Power Posteriors $+$

Merrilee Hurn

Department of Mathematical Sciences, University of Bath, UK

April 2016

Bayesian Model Choice

- Possible models $m_1, \ldots, m_I$ for data y.
- Posterior distribution given data y model m_i is

$$p(\theta_i|y, m_i) = \frac{p(y|\theta_i, m_i)p(\theta_i|m_i)}{p(y|m_i)}$$

  where θ_i are the parameters for model m_i.
- The evidence/marginal likelihood for data y given model m_i is the normalising constant of the posterior distribution within model m_i,

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

- The marginal likelihood is often then used to calculate Bayes factors to compare two competing models, m_i and m_j,

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

Estimating the Marginal Likelihood

Estimation of the evidence is non-trivial for most statistical models:

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

and there are a number of approaches proposed (Laplace's method, Chib's method, bridge sampling, annealed importance sampling, nested sampling, stepping stone sampling...)

Friel and Wyse (2012) for a recent review

This talk is going to concentrate on the form of thermodynamic integration known as Power Posteriors

Lartillot and Phillippe (2006), Friel and Pettitt (2008)

Power posteriors

Define the power posterior at inverse temperature t by

$$\begin{aligned} p_t(\theta|y) &\propto p(y|\theta)^t p(\theta), \quad t \in [0,1] \\ \text{with } z(y|t) &= \int_\theta p(y|\theta)^t p(\theta) d\theta. \end{aligned}$$

Key identity:

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

As a result

$$\begin{aligned} \int_0^1 \mathbf{E}_{\theta|y,t} \log(p(y|\theta))\ dt &= [\ \log(z(y|t))\ ]_0^1 \\ &= \log(z(y|1)) - \log(z(y|0)) \end{aligned}$$

$[t=0]$ $p_0(\theta|y)$ is the prior and $z(y|0)=1$ by assumption
$[t=1]$ $p_1(\theta|y)$ is the posterior and $z(y|1)$ is the evidence

$$\text{and so } \int_0^1 \mathbf{E}_{\theta|y,t} \log(p(y|\theta))dt = [\ \log(z(y|t))\ ]_0^1 = \log(z(y|1)))$$

The problem of calculating a $|\theta|$-dimensional integral over the parameter space to find the marginal likelihood has seemingly become that of evaluating a 1-dimensional integral over the unit interval to find the log marginal likelihood.

The catch? The integrand rather than the integrating...

Implementation and errors

- Discretise the inverse temperatures

$$0 = t_0 < t_1 < \ldots < t_n = 1.$$

Lartillot and Phillippe (2006) use uniformly spaced t_i

Friel and Pettitt (2008) recommend $t_i = (i/n)^5$

- For each t_i, estimate $\mathbf{E}_{\theta|y,t_i} \log(p(y|\theta))$ by sampling $p(\theta|y, t_i)$.

Expensive. Sampling error

- Approximate the integral using numerical integration.

Lartillot and Phillippe (2006) use Simpson's rule

Friel and Pettitt (2008) use the trapezium rule

Cheap. Discretisation error

How do we get the highest accuracy for the least cost?

Reducing discretisation error for the trapezium rule

Exploit the extra information that the gradient of the curve $\mathbf{E}_{\theta|y,t}\log(p(y|\theta))$ is the variance $\mathbf{Var}_{\theta|y,t}\log(p(y|\theta))$ since

