

Easily Computed Marginal Likelihoods from Posterior Simulation Using the THAMES Estimator

Martin Metodiev^{*,†}, Marie Perrot-Dockès[†], Sarah Ouadah[‡], Nicholas J. Irons[§] ,
Pierre Latouche^{*,†,¶}, and Adrian E. Raftery^{§,¶,||}

Abstract. We propose an easily computed estimator of the marginal likelihood from posterior simulation output, via reciprocal importance sampling, combining earlier proposals of DiCiccio et al (1997) and Robert and Wraith (2009). This involves only the unnormalized posterior densities from the sampled parameter values, and does not involve additional simulations beyond the main posterior simulation, or additional complicated calculations, provided that the parameter space is unconstrained. Even if this is not the case, the estimator is easily adjusted by a simple Monte Carlo approximation. It is unbiased for the reciprocal of the marginal likelihood, consistent, has finite variance, and is asymptotically normal. It involves one user-specified control parameter, and we derive an optimal way of specifying this. We illustrate it with several numerical examples.

MSC2020 subject classifications: Primary 62F15, 62-04; secondary 62F12.

Keywords: marginal likelihood estimation, reciprocal importance sampling.

1 Introduction

A key quantity in Bayesian model selection is the marginal likelihood, also known as the evidence, the normalizing constant of the posterior density, or the integrated likelihood. Consider a statistical model with parameter vector θ and data \mathcal{D} . Let $L(\theta) = p(\mathcal{D}|\theta)$ be the usual likelihood, and $\pi(\theta)$ be the prior distribution of θ . Then $Z = p(\mathcal{D}) = \int L(\theta)\pi(\theta)d\theta$ is the marginal likelihood.

The marginal likelihood plays a key role in defining Bayes factors. Consider two models M_1 and M_2 with marginal likelihoods Z_1 and Z_2 . Then the Bayes factor (or ratio of posterior to prior odds) for model M_1 against M_2 is $B_{1,2} = Z_1/Z_2$.

The marginal likelihood is also a critical quantity for Bayesian model averaging (BMA). Consider K models, M_1, \dots, M_K , with prior model probabilities Π_k (which

arXiv: [2305.08952](https://arxiv.org/abs/2305.08952)

^{*}Université Clermont Auvergne, Laboratoire de Mathématiques Blaise Pascal, martin.metodiev@doctorant.uca.fr; PIERRE.LATOUICHE@uca.fr

[†]Université Paris Cité, CNRS, MAP5, F-75006 Paris, France, marie.perrot-dockees@u-paris.fr

[‡]Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, sarah.ouadah@agroparistech.fr

[§]University of Washington, Department of Statistics, njirons@uw.edu

[¶]Equal contribution.

^{||}Corresponding author, raftery@uw.edu.

add up to 1), and marginal likelihoods Z_k . Suppose Q is a quantity of interest, such as a parameter or a future observation to be predicted. Then the BMA posterior distribution of Q is

$$p(Q|\mathcal{D}) = \sum_{k=1}^K p(Q|\mathcal{D}, M_k)p(M_k|\mathcal{D}), \quad (1)$$

where $p(M_k|\mathcal{D})$ is the posterior model probability of M_k , which satisfies $p(M_k|\mathcal{D}) \propto \Pi_k Z_k$ and $\sum_{k=1}^K p(M_k|\mathcal{D}) = 1$. So $p(Q|\mathcal{D}) = \sum_{k=1}^K p(Q|\mathcal{D}, M_k)\Pi_k Z_k / \sum_{k=1}^K \Pi_k Z_k$.

Finally, the most likely model *a posteriori* is the one that maximizes $\Pi_k Z_k$. Choosing it minimizes the model selection error rate on average over the prior (Jeffreys, 1961). Often the prior over the model space is chosen to be uniform, in which case $\Pi_k = 1/K$, $\forall k$. In this case, Bayesian model selection by choosing the most likely model *a posteriori* boils down to choosing the model with the largest Z_k , and hence involves only the marginal likelihoods.

Bayesian models are often estimated using Monte Carlo methods in which a sample of values of θ is simulated from the posterior distribution. The most common class of such methods is Markov chain Monte Carlo (MCMC). Perhaps surprisingly, estimating the marginal likelihood from the output of MCMC and other posterior simulation methods has turned out not to be straightforward. Many different methods have been proposed, and none of them is widely considered to be generally the best. Llorente et al. (2023) provide a comprehensive review of such methods, describing 16 different methods and, remarkably, cite over 20 *other* review articles!

We seek a method that is precise, generic and simple for estimating the marginal likelihood from posterior simulation output. We take this to mean that it gives low variance estimates of the marginal likelihood, uses posterior simulation output for just the one model being analyzed, uses only likelihoods and prior densities of the sampled values of θ , and does not need additional simulations or complicated calculations.

Some well-known methods do not satisfy our desiderata. These include Chib's method (Chib, 1995), which requires complicated additional calculations, bridge sampling (Meng and Wong, 1996), which requires simulations from two models, importance sampling, which requires additional simulations, nested sampling (Skilling, 2006), which involves other simulations, and more advanced methods such as adaptive annealed importance sampling (Liu, 2014). They also include the harmonic mean of the likelihoods (Newton and Raftery, 1994), which is unbiased and consistent, but has infinite variance and is unstable, as pointed out by the original authors.

Arguably, the only methods that are precise, generic and simple for estimating the marginal likelihood from MCMC by our definition are versions of reciprocal importance sampling (RIS) (Gelfand and Dey, 1994). These are based on the identity:

$$Z^{-1} = E_{\theta} \left[\frac{h(\theta)}{L(\theta)\pi(\theta)} \middle| \mathcal{D} \right], \quad (2)$$

where $h(\theta)$ is a (normalized) probability density function (pdf) over the posterior sup-

port. Remarkably, (2) holds for any pdf $h(\theta)$. This leads to the estimator

$$\hat{Z}^{-1} = \frac{1}{T} \sum_{t=1}^T \frac{h(\theta^{(t)})}{L(\theta^{(t)})\pi(\theta^{(t)})}, \quad (3)$$

where $\theta^{(1)}, \dots, \theta^{(T)}$ are simulated from the posterior using MCMC or another method. This estimator has good properties in general, provided that the tails of the distribution $h(\theta)$ are thin enough in all directions. It can be hard to choose $h(\theta)$ so that it both overlaps substantially with the posterior distribution (needed for efficiency) and has thin enough tails (needed for finite variance), especially in higher dimensions. We propose a choice of $h(\theta)$ that leads to easily computed estimates and is optimal or near optimal in a certain sense.

The paper is organized as follows. In Section 2 we discuss reciprocal importance sampling and its properties. In Section 3 we describe our proposed choice of $h(\theta)$ and derive some of its properties. In Section 4 we give several numerical examples, including a multivariate Gaussian example, a Bayesian regression example, a non-Gaussian case, and a Bayesian hierarchical model. We conclude in Section 5 with a discussion. The code for this paper is made available via Github for scientific dissemination at the following [link](#). The THAMES has been implemented in an R package (Irons et al., 2023).

2 Reciprocal importance sampling

In general, the RIS estimator of the marginal likelihood is defined by Equation (3). This has several good properties. It is unbiased, in the sense that $E[\hat{Z}^{-1}] = Z^{-1}$, where the expectation is over the posterior distribution of θ . It is also strongly simulation-consistent, in the sense that $\hat{Z}^{-1} \rightarrow Z^{-1}$ almost surely as $T \rightarrow \infty$.

In addition, the RIS estimator of the reciprocal marginal likelihood, \hat{Z}^{-1} , has finite variance and is asymptotically normally distributed as $T \rightarrow \infty$ if the tails of $h(\theta)$ are thin enough. Specifically, this requires that

$$\int \frac{h(\theta)^2}{L(\theta)\pi(\theta)} d\theta < \infty. \quad (4)$$

It is hard to choose $h(\theta)$ so that it both overlaps substantially with the area of the parameter space with high posterior density, which is needed for efficiency, and so that it also has thin enough tails, which is needed for finite variance. The difficulty grows as the dimension increases.

Two choices of $h(\theta)$ in the literature deserve attention. DiCiccio et al. (1997) proposed $h(\theta) = MVN(\theta; \hat{\theta}, \hat{\Sigma})$, where $\hat{\theta}$ is the posterior mean or mode, and $\hat{\Sigma}$ is an estimate of the posterior covariance matrix. This overlaps nicely with $L(\theta)\pi(\theta)$, but its tails may not be thin enough when the posterior is asymmetric or the parameter is high-dimensional.

To remedy the problem of the tails possibly being too thick, DiCiccio et al. (1997) proposed truncating it, using instead $h(\theta) = TMVN_A(\hat{\theta}, \hat{\Sigma})$, a multivariate normal distribution truncated to the set A , where

$$A = \{\theta : (\theta - \hat{\theta})^T \hat{\Sigma}^{-1} (\theta - \hat{\theta}) < c^2\}. \quad (5)$$

Thus A is an ellipsoid with radius c and volume

$$V(A) = c^d \pi^{d/2} |\hat{\Sigma}|^{1/2} / \Gamma\left(\frac{d}{2} + 1\right). \quad (6)$$

Truncating the distribution ensures that the estimator \hat{Z}^{-1} has finite variance. The authors found that the truncation improved the performance of the RIS estimator. However, with high-dimensional parameters, the result might be sensitive to the specification of c .

Robert and Wraith (2009) proposed setting $h(\theta)$ to be a uniform distribution on the convex hull of simulated MCMC parameters values in the α -HPD region, namely the highest posterior density region containing a proportion α of the sampled parameter values. They considered the values $\alpha = 0.1$ and 0.25 . They applied it to a two-dimensional toy example where it performed well.

However, as far as we know, that method has not yet been fully developed for realistic, higher-dimensional situations. For example, we know of no simple way to compute the volume of the convex hull of a set of points in higher dimensions, which is required for the method in general. It is also not clear how best to choose α nor how sensitive the method would be to α in higher dimensions. It has been used in a higher-dimensional application by Durmus et al. (2018), but this involved comparing competing models defined on the same parameter space, thus avoiding the need to calculate the volume of A , which canceled out in Bayesian model comparisons. Calculating the volume of A may be the most difficult part of this method in general.

