A Hybrid Approximation to the Marginal Likelihood

Eric Chuu
Texas A&M University

Debdeep Pati
Texas A&M University

Anirban Bhattacharya
Texas A&M University

Abstract

Computing the marginal likelihood or evidence is one of the core challenges in Bayesian analysis. While there are many established methods for estimating this quantity, they predominantly rely on using a large number of posterior samples obtained from a Markov Chain Monte Carlo (MCMC) algorithm. As the dimension of the parameter space increases, however, many of these methods become prohibitively slow and potentially inaccurate. In this paper, we propose a novel method in which we use the MCMC samples to learn a high probability partition of the parameter space and then form a deterministic approximation over each of these partition sets. This two-step procedure, which constitutes both a probabilistic and a deterministic component, is termed a Hybrid approximation to the marginal likelihood. We demonstrate its versatility in a plethora of examples with varying dimension and sample size, and we also highlight the Hybrid approximation's effectiveness in situations where there is either a limited number or only approximate MCMC samples available.

1 INTRODUCTION

Model selection and model averaging are among the most important inferential goals in Bayesian statistics. These goals inherently rely on evaluating model uncertainty, which in turn comes down to calculating the marginal likelihood of competing models. This makes accurate and efficient computation of the marginal likelihood an important problem.

Suppose we observe data y with a likelihood function $p(y \mid u)$ indexed by u coming from some parameter space $\mathcal{U}$. Provided that the prior distribution over the unknowns is specified, the marginal likelihood or evidence can be written as

$$p(y) = \int_{\mathcal{U}} p(y \mid u)\, p(u)\, du. \tag{1}$$

Barring specific conjugate settings, the marginal likelihood is analytically intractable in practice and poses a computationally challenging problem. Since numerical integration becomes infeasible beyond moderate dimension, Monte Carlo approximations present an alternative solution. In much of the literature devoted to estimating this quantity, the recurring idea is to form an asymptotically unbiased approximation of (1) using MCMC samples. However, running an MCMC algorithm in addition to forming a Monte Carlo approximation can quickly accrue error as the dimension of the parameter space increases. Consequently, these algorithms may require an exceedingly large number of samples in order to form accurate estimates. In many problems, however, obtaining MCMC samples is time-consuming and potentially unreliable. As such, the need for an approximation that does not too heavily rely on both the quantity and quality of the MCMC samples is evident.

Some commonly used algorithms include Laplace's method (Tierney and Kadane, 1986), which assumes that the posterior distribution can be approximated with a normal distribution, and the Harmonic Mean estimator (Newton and Raftery, 1994), which is easy to implement, but has been shown to be unstable and can have infinite variance (Newton and Raftery, 1994), (Raftery et al., 2007). On a similar vein, the Adjusted Harmonic Mean estimator (Lenk, 2009) and the Corrected Arithmetic Mean estimator (Pajor, 2017) leverage the harmonic mean and arithmetic mean identities, respectively, with the idea of sampling from high posterior probability regions of the parameter space to improve upon the original estimators. Annealed Importance Sampling (Neal, 2001) uses a dynamic importance sampling function that sequentially transitions through intermediate distributions to the target distribution.

Proceedings of the 24$^{\text{th}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

Other popular algorithms include Chib's method (Chib, 1995; Chib and Jeliazkov, 2001), Bridge Sampling (Meng and Wong, 1996), Warp Bridge Sampling (Meng and Schilling, 2002), and Nested Sampling (Skilling, 2006). See Friel and Wyse (2012) for a more comprehensive overview and discussion of these algorithms. There have also been recent developments in variational inference techniques that provide alternative ways to approximate and bound the marginal likelihood (Rezende and Mohamed, 2015; Salimans et al., 2015).

In contrast to these methods, we propose a novel approach which can be thought of as a hybrid between probabilistic and deterministic procedures. A high level view of our method can be broken down into two major steps: (i) the MCMC samples are used to learn a partition of the parameter space $\mathcal{U}$, and (ii) with this partition, we then make a deterministic approximation to the log posterior on each of the partition sets. In essence, we seek to exploit the assumption that the posterior distribution will be far from a uniform looking distribution and instead show concentration around some parameter. If the partition obtained from the MCMC samples can identify areas of high posterior mass by carving up these regions more finely, then we are better equipped to make an accurate approximation to the log posterior over each of these regions. Given the use of a probabilistic procedure in step (i), coupled with a deterministic calculation in step (ii), we refer to the resulting approximation to the marginal likelihood as the *Hybrid estimator*.

Our contribution fundamentally provides a way to bypass the need for a large number of posterior samples for accurate computation of the marginal likelihood. In many applications, evaluating the likelihood can be extremely time consuming, so in turn, collecting lots of posterior samples in such cases is prohibitively expensive in both time and computation. The typical guarantees for MCMC-based estimates of the marginal likelihood are asymptotic in the number of posterior samples. Our approach instead only uses the MCMC samples to learn a skeleton of the posterior distribution, which then simplifies the subsequent calculation; hence, the Hybrid estimator establishes a scalable framework for computing the evidence in high dimensional problems.

