UNIVERSITY OF
BATH

The original publication is available at www.springerlink.com

University of Bath

Improving power posterior estimation of statistical evidence

Nial Friel[*], Merrilee Hurn[†]and Jason Wyse[‡]

March 21, 2013

Abstract

The statistical evidence (or marginal likelihood) is a key quantity in Bayesian statistics, allowing one to assess the probability of the data given the model under investigation. This paper focuses on refining the power posterior approach to improve estimation of the evidence. The power posterior method involves transitioning from the prior to the posterior by powering the likelihood by an inverse temperature. In common with other tempering algorithms, the power posterior involves some degree of tuning. The main contributions of this article are twofold – we present a result from the numerical analysis literature which can reduce the bias in the estimate of the evidence by addressing the error arising from numerically integrating across the inverse temperatures. We also tackle the selection of the inverse temperature ladder, applying this approach additionally to the Stepping Stone sampler estimation of evidence. A key practical point is that both of these innovations incur virtually no extra cost.

Keywords: Marginal likelihood, Markov chain Monte Carlo, Power posteriors, Statistical evidence, Stepping Stone sampler, Tempering, Thermodynamic integration.

1 Introduction

The statistical evidence (sometimes called the marginal likelihood or integrated likelihood) is a vital quantity in Bayesian statistics for the comparison of models, $m_1, \ldots, m_l$. Under the Bayesian paradigm we consider the posterior distribution

$$p(\theta_i, m_i|y) \propto p(y|\theta_i, m_i)p(\theta_i|m_i)p(m_i), \quad \text{for } i = 1, \ldots, l, \tag{1}$$

for data y and parameters θ_i within model m_i, where $p(\theta_i|m_i)$ denotes the prior distribution for parameters within model m_i and where $p(m_i)$ denotes the prior model probability. The evidence for data y given model

[*]School of Mathematical Sciences, University College Dublin, Belfield, Dublin 4, Republic of Ireland; Email: nial.friel@ucd.ie

[†]Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, UK; Email: M.A.Hurn@bath.ac.uk

[‡]School of Computer Science and Statistics, Trinity College Dublin, College Green, Dublin 2, Republic of Ireland; Email: wyseja@scss.tcd.ie

1

m_i arises as the normalising constant of the posterior distribution within model m_i,

$$p(\theta_i|y, m_i) \propto p(y|\theta_i, m_i)p(\theta_i|m_i), \tag{2}$$

and thus results from integrating the un-normalised posterior across the θ_i parameter space,

$$p(y|m_i) = \int_{\theta_i} p(y|\theta_i, m_i)p(\theta_i|m_i) \, d\theta_i. \tag{3}$$

This of course assumes that the prior distribution for θ_i is proper. The marginal likelihood is often then used to calculate Bayes factors when one wants to compare two competing models, m_i and m_j,

$$BF_{ij} = \frac{p(y|m_i)}{p(y|m_j)} = \frac{p(m_i|y)}{p(m_j|y)} \frac{p(m_j)}{p(m_i)}. \tag{4}$$

Here, $p(m_i|y)$ is the posterior probability for model m_i and it can be evaluated, using the evidence for each of the collection of models under consideration,

$$p(m_i|y) \propto p(y|m_i)p(m_i), \quad \text{for } i = 1, \dots, l. \tag{5}$$

Estimation of the evidence is a non-trivial task for most statistical models and there has been considerable effort in the literature to find algorithms and methods for this purpose. Laplace's method (Tierney and Kadane, 1986) is an early approach and very widely used. Other notable and popular approaches include Chib's method (Chib 1995), annealed importance sampling (Neal 2001), nested sampling (Skilling 2006), bridge sampling (Meng and Wong, 1996) and power posteriors (Friel and Pettitt, 2008) which is the focus of this paper. For a recent review and perspective on these and other methods, see Friel and Wyse (2012).

This paper is organised as follows. Section 2 outlines the power posterior method, and the approach we propose to improve estimation of the evidence. Section 3 illustrates the potential gain from implementing the methodology which we propose. We offer some conclusions in Section 4.

2   The power posterior approach

In what follows we will drop the explicit conditioning on model m_i for notational simplicity. We follow the notation of Friel and Pettitt (2008) and denote the power posterior by

$$p_t(\theta|y) \quad \propto \quad p(y|\theta)^t p(\theta), \ t \in [0,1] \tag{6}$$

$$\text{with } z(y|t) \quad = \quad \int_\theta p(y|\theta)^t p(\theta) d\theta. \tag{7}$$

where $t \in [0,1]$ is thought of as an inverse temperature, which has the effect of tempering the likelihood, whereby at the extreme ends of the inverse temperature range, $p_0(\theta|y)$ and $p_1(\theta|y)$ correspond to the prior

and posterior, respectively. The power posterior estimator for the evidence relies on noting that

$$
\begin{aligned}
\frac{d}{dt}\log(z(y|t)) &= \frac{1}{z(y|t)}\frac{d}{dt}z(y|t) \\
&= \frac{1}{z(y|t)}\int_\theta \frac{d}{dt}p(y|\theta)^t p(\theta)d\theta \\
&= \frac{1}{z(y|t)}\int_\theta p(y|\theta)^t \log(p(y|\theta))p(\theta)d\theta \\
&= \int_\theta \frac{p(y|\theta)^t p(\theta)}{z(y|t)}\log(p(y|\theta))d\theta \\
&= \mathbf{E}_{\theta|y,t}\log(p(y|\theta)).
\end{aligned}
\tag{8}
$$

As a result

$$
\begin{aligned}
\int_0^1 \mathbf{E}_{\theta|y,t}\log(p(y|\theta))dt &= \left[\,\log(z(y|t))\,\right]_0^1 \\
&= \log(z(y|t=1)) \text{ (assuming that the prior is normalised)}
\end{aligned}
\tag{9}
$$

which is the log of the desired marginal likelihood.

In practice the inverse temperature range is discretised as $0 = t_0 < t_1, \ldots, t_n = 1$ to form an estimator based on (9). For each t_i, a sample from $p(\theta|y,t_i)$ can be used to estimate $\mathbf{E}_{\theta|y,t_i}\log(p(y|\theta))$. Finally, a trapezoidal rule is used to approximate

$$
\log p(y) \approx \sum_{i=1}^n (t_i - t_{i-1})\left(\frac{\mathbf{E}_{\theta|y,t_{i-1}}\log(p(y|\theta)) + \mathbf{E}_{\theta|y,t_i}\log(p(y|\theta))}{2}\right).
\tag{10}
$$

Discretising t introduces an approximation into this method and the two goals of this paper are to reduce the bias in the power posterior estimation method due to the approximation and also to find an adaptive method for choosing the inverse temperatures (which we also refer to as rungs, following a common analogy of an inverse temperature ladder between prior and posterior). For both of these we will exploit the fact that the gradient of the expected log deviance curve equals its variance, as we now outline.

Differentiating $\mathbf{E}_{\theta|y,t}\log(p(y|\theta))$ with respect to t yields

$$
\begin{aligned}
\frac{d}{dt}\mathbf{E}_{\theta|y,t}\log(p(y|\theta)) &= \int_\theta \log(p(y|\theta))\frac{d}{dt}p_t(\theta|y)d\theta \\
&= \int_\theta \log(p(y|\theta))\left[\log(p(y|\theta)) - \frac{1}{z(y|t)}\frac{d}{dt}z(y|t)\right]p_t(\theta|y)d\theta \\
&= \int_\theta \log(p(y|\theta))\left[\log(p(y|\theta)) - \frac{d}{dt}\log(z(y|t))\right]p_t(\theta|y)d\theta \\
&= \mathbf{E}_{\theta|y,t}\log(p(y|\theta))^2 - (\mathbf{E}_{\theta|y,t}\log(p(y|\theta)))^2 \\
&= \mathbf{V}_{\theta|y,t}(\log(p(y|\theta)))
\end{aligned}
\tag{11}
$$

where $\mathbf{V}_{\theta|y,t}(\log(p(y|\theta)))$ denotes the variance of the log deviance at inverse temperature t.

2.1 Reducing the bias by improving the numerical integration

Equation (11) immediately provides two useful pieces of information. First, the curve which we wish to integrate numerically is (strictly) increasing. Secondly, we can improve upon the standard trapezium rule used to numerically integrate the expected log deviance by incorporating derivative information at virtually no extra computational cost (the cost merely of calculating the variance of a set of simulations for fixed t). We do this by using the corrected trapezium rule which comes from an error analysis of the standard trapezium rule, see for example Atkinson and Han (2004), Section 5.2; when integrating a function f between points a and b

$$\int_a^b f(x)dx = (b-a)\left[\frac{f(b)+f(a)}{2}\right] - \frac{(b-a)^3}{12}f''(c) \tag{12}$$

where c is some point in $[a,b]$. The first term of the right hand side of this equation is the usual trapezium rule and the second can be approximated using

$$f''(c) \approx \frac{f'(b)-f'(a)}{b-a}$$

$$\text{so that } \int_a^b f(x)dx \approx (b-a)\left[\frac{f(b)+f(a)}{2}\right] - \frac{(b-a)^2}{12}\left[f'(b)-f'(a)\right]. \tag{13}$$

This latter form motivates the corrected trapezium rule which for unequally spaced x-axis points, taken together with the information derived above regarding the derivative of the log deviance gives

$$\begin{aligned}\log(z(y|t=1)) &\approx \sum_{i=0}^{n-1}(t_{i+1}-t_i)\left[\frac{\mathbf{E}_{\theta|y,t_i}\log(p(y|\theta)) + \mathbf{E}_{\theta|y,t_{i+1}}\log(p(y|\theta))}{2}\right] \\ &- \sum_{i=0}^{n-1}\frac{(t_{i+1}-t_i)^2}{12}\left[\mathbf{V}_{\theta|y,t_{i+1}}(\log(p(y|\theta))) - \mathbf{V}_{\theta|y,t_i}(\log(p(y|\theta)))\right]\end{aligned} \tag{14}$$

where both the expectations $\{\mathbf{E}_{\theta|y,t_i}\log(p(y|\theta))\}$ and variances $\{\mathbf{V}_{\theta|y,t_i}(\log(p(y|\theta)))\}$ are to be estimated using MCMC runs at a number of values of t_i. We will refer to the algorithm incorporating this correction as the modified power posterior.

