}^T\mathbf{B}^T\mathbf{y}\mathbf{y}^T\mathbf{B}\boldsymbol{\beta}\right] \\
&= Tr(\mathbf{B}^T\mathbf{y}\mathbf{y}^T\mathbf{B}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T\mathbf{B}^T\mathbf{y}\mathbf{y}^T\mathbf{B}\boldsymbol{\mu}
\end{aligned}
$$

Expression 2c may be written analytically also using identity A.3

$$
\begin{aligned}
E\left[\boldsymbol{\beta}^T\mathbf{B}^T\mathbf{B}\boldsymbol{\beta}\mathbf{y}^T\mathbf{y}\right] &= \mathbf{y}^T\mathbf{y}E\left[\boldsymbol{\beta}^T\mathbf{B}^T\mathbf{B}\boldsymbol{\beta}\right] \\
&= \mathbf{y}^T\mathbf{y}E\left[\boldsymbol{\beta}^T\mathbf{B}^T\mathbf{B}\boldsymbol{\beta}\right] \\
&= \mathbf{y}^T\mathbf{y}(Tr(\mathbf{B}^T\mathbf{B}\boldsymbol{\Sigma}) + \boldsymbol{\mu}^T\mathbf{B}^T\mathbf{B}\boldsymbol{\mu})
\end{aligned}
$$

Expression 2d admits an analytic form making use of identity A.4

$$
\begin{aligned}
E\left[\boldsymbol{\beta}^T\mathbf{B}^T\mathbf{B}\boldsymbol{\beta}\mathbf{y}^T\mathbf{B}\boldsymbol{\beta}\right] &= E\left[(\mathbf{B}\boldsymbol{\beta})^T(\mathbf{B}\boldsymbol{\beta})\mathbf{y}^T(\mathbf{B}\boldsymbol{\beta})\right] \\
&= E\left[\mathbf{y}^T\mathbf{B}\boldsymbol{\beta}(\mathbf{B}\boldsymbol{\beta})^T\mathbf{B}\boldsymbol{\beta}\right] \\
&= \mathbf{y}^T E\left[\mathbf{B}\boldsymbol{\beta}(\mathbf{B}\boldsymbol{\beta})^T\mathbf{B}\boldsymbol{\beta}\right] \\
&= \mathbf{y}^T\left(2\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T + \mathbf{B}\boldsymbol{\mu}(\mathbf{B}\boldsymbol{\mu})^T\right)\mathbf{B}\boldsymbol{\mu} + Tr(\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T) \times (\mathbf{B}\boldsymbol{\mu})
\end{aligned}
$$

Finally, the expectation of Expression 3 may be written analytically using identity A.5

$$
\begin{aligned}
& E\left[(\mathbf{y} - \mathbf{B}\boldsymbol{\beta})^T\mathbf{B}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{B}^T(\mathbf{y} - \mathbf{B}\boldsymbol{\beta})\right] \\
&= Tr\left(2\mathbf{B}\boldsymbol{\Sigma}(\mathbf{B}^T\mathbf{B})\boldsymbol{\Sigma}\mathbf{B}^T\right) \\
&\quad + \left[(-\mathbf{B}\boldsymbol{\mu} + \mathbf{y})^T\mathbf{B} - (\mathbf{B}\boldsymbol{\mu})^T\mathbf{B}\right]\boldsymbol{\Sigma}\left[\mathbf{B}^T(-\mathbf{B}\boldsymbol{\mu} + \mathbf{y}) - \mathbf{B}^T\mathbf{B}\boldsymbol{\mu}\right] \\
&\quad + \left[Tr(-\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T) + (-\mathbf{B}\boldsymbol{\mu} + \mathbf{y})^T(\mathbf{B}\boldsymbol{\mu})\right]\left[Tr(-\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T) + (\mathbf{B}\boldsymbol{\mu})^T(-\mathbf{B}\boldsymbol{\mu} + \mathbf{y})\right]
\end{aligned}
$$

# Appendix B

# Details for a 2-Variable Goodwin Oscillator Model

The posterior surfaces shown in Figures 3.6, 3.7, 3.8, 3.9, 3.10 and 3.11 were induced using the following Goodwin model, also described in ([22])

$$\begin{aligned} \frac{dx}{dt} &= \frac{k_1}{36 + k_2 y} - k_3 \\ \frac{dy}{dt} &= k_4 x - k_5 \end{aligned}$$

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

# Bibliography


[1] D.G.M. Beersma. Why and how do we model circadian rhythms? *Journal of Biological Rhythms*, 4:304–313, 2005.