The paper is organized as follows. In Section 2, we motivate each step in the algorithm and provide a formal statement of the Hybrid approximation scheme. In Section 3, we demonstrate the performance of the Hybrid estimator in a variety of simulation studies and compare the results with some of the aforementioned estimators. Finally, in Section 4, we briefly discuss some details for extensions and future work.

2 METHODOLOGY

We introduce some preliminary notation. Let γ be a probability density with respect to the Lebesgue measure on $\mathbb{R}^d$ given by

$$\gamma(u) = \frac{e^{-\Phi(u)}\,\pi(u)}{\mathcal{Z}}, \quad u \in \mathcal{U} \subseteq \mathbb{R}^d.$$

When $\Phi(\cdot)$ corresponds to a negative log-likelihood function and $\pi(\cdot)$ a prior distribution, $\gamma(\cdot)$ is the corresponding posterior distribution, although such an interpretation is not necessary for our approach. Then, the marginal likelihood has the following form,

$$\mathcal{Z} = \int_{\mathcal{U}} e^{-\Psi(u)}\,du, \tag{2}$$

where $\Psi(u) = \Phi(u) + (-\log \pi(u))$ is the negative log-posterior. As stated before, while we can evaluate Ψ, we are unable to compute the integral in (2). We can address this problem using the two sub-routines mentioned in the previous section. First, we find a partition of the parameter space that gives more attention to (i.e, more finely partitions) regions of the posterior that have high posterior mass. Next, we propose a suitable approximation for Ψ that allows for easier evaluation of the integral over each of the partition sets learned from the previous step. These steps used in conjunction with each other give us a way to approximate $\mathcal{Z}$ by computing a simplified version of the integral over partition sets of the parameter space that have ideally taken into account the assumed non-uniform nature of the posterior distribution.

2.1 Deterministic Approximation

We first elaborate on our strategy to replace Ψ with an approximation $\widehat{\Psi}$. Our starting point is the following observation: fix $q \in (0,1)$ small and let $A \subseteq \mathcal{U}$ be a compact subset with $\gamma(A) \geqslant (1-q)$. Rearranging this equation, one obtains $(1-q) \leqslant \gamma(A) = \mathcal{Z}^{-1}\int_A e^{-\Psi(u)}du \leqslant 1$, leading to the two-sided bound

$$\int_A e^{-\Psi(u)}\,du \leqslant \mathcal{Z} \leqslant \frac{1}{1-q}\int_A e^{-\Psi(u)}\,du. \tag{3}$$

We then make the following approximation

$$\log \mathcal{Z} \approx F_A := \log\left[\int_A e^{-\Psi(u)}\,du\right]. \tag{4}$$

From Eq. (3), it is immediate that $|\log \mathcal{Z} - F_A| \leqslant \log\{1/(1-q)\} \approx q$ for q small. Henceforth, we aim to estimate the quantity F_A. This initial approximation step can be thought of as compactifying the parameter space to reduce its entropy. Even if $\mathcal{U}$ itself is compact, γ can be highly concentrated in a

region A with $\text{vol}(A) \ll \text{vol}(\mathcal{U})$, particularly when the posterior exhibits concentration (Ghosal and Van Der Vaart, 2007), and it is judicious to eliminate such low posterior-probability regions.

Having compactified the integral domain, our general plan is to replace Ψ with a suitable approximation $\widehat{\Psi}$ on the compact set A. In this article, we specifically focus on a piecewise constant approximation of the form

$$\widehat{\Psi}(u) = \sum_{k=1}^{K} c_k^\star \cdot \mathbb{1}_{A_k}(u), \tag{5}$$

where $\mathcal{A} = \{A_1, \ldots, A_K\}$ is a partition of A, i.e., $A = \bigcup_{k=1}^K A_k$ and $A_k \cap A_{k'} = \varnothing$ for all $k \neq k'$, and $c_k^\star$ is a representative value of Ψ within the partition set A_k. To simplify the ensuing calculations, we further restrict ourselves to dyadic partitions in this article so that each of the partition sets is rectangular, $A_k = \prod_{l=1}^d [a_k^{(l)}, b_k^{(l)}]$. This leads to the approximation

$$\int_A e^{-\Psi(u)}\, du \approx \int_A e^{-\widehat{\Psi}(u)}\, du = \sum_{k=1}^{K} e^{-c_k^\star} \cdot \mu(A_k), \tag{6}$$

where $\mu(B) = \int_B 1\, du$ denotes the d-dimensional volume of a set B. We eventually define

$$\widehat{F}_A := \log\Big[\int_A e^{-\widehat{\Psi}(u)}\, du\Big] = \log\Big[\sum_{k=1}^{K} e^{-c_k^\star} \cdot \mu(A_k)\Big] \tag{7}$$

to be our estimator of F_A, and hence of $\log \mathcal{Z}$. The choice of the piecewise constant approximation is motivated both by its approximation capabilities (Binev et al., 2005) as well as the analytic tractability of the approximating integral in Eq. (7). We remark here that the integral remains tractable if a piecewise linear approximation is employed, suggesting a natural generalization of our estimator.