2.2 Adaptive choice of the inverse temperature placement

The next question which arises is how to choose the $\{t_i\}$ between $t_0 = 0$ and $t_n = 1$. Friel and Pettitt (2008) find that setting $t_i = (i/n)^5$ performs well. We refer to this as the powered fraction (PF) schedule. Lartillot and Philippe (2006) discuss very similar ideas in the phylogenetics literature, although using Simpson's rule for the numerical integration; they use equally spaced inverse temperatures between 0 and 1.

Here we will only consider the discretisation error associated with the numerical integration, rather than the stochastic error arising with sampling from the different $p_{t_i}(\theta|y)$. Calderhead and Girolami (2009)

show that this discretisation error depends upon the Kullback-Leibler distance between successive $p_{t_i}(\theta|y)$. Lefebvre, Steele and Vandal (2010) also consider a symmetrised Kullback-Leibler divergence in picking optimal schedules for path sampling. At first glance the Kullback-Leibler distance does not seem a particularly tractable quantity to manipulate. However, these papers and Behrens, Friel and Hurn (2012) all note that, in the notation of this paper,

$$\sum_{i=0}^{n-1}(KL[p_{t_i}(\theta|y), p_{t_{i+1}}(\theta|y)] + KL[p_{t_{i+1}}(\theta|y), p_{t_i}(\theta|y)]) = 2S_n(t_0, \ldots, t_n) \tag{15}$$

where KL denotes the Kullback-Leibler distance and

$$S_n(t_0, \ldots, t_n) = \sum_{i=0}^{n-1}(t_{i+1} - t_i)\mathbf{E}_{\theta|y,t_{i+1}}\log(p(y|\theta)) - \sum_{i=0}^{n-1}(t_{i+1} - t_i)\mathbf{E}_{\theta|y,t_i}\log(p(y|\theta)). \tag{16}$$

S_n can be interpreted graphically as the sum of the rectangular areas between a lower and an upper approximation to the integral of $\mathbf{E}_{\theta|y,t_i}\log(p(y|\theta))$ between $t_0 = 0$ and $t_1 = 1$. Behrens, Friel and Hurn (2012) use minimising S_n as a rationale for choosing the inverse temperatures in tempered transitions. We propose to use the same target in selecting the $\{t_i\}$ for power posteriors. However, unlike in tempered transitions where the tuning forms a small part of the overall computational load, here the cost is almost exclusively the estimation of $\mathbf{E}_{\theta|y,t_i}\log(p(y|\theta))$ and its gradient.

We propose the following approach. Begin by estimating the expected log deviance and its gradient at $t = 1$ then $t = 0$ (both points which are needed in all possible schemes). Where to site the next t? There is no analytic solution without knowing the curve and we do not want to use computational resources in performing a search for the optimal location. Instead we find the intersection of the two straight lines defined by our current knowledge of the curve: If the estimated function and gradient at t_k are denoted by $\hat{f}_k$ and $\hat{V}_k$ respectively, and those at t_{k+1} by $\hat{f}_{k+1}$ and $\hat{V}_{k+1}$, we set the new point to be

$$t = \frac{\hat{f}_{k+1} - \hat{f}_k + \hat{f}_k\hat{V}_k - \hat{f}_{k+1}\hat{V}_{k+1}}{\hat{V}_k - \hat{V}_{k+1}}. \tag{17}$$

If this intersecting point is outside the interval $[t_k, t_{k+1}]$, it suggests there is some sort of inflection within the interval and instead we use a simple weighted average.

$$t = t_k + \frac{\hat{V}_{k+1}}{\hat{V}_k + \hat{V}_{k+1}}(t_{k+1} - t_k). \tag{18}$$

This now gives us two rectangular contributions to S_n. We identify the larger of these terms and locate the next point in the corresponding interval, and so on iteratively. This scheme will almost certainly not identify the optimal placing of the $n - 1$ interior rungs. However it is quick, cheap and intuitively reasonable. (In practice, Monte Carlo error can mean that the function is not increasing and so the criterion is changed to picking the interval with the largest absolute contribution to S_n. In the event of picking such an interval, we set the new t to be the midpoint of the interval reflecting the significant levels of uncertainty.)

2.3 Computational details

We use sequential runs of MCMC, beginning with sampling from the posterior, $t = 1$. From each MCMC run we store the estimated expected log deviance, its estimated gradient and the values at the last iteration of the run. These final values are used as starting points for runs at a smaller value of t. In the case of the power fraction placements where the inverse temperatures are deterministic, this means that the run at t_k is initialised with the values from t_{k+1}. In the adaptive placements, the run at each new t is initialised with the values of the currently closest larger t_k.

3 Examples

We present three examples, the first two of which were included in the review paper by Friel and Wyse (2012) where the performance of power posteriors was compared to some other existing methods. The first is a non-nested linear regression comparison for which the marginal likelihoods can be calculated analytically. Here our experiments concentrate on the effects of the modified integration rule and the adaptive placement of the inverse temperatures. Example 2 is a larger problem, choosing between two logistic regression models, for which an analytic solution is not possible; we use this example to compare the improved algorithm with the Stepping Stone sampling approach of Xie *et al* (2011). The final example is by far the largest and exhibits the most interestingly shaped $\mathbf{E}_{\theta|y,t} \log(p(y|\theta))$, the focus here is on the stability of the approach to poor estimation of the gradient.

3.1 Example 1: Radiata pine

The first example compares two linear regression models for the Radiata pine data originally in Williams (1959). The response variable here is the maximum compression strength parallel to the grain, y_i, while the predictors are density, x_i, or density adjusted for resin content, z_i, for $n = 42$ specimens of radiata pine. Two possible Gaussian linear regression models are considered;

$$\text{Model 1:} \quad y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i, \quad \epsilon_i \sim N(0, \tau^{-1}), \quad i = 1, \ldots, n,$$
$$\text{Model 2:} \quad y_i = \gamma + \delta(z_i - \bar{z}) + \eta_i, \quad \eta_i \sim N(0, \lambda^{-1}), \quad i = 1, \ldots, n.$$

Priors are chosen to match the analyses of Friel and Wyse (2012) (baring a notational factor of 2). The regression parameters $(\alpha, \beta)^T$ and $(\gamma, \delta)^T$ were taken to be Normally distributed with mean $(3000, 185)^T$ and precision τQ_0 and λQ_0 respectively where $Q_0 = \text{diag}(r_0, s_0)$. The values of r_0 and s_0 were fixed to be 0.06 and 6. A gamma prior with shape $a_0 = 3$ and rate $b_0 = 2 \times 300^2$ was assumed for both τ and λ.

We consider estimating the evidence using $n = 10$, 20, 50 or 100 rungs in the tempering scheme where the rungs are chosen either according to the powered fraction (PF) rule or adaptively following the heuristic in Section 2.2. The parameters at all levels are updated using the Gibbs sampler, with 10000 iterations at each rung, discarding the first fifth of these as burn in. To quantify the performances, the bias, standard deviation and Root Mean Square Error (RMSE) are estimated by performing 100 replicates of the four schemes at 10, 20, 50 and 100 rungs. The results are given in Table 1. Comparing the first column with the second, and the third with the fourth, we can see that the modified power posterior approach dominates the usual one in all cases. For small numbers of rungs, where the RMSE is dominated by the bias term, this effect is particularly dramatic. The standard error is not greatly affected by the choice of scheme, although in this example the smallest values all occur for modified schemes. As the number of rungs increases, the RMSE becomes more affected by the standard error than the bias.

Comparing the first column of Table 1 with the third, we can isolate the effect of the adaptive placement. This is also most noticeable at small numbers of rungs, perhaps unsurprisingly given that the placing of an individual t_i makes less difference as their total number increases. To help to visualise the effect of both the correction and the adaptive placing of inverse temperatures, Figure 1 plots the 100 observed biases pairwise for the standard vs modified power posterior. Points are plotted for both the PF spacing and the adaptive spacings, and for each number of rungs. How much change occurs for each standard power posterior under the additive modification given by Equation (14) can be seen as the vertical distance between the point and the red $y = x$ line. A few points are immediately clear. Firstly the correction is fairly well estimated (the sets of points are roughly parallel to the $y = x$ line, even in the $n = 10$ case; this is a function of the MCMC run lengths). Second, there is an interesting difference between the adaptive and the PF versions, less correction is needed for the adaptive schedule, that is, it is already doing a better job of linearly approximating the function for the numerical integration. Thirdly, as was already observed in Table 1, as the number of rungs increases, the differences between the two types of spacings become less pronounced.

