Estimating Bayes Factors via Thermodynamic Integration and Population MCMC

Ben Calderhead [a,*] and Mark Girolami [a,b]

[a] *Department of Computing Science, University of Glasgow, UK*

[b] *Department of Statistics, University of Glasgow, UK*

Abstract

A Bayesian approach to model comparison based on the integrated or marginal likelihood is considered, and applications to linear regression models and nonlinear ordinary differential equation (ODE) models are used as the setting in which to elucidate and further develop existing statistical methodology. The focus is on two methods of marginal likelihood estimation. First, a statistical failure of the widely employed *Posterior Harmonic Mean estimator* is highlighted. It is demonstrated that there is a systematic bias capable of significantly skewing Bayes factor estimates, which has not previously been highlighted in the literature. Second, a detailed study of the recently proposed *Thermodynamic Integral* estimator is presented, which characterises the error associated with its discrete form. An experimental study using analytically tractable linear regression models highlights substantial differences with recently published results regarding optimal discretisation. Finally, with the insights gained, it is demonstrated how Population MCMC and thermodynamic integration methods may be elegantly combined to estimate Bayes factors accurately enough to discriminate between nonlinear models based on systems of ODEs, which has important application in describing the behaviour of complex processes arising in a wide variety of research areas, such as Systems Biology, Computational Ecology and Chemical Engineering.

Key words: Bayesian model comparison, Population Markov Chain Monte Carlo, Parameter estimation, Nonlinear ODE models, Systems biology

[*] Corresponding author.
Email address: `bc@dcs.gla.ac.uk` (Ben Calderhead).
URL: `http://www.dcs.gla.ac.uk/inference` (Ben Calderhead).

1 Introduction

Marginal likelihood estimation over a statistical model can be an extremely challenging task. A statistical model of a complex process can be considered a codification of an underlying hypothesis regarding the system under study. Such competing hypotheses can be objectively assessed using Bayes factors obtained from the marginal likelihoods based on each statistical model. Except in a small number of special cases where there is a conjugate prior available, the marginal likelihood does not generally admit a closed-form expression, and estimating this is seriously hampered by a number of practical difficulties. In particular, the complexity of some statistical models may induce likelihood surfaces that are highly correlated and multimodal, a good example of which are models over nonlinear differential equations (see Section 4). Such probability distributions are very difficult to sample from and indeed standard Markov Chain Monte Carlo (MCMC) methods very often fail catastrophically. This aspect makes obtaining unbiased low-variance estimates of marginal likelihoods, and hence Bayes factors, for such models a formidable challenge. However, even for simple linear regression models, which exhibit log-concave likelihood surfaces, we find that commonly used methods for estimating the marginal likelihoods still may produce substantially biased finite sample estimates. In this paper we address these problems and present a number of novel and useful contributions which are of importance to the development of Bayesian statistical methodology and application.

We provide a study of the statistical failure of the Posterior Harmonic Mean estimator for calculating Bayes factors (Section 3). This is quite different to the numerical problems regarding instability which are often cited in the statistics literature (Newton and Raftery, 1994; Raftery et al., 2007) and something which has not been highlighted in the literature previously. We investigate the use of thermodynamic integration methods for estimating marginal likelihoods (Sections 2 and 3). In particular, we provide a decomposition of the Thermodynamic Integral in terms of upper and lower bounds and characterise the error associated with the discrete form of the estimator in terms of the Kullback-Leibler divergence by using analytically tractable linear models to gain insight. We provide an analytic characterisation of the optimal discretisation strategy, in terms of minimising the variance of the estimates produced, using thermodynamic integration over linear models, and undertake an experimental study highlighting significant differences with recently published results (Section 3). Finally, we apply our insights to illustrate how Population Markov Chain Monte Carlo (MCMC) may be elegantly combined with thermodynamic integration to gain estimates of marginal likelihoods that are accurate enough to discriminate between competing model hypotheses described using nonlinear ODEs (Section 4). The benefits of Population MCMC in this setting are twofold. Firstly, it allows us to obtain samples from all the required thermodynamic distributions simultaneously, and secondly, the population structure enables sampling from the multimodal probability distributions that are induced by nonlinear

ODEs.

2 Estimating Marginal Likelihoods

The ability to calculate marginal likelihoods accurately is of vital importance for computing meaningful Bayes factors for model comparison (Robert and Casella, 2004). 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 $\mathbf{y} \in \mathbb{R}^m$ and two competing model hypotheses H_1 and H_2, we wish to calculate the probability of each model hypothesis given the data. The posterior odds are given by

$$\underbrace{\frac{p(H_1 \mid \mathbf{y})}{p(H_2 \mid \mathbf{y})}}_{\text{Posterior Odds}} = \underbrace{\frac{p(\mathbf{y} \mid H_1)}{p(\mathbf{y} \mid H_2)}}_{\text{Bayes Factor}} \underbrace{\frac{p(H_1)}{p(H_2)}}_{\text{Prior Odds.}} \tag{1}$$

In the case that there is no preference a priori for a particular model, the prior probabilities of the models may be set to be equal, which shall be the case for the experiments presented in this paper. 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.

Table 1 shows a standard interpretation of the Bayes factor B_{12} (Kass and Raftery, 1995), which compares the model H_1 with the model H_2. This is given in terms of evidence in favour of the first labeled model over the second.

Table 1
Interpretation of Bayes Factor (Kass and Raftery, 1995)

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

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

$$\begin{aligned} p(\mathbf{y} \mid H_j) &= \int p(\mathbf{y} \mid \theta_j, H_j)\pi(\theta_j \mid H_j) d\theta_j \\ &= E_{\pi(\theta_j)}\left[p(\mathbf{y} \mid \theta_j, H_j)\right], \end{aligned} \tag{2}$$

where θ_j is a vector describing the parameters for model H_j, $\pi(\theta_j \mid H_j)$ is the prior density of the parameters, and $p(\mathbf{y} \mid \theta_j, H_j)$ 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 (Robert and Casella, 2004).

We consider two methods based on importance sampling ideas, the Prior Arithmetic Mean estimator (McCulloch and Rossi, 1991) and the Posterior Harmonic Mean estimator (Newton and Raftery, 1994; Raftery et al., 2007). These approaches are straightforward to implement and do not require a huge amount of computational power, however we shall demonstrate that the biased results that both produce render them unsuitable for accurately comparing models.

Path sampling methods (Gelman and Meng, 1998) have been shown to perform well at the task of estimating marginal likelihoods (Friel and Pettitt, 2008; Lartillot and Philippe, 2006). Such methods rely on sampling from a sequence of distributions which form a "bridge" in the probability density space connecting the prior distribution to the posterior distribution, and integrating over them. The third method we examine is therefore based on approximating the thermodynamic integral (Friel and Pettitt, 2008; Lartillot and Philippe, 2006) which is probably the most general example of path sampling methods. Other "non-equilibrium" methods (Del Moral et al., 2006, 2007) are very similar in principle, for example Annealed Importance Sampling (Neal, 2001), however we do not consider these methods further since they are all based on the thermodynamic integral and appear to produce similar results to the thermodynamic integral approximation that we do consider (Vyshemirsky and Girolami, 2008). For a complete review of marginal likelihood estimation methods see e.g. (Friel and Pettitt, 2008).

2.1 Monte Carlo Methods

For the purpose of computing Bayes factors we wish to evaluate, for a particular model, the marginal likelihood (2). It is possible to obtain a Monte Carlo estimate of the marginal likelihood using $\frac{1}{S}\sum_{i=1}^{S} p(\mathbf{y} \mid \theta^{(i)})$ where $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(S)} \sim \pi(\theta)$. (Explicit conditioning on a model H is now dropped for reasons of clarity.) By the Law of Large numbers, this estimator converges to the true expectation as the number of independent samples, S, drawn from the prior, tends to infinity (Robert and Casella, 2004).