Since F_A is a non-linear functional of Ψ, it is reasonable to question the validity of the approximation in Eq. (6), or equivalently, the approximation of F_A with $\widehat{F}_A$ — even if $\widehat{\Psi}$ is a good approximation to Ψ, it is not immediately clear if the same should be true of $\widehat{F}_A$. Using an interpolation trick, we show below that the approximation error $|\widehat{F}_A - F_A|$ can be bounded in terms of a specific distance between $\widehat{\Psi}$ and Ψ. Define

$$F(t) = \log\Big[\int_A e^{-\left(t\Psi(u) + (1-t)\widehat{\Psi}(u)\right)}\, du\Big], \quad t \in [0,1].$$

Clearly, $F(0) = \widehat{F}_A$ and $F(1) = F_A$, so that

$$F_A - \widehat{F}_A = F(1) - F(0) = \int_0^1 F'(t) dt.$$

Computing F', we get

$$\begin{aligned} F'(t) &= \frac{-\int_A \left(\Psi(u) - \widehat{\Psi}(u)\right) e^{-\left(t\Psi(u) + (1-t)\widehat{\Psi}(u)\right)}\, du}{\int_A e^{-\left(t\Psi(u) + (1-t)\widehat{\Psi}(u)\right)}\, du} \\ &= -\mathbb{E}_{U \sim \pi_t}\big(\Psi(U) - \widehat{\Psi}(U)\big), \end{aligned}$$

where π_t is the probability density on A given by

$$\pi_t(u) \propto e^{-\left(t\Psi(u) + (1-t)\widehat{\Psi}(u)\right)}, \quad u \in A.$$

Using the integral representation, we can now bound the approximation error,

$$|F_A - \widehat{F}_A| \leqslant \sup_{t \in [0,1]} \big|\mathbb{E}_{U \sim \pi_t}\big(\Psi(U) - \widehat{\Psi}(U)\big)\big|.$$

Interestingly, note that $\pi_1 \propto \gamma \mathbb{1}_A$ is our target density restricted to A, and $\pi_0(u) \propto e^{-\widehat{\Psi}(u)}\, \mathbb{1}_A(u)$ has normalizing constant $\widehat{F}_A$. The collection of densities $\{\pi_t\}$ can therefore be thought of as continuously interpolating between π_0 and π_1. Piecing together the various approximations, we arrive at the following result.

Proposition 1. *For any compact subset $A \subseteq \mathcal{U}$, we have*

$$|\widehat{F}_A - \log \mathcal{Z}| \leqslant \sup_{t \in [0,1]} \big|\mathbb{E}_{U \sim \pi_t}\big(\Psi(U) - \widehat{\Psi}(U)\big)\big| + \log\left(\frac{1}{\nu(A)}\right).$$

Here, ν denotes the Lebesgue measure on $\mathbb{R}^D$. The first term in the right hand side above can be further bounded by $\|\Psi - \widehat{\Psi}\|_\infty := \sup_{u \in A} |\Psi(u) - \widehat{\Psi}(u)|$. This conclusion is not restricted to the piecewise constant approximation and can be used for other approximations, such as the piecewise linear one.

2.2 High Probability Partitioning of the Parameter Space

Next, we address the task of obtaining a suitable partition of the parameter space. Clearly, traditional quadrature methods would render this method ineffective, requiring the number of function evaluations to grow exponentially with d. Furthermore, with a posterior distribution that exhibits any degree of concentration, there will indubitably be regions of $\mathcal{U}$ where the posterior probability is close to 0. From a computationally mindful standpoint, it makes sense to then focus on more finely partitioned regions of $\mathcal{U}$ that have high posterior probability. With this in mind, we turn to using samples from γ to obtain such a partition. Specifically, let $u_1, \ldots, u_J$ be approximate samples from γ, e.g., the output of an MCMC procedure. We treat $\{(u_j, \Psi(u_j)\}_{j=1}^J$ as covariate-response pairs and feed them to a tree-based model such as CART (Breiman, 1984), implemented in the `R` package `rpart`