Given the good reductions in bias seen in Table 1, it is important to ask how much extra time is required. To assess this, a total of 20 runs for Model 1 using 10000 iterations and 100 inverse temperatures were timed. Four versions of the algorithm were considered, corresponding to Table 1. All the coding was in R and times are given relative to the PF standard power posterior version:

PF standard	PF modified	Adaptive standard	Adaptive modified
1.0000	1.0083	1.0076	1.0121

What we see is that the adaptive selection of inverse temperatures and the correction term in the numerical integration come at an acceptably small computational cost. Given this, our recommendation would be to use the modified integration rule when employing power posteriors. The benefits of the adaptive placement

	Rungs:	Power posterior:	Standard PF	Modified PF	Standard Adaptive	Modified Adaptive
MODEL 1	$n = 10$	Bias	-0.6569	0.0970	-0.4363	**0.0434**
		Standard error	0.0246	**0.0196**	0.0216	0.0199
		RMSE	0.6573	0.0990	0.4369	**0.0478**
	$n = 20$	Bias	-0.1628	**0.0044**	-0.1128	0.0057
		Standard error	0.0161	**0.0153**	0.0163	0.0154
		RMSE	0.1636	**0.0160**	0.1140	0.0164
	$n = 50$	Bias	-0.0258	**0.0000**	-0.0251	-0.0041
		Standard error	0.0098	**0.0097**	0.0104	0.0101
		RMSE	0.0276	**0.0097**	0.0272	0.0109
	$n = 100$	Bias	-0.0059	**0.0005**	-0.0101	-0.0041
		Standard error	0.0086	0.0085	0.0080	**0.0079**
		RMSE	0.0104	**0.0086**	0.0129	0.0089
MODEL 2	$n = 10$	Bias	-0.6354	0.1012	-0.4262	**0.0336**
		Standard error	0.0247	**0.0197**	0.0253	0.0228
		RMSE	0.6359	0.1031	0.4269	**0.0406**
	$n = 20$	Bias	-0.1585	0.0042	-0.1116	**0.0029**
		Standard error	0.0170	0.0160	0.0152	**0.0141**
		RMSE	0.1594	0.0165	0.1126	**0.0144**
	$n = 50$	Bias	-0.0249	**0.0002**	-0.0241	-0.0038
		Standard error	0.0106	0.0106	0.0103	**0.0101**
		RMSE	0.0270	**0.0106**	0.0262	0.0108
	$n = 100$	Bias	-0.0073	**-0.0011**	-0.0085	-0.0027
		Standard error	0.0084	0.0084	0.0062	**0.0061**
		RMSE	0.0112	0.0085	0.0105	**0.0066**

Table 1: Estimated bias, standard error and Root Mean Square Error for the two Radiata pine models using different numbers of rungs and different schemes (the standard power posterior vs the modified version, the Power Fraction scheme vs the adaptive scheme). For each model and number of rungs, the minimum absolute bias, standard error and RMSE are highlighted in bold. There are 100 replicates used for estimation, all use 10000 iterations at each rung discarding the first fifth as burn in.
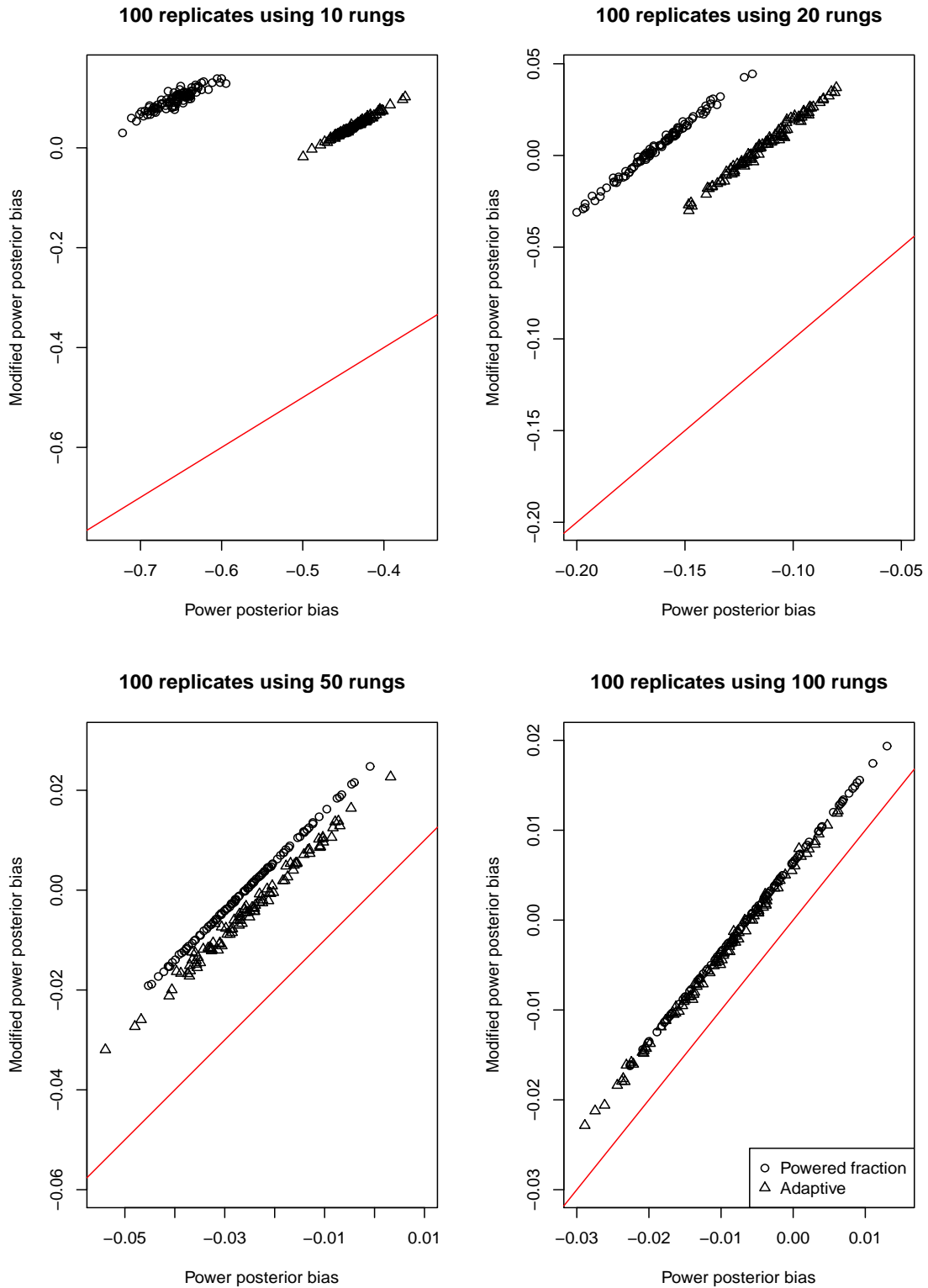
Figure 1: The 100 observed biases for Model 1 for the Radiata data using either PF or adaptive spacings plotted as standard vs modified pairs. The vertical distances from the red $y = x$ lines indicate the size of the correction in the modified power posterior.

are less dramatic except at small n but as it provides a completely automatic way to choose the inverse temperatures for very little cost, we would also recommend it.

3.2 Example 2: Pima indians

We turn next to the Pima Indian example considered by Friel and Wyse (2012), originally described by Smith *et al* (1988) . These data record diabetes incidence and possible disease indicators for $n = 532$ Pima Indian women aged over 20. The seven possible disease indicators are the number of pregnancies (NP), plasma glucose concentration (PGC), diastolic blood pressure (BP), triceps skin fold thickness (TST), body mass index (BMI), diabetes pedigree function (DP) and age (AGE), with all these covariates standardised.

The model assumed for the observed diabetes incidence, $y = (y_1, \ldots, y_n)$, is

$$p(y|\theta) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i} \tag{19}$$

where p_i is the probability of incidence for person i, and p_i is related to the ith person's covariates and a constant term, denoted by $x_i = (1, x_{i1}, \ldots, x_{id})^T$, and the parameters, $\theta = (\theta_0, \theta_1, \ldots, \theta_d)^T$, by

$$\log \left(\frac{p_i}{1 - p_i} \right) = \theta^T x_i \tag{20}$$

where d is the number of explanatory variables. An independent multivariate Gaussian prior is assumed for θ, with mean zero and non-informative precision of $\tau = 0.01$, so that

$$p(\theta) = (2\pi)^{-d/2} \tau^{d/2} \exp \left\{ -\frac{\tau}{2} \theta^T \theta \right\}. \tag{21}$$

There are 129 potential models (2^7 models with covariates plus a model with only a constant term). A long reversible jump run (Green, 1995) revealed the two models with the highest posterior probability:

Model 1: logit(p)= 1+NP+PGC+BMI+DP

Model 2: logit(p)= 1+NP+PGC+BMI+DP+AGE

For these two models, the power posterior is not amenable to the Gibbs sampler and so a Metropolis update is used instead. This raises the problem of proposal scaling at the different inverse temperatures. Since both the correction and the adaptive inverse temperature placements assume good estimates of the variance of the log deviance, mixing is an important issue; we return to the question of robustness to the quality of these estimates in Section 3.3. As an alternative to the sampler used in Friel and Wyse (2012), we work with a joint update of all the model parameters using a multivariate Normal proposal centred at the current value and with diagonal variance matrix, entries $\min(0.01/t, 1/\tau)$ where t is the inverse temperature and τ is the precision of the prior. The top panel of Figure 2 shows the estimated log deviance curves for these

10

two competing models using this sampler in the modified power posterior with 200 PF inverse temperatures each with 20000 iterations, discarding the first fifth as burn in. The bottom panel of the same figure shows two realisations of the adaptive placements, one for each model, against the PF scheme with $n = 50$.