[2] H. Beyer, T. Jansen, C.R. Reeves, and M.D. Vose, editors. *Theory of Evolutionary Algorithms*. Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany, 2006.

[3] J.M. Bower and H. Bolouri. *Computational Modeling of Genetic and Biochemical Networks*. MIT Press, 2001.

[4] O. Cappé, A. Guillin, J. Marin, and C. Robert. Population monte carlo. *Journal of Computational and Graphical Statistics*, 13 (4):907–929, 2004.

[5] N. Chopin. A sequential particle filter method for static models. *Biometrika*, 89:539–552, 2002.

[6] P. Del Moral, A. Doucet, and A. Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society B*, 68 (3):411–436, 2006.

[7] P. Del Moral, A. Doucet, and A. Jasra. *Bayesian Statistics*, chapter Sequential Monte Carlo for Bayesian Computation, pages 1–34. Oxford University Press, 2007.

[8] D.G.T. Denison, C.C. Holmes, B.K. Mallick, and A.F.M. Smith. *Bayesian Methods for Nonlinear Classification and Regression*. John Wiley & Sons Ltd., 2002.

[9] A. Doucet, N. de Freitas, and N. Gordon. *Sequential Monte Carlo Methods in Practice*. Springer Verlag, New York, 2001.

[10] M.R. Doyle, S.J. Davis, R.M. Bastow, H.G. McWatters, L. Kozma-Bognr, F. Nagy, A.J. Millar, and R.M. Amasino. The elf4 gene controls circadian rhythms and flowering time in arabidopsis thaliana. *Nature*, 419:74–77, 2002.

[11] J. Dunlap. Molecular bases of circadian clocks. *Cell*, 96:271–290, 1999.

[12] J.C. Dunlap, J.L. Loros, and P.J. DeCoursey. *Chronobiology: Biological TimeKeeping*. Sinauer, Sunderland, 2003.

[13] N. Friel and A.N. Pettitt. Marginal likelihood estimation via power posteriors. Technical report, Department of Statistics, University of Glasgow, 2005.

[14] D. Gamerman. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Chapman and Hall/CRC, 2002.

[15] A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, 2004.

[16] A. Gelman and X.L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13 (2):163–185, 1998.

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

[18] C. J. Geyer. Practical markov chain monte carlo. *Statistical Science*, 7:473–482, 1992.

[19] C.J. Geyer. Parallel tempering. In *Computing Science and Statistics Proceedings of the 23rd Symposium on the Interface*, page 156. American Statistical Association, New York, 1991.

[20] W.R. Gilks, G.O. Roberts, and E.I. George. Adaptive direction sampling. *The Statistician*, 43 (1):179–189, 1994.

[21] O.R. Gonzalez, C. Kper, K. Jung, P.C. Naval Jr, and E. Mendoza. Parameter estimation using simulated annealing for s-system models of biochemical networks. *Bioinformatics*, 23 (4):480–486, 2007.

[22] B.C. Goodwin. Oscillatory behavior in enzymatic control processes. *Adv. Enzyme Regul.*, 3:425–438, 1965.

[23] G. Goswami and J.S. Liu. On learning strategies for evolutionary monte carlo. *Stat Comput*, 17:23–38, 2007.

[24] C. Guihenneuc-Jouyaux, S. Knight, K.L. Mengersen, S. Richardson, and C. Robert. Mcmc convergence diagnostics in action. Technical report, CREST, INSEE, Paris, 1998.

[25] S.L. Harmer, J.B. Hogenesch, M. Straume, H.S. Chang, B. Han, T. Zhu, X. Wang, J.A. Kreps, and S.A. Kay. Orchestrated transcription of the key pathways in arabidopsis by the circadian clock. *Science*, 290:2110–2113, 2000.

[26] W. K. Hastings. Monte carlo sampling methods using markov chains, and their applications. *Biometrika*, 57:97109, 1970.

[27] A. L. Hodgkin and A. F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *Journal of Physiology*, 117:500–544, 1952.