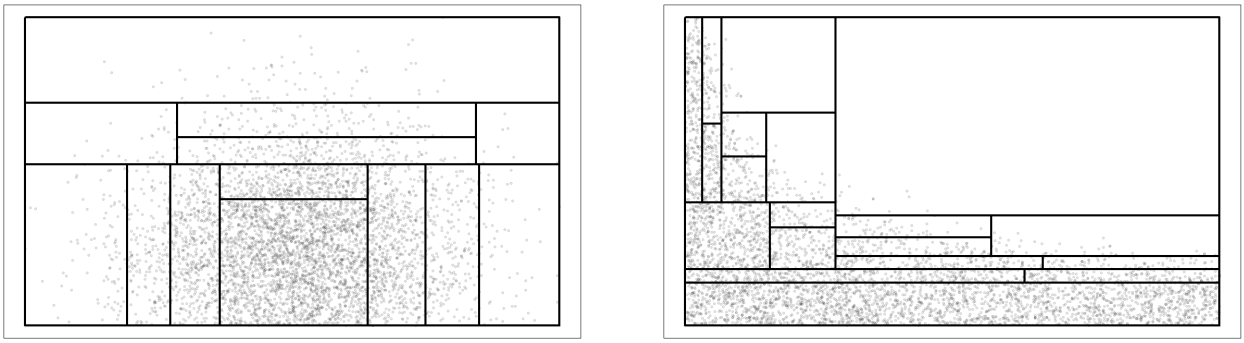

Figure 1: *Left: bivariate normal distribution truncated to the first orthant. Right: A density of the form* $\gamma(u) \propto \exp(-nu_1^2u_2^4)\pi(u)$*, where* $u \in [0,1]^2$ *and* $\pi(\cdot)$ *is the uniform measure on* $[0,1]^2$*. For this simulation,* $n = 1000$*. Both plots have 5000 MCMC samples with the partition returned from fitting a CART model.*

(Therneau and Atkinson, 2019), to obtain a dyadic partition. While the MCMC samples are typically used to construct Monte Carlo averages, we use them to construct a high probability partition of the parameter space. We assume the capability to evaluate Ψ, which is a very mild assumption since obtaining samples from γ using even a basic sampler like Metropolis–Hastings requires evaluating Ψ. Finally, the above procedure implicitly suggests the compactification A to be a bounding box using the range of posterior samples, $A = \otimes_l \left[\min\{u_j^{(l)}\}, \max\{u_j^{(l)}\}\right]$, $1 \leqslant j \leqslant J$, $1 \leqslant l \leqslant d$, where $u_j^{(l)}$ is the lth component of u_j.

2.3 Partitioning in Two Dimensions

Before moving into higher dimensions, we provide an illustration of the process described in the previous section in 2 dimensions, where the partitioning can be easily visualized. Suppose γ is a density on $\mathbb{R}^2$ supported on $\mathcal{U} \subseteq \mathbb{R}^2$, and $u_j \sim \gamma$ for $j = 1, \ldots, J$. Forming the pairs, $\{(u_j, \Psi(u_j))\}_{j=1}^J$, we then fit a CART model to these points and extract the decision rules, which form a dyadic partition of the aforementioned bounding box $A \subseteq \mathcal{U}$. Denote the partition as $\mathcal{A} = \{A_1, \ldots, A_K\}$. Plotting the sampled points and overlaying the partitions learned from the regression tree, we observe in Figure 1 that areas of $\mathcal{U}$ with a high concentration of points coincide with regions that are more finely partitioned by the regression tree. Taking γ to be a posterior distribution, we see that this behavior of partitioning areas of greater posterior mass is desirable in producing a better approximation. Equipped with the partition $\mathcal{A}$, we need only to determine the representative point of each partition set in order to form the approximation to Ψ.

Recall that CART fits a constant for each point within a given partition set. At any given stage, the CART model will search for the optimal predictor value, $u = (u_1, u_2)$, on which to partition the remaining points such that the sum of squares error (SSE) between the response, $\Psi(u)$, and the predicted constant is minimized. In particular, to partition data into two regions A_1 and A_2, the objective function is given as

$$SSE = \sum_{u_i \in A_1} (\Psi(u_i) - c_1)^2 + \sum_{u_i \in A_2} (\Psi(u_i) - c_2)^2. \quad (8)$$

Upon minimization of the SSE, the resulting partition sets A_1 and A_2 have fitted values c_1 and c_2, respectively. For each partition set $A_k \in \mathcal{A}$, a natural choice for the representative point $c_k^\star$ is the fitted value for A_k produced by the tree-fitting algorithm. Following this two-step process of using CART to obtain both the partition and the fitted values for each of the partition sets and then plugging these into Eq. (6), we obtain the Hybrid approximation to the marginal likelihood.

2.3.1 Conjugate Normal Model

We consider the following conjugate normal model: $y_{1:n} \mid \mu, \sigma^2 \sim \mathcal{N}(\mu, \sigma^2)$, $\mu \mid \sigma^2 \sim \mathcal{N}(m_0, \sigma^2/w_0)$, $\sigma^2 \sim \mathcal{IG}(r_0/2, s_0/2)$, where $\mathcal{IG}(\cdot, \cdot)$ denotes the inverse-gamma distribution. In order to compute the Hybrid estimator, we require samples from the posterior distribution and a way to evaluate Ψ. In this example, the posterior distribution of $u = (\mu, \sigma^2)$ is known, and since the likelihood and prior are specified, the evaluation of Ψ is straightforward. With this architecture in place, we feed the pairs, $\{(u_j), \Psi(u_j)\}_{j=1}^J$, through CART to obtain a partition over the parameter space and each partition set's representative point. Then, we use Eq. (6) to compute the final approximation.