Unlike in the Radiata example where the evidence can be evaluated analytically, here some estimation is required. Friel and Wyse (2012) use the Laplace approximation of the log evidence (Tierney and Kadane, 1986) as the "benchmark" in assessing bias. However this is not necessarily very accurate and so we replace the Laplace approximation by a very long run (2000 rungs, each using 20000 iterations) of the power posterior approach; the estimates we get are -257.2342 and -259.8519 for Models 1 and 2 respectively as opposed to the Laplace approximations of -257.2588 and -259.8906. We will use these values in a comparison of performance, with and without tuning, of power posteriors and another approach to estimating the evidence, the recent Stepping Stone sampler of Xie *et al* (2011).

Based on importance sampling, the Stepping Stones sampler uses the same idea of powered posteriors, Equation (6), as a series of intermediate distribution, the "stepping stones", between the prior and the posterior, but utilising them as importance distributions rather than performing numerical integration. It generates samples from each of the power posteriors from $t_0 = 0$ up to t_{n-1}, estimating the ratio of consecutive normalising constants $r_k = z(y|t_{k+1})/z(y|t_k)$, where $z(y|t)$ is given by Equation (7), using the t_k power posterior as an importance distribution:

$$\hat{r}_k = \frac{1}{m} \sum_{i=1}^{m} p(y|\theta^{(i)})^{t_{k+1}-t_k}, \ k = 0, \dots, n-1 \tag{22}$$

where m is the number of MCMC samples post burn in and the corresponding sampled values $\{\theta^{(i)}\}$ are drawn from $p_{t_k}(\theta|y)$. Assuming that the prior is normalised, the final estimate of the evidence, $z(y|t_n = 1)$, is the product of these n independent estimates, $\prod_{k=0}^{n-1} \hat{r}_k$. The Stepping Stone sampler is unbiased for the marginal likelihood, although slightly biased for the log marginal likelihood. Xie *et al* (2011) compare its performance in estimating the log marginal likelihood to that of power posteriors, finding that for a well chosen inverse temperature placement or for a large number of inverse temperatures, power posteriors and the Stepping Stone approach are comparable, with the latter slightly outperforming the former, but when there are few inverse temperatures or badly placed ones, power posterior estimates perform relatively poorly.

Table 2 shows the estimated biases, standard errors and RMSEs for both models and for four possible schemes: standard power posteriors with PF spacings, modified power posteriors with adaptive spacings, Stepping Stone with PF spacings, Stepping Stone with adaptive spacings. There are 10000 iterations at each rung discarding the first fifth as burn in. Notice that for the Stepping Stone sampler, the samples from the posterior (i.e. $t_n = 1$) are not required, other than indirectly to initialise the t_{n-1} sample; other than that, the two approaches use exactly the same MCMC samples. What we can see from Table 2 is

11

The estimated expected log deviance for both models

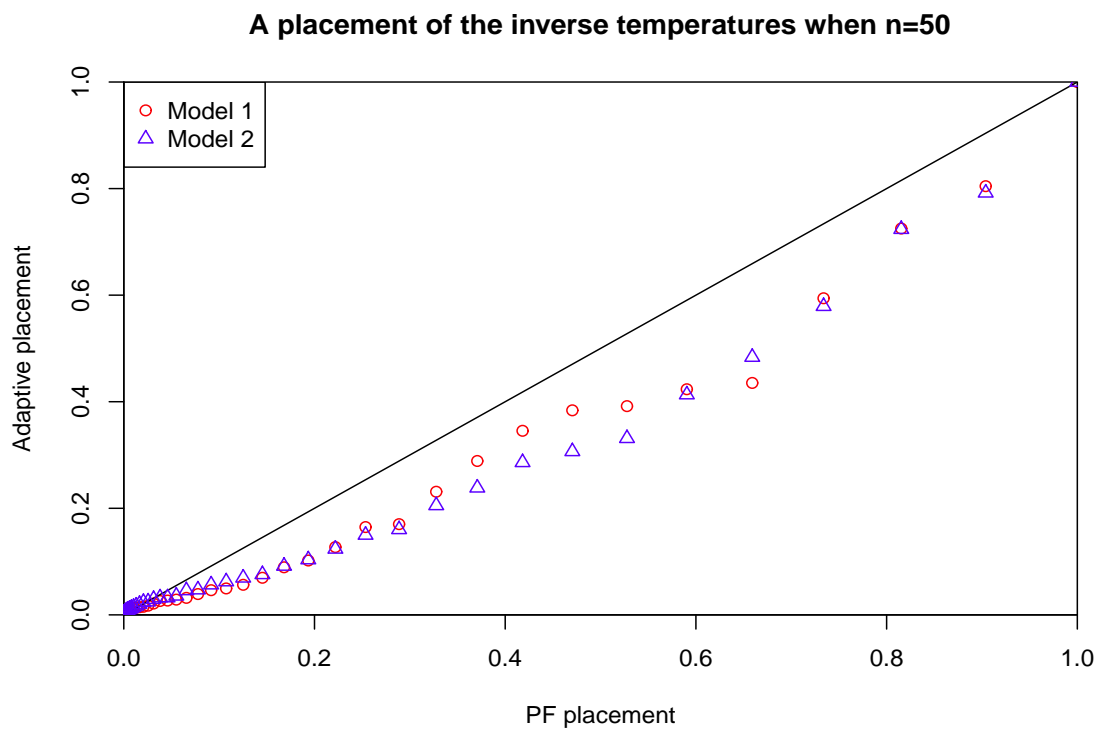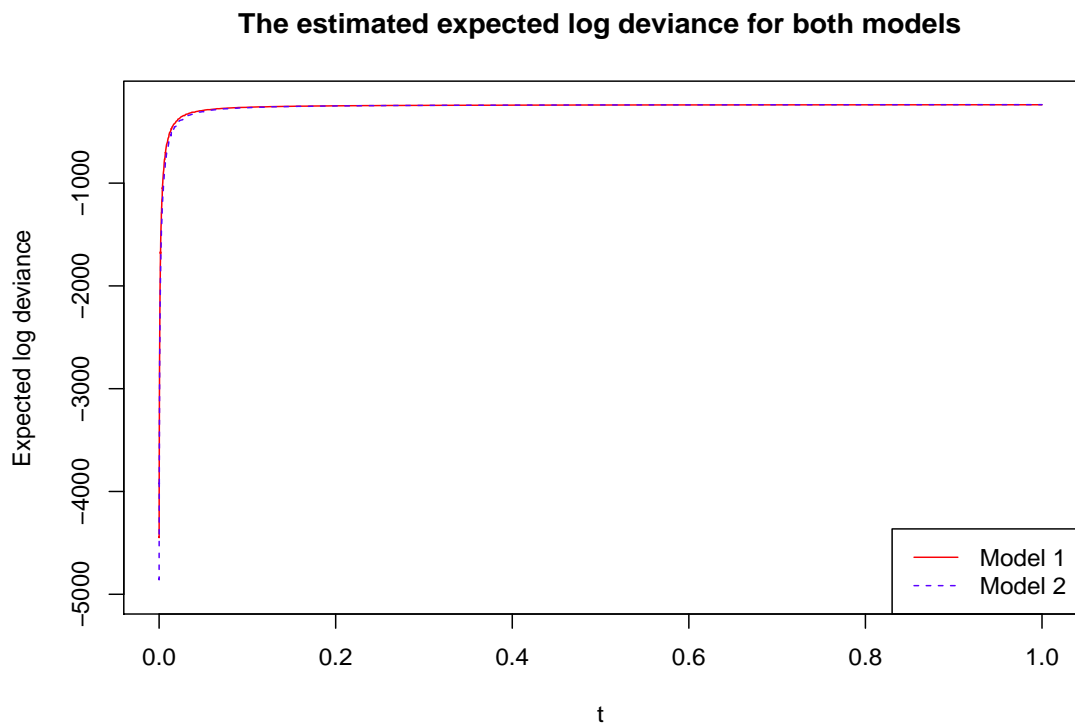**A placement of the inverse temperatures when n=50**

Figure 2: Top: The expected log deviance curves for the Pima models (estimated using 200 PF inverse temperatures with the modified power posterior each with 20000 iterations, discarding the first fifth as burn in). Bottom: A realisation of the adaptive temperature placements for the two models plotted against the PF scheme when $n = 50$.