3 Estimating the marginal likelihood

3.1 Estimating the marginal likelihood with THAMES

We propose combining the proposals of DiCiccio et al. (1997) and Robert and Wraith (2009) to obtain a method that we believe satisfies all our desiderata. We propose specifying $h(\theta)$ to be a uniform distribution, but to be uniform over the set A defined in Equation (5), rather than over a convex hull of points. This resolves the problem of computing the volume of A , since this is given analytically by Equation (6). If A is not a subset of the posterior support, for example if the posterior support is constrained, we adjust the volume of A by a simple Monte Carlo approximation. This yields the estimator

$$\hat{Z}^{-1} = \frac{1}{V(A)T} \sum_{\substack{t=1 \\ \theta^{(t)} \in A}}^T \frac{1}{L(\theta^{(t)})\pi(\theta^{(t)})}. \quad (7)$$

Thus \hat{Z} is a truncated harmonic mean of the unnormalized posterior densities, $L(\theta^{(t)})\pi(\theta^{(t)})$.¹ We call it the Truncated HARmonic Mean ESTimator, or THAMES.

The THAMES, \hat{Z}^{-1} , has several desirable properties. It is simple to compute, involving only the prior and likelihood values of the sampled parameter values. In fact it involves only the product of the prior and likelihood values, namely the unnormalized posterior densities of the sampled parameter values. It is unbiased as an estimator of Z^{-1} , as long as A is specified independently of the sample. It is also simulation-consistent, in the sense that $\hat{Z}^{-1} \rightarrow Z^{-1}$ almost surely as $T \rightarrow \infty$, by the strong law of large numbers. Its variance (over simulation from the posterior given the data \mathcal{D}) is finite provided that

$$\int_A (L(\theta)\pi(\theta))^{-1} d\theta < \infty, \quad (8)$$

which will usually hold since A is a bounded set in \mathbb{R}^d . In fact, it suffices that the likelihood and the prior are continuous with respect to θ and strictly positive on the closure of A . If Equation (8) holds, \hat{Z}^{-1} is asymptotically normal (again as the number of parameter values simulated increases), by the Lindeberg central limit theorem. Note that asymptotic normality holds on the scale of \hat{Z}^{-1} , and not exactly on other scales such as \hat{Z} or $\log(\hat{Z})$.

If the posterior simulation method yields independent draws, then $\text{Var}(\hat{Z}^{-1})$ can be estimated directly as the empirical variance of the values of $(L(\theta^{(t)})\pi(\theta^{(t)})\mathbb{1}(\theta^{(t)} \in A))^{-1}$, divided by $T \cdot V(A)^2$. If MCMC is used, successive simulations from the posterior will in general not be independent. A central limit theorem will still hold, but the variance needs to take account of the serial dependence. This can be done approximately by computing the variance based on serial independence and multiplying it by an estimate of the spectral density of the sequence at zero. For example, if the sequence of values of $1/(L(\theta)\pi(\theta))$ can be approximated by a first-order autoregressive model with parameter ϕ , then this would be approximately $1/(1 - \phi)^2$. An alternative would be to thin the sequence enough that the resulting subsequence is approximately uncorrelated and then use the variance based on assuming independence. A different approach was taken by Frühwirth-Schnatter (2004).

Note that an approximate normal confidence interval can be obtained for \hat{Z}^{-1} , because that is the scale on which a central limit theorem holds. This could be turned into a confidence interval for \hat{Z} by taking the reciprocals of the ends of the normal confidence intervals for \hat{Z}^{-1} ; the resulting confidence interval would not be symmetric. The same could be done for $\log(\hat{Z})$ in a similar manner.

3.2 Optimal choice of control parameter, c

We now address the question of how to choose the radius c of the ellipse that specifies the THAMES in Equation (5). Ignoring serial correlation between simulated values of the

¹Recall that the unstable harmonic mean estimator described by (Newton and Raftery, 1994) was quite different, not being truncated, and being a harmonic mean of the likelihoods rather than the unnormalized posterior density values.

parameters, we suggest choosing c to minimize the estimated variance of \hat{Z}^{-1} . This could be done empirically by computing \hat{Z}^{-1} for a range of values of c , estimating $\text{Var}(\hat{Z}^{-1})$ for each value of c , and optimizing it over c by a grid search or a one-dimensional numerical optimization method.

It is possible to obtain analytic results in the case where the posterior distribution is normal. This is of considerable interest as the posterior distribution is asymptotically normal in many common situations, including some where standard regularity conditions do not hold (Heyde and Johnstone, 1979; Ghosal, 2000; Shen, 2002; Miller, 2021). In this case the THAMES has finite variance since the posterior density, and thus the product of the likelihood and the prior, is continuous with respect to θ and strictly positive everywhere.

We want to minimize the variance of the THAMES. Due to our assumption of independence of all of the successive MCMC simulations, this variance can be simplified to

$$\text{Var}(\hat{Z}^{-1}|\mathcal{D}) = \frac{1}{T} \cdot \frac{1}{Z^2} \cdot \text{SCV}(d, c). \quad (9)$$

Here $\text{SCV}(d, c)$ denotes

$$\text{SCV}(d, c) := \frac{\text{Var}_{\theta^{(1)}} \left(\frac{\mathbb{1}_A(\theta^{(1)})/V(A)}{L(\theta^{(1)})\pi(\theta^{(1)})} \middle| \mathcal{D} \right)}{E_{\theta^{(1)}} \left(\frac{\mathbb{1}_A(\theta^{(1)})/V(A)}{L(\theta^{(1)})\pi(\theta^{(1)})} \middle| \mathcal{D} \right)^2}, \quad (10)$$

the squared coefficient of variation of the first term of the THAMES. Since the variance is a product of $\frac{1}{T}$, $\frac{1}{Z^2}$ and $\text{SCV}(d, c)$, minimizing $\text{SCV}(d, c)$ with respect to c is equivalent to minimizing the variance of the THAMES.

We derive a statement about the optimal choice of c by assuming that the posterior covariance matrix Σ and the posterior mean m can be provided by a stochastic oracle. The THAMES can then be defined using

$$A_{or} := \{\theta : (\theta - m)^T \Sigma^{-1} (\theta - m) < c^2\}. \quad (11)$$

We will show that the radius c that minimizes the variance of the THAMES depends on the dimension d , and is equal to $c_d = \sqrt{d + L_d}$ with L_d being close to one for large d . Interestingly, in this case the SCV depends neither on the data, \mathcal{D} , nor on the number of samples from the posterior, T . Of course, this is rarely exactly the case in practice. However, plugging in consistent estimators of (m, Σ) gives approximately the same results if the number of samples from the posterior is large enough, provided that a sample splitting procedure is used. This will be a consequence of Theorem 3.3. The sample splitting procedure that we suggest is described in Section 3.3. The proofs of these results are given in Supplement A (Metodiev et al., 2024a). Additional numerical results about the behaviour of the optimal radius are given in Supplement B (Metodiev et al., 2024b).

Assumption 1. For the following theorems it is assumed that we can ignore serial correlation (i.e. we assume independence of the successive MCMC iterations) and that the posterior distribution is normal with mean $m \in \mathbb{R}^d$ and positive definite covariance matrix $\Sigma \in \mathcal{M}_{d \times d}(\mathbb{R})$. We further assume that the THAMES is defined on A_{or} .

Theorem 3.1. *There exists a unique radius $c_d \in (0, \infty)$ such that the ellipsoid A_{or} with radius c_d minimizes the variance of the THAMES. This value c_d does not depend on the posterior mean or covariance matrix. It satisfies $c_d = \sqrt{d + L_d}$, where the optimal shifting parameter $L_d \geq 0$ is a sequence for which $\frac{L_d}{d} \xrightarrow{d \rightarrow \infty} 0$.*

Remark 1. Theorem 3.1 ensures that the optimal radius c_d is asymptotically equivalent to \sqrt{d} . In fact, our calculations suggest that $c_d = \sqrt{d + L_d}$ can be approximated by $\sqrt{d + 1}$ (See Supplement B (Metodiev et al., 2024b)).

Theorem 3.2. *The following statements hold for the SCV:*

1. *For any choice of the shifting parameter $L, L \in \mathbb{R}$ and for all $\varepsilon > 0$*

$$1 - \varepsilon \leq \frac{SCV(d, \sqrt{d + L_d})}{\sqrt{(d + 2)\pi/4}} \leq \frac{SCV(d, \sqrt{d + L})}{\sqrt{(d + 2)\pi/4}} \leq 2 + \varepsilon, \quad (12)$$

for all but finitely many d . Thus choosing the radius $\sqrt{d + L}$ results in an SCV that is both asymptotically at most twice as large as the optimal SCV and is of order \sqrt{d} .

2. *The following inequality for the SCV can be given for choosing the radius $c = \sqrt{d + 1}$:*

$$0.63\sqrt{(d + 2)\pi/4} - 1 \leq SCV(d, \sqrt{d + L_d}) \quad (13)$$

$$\leq SCV(d, \sqrt{d + 1}) \leq 2.1\sqrt{(d + 2)\pi/4} - 1 \quad (14)$$

This inequality holds for all $d \geq 1$.

Remark 2. Statement 1 of Theorem 3.2 shows that $SCV(d, c)$ is increasing with order \sqrt{d} as $d \rightarrow \infty$, both in our choice $c = \sqrt{d + 1}$ and the optimal choice c_d . Further, any choice of the shifting parameter L used to define the radius $\sqrt{d + L}$ is asymptotically at most twice as bad as any optimal solution in terms of the SCV. This suggests some robustness of our estimator with respect to the choice of L . For numerical results about the behaviour of the SCV for different values of L , we refer the reader to Supplement B (Metodiev et al., 2024b).

One can also calculate the bias of the THAMES on the scale of the marginal likelihood by considering a second-order Taylor approximation and using Equation (9):

$$E[\hat{Z}] = Z + Var[\hat{Z}^{-1}]/(Z^{-1})^3 = Z(1 + SCV(d, c)/T). \quad (15)$$

Considering Equation (15), the bias can be estimated by using the plug-in estimates \widehat{SCV} and \hat{Z}^{-1} . We also observe that the bias vanishes as T increases.

Remark 3. Statement 2 of Theorem 3.2 gives a very rough theoretical guarantee: For any dimension $d \geq 1$, the SCV obtained by choosing our recommendation for the radius, $\sqrt{d + 1}$, and the SCV obtained by choosing the optimal radius, c_d , can be bounded by an affine transform of $\sqrt{d + 2}$. However, our calculations suggest that the SCV at the point $c = \sqrt{d + 1}$ has an asymptotically optimal performance (See Supplement B (Metodiev et al., 2024b)).

So far, we have given results for the idealized situation where the posterior distribution is exactly normal. We now give a result for the more common and realistic situation where the posterior distribution is only asymptotically normal.