[28] M. Holder and P.O. Lewis. Phylogenetic estimation: Traditional and bayesian approaches. *Nat. Rev. Genet.*, 4:275–284, 2003.

[29] J. H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, 1975.

[30] Y. Iba. Population monte carlo algorithms. *Transactions of the Japanese Society of Artificial Intelligence*, 16:279–286, 2000.

[31] L. Ingber. Simulated annealing: Practice versus theory. *Mathl. Comput. Modelling*, 18 (11):29–57, 1993.

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

[33] E.T. Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[34] H. Jeffreys. *Scientific Inference*. Cambridge University Press, 1937.

[35] R.E. Kass and A.E. Raftery. Bayes factors. *American Statistical Association*, 90 (430):773–795, 1995.

[36] S. Kikuchi, D Tominaga, M. Arita, K. Takahashi, and M. Tomita. Dynamic modeling of genetic networks using genetic algorithm and s-system. *Bioinformatics*, 19(5):643–650, 2003.

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

[38] H. Kitano. Systems biology: A brief overview. *Science*, 295:1662, 2002.

[39] H. Kitano. Towards a theory of biological robustness. *Molecular Systems Biology*, 3:137, 2007.

[40] G. Kurosawa and A. Goldbeter. Amplitude of circadian oscillations entrained by 24-h light-dark cycles. *Journal of Theoretical Biology*, 242:478–488, 2006.

[41] G. Kurosawa, A. Mochizuki, and Y. Iwasa. Comparative study of circadian clock models, in search of processes promoting oscillation. *Journal of Theoretical Biology*, 216:193–208, 2002.

[42] N. Lartillot and H. Philippe. Computing bayes factors using thermodynamic integration. *Syst. Biol.*, 55 (2):195–207, 2006.

[43] K.B. Laskey and J. Myers. Population markov chain monte carlo. *Machine Learning*, 50:175–196, 2003.

[44] F. Liang and W.H. Wong. Evolutionary monte carlo: Applications to cp model sampling and change point problem. *Statisticca Sinica*, 10:317–342, 2000.

[45] F. Liang and W.H. Wong. Real-parameter evolutionary monte carlo with applications to bayesian mixture models. *American Statistical Association*, 96 (454):653–666, 2001.

[46] O. Lipan and W.H. Wong. Is the future biology shakespearean or newtonian? *Molecular Biosystems*, 2:411416, 2006.

[47] J.S. Liu, F. Liang, and W.H. Wong. The multiple-try method and local optimization in metropolis sampling. *Journal of the American Statistical Association*, 95 (449):121–134, 2000.

[48] J.C.W. Locke, A.J. Millar, and M.S. Turner. Modelling genetic networks with noisy and varied experimental data: the circadian clock in arabidopsis thaliana. *Journal of Theoretical Biology*, 234:383–393, 2005.

[49] A .P. Lyubartsev, A.A. Martsinovskii, S.V. Shevkunov, and P.N. Vorontsov-Velyaminov. New approach to monte carlo calculation of the free energy: Method of expanded ensembles. *Journal of Chemical Physics*, 96:1776–1783, 1992.

[50] D.J.C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, 2003.

[51] E. Marinari and G. Parisi. Simulated tempering: A new monte carlo scheme. *Europhysics Letters*, 19:451, 1992.

[52] P. Mendes and D. Kell. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics*, 14(10):869–883, 1998.

[53] K.L. Mengersen and C.P. Robert. *Bayesian Statistics 7*, chapter IID Sampling using Self-Avoiding Population Monte Carlo: The Pinball Sampler, page 277. Clarendon Press, Oxford, 2003.

[54] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculation by fast computing machines. *Journal of Chemical Physics*, 21(6):10871092, 1953.

[55] M. Mitchell. *An Introduction to Genetic Algorithms*. MIT Press, 1998.

[56] C.G. Moles, P. Mendes, and J.R. Banga. Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Research*, 13:2467–2474, 2003.

[57] N.A.M. Monk. Oscillatory expression of hes1, p53, and nf-kb driven by transcriptional time delays. *Current Biology*, 13 (16):1409–1413, 2003.

[58] I. Murray. *Advances in Markov Chain Monte Carlo Methods*. PhD thesis, University College London, 2007.