	Rungs:		Standard PP PF	Modified PP Adaptive	Stepping Stone PF	Stepping Stone Adaptive
MODEL 1	$n = 10$	Bias	-3.67946	0.64809	**-0.02251**	-0.03174
		Standard error	0.35152	0.25121	0.25666	**0.18835**
		RMSE	3.69621	0.69507	0.25765	**0.19101**
	$n = 20$	Bias	-0.85666	0.03552	0.01973	**0.00865**
		Standard error	0.23478	0.16722	0.18491	**0.14805**
		RMSE	0.88825	0.17095	0.18596	**0.14830**
	$n = 50$	Bias	-0.13628	-0.00767	**-0.00150**	0.03435
		Standard error	0.13008	**0.09737**	0.11322	0.09836
		RMSE	0.18840	**0.09767**	0.11323	0.10418
	$n = 100$	Bias	**-0.00383**	0.01845	0.02736	0.09785
		Standard error	0.10495	0.08248	0.09777	**0.08172**
		RMSE	0.10502	**0.08452**	0.10153	0.12749
MODEL 2	$n = 10$	Bias	-4.16969	0.77087	**-0.00362**	0.01845
		Standard error	0.33864	0.32020	0.28518	**0.25373**
		RMSE	4.18342	0.91444	0.28521	**0.25440**
	$n = 20$	Bias	-0.94958	0.02703	0.03612	**-0.01686**
		Standard error	0.25089	0.20139	0.19515	**0.18510**
		RMSE	0.98216	0.20212	0.19847	**0.18587**
	$n = 50$	Bias	-0.11702	-0.03084	0.02995	**0.01999**
		Standard error	0.15566	0.12323	0.13724	**0.11467**
		RMSE	0.19474	0.12418	0.14047	**0.11640**
	$n = 100$	Bias	**-0.00384**	0.02610	0.03580	0.11592
		Standard error	0.11019	**0.09285**	0.10378	0.09334
		RMSE	0.11025	**0.09353**	0.10978	0.14883

Table 2: Estimated bias, standard error and Root Mean Square Error for the two Pima Indian models using different numbers of rungs and different schemes (the standard power posterior with PF, the modified power posterior with adaptive rungs, and the Stepping Stone sampler with both PF and adaptive schemes). There are 100 replicates used for the estimation, all use 10000 iterations at each rung discarding the first fifth as burn in.

	Rungs:		Standard PP PF	Modified PP Adaptive	Stepping Stone PF	Stepping Stone Adaptive
MODEL 1	$n = 10$	Bias	-3.64242	0.72598	**0.01293**	0.02300
		Standard error	0.10541	0.09356	0.07597	**0.06201**
		RMSE	3.64395	0.73198	0.07706	**0.06614**
	$n = 20$	Bias	-0.87596	0.04038	0.01651	**0.01554**
		Standard error	0.06776	0.05361	0.05348	**0.04788**
		RMSE	0.87858	0.06712	0.05597	**0.05034**
	$n = 50$	Bias	-0.11517	**0.01508**	0.02568	0.02021
		Standard error	0.04333	0.03735	0.03705	**0.03500**
		RMSE	0.12305	**0.04028**	0.04508	0.04041
	$n = 100$	Bias	-0.01262	**0.01133**	0.02210	0.02138
		Standard error	0.02961	0.02511	0.02752	**0.02454**
		RMSE	0.03219	**0.02755**	0.03529	0.03254
MODEL 2	$n = 10$	Bias	-4.17639	0.80148	0.03336	**0.02949**
		Standard error	0.13585	0.10192	0.09553	**0.08534**
		RMSE	4.17860	0.80794	0.10119	**0.09029**
	$n = 20$	Bias	-0.97693	0.05061	0.03903	**0.02079**
		Standard error	0.08038	0.05887	0.06134	**0.05227**
		RMSE	0.98024	0.07763	0.07270	**0.05626**
	$n = 50$	Bias	-0.13474	**0.01974**	0.02745	0.02509
		Standard error	0.04836	0.04230	0.04502	**0.03978**
		RMSE	0.14316	**0.04668**	0.05273	0.04703
	$n = 100$	Bias	**-0.00197**	0.02507	0.03797	0.03753
		Standard error	0.03727	0.02933	0.03546	**0.02809**
		RMSE	**0.03732**	0.03859	0.05195	0.04688

Table 3: Estimated bias, standard error and Root Mean Square Error for the two Pima Indian models using different numbers of rungs and different schemes (the standard power posterior with PF, the modified power posterior with adaptive rungs, and the Stepping Stone sampler with both PF and adaptive schemes). There are 100 replicates used for the estimation, all use 100000 iterations at each rung discarding the first fifth as burn in (that is, ten times the run lengths and burn ins of Table 2).

that the Stepping Stone estimates have generally small biases independent of the number of rungs or their placement. However something unexpected occurs for the Stepping Stones estimates for larger n with adaptive spacings, in that their bias begins to increase. This does not occur if the adaptively placed rungs from a power posterior run are used as deterministic rungs in a separate Stepping Stones run. To explore this further, we ran additional experiments, replacing 10000 iterations at each rung (discarding the first 2000 as burn-in) by 100000 iterations (discarding the first 20000 as burn-in). The results are presented in Table 3. What we see is that there is no longer a discrepancy between the bias of the Stepping Stones algorithm using the deterministic and the bias using the adaptive rung placements. The fact that it is only the adaptive Stepping Stones biases which change noticeably with the increase in run length, suggests to us that Stepping Stones is perhaps more sensitive than power posteriors to the fact that for adaptive placement of the rungs, the initialisation of the MCMC is not necessarily at the next largest of the final set of $\{t_i\}$ as it is with a deterministic placement. Whilst it might seem odd to present unconverged results, the point we would like to make is that there was no hint of lack of convergence visually inspecting the MCMC chains and it is only the adaptive Stepping Stone for which this quiet lack of convergence has had an effect here.

Turning to the seemingly converged results in Table 3, without any modification and with only PF spacing, Stepping Stones outperforms power posteriors, agreeing with the Xie *et al* findings. In fact when $n = 10$, even with modification, the bias associated with power posteriors dominates the RMSE (perhaps not surprisingly given the shape of the curve in the top panel of Figure 2). For larger n, the modified power posterior with adaptive spacing generally does better in terms of RMSE than the PF Stepping Stones. Introducing adaptive spacings to Stepping Stones does not greatly affect the bias, as expected, but it does reduce the standard error of the importance sampling based estimates. Figure 3 plots the corresponding Model 1 biases for the four methods and the different numbers of rungs, matching up estimates calculated from the same MCMC output (that is, power posteriors and Stepping Stones with PF spacing, modified power posteriors and Stepping Stones with adaptive spacing). The most arresting feature here is perhaps the high degree of correlation between the estimates for all values of n.

3.3   Example 3: Galaxy data

To demonstrate a large application with a more challenging integral than the previous two, we use the much-studied Galaxy data set, see for example Richardson and Green (1997), which comprises measurements on the velocities of 82 galaxies. Denoting the 82 measurements by $y = \{y_1, \ldots, y_{82}\}$, we follow Richardson and Green (1997) in incorporating corresponding latent allocation variables $z = \{z_1, \ldots, z_{82}\}$. Given
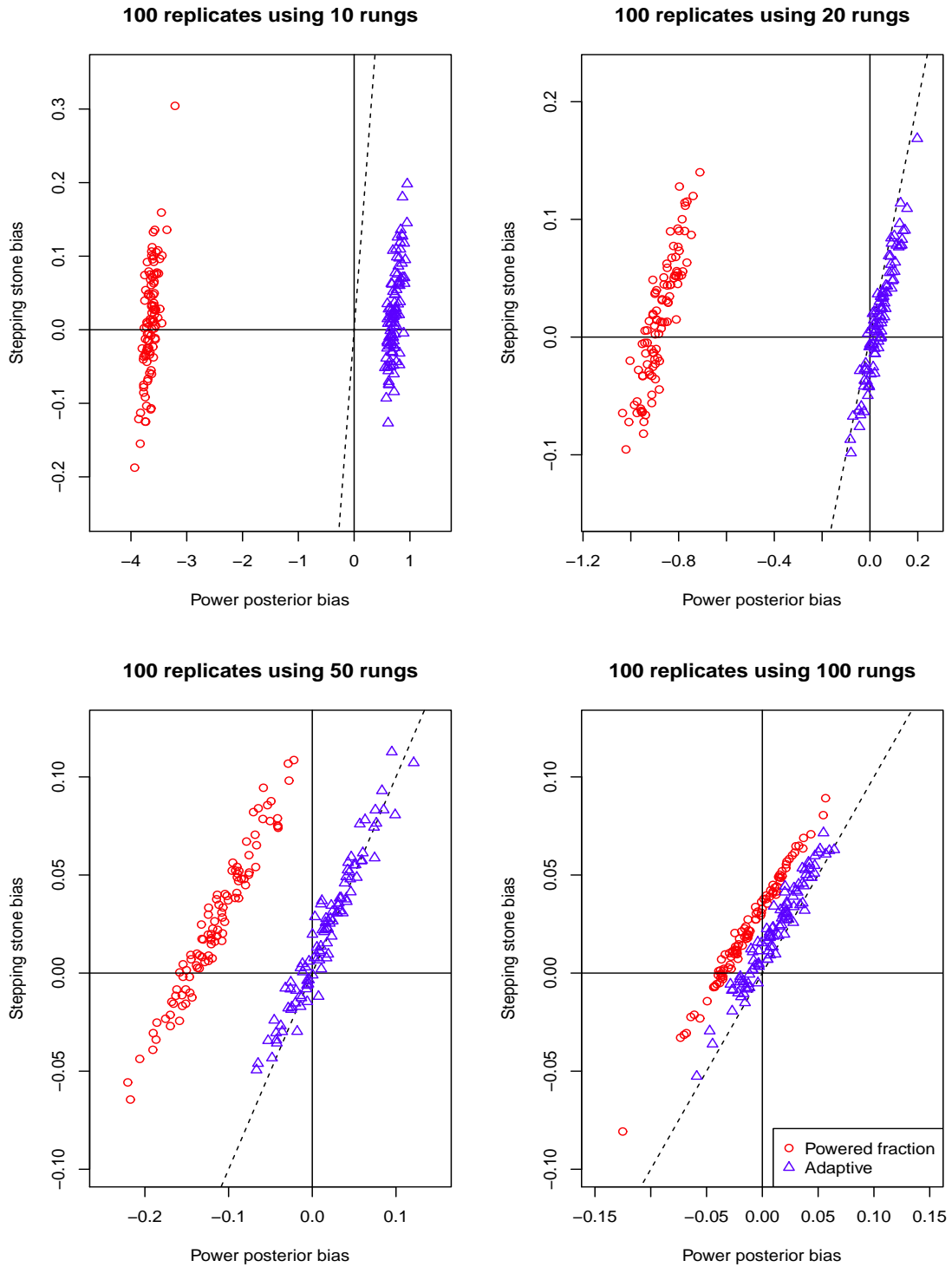
Figure 3: Biases for Model 1 for the Pima data using the four methods and different numbers of rungs, matching up estimates calculated from the same long MCMC output (that is, power posteriors and Stepping Stones with PF spacing, and modified power posteriors and Stepping Stones with adaptive spacing). The solid black lines indicate zero bias, the dashed black lines represent the line $y = x$.