Theorem 3.3. *Let $p_n(\theta|\mathcal{D}_n)$ be a sequence of posterior densities with data \mathcal{D}_n , posterior covariance matrix Σ_n , posterior mean m_n and an SCV denoted by SCV_n . Then, if*

$$|\Sigma_n|^{\frac{1}{2}} p_n \left(\Sigma_n^{\frac{1}{2}} \cdot \theta + m_n \mid \mathcal{D}_n \right) \xrightarrow{n \rightarrow \infty} |\Sigma|^{\frac{1}{2}} p \left(\Sigma^{\frac{1}{2}} \cdot \theta + m \mid \mathcal{D} \right) \quad (16)$$

uniformly in θ on all compact subsets of \mathbb{R}^d , it follows that

$$SCV_n(d, c) \xrightarrow{n \rightarrow \infty} SCV(d, c) \quad (17)$$

uniformly in c on all compact subsets of $(0, \infty)$. In particular, for any $b \geq c_d \geq a > 0$,

$$(c_d)_n \in \operatorname{argmin}_{c \in [a, b]} SCV_n(d, c) \forall n \Rightarrow \lim_{n \rightarrow \infty} (c_d)_n = c_d. \quad (18)$$

Remark 4. We have already stated that the normal case is important because the posterior distribution is often asymptotically normal when the size of the data, n , is large. Theorem 3.3 assures us that our results still hold in this limiting case, under some assumptions.

If the convergence of the normalized posterior pdf is uniform in θ (Equation (16)), our statements about the limit behaviour of the SCV (Theorem 3.2 and Remarks 2-3) still hold approximately when n is large (Equation (17)). If additionally any optimal radius $(c_d)_n$ does not converge to zero or infinity, any result about c_d (Theorem 3.1 and Remark 1) also holds approximately when n is large (Equation (18)).

Let H_0 denote the Fisher information matrix. Reformulating Equation (16) by replacing Σ_n by $\frac{1}{n}H_0^{-1}$, to which it is asymptotically equivalent, gives a statement that has been proven under a variety of assumptions, e.g. Miller (2021, Theorem 4), except that in these results the type of convergence is usually not uniform convergence, but a weaker type of convergence, such as convergence in distribution or convergence in total variation.

Additional assumptions can be made about the pdfs of the sequence of distributions such that convergence in distribution implies uniform convergence of the pdfs. For example, if the pdfs are asymptotically equicontinuous and we have convergence in distribution, the convergence of the pdfs is uniform (Sweeting, 1996, Theorem 1).

Note that in this case there is no problem if the parameter space is constrained. Uniform convergence of the pdfs implies that A_n is a subset of the posterior support if n is large enough. There are also no assumptions about (m_n, Σ_n) , other than that they converge to the moments of the posterior limit. In this sense, the estimators $(\hat{\mu}, \hat{\Sigma})$ take the place of these constants in practice and Theorem 3.3 holds even when we use these estimators.

Remark 5. Due to the assumption of normality it is the case that when choosing the optimal radius $c_d = \sqrt{d + L_d}$, the probability of a term of the THAMES in $\theta^{(t)}$ not being set to 0 is equal to

$$\mathbb{P}(\theta^{(t)} \in A_{or}) = \mathbb{P}((\theta^{(t)} - m)^T \Sigma^{-1} (\theta^{(t)} - m) < d + L_d) = \chi^2(d + L_d; d), \quad (19)$$

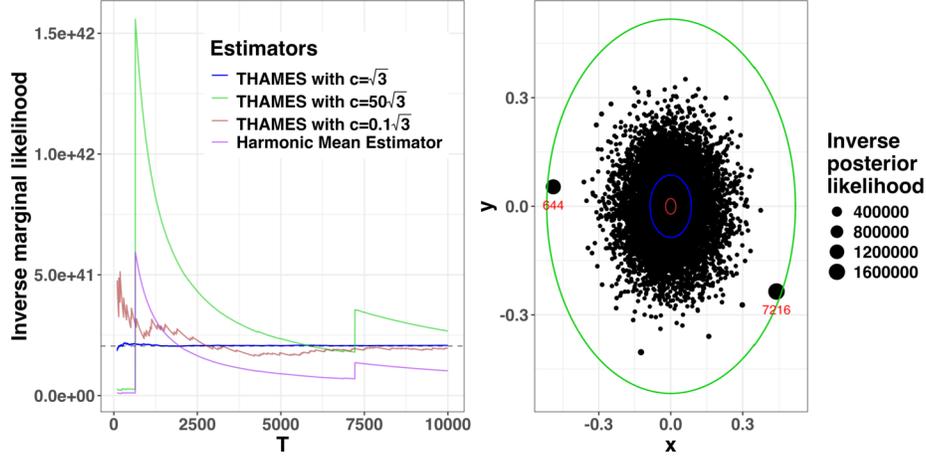


Figure 1: Left: The THAMES calculated by choosing the radii $c = \sqrt{d+1} = \sqrt{3}$, $c = 0.1\sqrt{3}$, $c = 50\sqrt{3}$ in the two-dimensional case, $d = 2$, with the true value $1/Z$ (dotted line) and the Harmonic Mean Estimator. Right: The posterior sample evaluated at the inverse of the unnormalized posterior density and the different ellipses used to define the THAMES. In this particular case the posterior covariance matrix is the scaled identity matrix, so the ellipses are spheres. The two samples occurring at points 644 and 7216 have a very low likelihood. They cause massive jumps in the Harmonic Mean Estimator when the radius of A is large and are excluded when the radius is equal to $\sqrt{d+1} = \sqrt{3}$. One can choose a smaller radius (e.g., $c = 0.1\sqrt{3}$), but then too much of the sample is excluded and convergence takes longer.

the CDF of the χ^2 -distribution with d degrees of freedom evaluated at $d + L_d$. This approaches 50% due to Theorem 3.1. Thus the algorithm sets about 50% of the highest terms in Equation (7) to 0. This means that for a large number of samples T and given the normality assumption, our algorithm is similar to the following method:

Instead of checking whether $\theta^{(t)} \in A_{or}$ directly, one can set roughly 50% of the highest terms of the THAMES, the terms not included in the Highest Posterior Density (HPD) region of size 50%, to 0.

Remark 6. It is assumed that the covariance matrix of the posterior distribution is positive definite. This assumption is necessary since otherwise a posterior density with respect to the Lebesgue-measure on \mathbb{R}^d would not exist. On the other hand this assumption is not restrictive, since the same estimation procedure can be applied to a lower dimensional subspace of \mathbb{R}^d on which a density is defined.

We can illustrate the relationship between the THAMES and the harmonic mean estimator defined by Newton and Raftery (1994) using the toy example from Figure 1. It was calculated using the same model as the one introduced in Section 4.1 with the dimensions of the parameter space $d = 2$, but by setting the data set to $\mathcal{D} \equiv 0$ to ensure stability of the estimator on the inverse likelihood scale.

The pdf of the Uniform distribution on the ellipse is essentially used as a rejection rule: Values with very low posterior density (and therefore high inverse posterior density) are rejected, while high-density values are accepted. A balance between the volume of the ellipse and the percentage of the rejected posterior sample needs to be found to ensure optimal performance. The harmonic mean estimator does not have this rejection rule, so sample points with low posterior densities can lead to massive jumps.

3.3 THAMES algorithm

Below is an algorithm for the implementation of the THAMES. Procedures for sample splitting, as well as the truncated ellipsoid correction used in the case that the parameter space is constrained have been included. These additions are described in page 11.

We recommend these additions, but we have also found that in some cases they make almost no difference. For example, sample splitting does not appear to have an impact when the dimension of the parameter space, d , is small (Section 4.1), while the truncated ellipsoid correction is negligible when the posterior mean is not close to the edge of the posterior support (Section 4.4 and Section 4.3).

Algorithm 1 \hat{Z}^{-1} calculation.

Input: Data \mathcal{D} and posterior samples $(\theta^{(i)})_{i \in \llbracket 1, T \rrbracket}$.

Sample splitting: Calculate the empirical mean $\hat{\theta}$ and sample covariance matrix $\hat{\Sigma}$ based on the first $T/2$ posterior samples $(\theta^{(i)})_{i \in \llbracket 1, T/2 \rrbracket}$.

Standardization: $\tilde{\theta}^{(i)} = \hat{\Sigma}^{-1/2}(\theta^{(i)} - \hat{\theta})$ for $i \in \llbracket T/2 + 1, T \rrbracket$.

Truncation subset: $\mathcal{S} = \{i : \|\tilde{\theta}^{(i)}\|_2^2 < d + 1\}$.

Calculate THAMES estimator:

$$\hat{Z}^{-1} = \frac{1}{T/2} \sum_{i=T/2+1}^T \frac{h(\theta^{(i)})}{L(\theta^{(i)})\pi(\theta^{(i)})},$$

where $h(\theta^{(i)}) = 1/V(A)$ if $i \in \mathcal{S}$ and 0 otherwise, with

$$V(A) = \sqrt{|\hat{\Sigma}|} \pi^{d/2} (d+1)^{d/2} / \Gamma(\frac{d}{2} + 1) \text{ and } A = \{\theta : (\theta - \hat{\theta})^T \hat{\Sigma}^{-1} (\theta - \hat{\theta}) < d + 1\}.$$

if the posterior support $\text{supp}(\theta|\mathcal{D})$ is constrained **then**

Simulate the sample $\nu^{(1)}, \dots, \nu^{(N)}$ from the uniform distribution on A .

Approximate the volume ratio $V(A \cap \text{supp}(\theta|\mathcal{D})) / V(A)$ via the Monte Carlo estimator

$$\hat{R} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}_{\{\theta|\pi(\theta)L(\theta)>0\}}(\nu^{(i)}).$$

Assign $\hat{Z}^{-1} \leftarrow \hat{R}^{-1} \hat{Z}^{-1}$.

end if

Output: THAMES estimator \hat{Z}^{-1} .