This estimator, however, is often very inefficient for a finite number of samples as many samples will fall outside regions of high likelihood, see e.g. (Gamerman, 2002). To get around the inefficiency of sampling from the prior a common approach is to employ importance sampling, as used by the Posterior Harmonic Mean estimator (Newton and Raftery, 1994). The Monte Carlo estimate using an importance sampling scheme is given by

$$\frac{\sum_{i=1}^{S} w_i p(\mathbf{y} \mid \theta^{(i)})}{\sum_{i=1}^{S} w_i}, \tag{3}$$

where $w_i = p(\theta)/\pi^*(\theta)$, and the density function $\pi^*(\theta)$ is the importance sampling function. (Note that $\pi^*(\theta)$ is not strictly required to be a normalised density function). Choosing the importance sampling function to be the posterior density, and substituting this into (3) gives

$$\left\{ \frac{1}{S} \sum_{i=1}^{S} p(\mathbf{y} \mid \theta^{(i)})^{-1} \right\}^{-1}, \tag{4}$$

which is the harmonic mean of the likelihood values, where the parameters are sampled from the posterior, $\theta \sim p(\theta \mid \mathbf{y})$. It is known that sometimes the variance of this estimator can become very large, since occasionally a sample may be taken into account with small likelihood, which has a large effect on the result due to the reciprocal present in (4). Raftery et al. (2007) have also more recently observed that this estimator can produce biased results for finite sample sizes, despite being asymptotically unbiased, due to numerical instabilities resulting in high variance estimates. We shall see, perhaps more worryingly, in Section 3 that even when harmonic mean estimates are numerically stable and exhibit low variance, they may still be strongly biased, leading to wrong interpretations of the observed evidence through the calculation of Bayes factors.

2.2 *Thermodynamic Integration*

It is widely accepted that estimating the marginal likelihood of a non-trivial statistical model is generally very challenging and methods employing some form of thermodynamic integration or path sampling (Gelman and Meng, 1998), although computationally more expensive than the importance sampling methods previously described, have been shown to perform well on a number of forms of statistical model, see for example (Friel and Pettitt, 2008), or (Lartillot and Philippe, 2006) for use in a phylogenetic context.

Such methods are based on the fact that the logarithm of the marginal likelihood can be represented in terms of the following integral

$$\log p(\mathbf{y}) = \int_0^1 E_{\theta|\mathbf{y},t} \left[\log p(\mathbf{y} \mid \theta)\right] dt. \tag{5}$$

The classical derivation of the thermodynamic integral is as follows. Given an unnormalised density, $q(\theta)$, the normalised probability density is given by $p(\theta) =$

$\frac{1}{Z}q(\theta)$, where $Z = \int_\theta q(\theta)d\theta$. In order to calculate the marginal likelihood using thermodynamic integration, however, we define the *power-posterior* as in (Friel and Pettitt, 2008) and (Lartillot and Philippe, 2006)

$$p_t(\theta|\mathbf{y}) = \frac{p(\mathbf{y} \mid \theta)^t p(\theta)}{z(\mathbf{y} \mid t)} \text{ , where } z(\mathbf{y} \mid t) = \int_\theta p(\mathbf{y} \mid \theta)^t p(\theta) d\theta. \tag{6}$$

We note at this point that $z(\mathbf{y} \mid t = 0)$ is the prior marginalised over θ, which is simply equal to 1, and that $z(\mathbf{y} \mid t = 1)$ is the marginal likelihood. By considering

$$\frac{d}{dt}\log(z(\mathbf{y} \mid t)) = \frac{1}{z(\mathbf{y} \mid t)}\frac{d}{dt}z(\mathbf{y} \mid t) = E_{\theta|\mathbf{y},t}\left[\log p(\mathbf{y} \mid \theta)\right],$$

where the expectation is taken with respect to the power-posteriors, (5) then follows by integrating with respect to t.

Additionally, we can derive an expression for the optimal temperature schedule in terms of minimising the Monte Carlo variance (Gelman and Meng, 1998). We firstly define a density $p(t)$ over the temperature values $t \in [0,1]$. Introducing $p(t)$ obtains

$$\log p(\mathbf{y}) = \int_0^1 \frac{E_{\theta|\mathbf{y},t}\left[\log p(\mathbf{y} \mid \theta)\right] p(t)}{p(t)} dt = E_{\theta,t|\mathbf{y}}\left[\frac{\log p(\mathbf{y} \mid \theta)}{p(t)}\right]. \tag{7}$$

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

$$\int_0^1 E_{\theta|\mathbf{y},t}\left[\frac{\{\log p(\mathbf{y} \mid \theta)\}^2}{p(t)}\right] dt + \lambda \int_0^1 p(t)dt, \tag{8}$$

whose solution is

$$p(t) = \frac{p^*(t)}{\int_0^1 p^*(t')dt'} \tag{9}$$

and

$$p^*(t) = \sqrt{E_{\theta|\mathbf{y},t}\left[\{\log p(\mathbf{y} \mid \theta)\}^2\right]}. \tag{10}$$

We shall return to this expression in Section 3 where we use it to guide our choice of temperature spacing for use in experiments.

Published practical methods for estimating this thermodynamic integral involve discretisation (Friel and Pettitt, 2008; Lartillot and Philippe, 2006). An estimate of the marginal likelihood can be obtained by numerically integrating over a finite number of temperatures within the range $t = 0$ to $t = 1$ and using the corresponding expectations $E_{\theta|\mathbf{y},t}\left[\log p(\mathbf{y}|\theta)\right]$ obtained at each discrete temperature. By running a Markov chain at each temperature until equilibrium and using these power-posterior samples, a Monte Carlo estimate of each required expectation can be obtained. Indeed, in Section 4, we demonstrate how population MCMC may be employed to efficiently obtain samples from all the required power-posteriors simultaneously.

2.2.1 A Discretisation of the Thermodynamic Integral

Friel and Pettitt (2008) consider the possibility of obtaining the marginal likelihood by estimating the expectation in (7). However, there are unfortunately as yet no practical methods available for estimating this expectation, and we are forced to introduce a discretisation. In order to obtain insight into the sources of error introduced by the discretisation of (5), we consider some simple manipulation of (6)

$$\frac{z(\mathbf{y} \mid t_n)}{z(\mathbf{y} \mid t_{n-1})} p(\theta \mid \mathbf{y}, t_n) = p(\mathbf{y} \mid \theta)^{\Delta t_n} p(\theta \mid \mathbf{y}, t_{n-1}), \tag{11}$$

where $1 = t_N > \cdots > t_1 = 0$ and $\Delta t_n \equiv t_n - t_{n-1}$. Taking logarithms, multiplying both sides of the equality by $p(\theta \mid \mathbf{y}, t_n)$, and integrating with respect to θ gives the following expression

$$\log \frac{z(\mathbf{y} \mid t_n)}{z(\mathbf{y} \mid t_{n-1})} = E_{\theta|\mathbf{y},t_n}\left[\log p(\mathbf{y} \mid \theta)\right] \Delta t_n - KL(p_n || p_{n-1}), \tag{12}$$

where $E_{\theta|\mathbf{y},t_n}$ denotes the expectation with respect to the power-posterior $p(\theta \mid \mathbf{y}, t_n)$ and

$$KL(p_n || p_{n-1}) = \int p(\theta \mid \mathbf{y}, t_n) \log \frac{p(\theta \mid \mathbf{y}, t_n)}{p(\theta \mid \mathbf{y}, t_{n-1})} d\theta \geq 0.$$

By summing over $n = 2 : N$, the logarithm of the marginal likelihood may be expressed in a discrete form

$$
\begin{aligned}
\log p(\mathbf{y}) &= \sum_n \log \frac{z(\mathbf{y} \mid t_n)}{z(\mathbf{y} \mid t_{n-1})} \\
&= \sum_n \big[\underbrace{E_{\theta|\mathbf{y},t_n}[\log p(\mathbf{y} \mid \theta)]\,\Delta t_n}_{Upper-bound} - \underbrace{KL(p_n||p_{n-1})}_{Bias}\big].
\end{aligned} \tag{13}
$$

We note that calculating the sum of just the expectations across each temperature interval, gives a strict upper bound on the estimate of the log of the marginal likelihood, with the bias from the true value being characterised by the sum of the KL divergences between posteriors across each temperature interval. In the limit of temperature differences the divergence term will tend to zero, so $KL(p_n||p_{n-1}) \to 0$ as $\Delta t_n \to 0$ and therefore

$$
\lim_{\Delta t_n \to 0} \sum_n E_{\theta|\mathbf{y},t_n}[\log p(\mathbf{y} \mid \theta)]\,\Delta t_n \to \int_0^1 E_{\theta|\mathbf{y},t}[\log p(\mathbf{y} \mid \theta)]\,dt
$$

recovers the continuous form of the thermodynamic integral in (5).