$z_i = j$, y_i follows the jth of the k component Gaussian distributions of the mixture,

$$p(y_i|z_i = j, \mu_j, \sigma_j^2) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(\frac{-(y_i - \mu_j)^2}{2\sigma_j^2}\right) \quad i = 1, \ldots, 82. \quad (23)$$

Conditional independence is assumed for the $\{y_i\}$ and we specify independent standard proper priors:

$$\mathsf{P}(z_i = j) = w_j, \quad \text{where} \sum_{j=1}^{k} w_j = 1 \quad (24)$$

$$\{w_1, \ldots, w_k\} \sim Dirichlet(1, k) \quad (25)$$

$$\mu_j \sim N(0, 1000), \quad j = 1, \ldots, k \quad (26)$$

$$\sigma_j^2 \sim InvGam(1, 1), \quad j = 1, \ldots, k. \quad (27)$$

The weights, means and variances are all updated using the Gibbs sampler but we use a Metropolis algorithm with a discrete uniform proposal for the allocation variables.

Behrens, Friel and Hurn (2012) considered the Galaxy data set and the above model when studying inverse temperature placement for the tempered transition algorithm, finding that its expected log deviance curve had some interesting features. The top panel of Figure 4 shows the estimated log deviance curves for $k = 3$ and $k = 4$. These have been plotted on a restricted vertical scale so that interesting differences can be seen for larger inverse temperature values (the functions are in the region of -430000 when $t = 0$). These curves have been estimated using 200 adaptively placed inverse temperatures, with 50000 MCMC iterations at each rung, discarding the first tenth as burn in. The sheer scale of the steepness towards zero is seen in the bottom panel of Figure 4 which plots the corresponding estimated gradients on a log scale. Based on these simulations, Figure 5 gives the usual and modified power posterior estimates along with the Stepping Stone estimate. It also plots an estimated upper and lower bound on the log evidence derived as the lower and upper step functions in the trapezium rule, Equation (10). The adaptive scheme has the benefit that if these estimated upper and lower discretisation bounds are still considered too wide after using the anticipated maximum number of inverse temperatures n, the process can simply be run on with additional inverse temperatures placed by the same algorithm (the same cannot be said of the PF scheme where it is not immediately clear how to add additional inverse temperatures). In this case, for $k = 3$, the modified power posterior estimate is -228.8428, the Stepping Stone estimate is -228.7486 with an estimated discretisation interval of $[-229.8626, -227.8737]$. The corresponding figures for $k = 4$ are -229.3113, -229.2291 and $[-230.5911, -228.0902]$.

The main aim of this section is to explore the effect uncertainty in estimating the gradients has on modified power posteriors using adaptive inverse temperature placements. To do this, we consider progressively shorter runs of MCMC at each rung using the 50000 iteration run as a benchmark and retaining the burn in
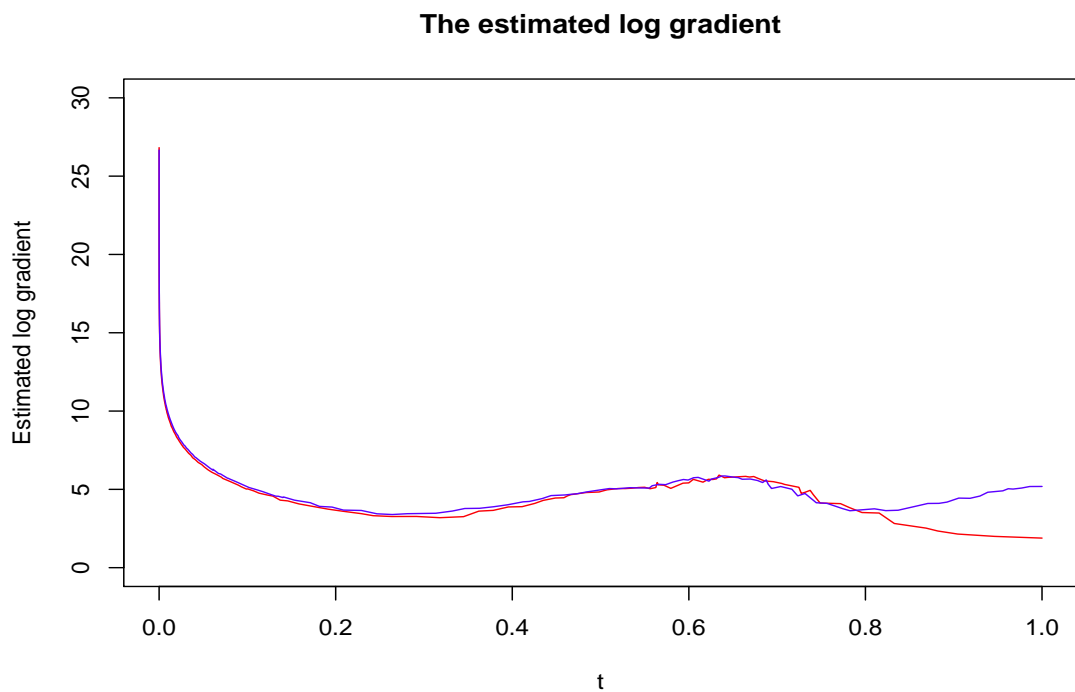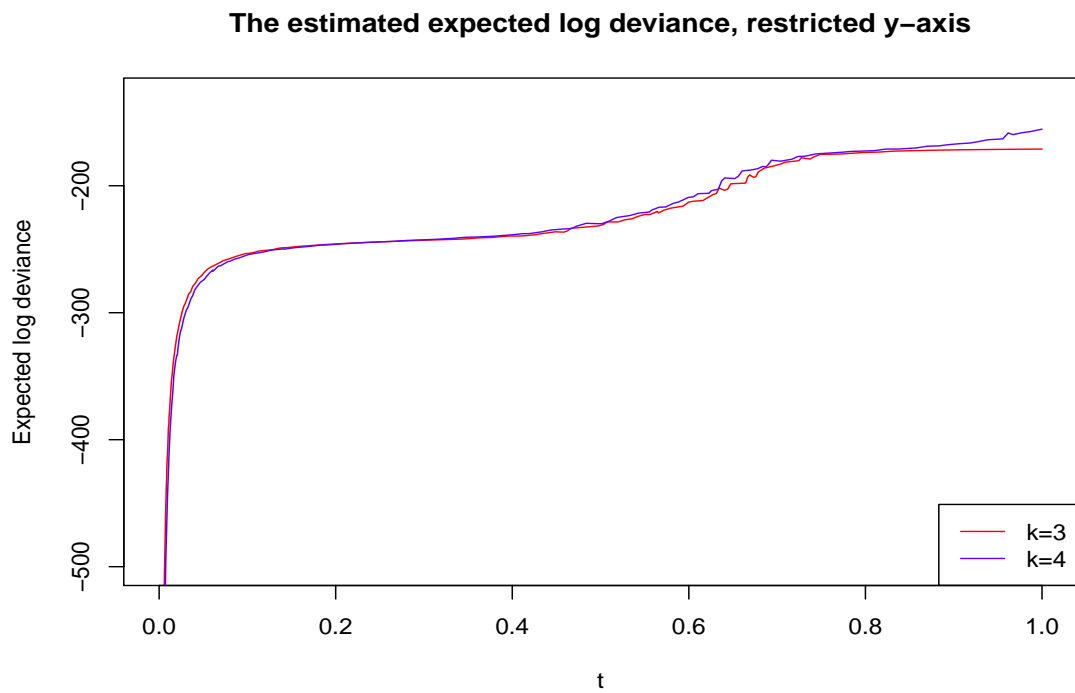
17