Sample splitting

The theoretical guarantees established in Section 3.2 operate under the assumption of an oracle ellipsoid A_{or} . In particular, this means that the ellipsoid determining the THAMES estimator \hat{Z}^{-1} is defined independently of $(\theta^{(i)})_{i \in \llbracket 1, T \rrbracket}$. In practice, we find that estimating A and Z^{-1} simultaneously using the same posterior sample can induce bias in \hat{Z}^{-1} when the parameter space is high-dimensional. We therefore implement a sample splitting procedure that involves estimating A and Z^{-1} using separate posterior draws. Specifically, we first estimate the posterior mean and covariance matrix via the empirical mean $\hat{\theta}$ and sample covariance $\hat{\Sigma}$ using the first $T/2$ posterior samples $(\theta^{(i)})_{i \in \llbracket 1, T/2 \rrbracket}$. Defining A as in Equation (5) based on $\hat{\theta}$ and $\hat{\Sigma}$, we then calculate the THAMES estimator \hat{Z}^{-1} using the last $T/2$ posterior samples $(\theta^{(i)})_{i \in \llbracket T/2+1, T \rrbracket}$. The same problem was noted by Gronau et al. (2020) in their popular implementation of bridge sampling. For this reason, the bridge sampling package uses the same sample splitting procedure just described, in its default setting.

Correcting for the presence of constrained parameters

Whenever the posterior support of the parameters is not \mathbb{R}^d , for example when the parameters are variances or probabilities, it is possible that our choice of h in Equation (3), the pdf of the uniform distribution on A , is not correctly normalized. This is due to the fact that A is not necessarily a subset of the posterior support and thus h is not a pdf over this space.

In this case, the expectation of the THAMES is distorted by a multiplicative constant:

$$E_{\theta}[\hat{Z}^{-1}|\mathcal{D}] = E_{\theta} \left[\frac{h(\theta)}{L(\theta)\pi(\theta)} \middle| \mathcal{D} \right] = Z^{-1} \cdot \frac{V(A \cap \text{supp}(\theta|\mathcal{D}))}{V(A)} =: Z^{-1}R, \quad (20)$$

where $V(A \cap \text{supp}(\theta|\mathcal{D}))$ denotes the volume of the intersection between A and the posterior support. One way to deal with this problem is to transform the parameter space, e.g., by setting $\vartheta := \log(\theta)$ if θ is a variance parameter. One can then continue with marginal likelihood estimation on ϑ , using the transformed prior distribution. In this case, it is of course important to include the Jacobian of the transformation when computing the prior density. It should be noted that the default proposal used in the bridge sampling package from Gronau et al. (2020) uses this transformation, because it suffers from the same problem when the parameter space is constrained. This solution is also a viable option for the THAMES, since the transformation removes the need for any adjustments. However, this solution requires deciding on a viable transformation for each new type of constraint (e.g., simplex constraints, interval constraints, etc.) and the posterior behaviour of the transformed parameters may be hard to interpret. For this reason, we suggest a different correction.

Another way is to adjust for the bias by calculating the ratio of these volumes, R , using a simple Monte Carlo approximation: We simulate $\nu^{(1)}, \dots, \nu^{(N)} \stackrel{\text{i.i.d.}}{\sim} \text{Unif}(A)$, $N \in$

\mathbb{N} and calculate

$$\hat{R} := \frac{1}{N} \sum_{i=1}^N \mathbb{1}_{\text{supp}(\theta|\mathcal{D})}(\nu^{(i)}) = \frac{1}{N} \sum_{i=1}^N \mathbb{1}_{\{\theta|\pi(\theta)L(\theta)>0\}}(\nu^{(i)}). \quad (21)$$

Given A , this is an unbiased and consistent estimator of R by the law of large numbers. The bias-adjusted THAMES is then $\hat{Z}_{adj}^{-1} = \hat{R}^{-1}\hat{Z}^{-1}$.

The problem of the parameter space being constrained is common not only for the THAMES, but for reciprocal importance sampling estimators in general. It has for example been addressed by Hajargasht and Woźniak (2018) and Sims et al. (2008). Hajargasht and Woźniak (2018) used variational Bayes techniques and showed that these ensure that the support of the chosen h is a subset of the posterior support, under mild conditions. Sims et al. (2008) used an ellipsoidal density truncated on a subset of the joint support, $\Theta_U := \{\theta|\pi(\theta)L(\theta) > U\}$, where $U > 0$. Since the support of $\pi(\theta)L(\theta)$ is equal to the posterior support, our truncation set is similar to the one chosen in Sims et al. (2008), except that we set $U = 0$.

The adjustment is usually very small. The problem arises only when the posterior mean is close enough to the edge of the parameter space. The edge of the parameter space often indicates a priori unlikely values. For this reason it is also rare that the data indicate posterior parameters being close to the edge. Thus the ratio between the volumes is close to one and the variance of \hat{R} is small. In fact, the adjustment did not have any sizeable impact on any of the examples simulated in Section 4. This may not be the case, however, if the actual data generating mechanism is very different from the model being considered. In this case, it can in practice happen that the posterior mean is indeed very close to the edge. We show one example of this in Supplement B Metodiev et al. (2024b).

In either case we have found that a small number of simulations, around $N = 100$, is usually enough. Confidence intervals obtained from the fact that \hat{R} is asymptotically normal can be used to check whether the variance of \hat{R} is large. In this case N should be increased to yield a more precise approximation. The computational cost of implementing this adjustment is typically small.

4 Examples

We now describe several simulated and real data examples to assess the THAMES estimator. In Sections 4.1, 4.2, and 4.3, three statistical models, for which exact expressions of the marginal likelihood are derived, are considered. This allows us to compare the THAMES estimated values to the exact ones for evaluation. In Section 4.4, we consider a real data example with models for which no analytical expressions for the marginal likelihood are available, to our knowledge, and where there is a need for reliable estimators. We compare our estimator to bridge sampling, which is more complicated than THAMES but is known to have performed well (Meng and Wong, 1996; Gronau et al., 2020).

4.1 Multivariate Gaussian data

We first consider the case where data $Y_i, i = 1, \dots, n$ are drawn independently from a multivariate normal distribution:

$$Y_i | \mu \stackrel{\text{iid}}{\sim} \text{MVN}_d(\mu, I_d), \quad i = 1, \dots, n,$$

along with a prior distribution on the mean vector μ :

$$p(\mu) = \text{MVN}_d(\mu; 0_d, s_0 I_d),$$

with $s_0 > 0$. As shown in Supplement A (Metodiev et al., 2024a), the posterior distribution of the mean vector μ given the data $\mathcal{D} = \{y_1, \dots, y_n\}$ is given by:

$$p(\mu | \mathcal{D}) = \text{MVN}_d(\mu; m_n, s_n I_d), \quad (22)$$

where $m_n = n\bar{y}/(n + 1/s_0)$, $\bar{y} = (1/n) \sum_{i=1}^n y_i$, and $s_n = 1/(n + 1/s_0)$.

Interestingly, while the observations $(Y_i)_i$ are independent given the vector μ , they are not independent marginally, and the marginal likelihood does not take the form of a product over marginal terms in i . Conversely, thanks to the isotropic Gaussian prior distribution which is considered for μ , where the $(\mu_j)_j$ are all iid, not only are the vectors $(Y_j)_j$ independent given μ , they are also independent marginally. From this property, we prove in Proposition 2 of Supplement A (Metodiev et al., 2024a) that the marginal likelihood of the model can be written analytically as

$$p(\mathcal{D}) = \prod_{j=1}^d \text{MVN}_n(y_{.j}; 0_n, s_0 \mathbf{1}_n \mathbf{1}_n^\top + I_n), \quad (23)$$

where $y_{.j} \in \mathbb{R}^n$ is the vector of all observations for variable j such that $[y_{.j}]_i = y_{ij}$ and $\mathbf{1}_n$ is the vector of 1 in \mathbb{R}^n .

Assessing the precision of the THAMES estimator as a function of T

We first considered the univariate case $d = 1$. We simulated a unique sample of size $n = 20$ with $\mu = 2$ and we set $s_0 = 1$, for illustration; other choices for s_0 led to similar conclusions regarding the quality of the estimation. Figure 2 shows the THAMES estimated values for the log marginal likelihood, for $T = 5, 1005, 2005, \dots, 9005$ samples of the posterior distribution (Equation (22)). Confidence intervals as well as the exact value of the log marginal likelihood computed using Equation (23) are also reported. It can be seen that the estimate converges to the correct value and that the confidence intervals contain the true value in all cases, even for $T = 5$ only.

Assessing the precision of the THAMES estimator as a function of d

For this second set of experiments, we considered different values of d , and aimed at testing the robustness of the THAMES approach on multiple data sets, with increasing

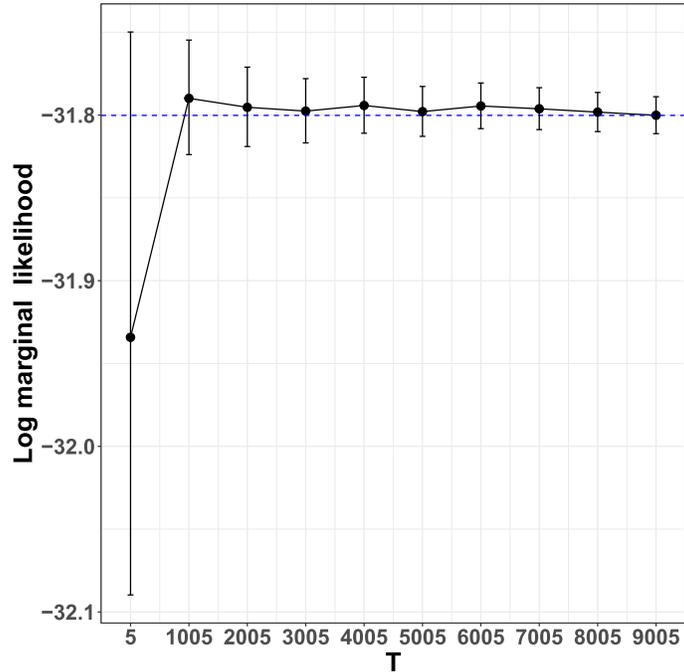


Figure 2: Estimation of the log marginal likelihood using THAMES for a unique univariate Gaussian sample with $n = 20$ as a function of T , the number of (cumulative) samples from the posterior distribution. The black dots indicate the values of the THAMES estimator of the log marginal likelihood. The vertical lines represent 95% confidence intervals, and the dashed blue line represents the exact value computed using Equation (23).

dimensionality. Thus, for each d , we generated 50 different data sets of size $n = 20$ using the multivariate Gaussian model. In practice, we set the true value of μ to 2, for all its components. Again, the prior parameter s_0 was set to 1 and similar conclusions were drawn for other values. Moreover, the value of T was set to 10,000 for all the experiments.