Table 1 shows results for the Hybrid estimator and a number of other competing methods. Here, the

Table 1: *Normal Inverse-Gamma Example. We report the mean, standard deviation, average error (AE, truth - estimated), and the root mean squared error (RMSE), taken over 100 replications. Each replication has 50 observations and 1000 posterior samples. The true log marginal likelihood is -113.143. Estimators include the Harmonic Mean estimator (HME), Corrected Arithmetic Mean estimator (CAME), Bridge Sampling estimator (BSE), and the Hybrid estimator (HybE).*

Method	Mean	SD	AE	RMSE
log HME	-104.762	0.733	-8.381	8.431
log CAME	-112.704	0.048	-0.439	0.441
log BSE	-113.143	0.006	0	0.006
log HybE	-113.029	0.025	-0.114	0.117

true log marginal likelihood can be computed in closed form, so we have direct comparisons to the ground truth. All estimators except for the Harmonic Mean estimator give accurate approximations to the log marginal likelihood.

2.4 Algorithm Description

Until this point, the representative point within each partition has simply been the fitted value for each partition returned from the CART model. When $\{(u_j, \Psi(u_j))\}_{j=1}^J$ is fed into the tree, it attempts to optimize the sum of squared errors as in Eq. (8). Note, however, that our eventual objective is to best approximate the functional $\log \int_A e^{-\Psi}$, and it is not unreasonable to suspect that the optimal value for A_k chosen by the regression tree model may not be a suitable choice for our end goal, especially for higher dimensions. Simulations in higher dimensions indeed confirm this. Before suggesting a remedy, we offer some additional understanding into the approximation mechanism that guides us toward an improved choice. To that end, write $\widehat{F}_A$ from Eq. (7) as

$$\widehat{F}_A = \log\left[\sum_{k=1}^K e^{-c_k^\star}\, p_k\right] + \log \mu(A) := \widehat{G} + \log \mu(A),$$

where recall $\mu(B) = \mathrm{vol}(B)$ is the Lebesgue measure of a Borel set B, and we define $p_k := \mu(A_k)/\mu(A)$. We can also write $F_A = G + \log \mu(A)$, with

$$\begin{aligned} G &:= \log\left[\frac{1}{\mu(A)}\int_A e^{-\Psi(u)}du\right] \\ &= \log\left[\sum_{k=1}^K p_k \frac{1}{\mu(A_k)}\int_{A_k} e^{-\Psi(u)}du\right] \\ &= \log\left[\sum_{k=1}^K e^{-c_k}\, p_k\right], \end{aligned}$$

where

$$e^{-c_k} = \frac{1}{\mu(A_k)}\int_{A_k} e^{-\Psi(u)}du = \mathbb{E}_{U_k \sim \mathrm{Unif}(A_k)}\left[e^{-\Psi(U_k)}\right].$$

Thus, for $\widehat{G}$ to approximate G, we would ideally like to have each $c_k^\star$ chosen so that $e^{-c_k^\star}$ targets e^{-c_k}. Importantly, the above exercise suggests the appropriate scale to perform the approximation – rather than working in the linear scale as in Eq. (8), it is potentially advantageous to work in the exponential scale.

2.4.1 Choosing the Representative Point

Based on the above discussion, we define a family of objective functions

$$Q_k(c) = \sum_{u \in A_k} \frac{|e^{-\Psi(u)} - e^{-c}|}{e^{-\Psi(u)}}, \quad c \in A_k, \tag{9}$$

one for each partition set A_k returned by the tree, and set $c_k^\star = \mathrm{argmin}_c Q_k(c)$. We experimented with a number of different metrics before zeroing in on the above relative error criterion in the exponential scale. Minimizing (9) is a weighted ℓ_1 problem and admits a closed-form solution.

Thus, our overall algorithm can be summarized as follows. We obtain samples $u_1, \ldots, u_J$ from γ, and feed $\{(u_j, \Psi(u_j)\}_{j=1}^J$ through a tree to partition the bounding box A of the samples. Then, rather than using the default fitted values returned by the tree, we take the representative value $c_k^\star$ within each A_k as the minimizer of Q_k. These $c_k^\star$s are then used to compute $\widehat{F}_A$ as in (7) – note that $\widehat{F}_A$ can be stably computed using the log-sum-exp trick. Finally, we declare $\widehat{F}_A$ as $\log \widehat{\mathcal{Z}}$, our estimator of $\log \mathcal{Z}$.