Similarly, a lower bound can be obtained by taking logarithms of (11) and multiplying both sides of the equality this time by $p(\theta \mid \mathbf{y}, t_{n-1})$. Integrating with respect to θ and summing over all n gives a lower bound on the logarithm of the marginal likelihood

$$
\begin{aligned}
\log p(\mathbf{y}) &= \sum_n \log \frac{z(\mathbf{y} \mid t_n)}{z(\mathbf{y} \mid t_{n-1})} \\
&= \sum_n \big[\underbrace{E_{\theta|\mathbf{y},t_{n-1}}[\log p(\mathbf{y} \mid \theta)]\,\Delta t_n}_{Lower-bound} + \underbrace{KL(p_{n-1}||p_n)}_{Bias}\big].
\end{aligned} \tag{14}
$$

We can average over these upper and lower bounds to form an expression for the logarithm of the marginal likelihood as follows

$$
\log p(\mathbf{y}) = \underbrace{\frac{1}{2}\sum_n \Delta t_n (E_{n-1} + E_n)}_{Approximation} + \underbrace{\frac{1}{2}\sum_n \big[KL(p_{n-1}||p_n) - KL(p_n||p_{n-1})\big]}_{Bias}, \tag{15}
$$

where $E_n = E_{\theta|\mathbf{y},t_n}[\log p(\mathbf{y} \mid \theta)]$, which is equivalent to using the trapezium rule for numerical integration with the associated error expressed in terms of the asymmetry of the KL divergence. We demonstrate in Section 3 that this approximation for estimating marginal likelihoods over the linear models which we consider gives a lower bias than using either of the upper or lower bound approximations on their own.

There are therefore two sources of error which appear when estimating the marginal likelihood. Firstly, there is the Monte Carlo error when estimating the power-posterior expectations 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 approximating the integral of the power-posteriors over t, represented by the KL divergence term. As discussed in (Friel and Pettitt, 2008) the discretisation of the unit line need not be uniform. Indeed, there are many ways in which the t_is may be chosen, and this can drastically affect the bias associated with the estimate, as we shall show in Section 3. An expression for the optimal temperature density function, $p(t)$, is given in (10). For linear regression models, considered in Section 3, the KL divergence conveniently has an analytic form, as does the minimum variance $p(t)$, and this therefore allows us to examine in detail the effect that varying the number and spacing of the temperature partitions has on this source of error.

3 Gaining Insight using Analytically Tractable Linear Models

In order to obtain deeper insights into the problem of estimating marginal likelihoods, we performed experiments using simple linear regression models. These models were used to determine the relationship between some continuous response variable y and a set of covariates $\mathbf{x} = (x_1, \ldots, x_d)$, where d is the dimension of the model. General models of the form $g(\mathbf{x}) = \sum_{i=1}^{k} \beta_i B_i(\mathbf{x})$ were used so that the function g comprised a linear combination of k basis functions $B_i(\mathbf{x})$ with coefficients β_i. In particular the responses were assumed to be related to the variables through the relationship $y = g(\mathbf{x}) + \varepsilon$, where ε has a Gaussian distribution with zero-mean and known standard deviation σ. This can also be written in matrix form $\mathbf{y} = \mathbf{B}\beta + \varepsilon$, where $\mathbf{y} = (y_1, \ldots, y_m)^T$, $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_m)^T$, and $\mathbf{B}$ is the design matrix.

For each pair of models, H_1, H_2, a 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, ε. The two models were then compared by using this “observed”dataset to calculate $P(\mathbf{y} \mid \mathbf{X}, H_j)$, where $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]$, from which the Bayes factors could be obtained.

3.1 Analytic Expressions

A conjugate prior distribution on the regression coefficients was used so that an analytic expression for the marginal likelihood could be obtained. 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 ζ^2 were placed on each of the unknown parameters $(\beta_1,\ldots,\beta_k)$, so that $\pi(\beta_i) = N(0,\zeta^2)$.

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

$$(2\pi\sigma^2)^{-m/2}\exp\left\{\frac{-(\mathbf{y}-\mathbf{B}\beta)^T(\mathbf{y}-\mathbf{B}\beta)}{2\sigma^2}\right\}. \tag{16}$$

Since both the priors and the likelihood function are Gaussian distributions and σ^2 and ζ^2 are fixed, the posterior is therefore also a Gaussian distribution for which there exists an analytic form. This Gaussian posterior is given by $p(\beta \mid \mathbf{X},\mathbf{y},\sigma^2,\zeta^2) = N(\mu,\Sigma)$, where

$$\mu = \left(\mathbf{B}^T\mathbf{B}+\frac{\sigma^2}{\zeta^2}\mathbf{I}\right)^{-1}\mathbf{B}^T\mathbf{y}, \quad \Sigma = \sigma^2\left(\mathbf{B}^T\mathbf{B}+\frac{\sigma^2}{\zeta^2}\mathbf{I}\right)^{-1}.$$

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

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

$$\begin{aligned} p(\mathbf{y} \mid \sigma^2,\zeta^2) &= \int p(\mathbf{y} \mid \beta,\sigma^2)\pi(\beta \mid \zeta^2)d\beta \qquad (17)\\ &= (2\pi)^{-m/2}|\Omega|^{-1/2}\exp\left\{-\frac{1}{2}\mathbf{y}^T\Omega^{-1}\mathbf{y}\right\}, \end{aligned}$$

where $\Omega = \sigma^2\mathbf{I}+\zeta^2\mathbf{B}\mathbf{B}^T$. 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.

3.1.1 Power Posteriors

The linear regression models we use also admit an analytic expression for the power posteriors required in the thermodynamic integration method. Noting that the posterior distribution is Gaussian, then the power posteriors, for a particular inverse temperature $t \in [0,1]$, are also simply Gaussian distributions $p(\beta \mid \mathbf{y}, t, \sigma^2, \zeta^2) = N_\beta(\mu_t, \Sigma_t)$ where the mean and covariance matrices are given by

$$\mu_t = \left(\mathbf{B}^T\mathbf{B} + \frac{\sigma^2}{t\zeta^2}\mathbf{I}\right)^{-1} \mathbf{B}^T\mathbf{y}, \quad \Sigma_t = \frac{\sigma^2}{t}\left(\mathbf{B}^T\mathbf{B} + \frac{\sigma^2}{t\zeta^2}\mathbf{I}\right)^{-1}.$$

The expectation of the log of the likelihood with respect to a power posterior can be obtained analytically as

$$E_{\beta|\mathbf{y},t,\sigma^2,\zeta^2}\left[\log p(\mathbf{y} \mid \beta, \sigma^2)\right] = -\frac{1}{2\sigma^2}\mathbf{e}^T\mathbf{e} - \frac{1}{2}Tr(\mathbf{B}^T\mathbf{B}\Sigma_t) - \frac{m}{2}\log(2\pi\sigma^2), \quad (18)$$

where $\mathbf{e} = \mathbf{y} - \mathbf{B}\mu_t$.

3.1.2 Discretised Thermodynamic Integral

For the discretised estimate of the thermodynamic integral (15) the bias represented by the difference in KL divergences admits an analytic expression since the power posteriors for the linear regression model are simply given by Gaussian distributions. The KL divergence for the temperature interval from t_{n-1} to t_n is given by

$$\begin{aligned} KL(p_n||p_{n-1}) = &\frac{1}{2}\log\frac{|\Sigma_{t_{n-1}}|}{|\Sigma_{t_n}|} + \frac{1}{2}Tr(\Sigma_{t_{n-1}}^{-1}\Sigma_{t_n}) \\ &+ \frac{1}{2}(\mu_{t_n} - \mu_{t_{n-1}})^T\Sigma_{t_{n-1}}^{-1}(\mu_{t_n} - \mu_{t_{n-1}}) - \frac{m}{2}, \end{aligned} \quad (19)$$

where $|\Sigma_t|$ denotes the determinant of the matrix Σ_t, and $KL(p_{n-1}||p_n)$ follows a similar form.

For the linear regression model we may also in fact compute the density function $p^*(t)$ analytically (see Appendix A), which is proportional to the normalised density function over the temperature (9). 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. It is clear from this that low values of temperature will have higher density $p(t)$ due to the large values of expected deviance at lower temperatures. In approximating the thermodynamic integral using deterministic numerical methods this suggests a logarithmic or power style partitioning (see Figure 1).