We also used this example to assess the sample splitting procedure for the posterior samples, as proposed in Section 3.3. The results are given in Figure 3. In the figure on the left, where *no* sample splitting of the posterior samples is used to compute THAMES, we observe that a bias appears as the dimensionality of the model considered increases, and the log marginal likelihood tends to be slightly underestimated. As illustrated by the figure on the right hand side, this bias is primarily related to the estimation of the posterior covariance matrix, and not to the THAMES estimation itself. Indeed, focusing on this figure on the right, we note that if the exact expression of the posterior covariance matrix given in Equation (22) is used to compute THAMES, then while the variance of the estimator increases with d , we do not observe any bias. Crucially, if the

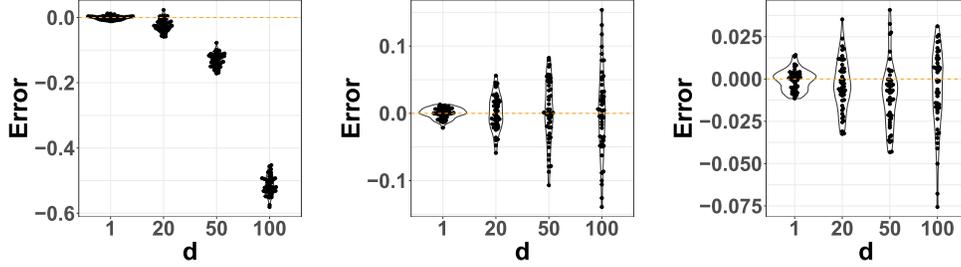


Figure 3: Difference between the estimated log marginal likelihood using THAMES and the true log marginal likelihood for a multivariate Gaussian model with $n = 20$ and $T = 10000$. The procedure is repeated on 50 different data sets for each d . Left: the THAMES approach *with no* sample splitting of the posterior samples. Middle: the THAMES approach *with* sample splitting of the posterior samples. Right: the results provided correspond to the case where the exact expression of the posterior covariance matrix given in Equation (22) is used to compute the THAMES.

sample splitting of the posterior samples is employed to compute THAMES, then again, we do not observe any bias.

Overall, we found that the sample splitting procedure of the posterior samples was not necessary to compute THAMES for low values of d . The estimated values are close to the exact ones. However, for large values of d , we recommend using the sample splitting procedure to remove the bias.

4.2 Bayesian regression

We consider a data set $(x_i, Y_i), i = 1, \dots, n$ to train a linear regression model of the form

$$Y_i | x_i, \beta, \sigma^2 \sim \mathcal{N}(x_i^\top \beta, \sigma^2), i = 1, \dots, n.$$

In this section, the goal is to assess the quality of our proposed estimator. As such, we choose a prior on (β, σ^2) for which an exact expression for the marginal likelihood, Z , exists. We compare our estimator, the THAMES, to the bridge sampling estimator implemented in Gronau et al. (2020) and a simple Monte Carlo (MC) estimator, calculated by averaging the likelihood for parameter values simulated from the prior.

Denoting $Y \in \mathbb{R}^n$, the vector of target variables Y_i , and $X \in \mathcal{M}_{n \times (d-1)}(\mathbb{R})$ the design matrix where the input vectors $x_i \in \mathbb{R}^{d-1}$ are stacked as row vectors, the linear regression model becomes:

$$Y | X, \beta, \sigma^2 \sim \text{MVN}_n(X\beta, \sigma^2 I_n).$$

We rely on a centered isotropic Gaussian prior distribution for the regression vector β

and the variance σ^2 :

$$p(\beta|\sigma^2) = \text{MVN}_{d-1}(\beta; 0_{d-1}, g\sigma^2(X^T X)^{-1}), \quad p(\sigma^2) = \text{InvGamma}\left(\sigma^2; \frac{1}{2}\nu_0, \frac{1}{2}\sigma_0^2\nu_0\right),$$

with $g, \sigma_0^2, \nu_0 > 0$. Introduced by Zellner (1971, 1986), this framework offers a conjugate prior with the attractive property of scale-invariance with respect to the regressor (Hoff, 2009). Then the posterior distribution of (β, σ^2) , given the training data set $\mathcal{D} = \{(x_1, y_1), \dots, (x_n, y_n)\}$, is tractable:

$$p(\beta|\sigma^2, \mathcal{D}) = \text{MVN}_{d-1}\left(\beta; \frac{g}{g+1}m_n, \frac{g}{g+1}\sigma^2(X^T X)^{-1}\right),$$

$$p(\sigma^2|\mathcal{D}) = \text{InvGamma}\left(\sigma^2; \frac{1}{2}(\nu_0 + n), \frac{1}{2}(\nu_0\sigma_0^2 + s_n)\right),$$

with $s_n = \mathbf{y}^T \mathbf{y} - \frac{g}{g+1} \mathbf{y}^T X (X^T X)^{-1} X^T \mathbf{y}$ and $m_n = (X^T X)^{-1} X^T \mathbf{y}$, where $\mathbf{y} \in \mathbb{R}^n$ is the *observed* vector of target variables associated with Y . Moreover, the marginal likelihood also has an analytical expression:

$$p(\mathbf{y}|X) = \frac{(g+1)^{-(d-1)/2}}{\pi^{n/2}} \cdot \frac{\Gamma(\frac{1}{2}(\nu_0 + n))}{\Gamma(\frac{1}{2}\nu_0)} \cdot \left(\frac{\nu_0\sigma_0^2}{\nu_0\sigma_0^2 + s_n}\right)^{\nu_0/2}.$$

Proofs for the exact expressions of the posterior and the marginal are given in Hoff (2009, Chapter 9). The data for this example are described by Hastie et al. (2009) and come from a study by Stamey et al. (1989). They examined the correlation between the level of prostate-specific antigen (lpsa) and eight clinical measures in men who were about to receive a radical prostatectomy. The variables are log cancer volume (lcavol), log prostate weight (lweight), age, log of the amount of benign prostatic hyperplasia (lbph), seminal vesicle invasion (svi), log of capsular penetration (lcp), Gleason score (gleason), and percent of Gleason scores 4 or 5 (pgg45). The target variable is the level of prostate-specific antigen (lpsa).

The choice of the hyperparameter g is a topic of much discussion (Fernández et al., 2001; Porwal and Raftery, 2022). In Porwal and Raftery (2022), $g = \sqrt{n}$ showed good performance when compared to a variety of different options, albeit in a slightly different setting where the prior on σ^2 is improper. For this reason, we chose $g = \sqrt{n}$. We chose $(\nu_0, \sigma_0^2) = (4, 1)$ for the other hyperparameters, but other choices for (g, ν_0, σ_0^2) led to similar conclusions regarding the quality of the estimation.

Seven different regression models M_2, M_3, \dots, M_8 , each with a different number of selected variables, ranging from 2 to 8, are considered for illustration. The variables are added in the order given above. Thus, M_2 includes the predictor variables lcavol and lweight, while Model M_3 considers the variables lcavol, lweight, as well as age for prediction. Finally, model M_8 takes all 8 input variables into account. Figure 4 shows the estimators of the log marginal likelihood for different number of samples from the posterior distribution in (β, σ^2) , for the different models, as well as the approximate confidence intervals.

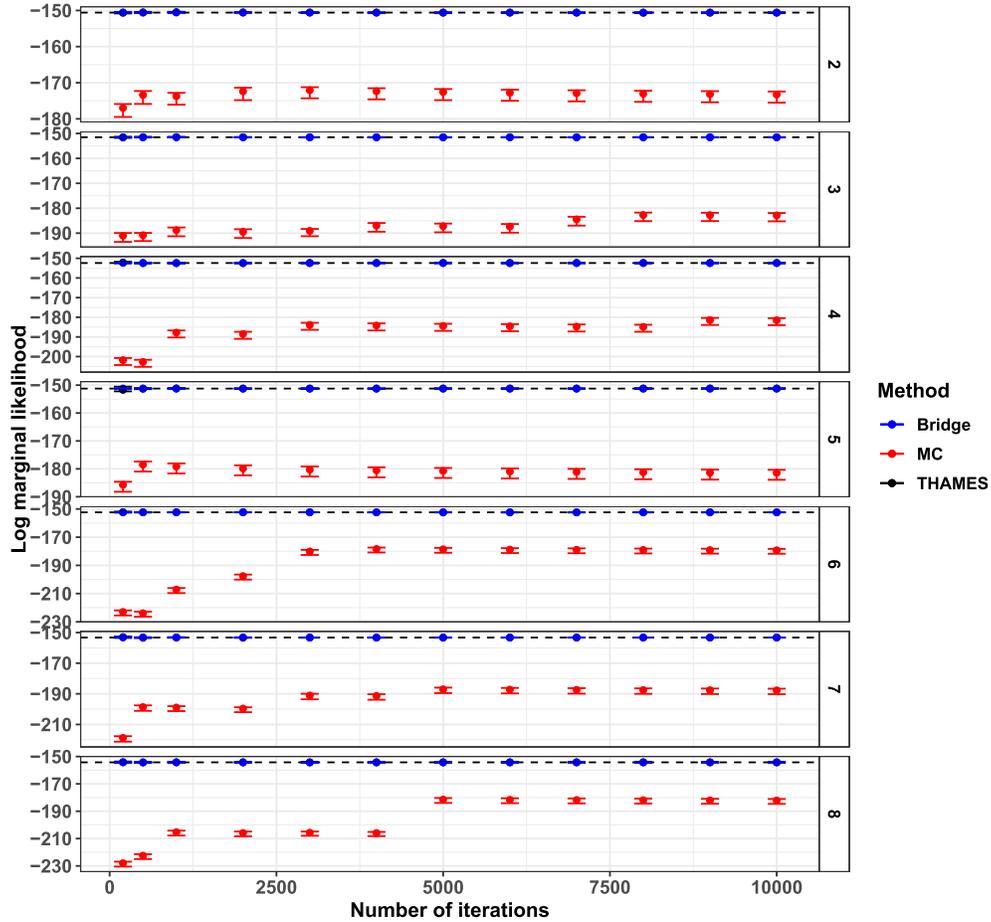


Figure 4: Log marginal likelihood (dotted line) and its estimators (dots) for the prostate data set. The approximate confidence intervals of the estimators are also indicated. Bridge sampling and THAMES are on point, while the simple Monte Carlo estimator does not seem to have converged.

Sample splitting was used and there was no noticeable bias in the results, even though we did not correct the THAMES for the bounded parameter σ^2 due to the fact that the posterior mode of this parameter is far away from 0. We also calculated the bias correction from Remark 2 which had no impact numerically, even for a posterior sample size as small as $T = 50$.