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

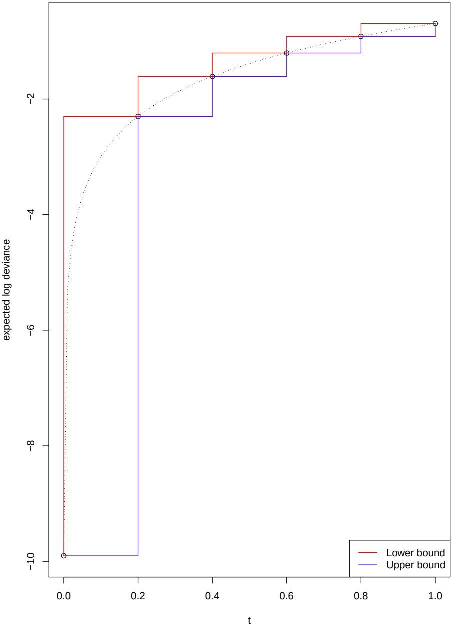
$$\begin{aligned}
\text{So } \frac{d}{dt}\mathbf{E}_{\theta|y,t}\log(p(y|\theta)) &= \frac{d}{dt}\left(\frac{1}{z(y|t)}\frac{dz(y|t)}{dt}\right) \\
&= -\left(\frac{1}{z(y|t)}\frac{dz(y|t)}{dt}\right)^2 + \frac{1}{z(y|t)}\frac{d^2 z(y|t)}{dt^2}
\end{aligned}$$

Lartillot and Phillippe (2006), Friel et al (2014)

IF we knew the curve exactly, we would have upper and lower bounds on the integral based on the function evaluations at the $\{t_i\}$. (The trapezium rule is the average of these two.)

More formally, minimising the total area of this sum of rectangles minimises the sum of symmetrised Kullback-Leibler distances between the successive pairs of $p(\theta|y, t_i)$.

Calderhead and Girolami (2009), Lefebvre et al (2010), Behrens et al (2012)



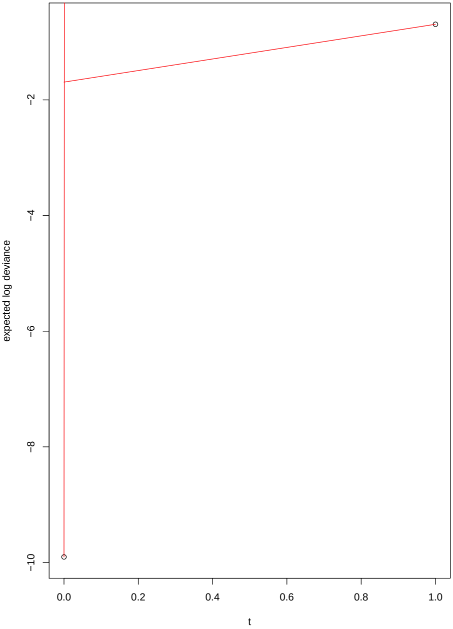
Can we use the (estimated) gradient of the curve adaptively to place the t_i to minimise the area between the two?

Friel et al (2014)

Can we use the (estimated) gradient of the curve adaptively to place the t_i to minimise the area between the two?

Friel et al (2014)

Start with $t = 0$ and $t = 1$. We have estimates of the function and its derivative at these two points. Site the next t where the two tangents meet.

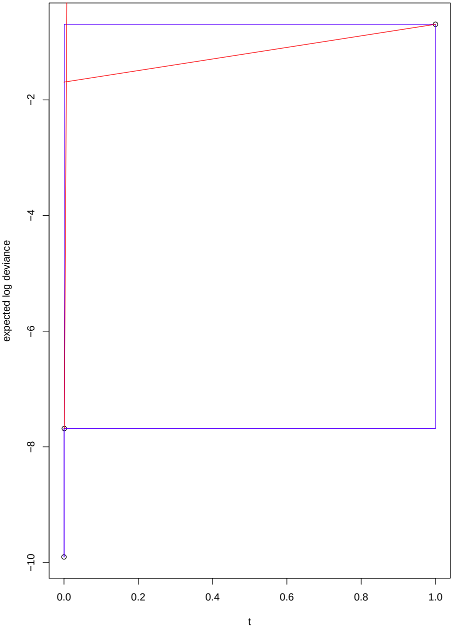


Can we use the (estimated) gradient of the curve adaptively to place the t_i to minimise the area between the two?

Friel et al (2014)

Start with $t = 0$ and $t = 1$. We have estimates of the function and its derivative at these two points. Site the next t where the two tangents meet.