Figure 4: Top: The expected log deviance curves for the Galaxy models using 200 rungs in the modified power posterior each with 50000 iterations, discarding the first tenth as burn in (plotted on a restricted vertical scale). Bottom: The log of the corresponding estimated gradients.
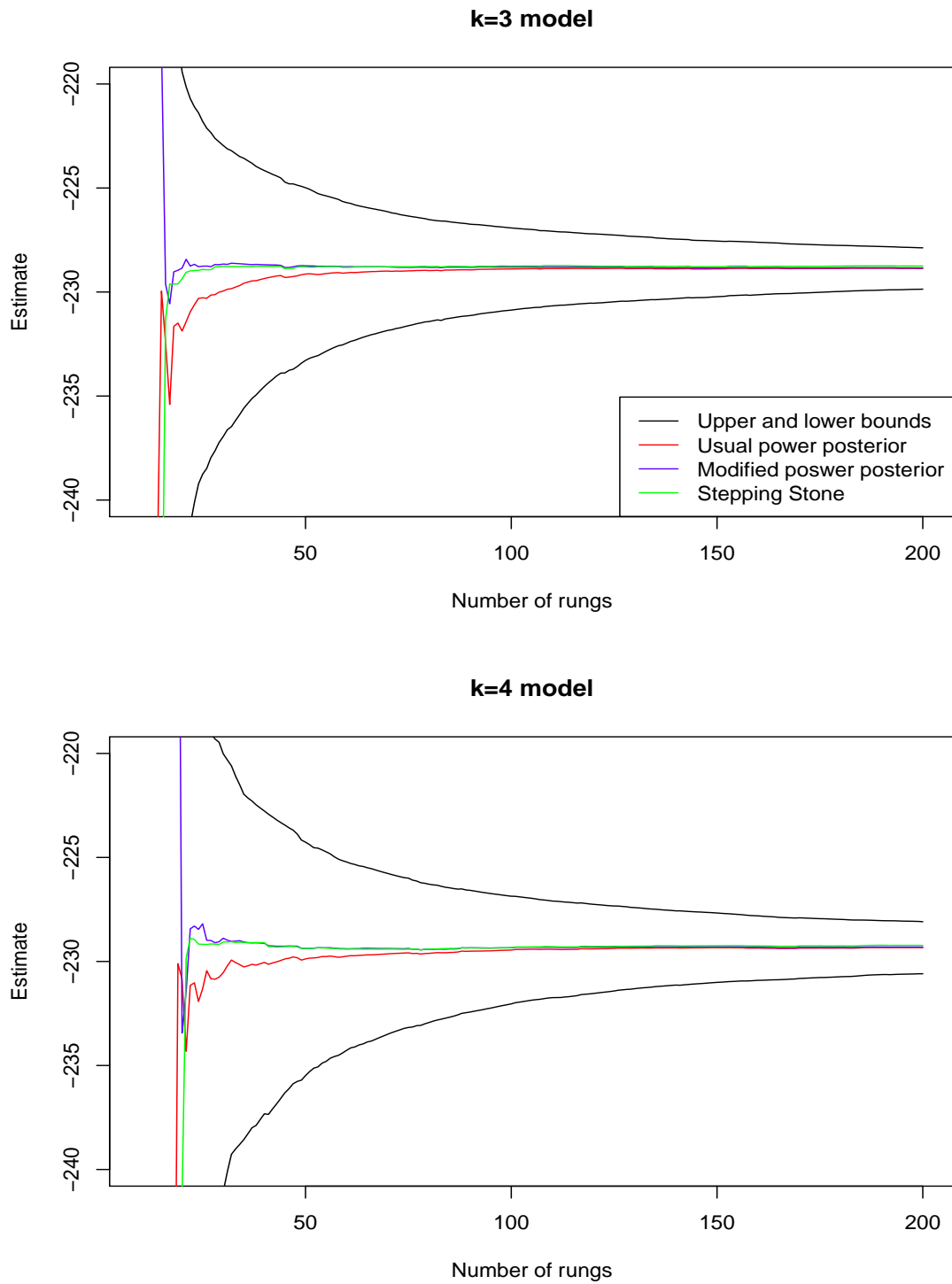
k=3 model

k=4 model

Figure 5: Estimates of the log evidence for the Galaxy data based on the estimated curves and gradients in Figure 4 for both $k = 3$ and $k = 4$ models, plotted for $n \geq 10$. Th e y-axes have been restricted to improve clarity.

rule of the first tenth of the run. Figure 6 plots the log of the estimated gradient, Equation (11), the usual and modified power posterior estimates and the Stepping Stone estimates for 5000, 2500 and 1000 MCMC iterations for the $k = 3$ model. Figure 7 plots the same quantities but now for $k = 4$ and the rather short MCMC runs of 750, 500 and 250 iterations. Since most of the inverse temperatures are close to zero, we plot the log variance against the log of t_i (which does mean we do not see the effect at $t = 0$). Considering first Figure 6, reducing the number of MCMC iterations at each rung to 5000 or 2500 does not greatly affect the results with all three estimates lying within the upper and lower bounds, albeit not as tightly converging to the 50000 iteration final estimate. As the run length decreases to 1000 or less (continuing on to Figure 7), we begin to see some discrepancies in the log variance curve and in the estimates themselves. Since we know that the effects of the adaptive placements and of the modification reduce as the number of rungs increases, we focus on $n \leq 100$ (bearing in mind that reducing the run length also affects the estimated log deviance values themselves). The gaps between the three estimates grow (we are particularly interested in the gap between standard and modified power posteriors). The Stepping Stone estimate does occasionally stray outside the lower and upper bounds, and there is again some evidence that it is veering towards higher values than the power posteriors. The observation that when there are 50000 iterations the Stepping Stones and power posterior estimates are closer together than in these figures adds some more weight to our assertion that Stepping Stones is quite sensitive to poor convergence. However, although there is some indication that the estimates have not stabilised by the $n = 200$ point, there are no catastrophic failures, just greater uncertainty. Given that power posteriors rely on good estimates of of log deviance as well as its variance, it may be better to use a moderate size of n with long MCMC runs at each t_i, rather than dividing up the same total number of iterations into short runs with a large n.

4 Conclusions

This article has, we hope, illustrated the potential gains that can be made when estimating the evidence using power posteriors by correcting the numerical integration error and by adaptively choosing the inverse temperature ladder. The methods that we have outlined come at virtually no extra computational cost, and we would therefore recommend that these are routinely used when implementing the power posterior approach.