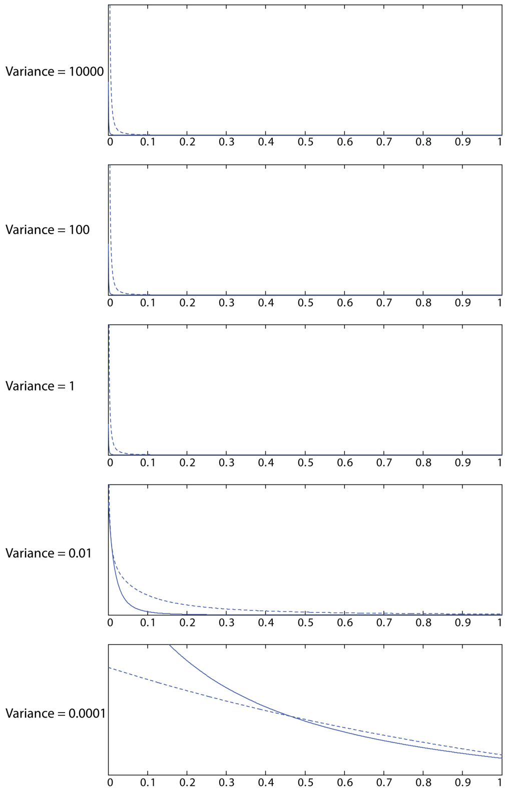

Fig. 1. The optimal density function $p^*(t)$ is plotted (on the y-axis) against temperature (on the x-axis) for the linear regression model. The continuous line represents $p^*(t)$ for a 2D model and the dotted line $p^*(t)$ for a 20D model. Notice that as the variance of the prior distribution decreases (i.e. as confidence in the prior increases), the introduction of new information (equivalent to increasing t) has less of an effect on the optimal density, which should be used to define the optimal temperature schedule.

3.2 Experimental Results

3.2.1 Choice of Temperature Schedule

Before comparing methods of estimating marginal likelihoods for linear regression models, we firstly consider which types of temperature schedules should be employed with thermodynamic integration to achieve optimal results in terms of minimising the bias and variance of Monte Carlo estimates of marginal likelihoods. These results complement and extend the insights offered by (Jasra et al., 2007) 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 are approximated for each mixture component. In (Lartillot and Philippe, 2006) a uniform spacing is employed and the issue of temperature schedules is not considered at all, while Friel and Pettitt (2008) give some discussion on the subject. To complement the current literature, we measure the accuracy by calculating the bias in the discrete marginal likelihood estimates given by the KL divergence which, as we have shown, may be computed analytically for linear regression models.

Tables 3 and 4 show the values associated with the relative bias introduced when estimating the marginal likelihood using (15) and employing different temperature schedules. The number of partitions used was also varied, to see to what extent the bias decreases as the number of partitions used increases. The relative bias is defined in log space as the ratio of the bias to the analytic marginal likelihood. Results are given for linear regression models of 2 and 20 dimensions respectively. Table 2 shows the geometric-based temperature schedules, defining $t_{1,\dots,N}$, that are used for the comparison. A uniform distribution is included since this was used by Lartillot and Philippe (2006). The optimal density function suggests the use of a power law distribution and so geometric schedules are included which cluster intermediate temperature levels towards the prior (see Figure 1) and also the posterior for comparison.

Table 2
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* clusters the temperature points around 0.5 and *Extremes* clusters the temperature steps towards both 0 and 1 and away from the middle. Both of these schedules are generated based on scaling and combining points produced by the prior and posterior schedules shown in Table 2. Higher powers, p correspond to a more acute clustering of points.

From Table 3 it can be seen that methods which cluster more partitions towards $t = 0$, corresponding to the prior, produce lower biases analytically than those which cluster partitions towards $t = 1$, corresponding to the posterior, as suggested by calculating $p^*(t)$. Partitions skewed towards the posterior end of the scale performed very badly, indeed much worse than a uniform distribution. Table 4 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 bias of just 0.86% using only 30 partitions of the unit line.

Figure 2 shows, for 2 and 20 dimensional linear regression models, how the biases in the estimates decrease as the number of partitions used in the temperature schedule increases. The drastically worse bias induced using a uniform spacing may be seen in Figure 3. Note that even using 100 uniform partitions the bias is much greater than when using just 10 partitions spaced according to a power law (with $p = 5$). It is also clear that using a trapezoidal estimate of the thermodynamic integral gives a much closer approximation for these linear regression models than using either the upper or lower bound approximations (13) and (14) by themselves.

Table 3
Relative bias introduced when using temperature schedules estimating the power posterior integral via the trapezium rule for a 2 dimensional linear regression model varying the number of partitions N.

Method Used	Power Raised	N=10	N=20	N=30	N=60	N=100
Uniform	1	311.7%	151.9%	99%	46.7%	26.2%
Centered	2	568.4%	295.6%	197.9%	97.2%	56.4%
Centered	3	711.5%	425.5%	290.8%	146.6%	86.6%
Extremes	2	12.8%	3.22%	4.83%	1.09%	0.13%
Extremes	5	11.8%	2.51%	1.09%	0.27%	0.09%
Prior	2	27.3%	5.54%	2.19%	0.56%	0.21%
Prior	5	3.41%	0.81%	0.36%	0.08%	0.03%
Prior	6	3.47%	0.84%	0.38%	0.08%	0.03%
Posterior	2	600.5%	303.7%	201.5%	98.1%	56.7%
Posterior	3	860.9%	448.3%	301.2%	149.3%	87.5%

Table 4
Relative bias introduced when using temperature schedules estimating the power posterior integral via the trapezium rule for a 20 dimensional linear regression model.

Method Used	Power Raised	N=10	N=20	N=30	N=60	N=100
Uniform	1	600.2%	290.1%	188.2%	88.6%	50.3%
Centered	2	1101.5%	568.7%	379.0%	184.8%	107.1%
Centered	3	1503.5%	821.9%	559.3%	279.7%	164.4%
Extremes	2	113.2%	26.2%	11.1%	2.45%	0.81%
Extremes	5	24.4%	5.73%	2.53%	0.63%	0.22%
Prior	2	54.6%	12.6%	5.25%	1.14%	0.39%
Prior	5	7.88%	1.92%	0.86%	0.21%	0.07%
Prior	6	8.74%	2.13%	0.94%	0.23%	0.09%
Posterior	2	1164.3%	584.5%	386.0%	186.5%	107.6%
Posterior	3	1674.6%	866.7%	579.5%	284.9%	166.2%

We may also use the analytic expression of the minimum variance density $p(t)$ (10) for the linear regression model to visualise where the bulk of the density lies and where significant changes of density occur. Plots proportional to the optimal density functions for linear regression models of varying dimension are shown in Figure 1. The shape of these support our findings based on the bias values and 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 (as used in our experiments), since this is where the density function has its concentration of mass and most rapidly changes shape.

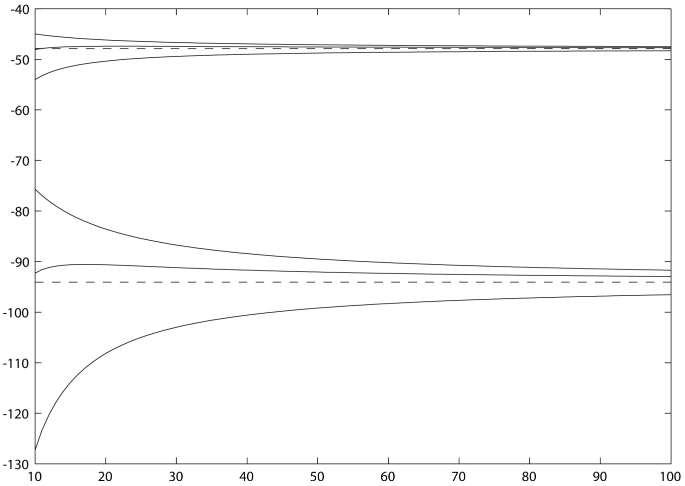

Fig. 2. Marginal log-likelihood is plotted on the y-axis against the number of partitions used on the x-axis. The temperature schedule here is chosen according to $t_i = (i/N)^5$. The upper and lower dashed lines show the analytic log marginal likelihoods for 2 and 20 dimensional linear regression models, respectively. The top and bottom lines for each dimension are the upper and lower bounds, respectively, given by the KL divergence, and the middle lines are the estimates using trapezoidal integration.