At each step, subdivide the largest rectangle using the intersection of tangents.

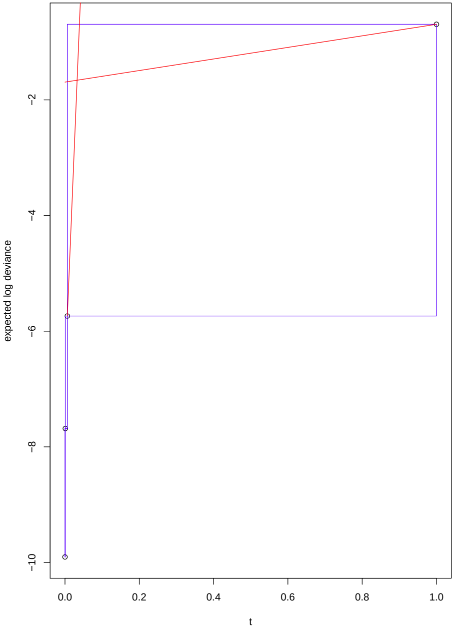


Can we use the (estimated) gradient of the curve adaptively to place the t_i to minimise the area between the two?

Friel et al (2014)

Start with $t = 0$ and $t = 1$. We have estimates of the function and its derivative at these two points. Site the next t where the two tangents meet.

At each step, subdivide the largest rectangle using the intersection of tangents.

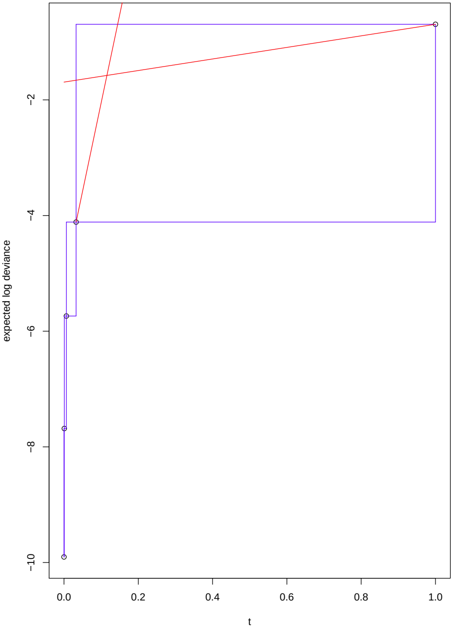


Can we use the (estimated) gradient of the curve adaptively to place the t_i to minimise the area between the two?

Friel et al (2014)

Start with $t = 0$ and $t = 1$. We have estimates of the function and its derivative at these two points. Site the next t where the two tangents meet.

At each step, subdivide the largest rectangle using the intersection of tangents.

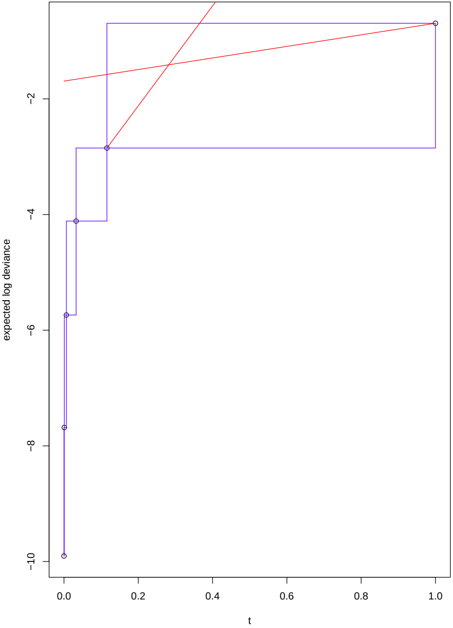


Can we use the (estimated) gradient of the curve adaptively to place the t_i to minimise the area between the two?

Friel et al (2014)

Start with $t = 0$ and $t = 1$. We have estimates of the function and its derivative at these two points. Site the next t where the two tangents meet.