Algorithm 1 Hybrid Approximation

Input: Sampler for γ, method for evaluating Ψ
Output: Estimate of the log marginal likelihood

1: Sample $\{u_j\}_{j=1}^J$ from γ
2: Fit $\{(u_j, \Psi(u_j))\}_{j=1}^J$ using a regression tree
3: From the fitted tree, extract the dyadic partition $\mathcal{A} = \{A_1, A_2, \ldots, A_K\}$ of the bounding box $A = \otimes_l \left[\min\{u_j^{(l)}\}, \max\{u_j^{(l)}\}\right]$ determined by the samples, with $A_k = \prod_{l=1}^d [a_k^{(l)}, b_k^{(l)}]$
4: **for** $k \in \{1, \ldots, K\}$ **do**
5: $\quad c_k^\star \leftarrow \mathrm{argmin}_{c \in A_k} \log Q_k(c)$
6: $\quad \widehat{\mathcal{Z}}_k \leftarrow e^{-c_k^\star} \prod_{l=1}^d \left(b_k^{(l)} - a_k^{(l)}\right)$
7: **end for**
8: Use the log-sum-exp trick to compute the final estimator, $\log \widehat{\mathcal{Z}} = \texttt{log-sum-exp}\left(\log \widehat{\mathcal{Z}}_1, \ldots, \log \widehat{\mathcal{Z}}_K\right)$

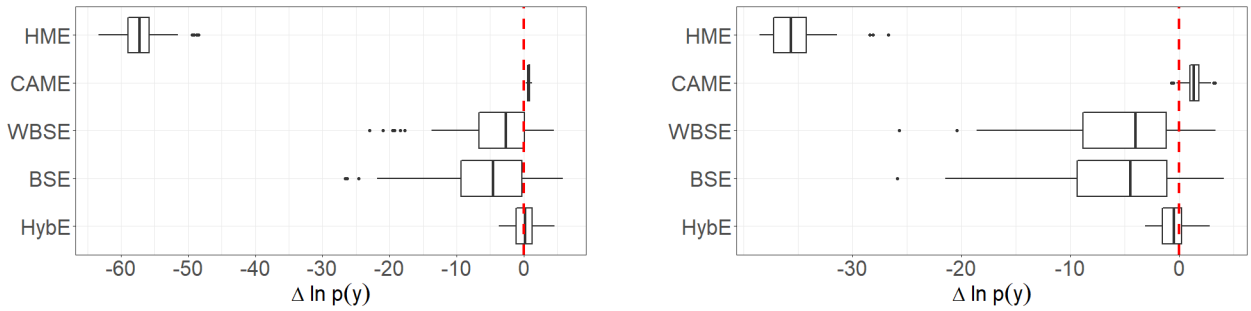

Figure 2: *Boxplots of the error (truth - estimate) for the log marginal likelihood in the MVN-IG (left, true* $\log p(y)$*: -303.8482) and truncated MVN (right, true* $\log p(y)$*: -250.2755) examples. Both examples correspond to 20-dimensional parameter spaces. Results are reported over 100 simulations, with 100 observations. Estimates are based on 45 MCMC samples. The results correspond to the natural logarithm of each of the estimators.*

3 RESULTS

In the following experiments, we present a variety of problem settings. First, we consider the linear regression model under different prior specifications where the true marginal likelihood is known, so we can easily verify the accuracy of any subsequent approximations. We then extend the application of the Hybrid estimator to examples for which the parameter is a $d \times d$ covariance matrix, thus showcasing its versatility even when the parameter space is non-Euclidean. We examine the performance of the Hybrid estimator alongside competing methods and focus primarily on situations where the posterior samples are either few in number or non-exact. In addition to the Hybrid estimator (HybE), we examine the following additional estimators: Bridge Sampling estimator (BSE), Warp Bridge Sampling estimator (WBSE), Harmonic Mean estimator (HME), and Corrected Arithmetic Mean estimator (CAME). The BSE and WBSE results are obtained using the `bridgesampling` package (Gronau et al., 2020). Corresponding calculations and formulae for posterior parameters and analytical marginal likelihoods are given in the Supplement.

We emphasize that in the experiments provided in this section we seek to mimic scenarios where posterior sampling is highly expensive and/or mixing is poor. By considering a small number of samples as the input for these marginal likelihood estimation algorithms, we provide a realistic scenario for the regime in which we wish to operate. In the examples in Sections 3.1 and 3.3, we sample directly from exact posterior distribution, while in Section 3.4, we use samples from an approximate posterior distribution.

3.1 Bayesian Linear Regression

Consider the following setup of the linear regression model, $y = X\beta + \varepsilon$, where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times d}$, $\beta \in \mathbb{R}^d$, and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. In the next two examples, we consider different prior distributions on β and σ^2.

3.1.1 Multivariate Normal Inverse-Gamma Model

We assume a multivariate normal inverse-gamma (MVN-IG) prior on (β, σ^2), where $\beta \mid \sigma^2 \sim \mathcal{N}_d(\mu_\beta, \sigma^2 V_\beta)$, $\sigma^2 \sim \mathcal{IG}(a_0, b_0)$. Given this choice of the prior, the posterior distribution is known to be $\beta \mid \sigma^2, y \sim \mathcal{N}\left(\mu_n, \sigma^2 V_n\right)$, $\sigma^2 \mid y \sim \mathcal{IG}\left(a_n, b_n\right)$. In our simulation, we take $d = 19$, so that $u = (\beta, \sigma^2) \in \mathbb{R}^{20}$. Since the log marginal likelihood in this example is well known, we evaluate each of the estimates against the true value. In Figure 2, we plot the errors for each of the estimators when only 45 MCMC samples are used for each approximation. The accuracy and standard error of the Hybrid estimator are clearly superior compared to the well-established estimators.