[59] R.M. Neal. Sampling from multimodal distributions using tempered transitions. Technical report, Department of Statistics, University of Toronto, 1994.

[60] R.M. Neal. Sampling from multimodal distributions using tempered transitions. *Statistics and Computing*, 6:353–366, 1996.

[61] R.M. Neal. Annealed importance sampling. *Statistics and Computing*, 11:125–139, 2001.

[62] R.M. Neal. Slice sampling. *The Annals of Statistics*, 31 (3):705–767, 2003.

[63] M.A. Newton and A.E. Raftery. Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of th eRoyal Statistical Society Series B*, 56 (1):3–48, 1994.

[64] A.E. Raftery, M.A. Newton, J.M. Satagopan, and P.N. Krivitsky. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Bayesian Statistics*, 8:1–45, 2007.

[65] J. Ramsay, G. Hooker, D. Campbell, and J. Cao. Parameter estimation for differential equations: A generalized smoothing approach. *Journal of the Royal Statistical Society*, (To Appear), 2007.

[66] D. Rand, B.V. Shulgin, D. Salazar, and A.J. Millar. Uncovering the design principles of circadian clocks: Mathematical analysis of flexibility and evolutionary goals. *Journal of Theoretical Biology*, 238:616–635, 2006.

[67] D.A. Rand, B.V. Shulgin, D. Salazar, and A.J. Millar. Design principles underlying circadian clocks. *J. R. Soc. Lond. Interface*, 1:119–130, 2004.

[68] J.M. Raser and E.K. O'Shea. Noise in gene expression: Origins, consequences, and control. *Science*, 309 (5743):2010–2013, 2005.

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

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

[71] G.O. Roberts and W.R. Gilks. Convergence of adaptive direction sampling. *Journal of Multivariate Analysis*, 49:287–298, 1994.

[72] D. Rubin. Using the sir algorithm to simulate posterior distributions. *Bayesian Statistics*, 3:395–402, 1988.

[73] P. Ruoff and L. Rensing. The temperature-compensated goodwin model simulates many circadian clock properties. *Journal of Theoretical Biology*, 179:275–285, 1996.

[74] U. Sauer, M. Heinemann, and N. Zamboni. Getting closer to the whole picture. *Science*, 316 (5824):550–551, 2007.

[75] H.M. Sauro, A.M. Uhrmacher, D. Harel, M. Hucka, M. Kwiatkowska, P. Mendes, C.A. Shaffer, L. Stromback, and J.J. Tyson. Challenges for modeling and simulation methods in systems biology. In *Proceedings of the 2006 Winter Simulation Conference*, 2006.

[76] J. Skilling. Nested sampling for general bayesian computation. *Bayesian Analysis*, 1 (4):833–860, 2006.

[77] I. M. Sobol. The distribution of points in a cube and the approximate evaluation of integrals. *USSR Comput. Math. Math. Phys.*, 7 (4):86–112, 1967.

[78] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, B. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, 9:3273–3297, 1998.

[79] G. Tiana, S. Krishna, S. Pigolotti, M.H. Jensen, and K. Sneppen. Oscillations and temporal signalling in cells. *Phys Biol*, 4:R1–R17, 2007.

[80] M.B. Viani, L.I. Pietrasanta, J.B. Thompson, A. Chand, I.C. Gebeshuber, J.H. Kindt, M. Richter, H.G. Hansma, and P.K. Hansma. Probing protein-protein interactions in real time. *Nature Structural Biology*, 7:644 – 647, 2000.

[81] J. Wako, K. Ando, M. Miki, and T. Hiroyasu. Parallel simulated annealing with adaptive temperature determined by genetic algorithm. *Transactions of Information Processing Society of Japan*, 47:1–11, 2006.

[82] Z.Y. Wang and E.M. Tobin. Constitutive expression of the circadian clock associated 1 (cca1) gene disrupts circadian rhythms and suppresses its own expression. *Cell*, 93 (7):1207–1217, 1998.

[83] Y. Yu, W. Dong, C. Altimus, X. Tang, J. Griffith, M. Morello, L. Dudek, J. Arnold, and H.B. Schttler. A genetic network for the clock of neurospora crassa. *PNAS*, 104 (8):2809–2814, 2007.