While the simple Monte Carlo estimator did not converge, the bridge sampling estimator and the THAMES behaved very similarly. Indeed, it can be seen that both these estimators converged rapidly to the correct value and that the intervals covered the correct values in most cases, even when the number of samples used was small, for

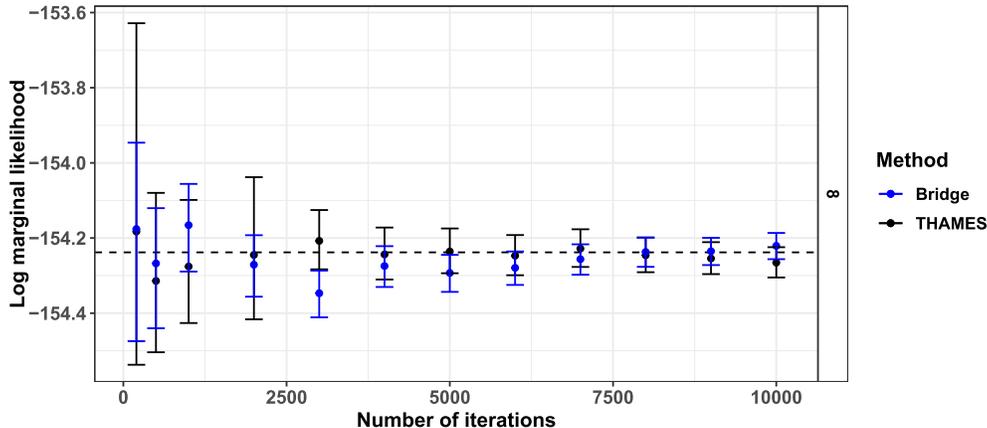


Figure 5: Log marginal likelihood (dotted line) and its estimators (dots) for the prostate data set with the full model (i.e., all eight clinical measures are selected). The approximate confidence intervals of the estimators are also indicated. The confidence intervals of the THAMES are more conservative, while the ones obtained from bridge sampling are more narrow.

all models investigated. Figure 5 shows the estimates produced by these methods when zoomed in on a finer scale in the full model, meaning that the number of clinical measures is equal to 8. Notably, the confidence intervals obtained by the THAMES were more conservative and wider than those obtained for the bridge sampling estimator, while the estimators themselves converged in a similar speed and manner.

For all models, those two estimators are particularly precise for 1000 samples of the posterior only. While the main goal of this section is to illustrate the precision of our estimation strategy for a series of models, we can also report that the model with the highest marginal likelihood, among those considered for this data set, is Model M_2 . In other words, the variables `lcavol` and `lweight` are seen as key for the prediction of the level of prostate-specific antigen.

4.3 Dirichlet-multinomial model

Extensions of the Dirichlet-multinomial model are widely used in the context of topic modelling, see, e.g., Blei et al. (2003). The expression for the marginal likelihood in this model is known, as in the previous two sections. This allows us to assess the performance of our estimator in another simulation study, in a non-Gaussian context.

A simulation study in this setting is useful for two reasons: First, this is a high-dimensional setting in which the posterior distribution of the parameters is highly non-Gaussian. In fact, the parameter space is bounded. This allows us to assess how well the THAMES performs in a very different setting, and also how much of an impact the correction for a bounded parameter space from Section 3.3 has. We check this

numerically in Supplement B (Metodiev et al., 2024b).

Second, there do exist similar models to this one for which the marginal likelihood is not tractable, e.g. Blei and Lafferty (2007). These models are therefore a possible application of the THAMES. The simulation study might give an idea of how well the THAMES would perform in these applications.

The Dirichlet-multinomial model is defined as follows: Each data point $Y_i \in \{0, \dots, l\}^K$ is drawn from a multinomial distribution given a Dirichlet-distributed random variable μ :

$$\mu \sim \text{Dirichlet}(\mu; (a_0, \dots, a_0)), \quad Y_i | \mu \stackrel{\text{i.i.d.}}{\sim} \mathcal{M}(l, \mu), \quad \forall i = 1, \dots, n.$$

Here, μ is positive and K -dimensional with components summing to 1. The covariance matrix of μ is thus necessarily singular. As noted in Remark 6, the THAMES needs to be used on posterior simulations from the subspace of \mathbb{R}^K on which a density is defined. In this case, this is $\mathbb{R}^{K-1} =: \mathbb{R}^d$. The prior density is thus

$$\pi(\mu_1, \dots, \mu_d) = \text{Dirichlet} \left(\mu_1, \dots, \mu_d, 1 - \sum_{j=1}^d \mu_j; (a_0, \dots, a_0) \right).$$

The posterior support is $\{\mu \in \mathbb{R}^d \mid \sum_{j=1}^d \mu_j \leq 1, \mu_1, \dots, \mu_d > 0\}$. The posterior distribution given the data $\mathcal{D} = \{y_1, \dots, y_n\}$ is tractable:

$$p(\mu_1, \dots, \mu_d | \mathcal{D}) = \text{Dirichlet} \left(\mu_1, \dots, \mu_d, 1 - \sum_{j=1}^d \mu_j; \alpha_1, \dots, \alpha_K \right),$$

with $\alpha_j = a_0 + \sum_{i=1}^n y_{ij}$. The marginal likelihood is thus also tractable, using Bayes' theorem.

Results

The marginal likelihood was estimated in the setting $(n, l, T, a_0) = (400, 150, 10000, 1)$ with d varying between 1, 20, 50 and 100. The quantities n and l were intentionally chosen to be large, since this model has very high-dimensional applications. For example, Blei et al. (2003) used a data set with 8000 documents, $n = 15,818$ words and used up to $K = 200$ different topics.

As mentioned, an alternative to the correction proposed in Section 3.3 is to reparametrize μ such that the support of the parameter is unconstrained. We did this by setting

$$(\mu_1^{(t)}, \dots, \mu_d^{(t)}) =: \frac{(\exp(\vartheta_1^{(t)}), \dots, \exp(\vartheta_d^{(t)}))}{\exp(-\sum_{k=1}^d \vartheta_k^{(t)}) + \sum_{k=1}^d \exp(\vartheta_k^{(t)})} = \text{softmax}(\vartheta^{(t)}), \quad t = 1, \dots, T.$$

We are using a bijective version of the softmax function (we take the first d elements of the version of the softmax which maps a sample from the Dirichlet to a parameter that

	d=1			d=20		
	MAE	SD	Time	MAE	SD	Time
Bridge	0.0001	0.0002	6.0026	0.0019	0.0024	36.9822
MC	0.093	0.1145	0.0020	6903.4	942.444	0.0008
THAMES	0.0064	0.0087	0.0158	0.0197	0.025	0.2486
	d=50			d=100		
	MAE	SD	Time	MAE	SD	Time
Bridge	0.0037	0.0046	58.9884	0.0086	0.0108	116.5902
MC	13879.1	1231.7635	0.0004	18418.1	844.7528	0.0046
THAMES	0.0315	0.039	0.3094	0.0473	0.0617	1.0532

Table 1: Average CPU times (in seconds per 10,000 posterior draws) as well as mean absolute errors and standard deviations for bridge sampling, the THAMES and the naive Monte Carlo (MC) estimator (errors of the latter were rounded to 1 decimal place). Estimates using MC are quickest to compute, but also the least precise. The THAMES is much faster than bridge sampling, although point estimates from the latter are more accurate.

sums to 0), and the induced prior $\pi_2(\vartheta) := \pi(\text{softmax}(\vartheta))|\text{softmax}_{\text{Jacobian}}(\vartheta)|$. We stress that this procedure is not necessary to calculate the THAMES, since the THAMES can be calculated on any parameter space. It is however necessary to calculate the bridge sampling estimator implemented in Gronau et al. (2020).

Table 1 shows the results when calculating the THAMES and the bridge sampling estimator on $(\vartheta^{(1)}, \dots, \vartheta^{(T)})$.² A fixed parameter μ was set to $\mu = (1/K, \dots, 1/K)$ and 50 different samples were generated using the parameters l and μ . The MC estimator was also computed for comparison.

Both bridge sampling and the THAMES outperformed the MC estimator. Additionally, while the bridge sampling estimator performed better in terms of mean absolute error, it should be noted that the THAMES is not only easier, but quicker to compute, with their differences in computation time growing as the dimension of the parameter space increases. This is likely due to the fact that the THAMES does not require additional evaluations of the likelihood, beyond the precomputed likelihood values of the posterior sample, so its computation time grows much more slowly with increasing d . Meanwhile bridge sampling does require additional evaluations, which take up an increasing amount of computation time. The average computation time for the bridge sampling estimator is about 361 times as high for $d = 1$, and 118 times as high for $d = 100$. However, we would like to emphasize that, in our opinion, the real strength of our estimator lies in the fact that it is not only quick, but also easy to implement.

²Computations were performed on an Intel(R) Core(TM) i7-7700HQ CPU at 2.80GHz with 16 GB RAM.

4.4 Mixed effects model

Netherlands schools data

To demonstrate the performance of THAMES on a random effects model, we consider the Netherlands (NL) schools dataset of Snijders and Bosker (1999). For our purposes, the data consist of language test scores of 2,287 eighth-grade pupils from 133 classes (in 131 schools) in the Netherlands. We denote by $y_{ij} \in \mathbb{R}$ the language test score of pupil i in class j , where $j \in \{1, \dots, J\}$ with $J = 133$ and $i \in \{1, \dots, n_j\}$ with n_j the size of class j . Let $n = \sum_{j=1}^J n_j = 2,287$ denote the full sample size.

We aim to determine if there is clustering of language test scores by class, with some classes performing significantly better than others on average. To do this, we fit both a simple mean model (which treats test scores of students in the same class as independent) and a random intercept model (which accounts for correlation of test scores within each class) to the data. The former (null) model H_0 posits that all classes perform the same, on average, while the latter (alternative) model H_1 allows for variation in performance at the class level. We estimate the log marginal likelihoods for the two models, Z_0 and Z_1 , respectively, using the THAMES. For comparison, we also compute estimates using bridge sampling (Gronau et al., 2020) and a simple Monte Carlo (MC) estimator that averages the likelihood against draws from the prior. With estimates of $\log(Z_0)$ and $\log(Z_1)$, we estimate the Bayes factor B_{01} to conduct a Bayesian hypothesis test of H_0 versus H_1 . Note that posterior simulation and marginal likelihood calculation are not analytically tractable for this model. As such, the use of approximate posterior sampling (e.g., via MCMC) and marginal likelihood estimation (e.g., via the THAMES) is required.