At each step, subdivide the largest rectangle using the intersection of tangents.

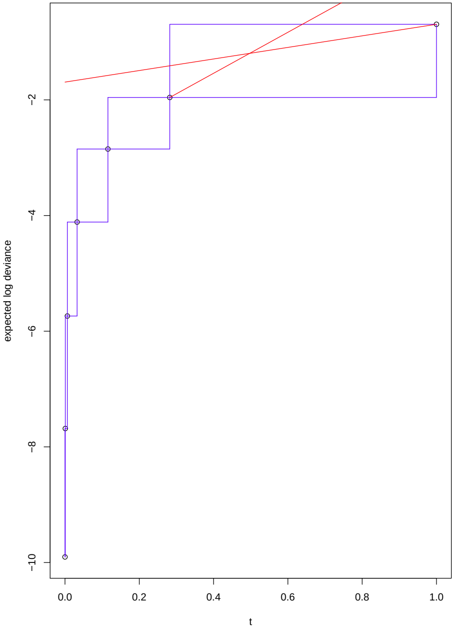


Can we use the (estimated) gradient of the curve adaptively to place the t_i to minimise the area between the two?

Friel et al (2014)

Start with $t = 0$ and $t = 1$. We have estimates of the function and its derivative at these two points. Site the next t where the two tangents meet.

At each step, subdivide the largest rectangle using the intersection of tangents.

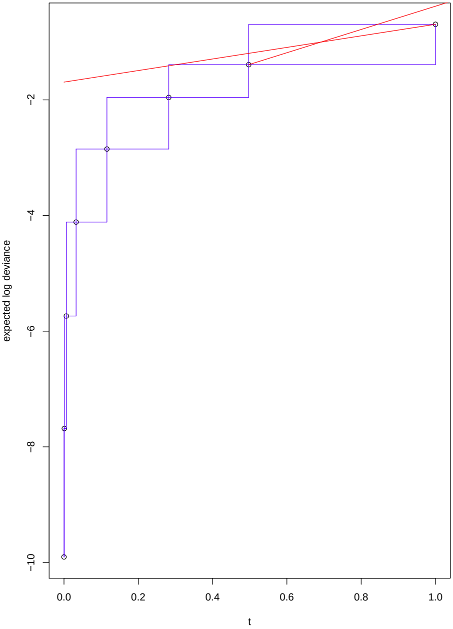


Can we use the (estimated) gradient of the curve adaptively to place the t_i to minimise the area between the two?

Friel et al (2014)

Start with $t = 0$ and $t = 1$. We have estimates of the function and its derivative at these two points. Site the next t where the two tangents meet.

At each step, subdivide the largest rectangle using the intersection of tangents.

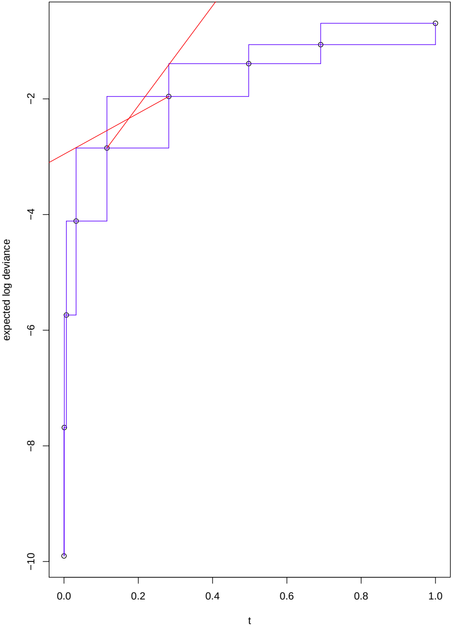


Can we use the (estimated) gradient of the curve adaptively to place the t_i to minimise the area between the two?

Friel et al (2014)

Start with $t = 0$ and $t = 1$. We have estimates of the function and its derivative at these two points. Site the next t where the two tangents meet.