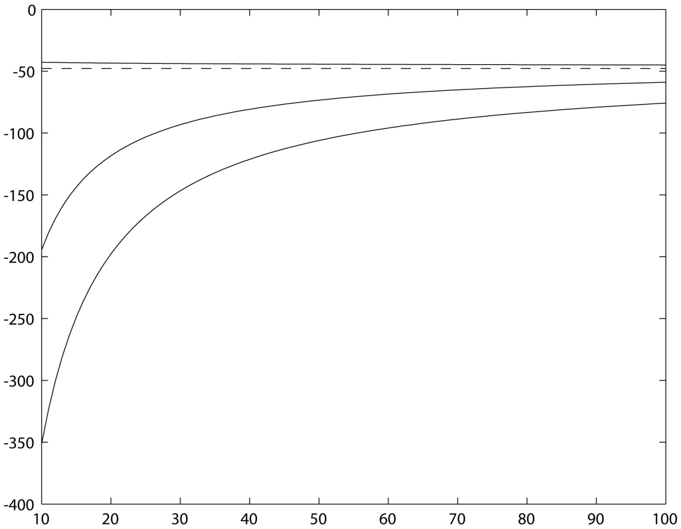

Fig. 3. Marginal log-likelihood is plotted on the y-axis against the number of partitions used on the x-axis. The temperature schedule here is chosen according to a uniform spacing. The dotted line gives the analytic log marginal likelihood for a 2 dimensional linear regression model, the bottom and top lines are the upper and lower bounds, respectively, given by the KL divergence, and the middle line is the estimate using trapezoidal integration. Very large biases appear in the estimates due to the use of a uniform temperature schedule, especially compared to the small bias from a power law temperature distribution.

In Figure 1, when the variance is greater than 1, the 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. 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 higher variance priors will be employed, and so it seems sensible to construct any temperature schedule using a power law distribution with temperature points strongly skewed towards the prior (see also related discussion in (Friel and Pettitt, 2008)).

This also makes sense when considered from a sampling point of view, since when using more advanced population MCMC methods to sample from a ladder of temperature distributions we want the transitions to be as smooth as possible to allow for a good mixing of chains (discussed later in Section 4.3).

It is interesting to see that using a simple uniform distribution of points to define the temperature partitions produces 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. (2007), 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 (Jasra et al., 2007) the results are drawn on the basis of the resulting estimated component means, whereas here 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 $t = 0$. From now on we therefore use a quintic power law spacing to define the temperature schedule when estimating the thermodynamic integral. It has not passed our attention that (10) provides a means of adaptively setting the discretisation of the unit line, however we leave this for future work.

3.2.2 Marginal Likelihoods over Linear Regression Models

We return to the main challenge of estimating marginal likelihoods in practice. As a simple illustrative example consider a linear regression model with a zero-mean unit-variance Gaussian prior on the regression coefficients. Given $m = 30$ samples of d covariates $\mathbf{X} \in \mathbb{R}^{m \times d}$; $d = \{2, 10, 20\}$, and target values $\mathbf{y} \in \mathbb{R}^m$ the marginal likelihood $p(\mathbf{y}|\mathbf{X})$ and power-posteriors $p(\beta|\mathbf{X},\mathbf{y},t)$ can be obtained analytically. Three estimators for the marginal likelihood are used: a Monte Carlo estimate employing samples drawn from the prior over the regression coefficients, a Monte Carlo estimate employing samples drawn from the posterior, also known as the Harmonic Mean estimate, and the discretised numerical approximation of the thermodynamic integral, based on estimating the sum of the expectations of the power-posterior across each temperature interval (15). Each procedure is repeated 100 times to obtain the variance of the estimates. We vary the number of samples used to estimate any expectations to examine the effect of this on the accuracy of the marginal likelihood estimates.

The results are shown in Table 5 where we note that at 10 dimensions sampling from the prior fails completely and even raising the number of samples to 10,000 makes little difference. On the other hand the Harmonic Mean estimator provides a superior, albeit biased, lower variance estimate. However this level of bias is particularly dangerous when relying on Bayes factors to assess the odds in favour of one model over another. In contrast, the quality of the estimates using the thermodynamic integral is quite spectacular in terms of both bias and variance even at the higher dimensions considered.

In the linear regression example (Table 5) each of the 30 temperature steps was such that $t_n = (n/30)^5$. Interestingly, for a 2 dimensional model, a uniform spacing produces a bias of -47.39 compared to the much smaller bias of -0.17 obtained when $t_n = (n/30)^5$ is used.

Table 5
Marginal Log-Likelihood Estimates for Linear Regression Model. The analytic marginal log-likelihoods for 2, 10 and 20 dimensions are $P(\mathbf{y} \mid \mathbf{x}) = -47.87, -67.20$ and -94.05, respectively. The temperature ladder used in computing the power posterior consisted of thirty discrete temperatures with $t_n = (n/30)^5$. The mean estimates and standard errors are shown.

Samples	d	Prior	Posterior	Power Posterior
1000	2	-49.68 ± 6.39	-42.21 ± 0.38	-48.04 ± 0.0013
	10	-417 ± 12088	-45.28 ± 1.62	-67.64 ± 0.0049
	20	-698 ± (-)	-50.28 ± 2.86	-94.86 ± 0.0089
10000	2	-47.97 ± 0.18	-42.35 ± 0.19	-48.04 ± 0.0001
	10	-271 ± 3480	-46.03 ± 1.60	-67.64 ± 0.0005
	20	-698 ± 125	-51.63 ± 1.63	-94.86 ± 0.0008

3.2.3 Bayes Factors over Linear Models

In this example we define two models which are linear in the parameters, generate "experimental" data from the first model, and calculate the Bayes factor 100 times in order to see how accurately we can predict which model produced the data. The experiments are then repeated using data generated from the second model. The Bayes factors are calculated using both importance sampling and thermodynamic integration methods, and the results are compared to the analytically calculated Bayes factors. The marginal likelihoods are calculated under the same experimental conditions as previously for the linear regression models. During the experiments we again vary the number of samples used to investigate the effect of this on the estimates. Note that when thermodynamic integration and sampling from the posterior are employed to calculate Bayes factors, only up to 10,000 samples are used due to computational time limitations. The two models are defined as

$$\text{Model 1:} \qquad y = \beta_2 x_1 + \beta_3 x_2. \tag{20}$$

$$\text{Model 2:} \qquad y = \beta_1 x_1^2 + \beta_2 x_1 + \beta_3 x_2. \tag{21}$$

Bayes factors are calculated using data generated from the first model given by (20), and then using data generated from the second model given by (21). The parameter values used for generating the data are sampled from their Gaussian prior distributions. When model 2 is used to generate data however β_1 is manually varied in order to simulate a more strongly (or weakly) non-linear model response. A β_1 value of 0.1 is used to show the case where the evidence in favour of model 2 is “not worth more than a bare mention”. β_1 values of 0.15 and 0.16 are also used, as these produce Bayes factors which are classed as “substantial” and “strong”, respectively (but not “decisive”) and therefore represent cases where the accuracy of the estimate could most affect the interpretation of the evidence. A summary of how Bayes factors should be interpreted is given in Table 1.

Table 6
Bayes factor results, $B_{1,2}$ using data generated from model 1. The analytic Bayes factor is $B_{1,2} = 28.3$.

Samples	Prior	Posterior	Power Posterior
1000	3.2E+16 ± 9.9E+28	2.39 ± 0.06	33.46 ± 3.26
10,000	968 ± 4.6E+7	2.52 ± 0.04	33.72 ± 0.42
100,000	30.5 ± 118	-	-

The results for data generated from the first model are given in Table 6. 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 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 very high. 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 observe that the Bayes factor is massively underestimated and, worryingly, the variance appears to be very small despite the huge bias.

Indeed, when interpreted using the standard scale, the Bayes factor estimates based on sampling from the posterior would suggest that the difference between the two models is very definitely "Not worth more than a bare mention", whereas the analytic Bayes factor suggests that the difference between models is in fact "Strong". The Bayes factor estimates based on posterior sampling are in this case unable to distinguish between these simple linear models. In contrast, for the thermodynamic integration method, the variance decreases rapidly as the number of Monte Carlo samples used increase, and the estimates correctly state a high confidence that the evidence in favour of model 1 is strong.

Table 7
Bayes factor results, $B_{2,1}$, using data from model 2 with $\beta_1 = 0.10$. The analytic Bayes factor is $B_{2,1} = 0.156$.

Samples	Prior	Posterior	Power Posterior
1000	0.174 ± 0.201	3.00 ± 0.09	0.132 ± 0.00004
10,000	0.166 ± 0.018	2.69 ± 0.05	0.130 ± 0.00001
100,000	0.150 ± 0.002	-	-

Table 8
Bayes factor results, $B_{2,1}$, using data from model 2 with $\beta_1 = 0.15$. The analytic Bayes factor is $B_{2,1} = 6.92$.

Samples	Prior	Posterior	Power Posterior
1000	9.96 ± 387	133.8 ± 180.9	6.2 ± 0.07
10,000	5.82 ± 14.88	117.9 ± 98.2	6.15 ± 0.01
100,000	6.85 ± 2.07	-	-

Table 9
Bayes factor results, $B_{2,1}$, using data from model 2 with $\beta_1 = 0.16$. The analytic Bayes factor is $B_{2,1} = 52.0$.