Linear model (LM)

We first consider a simple mean model (denoted LM), which posits that

$$y_{ij} = \mu + \varepsilon_{ij}, \quad j \in \{1, \dots, J\}, i \in \{1, \dots, n_j\}, \sum_j n_j = n,$$

$$\varepsilon_{ij} \stackrel{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2),$$

$$\mu \sim N(\hat{\mu}, \hat{\sigma}_\mu^2),$$

$$\sigma_\varepsilon^2 \sim \text{InverseGamma}(\hat{\nu}_\varepsilon, \hat{\beta}_\varepsilon).$$

The fixed hyperparameters $\hat{\mu}, \hat{\sigma}_\mu^2, \hat{\nu}_\varepsilon, \hat{\beta}_\varepsilon$ are specified so as to ensure that the prior distribution is dispersed relative to the likelihood, but on the same scale, as

$$\hat{\mu} = \text{mean}(y_{ij}) = 40.93, \quad \hat{\sigma}_\varepsilon^2 = \sqrt{2} \cdot \text{sd}(y_{ij}) = 12.73,$$

$$\hat{\nu}_\varepsilon = 0.5, \quad \hat{\beta}_\varepsilon = 0.5 \cdot \text{var}(y_{ij}) = 40.53.$$

The hyperparameters $(\hat{\nu}_\varepsilon, \hat{\beta}_\varepsilon)$ are chosen so that the prior mean of the precision $1/\sigma_\varepsilon^2$ equals $1/\text{var}(y_{ij})$. The set of parameters to be estimated in this model $(\mu, \sigma_\varepsilon^2)$ has dimension $d = 2$. As we are not using a conjugate prior for the linear model, the marginal likelihood does not admit an analytic expression in this case.

Full linear mixed model (full LMM)

We consider the random intercept model (denoted full LMM):

$$\begin{aligned}
y_{ij} &= \mu + \alpha_j + \varepsilon_{ij}, \\
\varepsilon_{ij} &\stackrel{\text{iid}}{\sim} N(0, \sigma_\varepsilon^2), \\
\alpha_j &\stackrel{\text{iid}}{\sim} N(0, \sigma_\alpha^2), \\
\mu &\sim N(\hat{\mu}, \hat{\sigma}_\mu^2), \\
\sigma_\varepsilon^2 &\sim \text{InverseGamma}(\hat{\nu}_\varepsilon, \hat{\beta}_\varepsilon), \\
\sigma_\alpha^2 &\sim \text{InverseGamma}(\hat{\nu}_\alpha, \hat{\beta}_\alpha).
\end{aligned}$$

Here $(\hat{\mu}, \hat{\sigma}_\mu^2, \hat{\nu}_\varepsilon, \hat{\beta}_\varepsilon)$ are as above and we specify $\hat{\nu}_\alpha = 0.5, \hat{\beta}_\alpha = 0.5 \cdot \text{var}(\hat{\mu}_j) = 13.77$, where $\hat{\mu}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$ is the sample mean for class $j \in \{1, \dots, J\}$. The hyperparameters $(\hat{\nu}_\alpha, \hat{\beta}_\alpha)$ are chosen so that the prior mean of the precision $1/\sigma_\alpha^2$ equals $1/\text{var}(\hat{\mu}_j)$. The set of parameters to be estimated in this model $(\mu, \sigma_\varepsilon^2, \sigma_\alpha^2, \alpha)$ has dimension $d = 136$.

Reduced linear mixed model (reduced LMM)

Note that the intercept parameters of the full LMM are not identifiable, as there is give-and-take between estimating the grand mean μ and the random intercepts α_j . By absorbing α_j into the error term structure ε_{ij} , we can specify an equivalent model (having the same marginal likelihood) with $d = 3$ identifiable parameters $(\mu, \sigma_\varepsilon^2, \sigma_\alpha^2)$. This amounts to marginalizing the α_j 's out of the model. The model (which we call reduced LMM) is given by

$$\begin{aligned}
y_{ij} &= \mu + \varepsilon_{ij}, \\
\varepsilon_{ij} &\sim N(0, \sigma_\varepsilon^2 + \sigma_\alpha^2), \\
\text{Cov}(\varepsilon_{ij}, \varepsilon_{i'j}) &= \sigma_\alpha^2, \quad i, i' \in \{1, \dots, n_j\}, i \neq i', \\
\text{Cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) &= 0, \quad j \neq j', \\
\mu &\sim N(\hat{\mu}, \hat{\sigma}_\mu^2), \\
\sigma_\varepsilon^2 &\sim \text{InverseGamma}(\hat{\nu}_\varepsilon, \hat{\beta}_\varepsilon), \\
\sigma_\alpha^2 &\sim \text{InverseGamma}(\hat{\nu}_\alpha, \hat{\beta}_\alpha).
\end{aligned}$$

Here $(\hat{\mu}, \hat{\sigma}_\mu^2, \hat{\nu}_\varepsilon, \hat{\beta}_\varepsilon, \hat{\nu}_\alpha, \hat{\beta}_\alpha)$ are as above.

Results

Figure 6 shows the log marginal likelihood of the NL schools data for each model computed using the THAMES, bridge sampling, and simple Monte Carlo estimators with approximate 95% confidence intervals as a function of the number of posterior MCMC or prior MC draws. Bridge sampling is a popular state-of-the-art method to

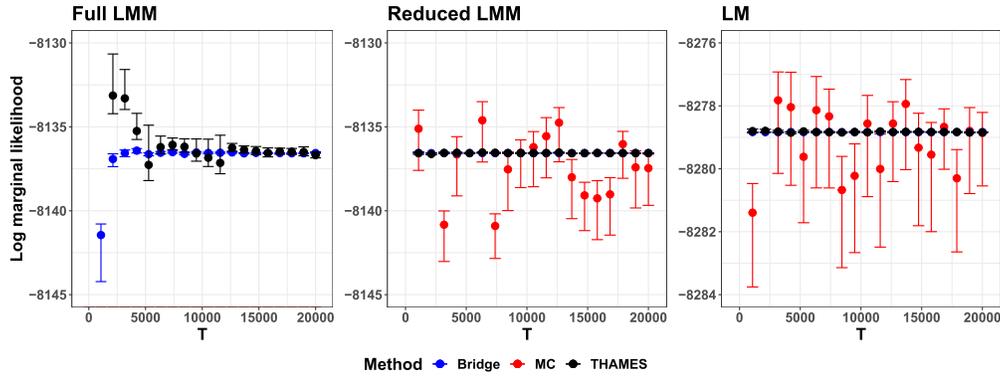


Figure 6: Log marginal likelihood estimates for models fitted to the NL schools data.

estimate log marginal likelihoods from posterior MCMC samples, which is substantially more complicated computationally than the THAMES. Posterior MCMC sampling is carried out in R using Stan (R Core Team, 2023; Stan Development Team, 2022). We use values of the sample size T evenly spaced between 1,000 and 20,000. For each T , we run 4 chains in parallel for $T/2$ iterations and remove the first $T/4$ as burn-in, yielding $T/4$ MCMC samples from each of the 4 chains, which are used to compute the THAMES and bridge sampling estimates.

THAMES provides consistent estimates of the log marginal likelihood with greater precision as the posterior sample size grows. As we would expect, THAMES converges much faster for the LM (with $d = 2$) and the reduced LMM (with $d = 3$) than for the full LMM (with $d = 136$), although the estimates of the reduced and full LMM converge to the same value. While the posterior support of this model is constrained due to the variance parameters ($\sigma_\varepsilon^2, \sigma_\alpha^2$), we found that the truncation correction defined in Section 3.3 had no impact on the results. For a given posterior sample size, we find that bridge sampling generally produces more precise estimates than the THAMES. However, the THAMES has the advantage of being much simpler to implement and more computationally efficient in practice. On average over the samples in Figure 6, bridge sampling required 6.4 times as much computation time as the THAMES for the full LMM, 556.5 times as much for the reduced LMM, and 26.8 times as much for the LM when estimated from the same number of posterior draws, as reported in Table 2.³ The MC estimator, while fast and theoretically unbiased and consistent, suffers from substantial variance. In the left panel of Figure 6, the MC estimates are not shown as they lie outside the range of the plot.

Using the THAMES estimates of the log marginal likelihoods for the LM ($\log(Z_0)$) and the (reduced) LMM ($\log(Z_1)$) with 20,000 posterior draws, the log Bayes factor ($\log(B_{01})$) is estimated as

$$\log(B_{01}) = \log(Z_0) - \log(Z_1) = -8278.842 + 8136.561 = -142.281,$$

³Computations were performed on an Apple M1 chip with 3.20GHz processor and 16 GB RAM.

	Full LMM	Reduced LMM	LM
Bridge	0.3815	1.6482	0.0723
MC	0.0002	0.0002	0.0002
THAMES	0.0581	0.0030	0.0027

Table 2: Average CPU times (in seconds per 1,000 posterior draws) to produce the estimates in Figure 6. The THAMES is faster than bridge sampling. Both take more time for the same number of iterations than the naive Monte Carlo (MC) estimator in terms of CPU time, even though the latter is far less precise (see Figure 6).

indicating decisive evidence in favor of the random intercept model (Kass and Raftery, 1995).

5 Discussion

We have proposed an estimator of the reciprocal of the marginal likelihood, called the THAMES, which is simple to compute, unbiased, consistent, has finite variance and is asymptotically normal, with available confidence intervals. It is a version of reciprocal importance sampling. The estimator has one user-specified control parameter, and we have derived an optimal value for this in the situation where the posterior distribution is normal, which is of great interest because posterior distributions are asymptotically normal in many situations. We have carried out several numerical experiments in which the estimator performs well.

A similar proposal was made independently in McEwen et al. (2022) under the name “Learned harmonic mean estimator”, where a variety of different sample models were suggested to work in conjunction. One of these models, the “Hypersphere”, corresponds to the THAMES, the difference being that no theoretical results were given for the optimal control parameter, c . Instead, c was optimized numerically as the minimum of the second harmonic moment, via the Brent hybrid root-finding algorithm. In the only high-dimensional application, which was in fact a Gaussian posterior, c was not optimized and it was noted that “alternative more effective target models can be developed that better scale to higher-dimensional settings”. We believe that with the THAMES, using the suggested optimal controlling parameter, this is the case.