What this article does not do is to give guidance as to how to allocate computational resources between the different inverse temperatures. We have seen in our examples that the gradient and thus also the variance of $\mathbf{E}_{\theta|y,t} \log(p(y|\theta))$ is largest as $t \to 0$, suggesting that we should allocate more MCMC iterations here rather than as $t \to 1$ to get good estimates. On the other hand, when t is small, the power posteriors $p_t(\theta|y)$ will probably be easy to sample compared to when $t = 1$ and so we should also take into account the MCMC
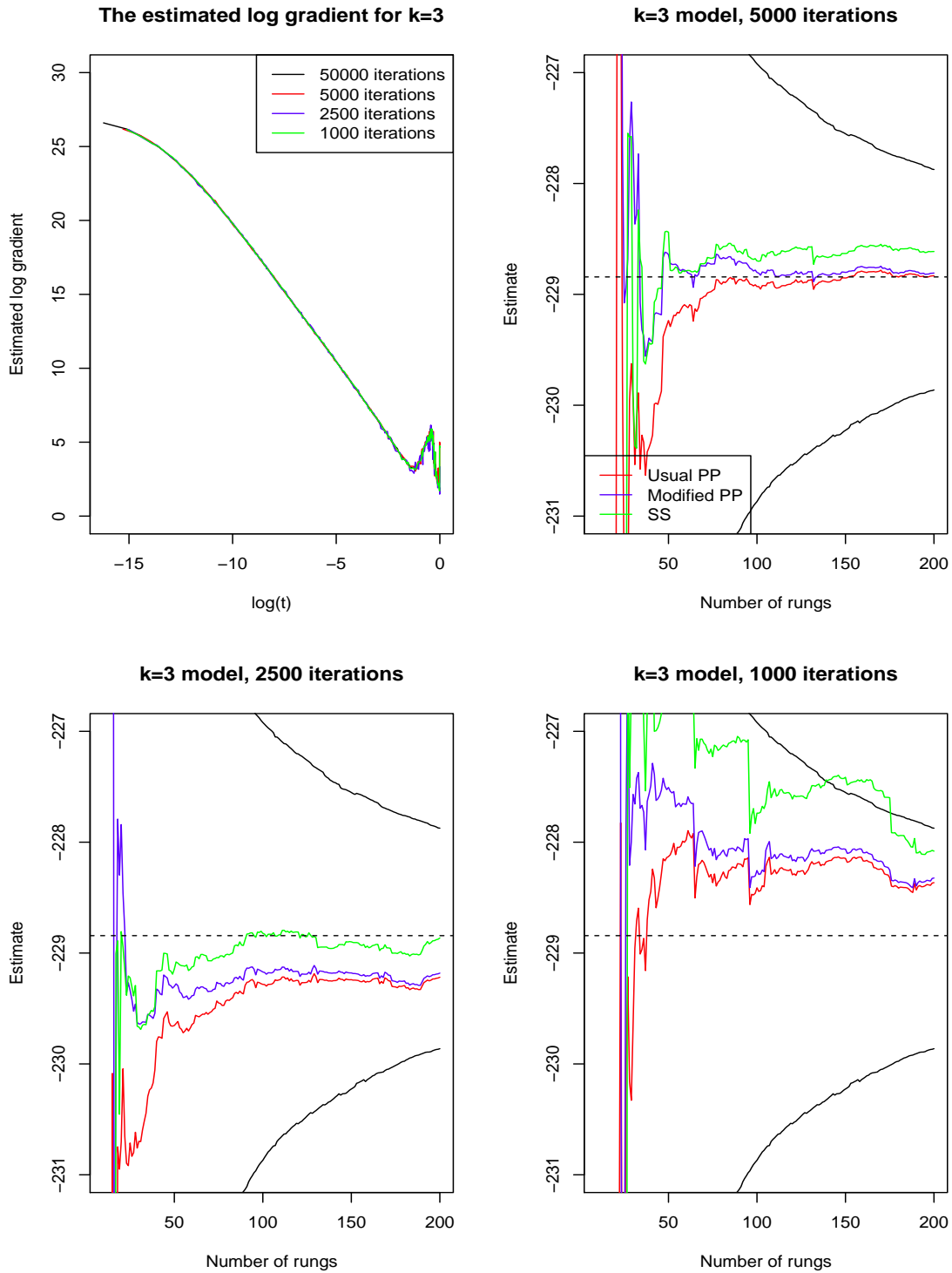
Figure 6: Top left: The estimated log gradients against the log of the adaptive inverse temperatures for $k = 3$. Remaining three panels: the usual and modified power posterior estimates and Stepping Stone estimates plotted against the number of rungs. In all three cases, the black solid lines and black dashed line represent the estimated upper and lower bounds and final modified power posterior estimate from the 50000 iteration run.
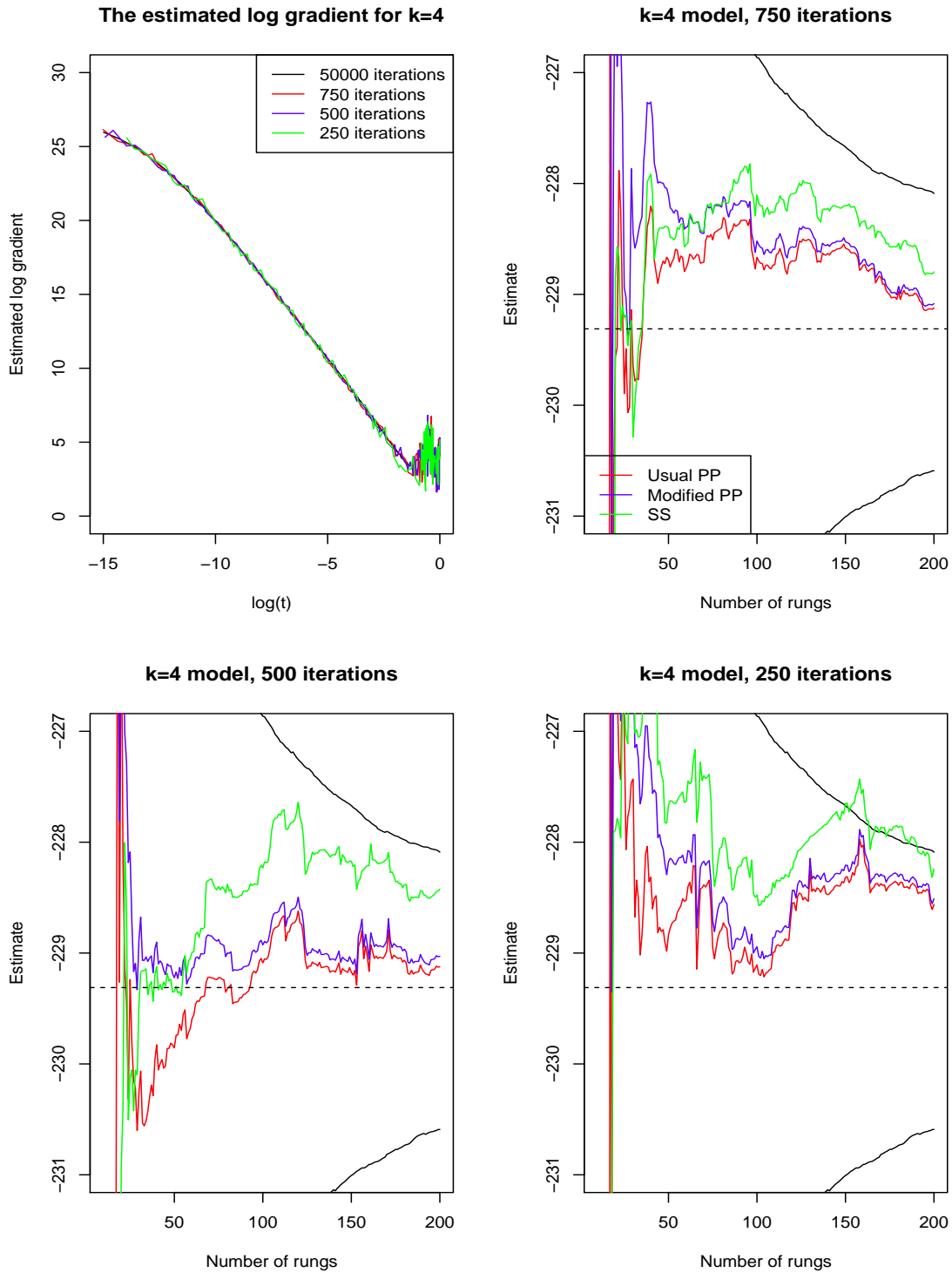
Figure 7: Top left: The estimated log gradients against the log of the adaptive inverse temperatures for $k = 3$. Remaining three panels: the usual and modified power posterior estimates and Stepping Stone estimates plotted against the number of rungs. In all three cases, the black solid lines and black dashed line represent the estimated upper and lower bounds and final modified power posterior estimate from the 50000 iteration run.

effective samples sizes. Neither the gradients nor the effective sample sizes can be known before sampling is carried out! There probably is some scope for an adaptation scheme here too, perhaps allocating some fraction of the total number of MCMC iterations evenly over the inverse temperatures before allocating the remainder based on what we have learned in this initial phase. This point also reinforces our caveat: what we have addressed here is discretisation error, the bounds we give are (noisy) bounds on this error and not credible intervals in the usual sense.

In our examples, applying the correction term has effectively smoothed over the benefits of the adaptive scheme over the PF one (just working a little harder in the latter case). This numerical analysis trick is peculiar to this particular use of tempered distributions. In general we suspect though that the adaptation ideas developed here might find wider use in other tempered schemes described in the literature.

References

[1] K. Atkinson and W. Han. *Elementary numerical analysis*. John Wiley and sons, third edition, 2004.

[2] G. Behrens, N. Friel, and M. Hurn. Tuning tempered transitions. *Statistics and Computing*, 22(1):65–78, 2012.

[3] B. Calderhead and M. Girolami. Estimating Bayes factors via thermodynamic integration and population mcmc. *Computational Statistics & Data Analysis*, 53(12):4028–4045, 2009.

[4] S. Chib. Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995.

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

[6] N. Friel and J. Wyse. Estimating the evidence - a review. *Statistica Neerlandica*, 66(3):288–308, 2012.

[7] P.J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

[8] N. Lartillot and H. Philippe. Computing Bayes factors using thermodynamic integration. *Systematic Biology*, 55(2):195–207, 2006.

[9] G. Lefebvre, R.J. Steele, and A.C. Vandal. A path sampling identity for computing the Kullback-Leibler and J-divergences. *Computational Statistics and Data Analysis*, 54(7):1719–1731, 2010.

[10] X.L. Meng and W.H. Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, 6(4):831–860, 1996.

[11] R.M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.

[12] S. Richardson and P.J. Green. On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society B*, 59(4):731–792, 1997.

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

[14] J.W. Smith, J.E. Everhart, W.C. Dickson, W.C. Knowler, and R.S. Johannes. Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 261. American Medical Informatics Association, 1988.

[15] L. Tierney and J.B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986.

[16] E. Williams. *Regression Analysis*. Wiley, Chichester, 1959.

[17] W. Xie, P.O. Lewis, Y. Fan, L. Kuo, and M.H. Chen. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology*, 60(2):150–160, 2011.