Samples	Prior	Posterior	Power Posterior
1000	75.9 ± 5.4E+4	1343 ± 1.9E+4	44.1 ± 4.09
10,000	48.6 ± 2907	1154 ± 9822	43.7 ± 0.58
100,000	52.4 ± 272	-	-

Using data generated by model 2, we see again that thermodynamic integration appears to offer the most accurate results in terms of exhibiting smallest bias and variance (Tables 7, 8, 9). It correctly predicts the strength of evidence in favour of model 2 for all values of β_1. Sampling from the prior produced reasonable results, but only when using a very large number of samples, and it should be noted that as β_1 increased the variance associated with its estimates also increased dramatically. Sampling from the posterior produced very poor results. The estimated Bayes factors had large biases even when using a large number of samples. For $\beta_1 = 0.1$, 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 $\beta_1 = 0.15$, the posterior-based estimates describe the difference between models as “Decisive” instead of merely “Substantial” and, for $\beta_1 = 0.16$, as “Decisive” instead of just “Strong”. This reinforces our impression that estimates based on sampling from the posterior should not be blindly trusted.

4 Nonlinear ODE Models

We now apply the insights we have gained regarding the discretisation strategy in thermodynamic integration, to the challenging application of estimating Bayes factors over ODE based models. The quality estimates obtained from the thermodynamic integral in Section 3.2.3 are perhaps not surprising given the well-behaved (log-concave) nature of the densities associated with the linear regression model. When the power-posterior is multimodal, and proper mixing difficult to achieve, the standard Metropolis method of sampling the power-posteriors $p(\theta|\mathbf{Y},\tau,t)$ presents the danger of obtaining poor estimates for each $E_{\theta|\mathbf{Y},\tau,t}\{\log p(\mathbf{Y}|\theta,\tau)\}$, since without extreme care the Markov chains may easily converge to local maxima. Recent advances in MCMC methodology suggest solutions to this problem of multimodality in the form of population-based MCMC methods (Jasra et al., 2007), which we therefore implement to sample the structural parameters of our models.

Linear ODE models are not so useful for investigating this potential problem as they induce log-concave posterior densities which, like the posteriors induced by the linear regression models in the previous section, are straightforward to sample from. We therefore turn our attention to nonlinear ODE models that induce multimodal posterior densities.

4.1 Bayesian Inference over Nonlinear ODE Models

We briefly introduce ordinary differential equation (ODE) models and give a brief overview of how free model parameters may be inferred from experimental time-series data using the Bayesian framework. A dynamical system may be described by a collection of G ODEs and model parameters θ, which define a functional relationship between the process state, $\mathbf{x}(\tau)$ (where $\mathbf{x}(\tau)$ is G dimensional, and τ represents a point in time), and its time derivative $\dot{\mathbf{x}}(\tau)$. Such a system of ODEs may be written compactly as $\dot{\mathbf{x}}(\tau) = \mathbf{f}(\mathbf{x},\theta,\tau)$ (where $\mathbf{f}$ is an G-dimensional vector field). A sequence of observations, $\mathbf{y}(\tau)$, of the process we wish to model are usually contaminated with some measurement error which is modeled as $\mathbf{y}(\tau) = \mathbf{x}(\tau) + \varepsilon(\tau)$ where $\varepsilon(\tau)$ defines an appropriate multivariate noise process, e.g. a zero-mean Gaussian noise process with variance σ_g^2 for each of the G states. If observations are made at T distinct time points, then the $G \times T$ matrices summarise the overall observed system as $\mathbf{Y} = \mathbf{X} + \mathbf{E}$. In order to obtain values for $\mathbf{X}$, the system of ODEs can be solved numerically, so that in the case of an initial value problem $\mathbf{X}(\theta, \mathbf{x}_0)$ denotes the solution to the system of equations at the specified time points for the parameters θ and initial conditions $\mathbf{x}_0$. The posterior density follows by employing appropriate priors such that $p(\theta, \mathbf{x}_0, \sigma|\mathbf{Y}) \propto \pi(\theta)\pi(\mathbf{x}_0)\pi(\sigma)\prod_g N_{\mathbf{Y}_{g,\cdot}}(\mathbf{X}(\theta, \mathbf{x}_0)_{g,\cdot}, \mathbf{I}\sigma_g^2)$, where N is a T-dimensional Gaussian distribution evaluated at all observed time points, and the desired marginal $p(\theta|\mathbf{Y})$ can be obtained from this joint posterior. A Metropolis style sampling scheme can be devised to sample from the joint posterior. However, as a consequence of the dynamics induced by the system, the corresponding likelihood surface defined by $p(\mathbf{Y}|\theta, \mathbf{x}_0, \sigma)$ can present formidable challenges to standard sampling methods, as will be demonstrated in the following example.

4.2 The Goodwin Model of Biochemical Oscillatory Control

As an illustrative example of the challenges of performing Bayesian inference over nonlinear ODE model parameters and assessing the validity of alternative model structures, we employ models of oscillatory enzymatic control, specifically the Goodwin model (Goodwin, 1965). This model has become the standard basic mechanism for periodic protein expression, driven by a negative feedback loop which inhibits mRNA transcription. Indeed, recent experimental evidence has shown that essential elements of the circadian clock in many organisms consist of negative feedback loops (Locke et al., 2005), similar to those in Goodwin's original model. The classical g-variable Goodwin model is defined as,

$$\begin{aligned} \frac{dx_1}{d\tau} &= \frac{a_1}{1 + a_2 x_g^{\rho}} - \alpha x_1 \\ \frac{dx_2}{d\tau} &= k_1 x_1 - \alpha x_2 \\ &\vdots \\ \frac{dx_g}{d\tau} &= k_{g-1} x_{g-1} - \alpha x_g, \end{aligned} \tag{22}$$

where τ is time, x_1 and x_2 correspond to the levels of mRNA and protein in the system, respectively, while x_3 to x_g correspond to other forms of proteins, with x_g ultimately inhibiting mRNA production. Output depends on the relationship between the synthesis rate constants, a_1 and $k_{1,\ldots,g-1}$, and the degradation rate constants, a_2 and α. It has been shown that this simple Goodwin model has unstable steady states only when $\rho > 8$, and we therefore set $\rho = 10$ as a fixed parameter so that we may be certain of oscillatory responses for a wide variety of parameter values. As g increases, so does the time taken for the negative feedback to propagate through the system, enabling a more dynamic range of responses. A g-variable Goodwin model therefore has $g+2$ tunable parameters. We note that while this simple model does not precisely describe the actual biology occurring in nature, it does induce a posterior distribution with the characteristic of multimodality common to the more complex models currently in use, for example (Locke et al., 2005). We therefore use it to elucidate the general problem of sampling from multimodal distributions which commonly occur. (We note that such nonlinear genetic networks may alternatively be modeled using stochastic differential equations within a Bayesian framework, see e.g. (Golightly and Wilkinson, 2007)).

The set of parameters to be inferred is therefore $\theta = \{\alpha, a_{1,2}, k_{1,...,g-1}\}$. We also denote any observed time series data by $\mathbf{Y} = [\mathbf{y}_1, ..., \mathbf{y}_g]$, where $\mathbf{y}_g$ is the vector of observed data for species g at the time points specified by the vector τ.

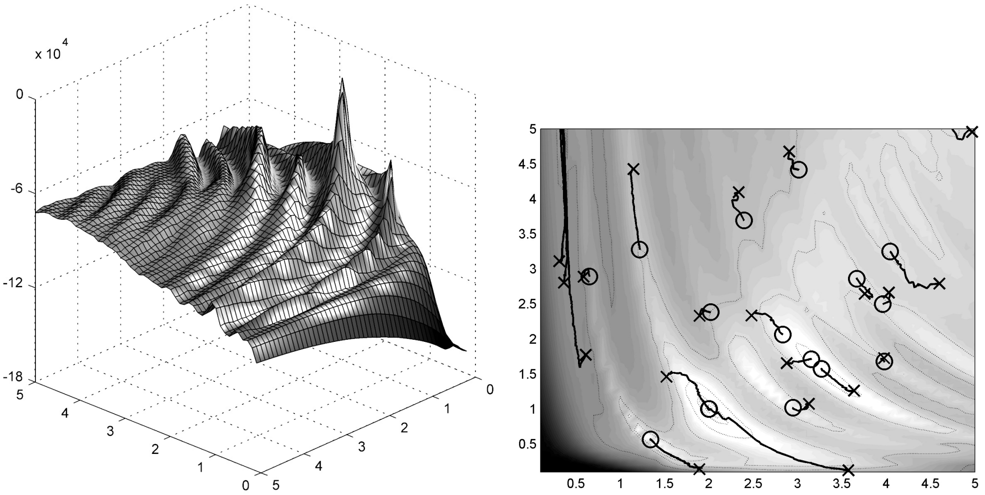

Fig. 4. Left plot: Posterior surface of two parameters of a Goodwin oscillator model (with z-axis in log-scale). Right plot: The progress of twenty independent Metropolis samplers, showing the starting positions (denoted by ×), path taken and finishing positions (denoted by ∘). The trapping of chains in local modes is most apparent.