The THAMES relies on estimating the posterior covariance matrix and mean. In our experience it is important that the estimator chosen for the covariance matrix be accurate for estimating each matrix entry. Elementwise accuracy appears to be important because the covariance matrix is used to precisely define a quadratic inequality. For example, using a shrinkage estimator for the covariance matrix, which can produce large errors in a small proportion of its elements, has in our experience degraded the performance of the THAMES in some situations.

One possible alternative to covariance matrix estimation would be to select a minimum-volume covering ellipse which includes a certain percentage of those points of the posterior sample which have the largest value with respect to the (unnormalized) posterior density evaluated at those points. This would ensure that an HPD-region is

well approximated, independent of the underlying posterior distribution. Determining a minimum-volume covering ellipse given a set of points can be difficult computationally, but this problem has been addressed in the literature in different settings and could possibly be adapted to the THAMES.

Acknowledgments

The authors would like to thank the anonymous referees, an Associate Editor and the Editor for their constructive comments. The authors would further like to thank Christoph Richard from the Friedrich-Alexander-Universität Erlangen-Nürnberg for his helpful comments. Their comments dramatically improved the content and quality of this paper.

Funding

Irons's research was supported by a Shanahan Endowment Fellowship and a Eunice Kennedy Shriver National Institute of Child Health and Human Development training grant, T32 HD101442-01, to the Center for Studies in Demography & Ecology at the University of Washington. Raftery's research was supported by NIH grant R01 HD070936 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD), by the Fondation des Sciences Mathématiques de Paris (FSMP), and by Université Paris-Cité.

Supplementary Material

Supplement A: Proofs and Calculations (DOI: [10.1214/24-BA1422SUPPA](https://doi.org/10.1214/24-BA1422SUPPA); .pdf). We prove the analytic results from Section 3 and derive the exact expression of the posterior and marginal density used for the multinomial likelihood in Section 4.

Supplement B: Additional Simulations (DOI: [10.1214/24-BA1422SUPPB](https://doi.org/10.1214/24-BA1422SUPPB); .pdf). We give some additional, numeric results about the approximate behaviour of the THAMES in the normal case as well as the case where the posterior support is constrained.

References

- Blei, D. M. and Lafferty, J. D. (2007). “A correlated topic model of Science.” *Annals of Applied Statistics*, 1(1): 17–35. [MR2393839](#). doi: <https://doi.org/10.1214/07-AOAS114>. 19
- Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). “Latent Dirichlet Allocation.” *Journal of Machine Learning Research*, 3: 993–1022. 18, 19
- Chib, S. (1995). “Marginal likelihood from the Gibbs output.” *Journal of the American Statistical Association*, 90: 1313–1321. [MR1379473](#). 2
- DiCiccio, T. J., Kass, R. E., Raftery, A. E., and Wasserman, L. (1997). “Computing Bayes factors by combining simulation and asymptotic approximations.” *Journal of the American Statistical Association*, 92: 903–915. [MR1482122](#). doi: <https://doi.org/10.2307/2965554>. 3, 4

- Durmus, A., Moulines, E., and Pereyra, M. (2018). “Efficient Bayesian computation by proximal Markov chain Monte Carlo: when Langevin meets Moreau.” *SIAM Journal on Imaging Sciences*, 11: 473–506. MR3763089. doi: <https://doi.org/10.1137/16M1108340>. 4
- Fernández, C., Ley, E., and Steel, M. F. (2001). “Benchmark priors for Bayesian model averaging.” *Journal of Econometrics*, 100(2): 381–427. URL <https://www.sciencedirect.com/science/article/pii/S0304407600000762> MR1820410. doi: [https://doi.org/10.1016/S0304-4076\(00\)00076-2](https://doi.org/10.1016/S0304-4076(00)00076-2). 16
- Frühwirth-Schnatter, S. (2004). “Estimating marginal likelihoods for mixture and Markov switching models using bridge sampling techniques.” *Econometrics Journal*, 7: 143–167. MR2076630. doi: <https://doi.org/10.1111/j.1368-423X.2004.00125.x>. 5
- Gelfand, A. E. and Dey, D. K. (1994). “Bayesian model choice: asymptotics and exact calculations.” *Journal of the Royal Statistical Society: Series B (Methodological)*, 56: 501–514. MR1278223. 2
- Ghosal, S. (2000). “Asymptotic normality of posterior distributions for exponential families when the number of parameters tends to infinity.” *Journal of Multivariate Analysis*, 74: 49–68. MR1790613. doi: <https://doi.org/10.1006/jmva.1999.1874>. 6
- Gronau, Q. F., Singmann, H., and Wagenmakers, E.-J. (2020). “bridgesampling: An R Package for Estimating Normalizing Constants.” *Journal of Statistical Software*, 92(10): 1–29. 11, 12, 15, 20, 21
- Hajargasht, G. and Woźniak, T. (2018). “Accurate Computation of Marginal Data Densities Using Variational Bayes.” *arXiv: Applications*. <https://arxiv.org/pdf/1805.10036.pdf>. 12
- Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2nd edition. MR2722294. doi: <https://doi.org/10.1007/978-0-387-84858-7>. 16
- Heyde, C. C. and Johnstone, I. M. (1979). “On asymptotic posterior normality for stochastic processes.” *Journal of the Royal Statistical Society: Series B (Methodological)*, 41: 184–189. MR0547243. 6
- Hoff, P. (2009). *A First Course in Bayesian Statistical Methods*. Springer Texts in Statistics. Springer New York. URL <https://books.google.de/books?id=DykcMwEACAAJ> MR2648134. doi: <https://doi.org/10.1007/978-0-387-92407-6>. 16
- Irons, N. J., Perrot-Dockès, M., and Metodiev, M. (2023). “thames: Easily Computed Marginal Likelihoods from Posterior Simulation Using the THAMES Estimator.” R package version 0.1.0. 3
- Jeffreys, H. (1961). *Theory of Probability*. Oxford, U. K.: Oxford University Press, 3rd edition. MR0187257. 2
- Kass, R. E. and Raftery, A. E. (1995). “Bayes factors.” *Journal of the American Sta-*

- tistical Association*, 90(430): 773–795. MR3363402. doi: <https://doi.org/10.1080/01621459.1995.10476572>. 24
- Liu, B. (2014). “Adaptive annealed importance sampling for multimodal posterior exploration and model selection with application to extrasolar planet detection.” *The Astrophysical Journal Supplement Series*, 213(1): 14. 2
- Llorente, F., Martino, L., Delgado, D., and Lopez-Santiago, J. (2023). “Marginal likelihood computation for model selection and hypothesis testing: An extensive review.” *SIAM Review*, 65: 3–58. MR4545927. doi: <https://doi.org/10.1137/20M1310849>. 2
- McEwen, J. D., Wallis, C. G. R., Price, M. A., and Docherty, M. M. (2022). “Machine learning assisted Bayesian model comparison: learnt harmonic mean estimator.” *arXiv*. <https://arxiv.org/pdf/2111.12720.pdf>. 24
- Meng, X.-L. and Wong, W. H. (1996). “Simulating ratios of normalizing constants via a simple identity: A theoretical exploration.” *Statistica Sinica*, 6: 831–860. MR1422406. 2, 12
- Metodiev, M., Perrot-Dockès, M., Ouadah, S., Irons, N. J., Latouche, P., and Raftery, A. E. (2024a). “Supplement A to “Easily Computed Marginal Likelihoods from Posterior Simulation Using the THAMES Estimator”” doi: <https://doi.org/10.1214/24-BA1422SUPPA>. 6, 13
- Metodiev, M., Perrot-Dockès, M., Ouadah, S., Irons, N. J., Latouche, P., and Raftery, A. E. (2024b). “Supplement B to “Easily Computed Marginal Likelihoods from Posterior Simulation Using the THAMES Estimator”” doi: <https://doi.org/10.1214/24-BA1422SUPPB>. 6, 7, 12, 19
- Miller, J. W. (2021). “Asymptotic normality, concentration, and coverage of generalized posteriors.” *The Journal of Machine Learning Research*, 22: 7598–7650. MR4318524. 6, 8
- Newton, M. A. and Raftery, A. E. (1994). “Approximate Bayesian inference with the weighted likelihood bootstrap.” *Journal of the Royal Statistical Society: Series B (Methodological)*, 56: 3–26. MR1257793. 2, 5, 9
- Porwal, A. and Raftery, A. E. (2022). “Comparing methods for statistical inference with model uncertainty.” *Proceedings of the National Academy of Sciences*, 119(16): e2120737119. URL <https://www.pnas.org/doi/abs/10.1073/pnas.2120737119> 16
- R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL <https://www.R-project.org/> 23
- Robert, C. P. and Wraith, D. (2009). “Computational methods for Bayesian model choice.” In *AIP conference proceedings*, volume 1193, 251–262. American Institute of Physics. 4
- Shen, X. (2002). “Asymptotic Normality of Semiparametric and Nonparametric Pos-

- terior Distributions.” *Journal of the American Statistical Association*, 97: 222–235. MR1947282. doi: <https://doi.org/10.1198/016214502753479365>. 6
- Sims, C. A., Waggoner, D. F., and Zha, T. (2008). “Methods for inference in large multiple-equation Markov-switching models.” *Journal of Econometrics*, 146(2): 255–274. MR2465172. doi: <https://doi.org/10.1016/j.jeconom.2008.08.023>. 12
- Skilling, J. (2006). “Nested sampling for general Bayesian computation.” *Bayesian Analysis*, 1: 833–859. MR2282208. doi: <https://doi.org/10.1214/06-BA127>. 2
- Snijders, T. A. B. and Bosker, R. J. (1999). *Multilevel Analysis. An Introduction to Basic and Advanced Multilevel Modelling*. Sage. MR3137621. 21
- Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E., and Yang, N. (1989). “Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II radical prostatectomy treated patients.” *Journal of Urology*, 16: 1076–1083. 16
- Stan Development Team (2022). “RStan: the R interface to Stan.” R package version 2.21.7. URL <https://mc-stan.org/> 23
- Sweeting, T. (1996). “On a Converse to Scheffe’s Theorem.” *The Annals of Statistics*, 14: 1252–1256. MR0856821. doi: <https://doi.org/10.1214/aos/1176350065>. 8
- Zellner, A. (1971). *An Introduction to Bayesian Inference in Econometrics*. Krieger. URL <https://books.google.de/books?id=paqiswEACAAJ> MR0433791. 16
- Zellner, A. (1986). “Bayesian Estimation and Prediction Using Asymmetric Loss Functions.” *Journal of the American Statistical Association*, 81(394): 446–451. URL <http://www.jstor.org/stable/2289234> MR0845882. 16