3.1.2 Truncated Multivariate Normal Model

Next, we place a multivariate normal prior on β truncated to the first orthant. In particular, $\beta \sim \mathcal{N}_d(0, \sigma^2\lambda^{-1} I_d) \cdot \mathbb{1}_{[0,\infty)^d}$, where σ^2, λ are known. This produces a posterior distribution of the form,

$$\beta \mid y \sim \mathcal{N}_d(\beta \mid Q^{-1}b, Q^{-1}) \cdot \mathbb{1}_{[0,\infty)^d},$$

where Q, b are defined in the Supplement. Then, the marginal likelihood can be written as

$$\begin{aligned} p(y) &= \int_R \mathcal{N}(y \mid X\beta, \sigma^2 I_n) 2^{-d} \mathcal{N}(\beta \mid 0, \sigma^2\lambda^{-1} I_d)\, d\beta \\ &= C \cdot \int_R \det(Q)^{\frac{1}{2}} e^{-\frac{1}{2}\left(\beta - Q^{-1}b\right)' Q \left(\beta - Q^{-1}b\right)} d\beta. \end{aligned}$$

Here, $R = [0, \infty)^\infty$ and C is a known constant term. Note that in this case, however, the integral is not analytically available and prevents the marginal likelihood

from being easily computed. Botev (2016) uses a minimax tilting method to calculate the normalizing constant of truncated normal distributions and shows that the proposed estimator has the vanishing relative error property (Kroese et al., 2011). In light of this, we accept Botev's estimator as the true marginal likelihood in the following experiments. The `TruncatedNormal` package (Botev and Belzile, 2019) provides samples from truncated normal distributions, so posterior samples from $\beta \mid y$ are readily available.

In Figure 2, we present the simulation results for the case when $d = 20$. Each approximation uses 45 MCMC samples, and we compare the results against the true log marginal likelihood. Once again, the HybE outperforms the other estimators and reinforces its ability to deal with a scarce number of samples. Provided with a sufficiently large number of samples, however, the BSE and CAME are both eventually able produce more accurate results than the HybE.

3.2 Unrestricted Covariance Matrices

The examples have thus far dealt with parameters in Euclidean space. In the next set of examples, we move beyond the usual Euclidean space and consider parameters in $\mathbb{R}^{d\times d}$. In particular, let $x_1, \ldots, x_n \overset{\text{iid}}{\sim} N_d(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{d\times d}$. Then the likelihood can be written as follows,

$$L(\Sigma) = (2\pi)^{-nd/2} \det(\Sigma)^{-n/2} e^{-\operatorname{tr}(\Sigma^{-1}S)/2}, \qquad (10)$$

where $S = \sum_{i=1}^{n} x_i x_i'$. For simplicity, we consider a conjugate inverse-Wishart (IW) prior, $\mathcal{W}^{-1}(\Lambda, \nu)$, for Σ, where Λ is positive definite $d \times d$ matrix and $\nu > d-1$ is the degrees of freedom. Consequently, the posterior distribution of Σ is $\mathcal{W}^{-1}(\Lambda + \mathrm{S}, \nu + n)$, and we can compute the marginal likelihood in closed form.

Note that despite being able to sample from the posterior distribution, we cannot yet carry out the Hybrid approximation algorithm. Since posterior samples are drawn from a sub-manifold of $\mathbb{R}^{d\times d}$, if we were to proceed as usual to obtain a partition over $\mathbb{R}^{d\times d}$, there would be no guarantee that a given point within the partition could be used to reconstruct a valid covariance matrix. As such, we circumvent this issue by taking the Cholesky factorization of Σ, so that $\Sigma = TT'$, where T is a lower triangular matrix with positive diagonal entries, t_{jj} for $j = 1, \ldots, d$.

Under this transformation, we can define $\Psi(T) = -\log L(T) - \log \pi(T)$, where Eq. (10) gives us

$$L(T) = (2\pi)^{-nd/2} \det(T)^{-n} e^{-\operatorname{tr}((TT')^{-1}S)/2}.$$

Conveniently, the determinant of the Jacobian matrix J of this transformation is well-known; $|J| = 2^d \prod_{j=1}^{d} t_{jj}^{d+1-j}$. By the change of variable formula, the induced prior on T is

$$\pi(T) = C_{\Lambda,\nu} \det(T)^{-(\nu+d+1)} e^{-\operatorname{tr}((TT')^{-1}\Lambda)/2} \times 2^d \prod_{j=1}^{d} t_{jj}^{d+1-j},$$

where $C_{\Lambda,\nu} = \det(\Lambda)^{\nu/2}/(2^{\nu d/2}\,\Gamma_d(\nu/2))$, and $\Gamma_d(\cdot)$ is the multivariate gamma function. Obtaining posterior samples of T is trivial, as we can simply draw Σ from $\mathcal{W}^{-1}(\Lambda + S, \nu + n)$, and then take the lower Cholesky factor. With this general setup in place, it is worth noting that even with another prior on Σ, we can carry out the entire algorithm, provided that we have a way to sample from the posterior of Σ and a way to compute the Jacobian of the transformation.