Consider the conditional posterior surface over two parameters of a Goodwin circadian oscillator model (Goodwin, 1965) shown in Figure 4. We can see rather severe ridges of high posterior values which cause a Metropolis sampler to get caught in these local modes. The right plot in Figure 4 shows the trace of samples obtained from 20 independent Metropolis samplers initialised at random parts of the parameter space, indicated by a ×, with the final sample denoted by a ∘. The localisation of the chains on the ridges is all too apparent and we will see shortly how this has a detrimental impact on the estimation of Bayes factors for model comparison. As previously mentioned, a possible solution to this sampling problem is available through the use of population MCMC methods, see e.g. (Iba, 2000; Liang and Wong, 2001; Laskey and Myers, 2003; Jasra et al., 2007). Such population MCMC methods can be very efficient in the context of model comparison because not only do they allow sampling from highly nonlinear multimodal posterior distributions, but the usually redundant samples taken from intermediate temperatures may also be reused in the estimation of the marginal likelihood using thermodynamic integration (Friel and Pettitt, 2008).

4.3 Population-based MCMC

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

$$\widetilde{p}(\theta|\mathbf{t}) = \prod_{n=1}^{N} p(\theta|t_n), \tag{23}$$

and the desired target density $p(\theta)$ is defined for one value of t_n. It is convenient to fix a geometric path between the prior and posterior, which we do in our implementation such that

$$\widetilde{p}(\theta|\mathbf{y},\mathbf{t}) \propto \prod_{n=1}^{N} \pi(\theta)p(\mathbf{y}|\theta)^{t_n}, \tag{24}$$

where $\pi(\theta)$ is the prior and $p(\mathbf{y}|\theta)^{t_n}$ is the likelihood, for $t_n \in [0,1]$. We note that although other sequences are also possible (Gelman and Meng, 1998), this particular formulation allows us to sample from all the required power posteriors simultaneously, which may be later employed in thermodynamic integration. A time homogeneous Markov transition kernel which has $p(\theta)$ as its stationary distribution can be constructed from both local proposal moves and global moves between the tempered chains of the population (Liang and Wong, 2001; Laskey and Myers, 2003; Jasra et al., 2007) thus allowing freer movement within the parameter space. Local moves are made by selecting a chain at a some temperature t_n with parameters θ_{t_n} and adding a normally distributed random vector to create a new proposed set of parameters, θ'_{t_n}. This new set of parameters is accepted according to the standard Metropolis-Hastings acceptance ratio, with probability $\min(1,r)$, where $r = p(\mathbf{y}|\theta'_{t_n})^{t_n}/p(\mathbf{y}|\theta_{t_n})^{t_n}$. Global moves are made by randomly selecting two adjacent temperatures t_n and t_{n+1}, and swapping over the parameter values of each chain, so that the proposed parameters are $\theta'_{t_n} = \theta_{t_{n+1}}$ and $\theta'_{t_{n+1}} = \theta_{t_n}$. These parameters are then accepted with ratio $\min(1,r)$, where now $r = [p(\mathbf{y}|\theta'_{t_n})^{t_n} p(\mathbf{y}|\theta'_{t_{n+1}})^{t_{n+1}}]/[p(\mathbf{y}|\theta_{t_n})^{t_n} p(\mathbf{y}|\theta_{t_{n+1}})^{t_{n+1}}]$.

The right hand plot of Figure 4 shows how each of the independent chains of a Metropolis sampler, having only local moves, get stuck at various local modes in the posterior density. Whereas in Figure 5 we see three tempered chains at $t_n = \{0.0001, 0.5, 1\}$, i.e. ranging from effectively the prior, to an intermediate

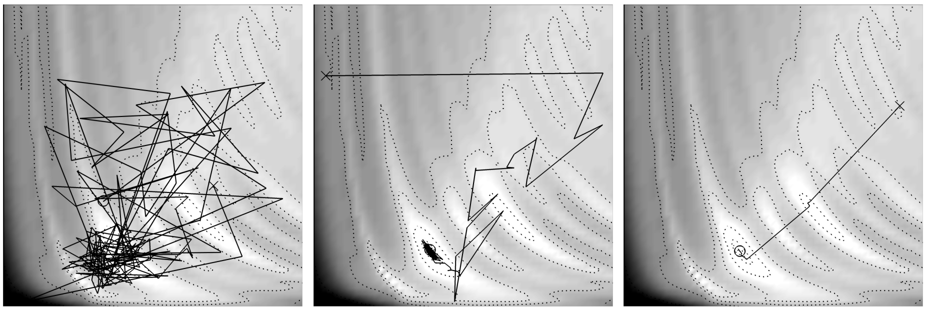

Fig. 5. Left Plot: Samples obtained from a chain with t close to 0, which is sampling from a smooth distribution strongly influenced by the prior, the free movement within the parameter space is quite clear to see. The iso-contours of the posterior are also plotted in this case. Middle Plot: Progress 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 due to the exchange proposals between temperatures. Right Plot: Samples drawn from the posterior when $t = 1$; compare this with the highly localised *sticky* exploration in Figure 4. The ability to move between modes is clear.

power-posterior, and finally to the posterior itself at $t_n = 1$. For t_n close to 0 the samples are being drawn from several of the modes in the posterior. Note that these moves are due to local Metropolis moves as well as proposals which sample between different temperatures. At the intermediate temperature a much freer traversal of the parameter space is possible, with large global *mode-hopping* steps being made at $t_n = 1$, which is the posterior distribution. Clearly the estimates of $E_{\theta|\mathbf{Y},\tau,t_n}\{\log p(\mathbf{Y}|\theta,\tau)\}$ at each temperature will be superior than those obtained from a Metropolis sampler at every temperature and this will be highlighted in section 4.4.

4.4 Parameter Identification via Posterior Inference

An oscillatory system response, consisting of 120 noisy observations of the first two chemical species made at equally spaced time intervals, was obtained from a g-variable Goodwin Model, for $g = \{3,5\}$ with $x_{1,\dots,g} = 0$ at time $\tau = 0$. Gamma priors were placed on the free parameters, such that $a_{1:2}, k_{1:g-1}, \alpha \sim \Gamma(2,1)$. The specific values of the parameters for both models were drawn from their chosen prior Gamma distributions and Gaussian distributed noise with variance $\sigma^2 = 0.2$ 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 $\sigma^2 = 0.2$ (Note that when using real experimental data the noise variance σ^2 would be unknown and could be inferred as an additional parameter). The overall likelihood was therefore the product of these errors over all data points. Note however that only the last 80 data points were used for inference, to allow the models to settle into a steady oscillatory state from their initial values, which were fixed at 0.

For the $g = 5$ model, we show the conditional posterior for two of the model parameters in the left hand plot of Figure 4. The jagged nature of the posterior surface hints at the challenge of appropriately sampling from the full $g = 5$-dimensional posterior. The right hand plot of Figure 4 clearly shows that multiple Metropolis samplers with adaptive proposal distributions suffer badly from poor mixing, which motivates the adoption of Population MCMC methods.

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 were used, distributed along a temperature schedule given by $t_n = (n/10)^5$. 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 6 shows the estimated marginal posteriors for the $g = 3$ 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 severely biased estimates of the posteriors, as can be seen from Figure 7.