At each step, subdivide the largest rectangle using the intersection of tangents.



We can also use the gradient to improve the numerical integration

Friel et al (2014)

When integrating a function f between points a and b

$$\int_a^b f(t)dt = (b-a)\left[\frac{f(b)+f(a)}{2}\right] - \frac{(b-a)^3}{12} f''(c)$$

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

$$\begin{aligned} f''(c) &\approx \frac{f'(b)-f'(a)}{b-a} \\ \text{so } \int_a^b f(t)dt &\approx (b-a)\left[\frac{f(b)+f(a)}{2}\right] - \frac{(b-a)^2}{12}\left[f'(b)-f'(a)\right] \end{aligned}$$

Atkinson and Han (2004)

and, in this case, we have (estimated) gradient information cheaply available via the variance at each t. Quite effective.

Reducing discretisation error for Simpson's rule

Simpson's rule is a better (higher order) numerical integration rule than the trapezium but usually relies on equally spaced function evaluations

$$\int_a^b f(t)dt = \frac{(b-a)}{6}\left[f(b) + 4f\left(\frac{a+b}{2}\right) + f(a)\right] - \frac{(b-a)^5}{180} f^{(4)}(c)$$

where the error term is evaluated at some c in $[a, b]$.

Atkinson and Han (2004)

Two papers propose adaptive spacings, both via iteratively bisecting t intervals for which the integral approximation is not sufficiently unchanged when the discretisation is increased.

Lefebvre et al (2009), Hug et al (2016)

In more detail.....

- Start with an initial discretisation $0 = t_0 < t_1 < \ldots < t_m = 1$ (can be coarse, need not be uniform spacing).
- For each of these intervals, say $[a, b]$
    - Estimate the function at a, at b and at the quarter-, mid- and three-quarter- points of $[a, b]$.
    - Approximate the integrals on $[a, b]$, $[a, (a + b)/2]$ and $[(a + b)/2, b]$ using Simpson's rule.
    - Stop if the first approximation is sufficiently close to the sum of the second two.
    - Otherwise, bisect $[a, b]$ and repeat the process on $[a, (a + b)/2]$ and $[(a + b)/2, b]$ separately.

It is also possible (and, in their experiments, important) to work on a transformed scale so that the adaptive algorithm works with a power law type spacing, using integration by substitution with $\lambda = t^{1/q}$ ($q = 5$ seems to work as well as a tuned q).

Hug et al (2016)

Could we try the same trick as with the trapezium rule and use derivative information to improve the numerical integration?

$$\int_a^b f(t)dt = \frac{(b-a)}{6}\left[f(b) + 4f\left(\frac{a+b}{2}\right) + f(a)\right] - \frac{(b-a)^5}{180} f^{(4)}(c)$$

where c is some point in $[a, b]$, so plug in

$$f^4(c) \approx \frac{f^3(b) - f^3(a)}{b-a} \text{ and in this case}$$

$$\begin{aligned}\frac{d^3}{dt^3}\mathbf{E}_{\theta|y,t}\log(p(y|\theta)) &= \mathbf{E}_{\theta|y,t}(\log(p(y|\theta))^4) - \\ &\quad 4\mathbf{E}_{\theta|y,t}(\log(p(y|\theta))^3)\mathbf{E}_{\theta|y,t}(\log(p(y|\theta))) - \\ &\quad 3(\mathbf{E}_{\theta|y,t}(\log(p(y|\theta))^2))^2 + \\ &\quad 12\mathbf{E}_{\theta|y,t}(\log(p(y|\theta))^2)(\mathbf{E}_{\theta|y,t}(\log(p(y|\theta))))^2 - \\ &\quad 6(\mathbf{E}_{\theta|y,t}(\log(p(y|\theta))))^4\end{aligned}$$

Not pretty, but we could estimate these four moments of $\log(p(y|\theta))$ although it does depend on the quality of the estimation...

Reducing sampling error

Over and above choosing an appropriate and efficient sampler for each t_i, what other options are there?