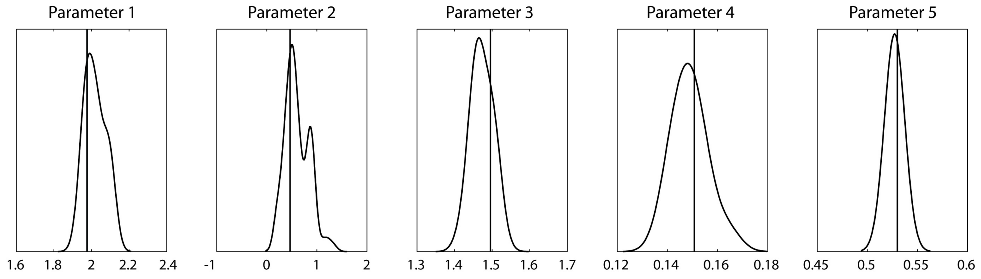

Fig. 6. The marginal posteriors obtained from population MCMC for each of the parameters of a Goodwin oscillator model. The values of the true parameter values are indicated by a black vertical line which coincides very well with the highest density regions of the posteriors.

4.5 Model Comparison using Bayes Factors

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

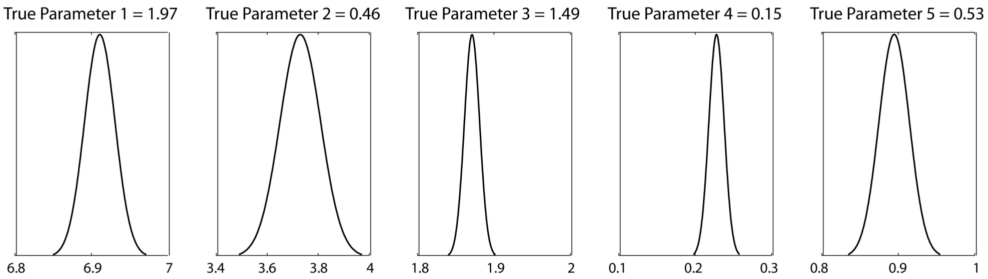

Fig. 7. The posteriors obtained from a Metropolis sampler with adaptive proposal distributions and convergence monitoring with the $\hat{R}$ statistic (Gelman and Rubin, 1992; El Adlouni et al., 2006). The woeful bias in the estimates of the posteriors are most apparent when compared to Figure 6.

likelihoods were calculated 10 times using each method for each combination of model and data used. Averages and variances were then calculated.

Table 10
Marginal Log-Likelihoods & Bayes Factors for Goodwin Models Using Metropolis (Mean ± S.E.)

	Simple Data	Complex Data
Simple Model	$-586 \pm 22,715$	$-1,623 \pm 40,710$
Complex Model	$-782 \pm 116,869$	$-600 \pm 891,103$
$\log B_{S,C}$	$195 \pm 205,745$	-
$\log B_{C,S}$	-	$1,022 \pm 802,184$

Table 11
Marginal Log-Likelihoods & Bayes Factors for Goodwin Models Using Population MCMC (Mean ± S.E.)

	Simple Data	Complex Data
Simple Model	-426 ± 31	$-1,432 \pm 37$
Complex Model	-536 ± 67	-190 ± 47
$\log B_{S,C}$	110 ± 93	-
$\log B_{C,S}$	-	$1,242 \pm 117$

Convergence of the Markov chains to a stationary distribution was carefully assessed for each sampling method using the $\hat{R}$ statistic (Gelman and Rubin, 1992). This statistic was calculated with samples from parallel running chains, produced from 3 parallel population MCMC simulations, to evaluate when the chains have reached an equilibrium, by comparing the in-chain and between-chain variances. 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.

In Tables 10 and 11, 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 10) making these estimates of little practical value 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 (Table 11) 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.

5 Conclusions

In this paper we have reviewed three methods for estimating marginal likelihoods and have gained important insights into the difficulties of calculating accurate Bayes factors by considering simple linear regression models. We have highlighted the dangers of employing the commonly used Posterior Harmonic Mean estimator and shown that methods involving thermodynamic integration provide much more stable estimates. We have also characterised the error associated with the discretised approximation of the thermodynamic integral in terms of the KL divergences between the posterior distributions across each temperature interval, and we have shown that by using a temperature schedule with partitions clustered towards $t = 0$ it is possible to obtain estimates of the marginal likelihood with extremely small bias.

We conclude that standard MCMC methodology is inappropriate for marginal likelihood estimation over highly nonlinear models, such as those based on nonlinear ODEs, since even employing thermodynamic integration it may produce such high variance estimates of Bayes factors as to render them completely uninformative. Population MCMC methodology, on the other hand, may be elegantly combined with the Thermodynamic Integral not only to sample simultaneously from a range of tempered nonlinear posterior distributions, but also to produce low variance estimates of Bayes factors for informative model comparison.

Acknowledgements

Ben Calderhead is supported by Microsoft Research through its PhD Scholarship Programme. Mark Girolami is supported by an Engineering and Physical Sciences Research Council (EPSRC) Advanced Research Fellowship EP/EO52029/1, and the Biotechnology and Biological Sciences Research Council (BBSRC) project grant BB/G006997/1.

A Derivation of Optimal Density for Temperature Schedule

Here we derive the analytic expression of (10) for a linear model. 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 the regions of highest mass. We make use of the following identities for the expectation operator (see The Matrix Reference Manual, http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/intro.html), where β is a stochastic vector drawn from a Gaussian distribution with mean μ, and covariance Σ.

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

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

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

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

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

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 Section 3, the power posterior distributions are Gaussian, with mean μ_t, and covariance Σ_t (section 3.1.1). We proceed by first multiplying out the brackets and noting that the expectation operator is linear

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

where E denotes the expectation with respect to the Gaussian distribution $p(\beta \mid \mathbf{y},t,\sigma^2,\zeta^2)$, as will be used from now on. An analytic expression for the second term in (A.6) may be found using identity (A.2)

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

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

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

The expectation of Expression 1 may be calculated by multiplying out the brackets and using (A.1) and (A.3)

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

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

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

Expression 2a admits an analytic form trivially as follows

$$E\left[\mathbf{y}^T\mathbf{B}\beta\mathbf{y}^T\mathbf{y}\right]=\mathbf{y}^T\mathbf{y}E\left[\mathbf{y}^T\mathbf{B}\beta\right]=\mathbf{y}^T\mathbf{y}\mathbf{y}^T\mathbf{B}\mu.$$

Expression 2b admits an analytic form using (A.3)

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

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

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

Expression 2d admits an analytic form making use of (A.5)

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

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

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

References

Del Moral, P., Doucet, A., Jasra, A., 2006. Sequential Monte Carlo samplers. Journal of the Royal Statistical Society B 68 (3), 411–436.

Del Moral, P., Doucet, A., Jasra, A., 2007. Bayesian Statistics. Oxford University Press, Ch. Sequential Monte Carlo for Bayesian Computation, pp. 1–34.

El Adlouni, S., Favre, C., Bobee, B., 2006. Comparison of methodologies to assess the convergence of Markov Chain Monte Carlo methods. Computational Statistics and Data Analysis 50 (10), 2685–2701.

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

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

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

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

Golightly, A., Wilkinson, D., 2007. Bayesian inference for nonlinear multivariate diffusion models observed with error. Computational Statistics and Data Analysis 52 (3), 1674–1693.

Goodwin, B., 1965. Oscillatory behavior in enzymatic control processes. Adv. Enzyme Regul. 3, 425–438.

Iba, Y., 2000. Population Monte Carlo algorithms. Transactions of the Japanese Society of Artificial Intelligence 16, 279–286.

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

Kass, R., Raftery, A., 1995. Bayes factors. American Statistical Association 90 (430), 773–795.

Lartillot, N., Philippe, H., 2006. Computing Bayes factors using thermodynamic integration. Syst. Biol. 55 (2), 195–207.

Laskey, K., Myers, J., 2003. Population Markov Chain Monte Carlo. Machine Learning 50, 175–196.

Liang, F., Wong, W., 2001. Real-parameter evolutionary Monte Carlo with applications to Bayesian mixture models. American Statistical Association 96 (454), 653–666.

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

McCulloch, R. E., Rossi, P. E., 1991. Bayes factors for nonlinear hypotheses and likelihood distributions. Tech. Rep. Technical Report 101, Statistics Research Center, University of Chicago, Graduate School of Business.

Neal, R., 2001. Annealed importance sampling. Statistics and Computing 11, 125–139.

Newton, M., Raftery, A., 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. Journal of the Royal Statistical Society: Series B 56 (1), 3–48.

Raftery, A., Newton, M., Satagopan, J., Krivitsky, P., 2007. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. Bayesian Statistics 8, 1–45.

Robert, C., Casella, G., 2004. Monte Carlo Statistical Methods. Springer.

Vyshemirsky, V., Girolami, M. A., 2008. Bayesian ranking of biochemical system models. Bioinformatics 24(6), 833–839.