Control variates? Want to estimate some $\mathbf{E}(X)$ and there are $\{W_1, \ldots, W_p\}$ which are known to vary with X: Use

$$\psi(X) = X + \beta_1(W_1 - \mathbf{E}(W_1)) + \ldots + \beta_p(W_p - \mathbf{E}(W_p))$$

with the $\{\beta_i\}$ selected so that $\mathbf{Var}(\psi(X)) < \mathbf{Var}(X)$ (e.g. in the $p = 1$ case, take $\beta_1 = -\mathbf{Cov}(X, W_1)/\mathbf{Var}(W_1)$).

Extension to this context? The Controlled thermodynamic integral

$$\int_0^1 \mathbf{E}_{\theta|y,t}\left(\log(p(y|\theta)) + h(\theta|y,t)\right) \, dt$$

Oates et al (2016a)

Components for constructing $h(\theta|y,t)$ come from the ZV world

- a (low order) polynomial P in θ with coefficients $\phi(y,t)$ plus derivatives for given $\phi(y,t)$: $\nabla_\theta(P|\phi(y,t))$, $\triangle_\theta(P|\phi(y,t))$
- the derivatives of the log power posterior with respect to θ, $\nabla_\theta(log(p_t(\theta|y)))^{**}$ (using derivatives of prior and likelihood)

$$h(\theta|y,t) \;\; = \;\; -\frac{1}{2}\triangle_\theta(P|\phi(y,t)) - \frac{1}{2}\nabla_\theta(P|\phi(y,t)).\nabla_\theta(\log(p_t(\theta|y)))$$

It should in theory be possible to choose variance-minimising sets of coefficients $\{\phi(y,t)$ AND $\{t_i\}$ jointly, but not adaptively. The former though can be done post-simulation-pre-estimation using plug-in estimates of various covariance terms.

Very effective in some of the examples (plus there's an extension for the others)

Oates et al (2016b)

** already needed IF using a differential-geometric MCMC sampling scheme. But if not....

Discussion

- Estimating the log marginal likelihood via Power Posteriors in its basic form is relatively straight-forward but computationally costly.
- To minimise the cost, we want to use as few t values as possible.
- How to choose those t, preferably adaptively?
- How to improve on the numerical integration?
- How to reduce the sampling variability?
- Now let's up the computational cost and do this for lots of models, not just one. Are there any savings to be made for related models, e.g. importance sampling or ...?

References I

Atkinson, K. and Han, W., (2004) *Elementary numerical analysis*, Third edition, John Wiley and sons.

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

Calderhead, B. and Girolami, M., (2009) Estimating Bayes factors via thermodynamic integration and population MCMC, *Computational Statistics & Data Analysis*, **53**, 4028–4045.

Friel, N., Hurn, M. and Wyse, J., (2014). Improving power posterior estimation of statistical evidence. *Statistics and Computing*, **24**, 709-723.

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

Friel, N. and Wyse, J., (2012), Estimating the evidence – a review, *Statistica Neerlandica*, **66**, 288–308.

References II

Hug, S. Schwarzfischer, M., Hasenauer, J., Marr, C. and Theis, F.J., (2016), An adaptive scheduling scheme for calculating Bayes factors with thermodynamic integration using Simpson's rule, *Statistics and Computing*, to appear.

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

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

Lefebvre, G., Steele, R.J., Vandal, A.C., Narayanan, S. and Arnold, D.L., (2009), Path sampling to compute integrated likelihoods: An adaptive approach, *Journal of Computational and Graphical Statistics*, **18**, 415–437

References III

Oates, C.J., Girolami, M. and Chopin, N., (2016a), Control functionals for Monte Carlo integration, *Journal of the Royal Statistical Society, Series B*, To appear.

Oates, C.J., Papamarkou, T. and Girolami, M., (2016b), The controlled thermodynamic integral for Bayesian model comparison, *Journal of the American Statistical Society*, To appear.