In Figure 3, we present the results for which each approximation uses 25 MCMC samples. The boxplot of each approximation's errors solidify the robustness of the Hybrid estimator, which produces accurate and low-variance estimates. Although the BSE and WBSE both cover the true log marginal likelihood value, it is apparent that these estimators suffer from stability and convergence issues that are not present in the Hybrid estimator.

3.3 Graphical Models

In the following examples, we extend the previous analysis of covariance matrices in the graphical modeling context. Gaussian graphical models are a popular tool to learn the dependence structure among variables of interest. Consider independent and identically distributed vectors $x_1, x_2, \ldots, x_n$ drawn from a d-variate normal distribution with mean vector 0 and a sparse inverse covariance matrix Ω. If the variables i and j do not share an edge in a graph G, then $\Omega_{ij} = 0$. Hence, an undirected (or concentration) graphical model corresponding to G restricts the inverse covariance matrix Ω to a linear subspace of the cone of positive definite matrices. A probabilistic framework for learning the dependence structure and the graph G requires specification of a prior distribution for (Ω, G). Conditional on G, a hyper-inverse Wishart (HIW) distribution (Dawid and Lauritzen, 1993) on $\Sigma = \Omega^{-1}$ and the corresponding induced class of distributions on Ω (Roverato, 2000) are attractive choices of priors.

3.3.1 HIW Induced Cholesky Factor Density

Denoted by $\mathrm{HIW}_G(\delta, B)$, the hyper-inverse Wishart distribution is a distribution on the cone of $d \times d$ positive definite matrices with parameters $\delta > 0$ and a fixed $d \times d$ positive definite matrix B. Refer to equations (4) and (5) of (Roverato, 2000) for the form of

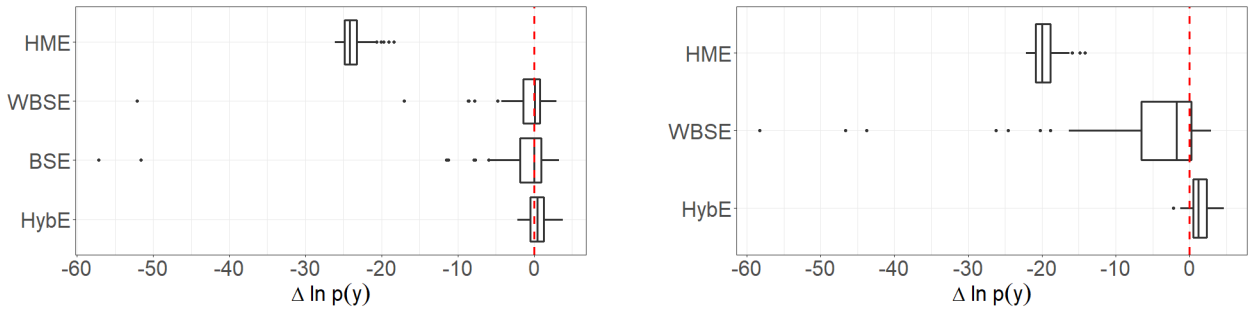

Figure 3: *Boxplots of the error (truth - estimate) for the unrestricted covariance (left, true* $\log p(y)$*: -673.7057) and graphical model (right, true* $\log p(y)$*: -506.3061) examples. Results are reported over 100 simulations, with 100 observations and 25 MCMC samples. For the IW example, we consider* 4×4 *covariance matrices with 10 free parameters. For the HIW example, we consider* 5×5 *precision matrices with 10 free parameters. Note that we do not include BSE results in the HIW example because the Bridge Sampling algorithm fails to converge with only 25 MCMC samples.*

the density. When G is decomposable, an alternative parameterization is given by the Cholesky decomposition TT' of $\Omega = \Sigma^{-1}$. Provided that the vertices of $G = (V, E)$ are enumerated according to a *perfect vertex elimination scheme*, the upper triangular matrix T' has the same zero pattern as Ω. Since the likelihood function is identical to the one given in Eq. (10), we need only compute the induced prior on T to complete the definition of $\Psi(T)$. Following Roverato (2000), the determinant of the Jacobian matrix J of this transformation is given by $|J| = 2^d \prod_{i=1}^d t_{ii}^{\nu_i+1}$, where the ith row of T' has exactly $\nu_i + 1$ many nonzero elements. More specifically, let $\text{ne}(v_i) = \{j : (v_i, v_j) \in E\}$. Then $\nu_i = |\text{ne}(v_i) \cap \{i+1, \ldots, d\}|$. The induced joint density of the elements of T', i.e., t_{sr} for $s < r$ with the edge $(v_s, v_r) \in E$, and $t_{ii}, i = 1, \ldots, d$, is given by

$$\pi(T) = \left[\prod_{i=1}^{d} \frac{2^{-(\delta+\nu_i)/2}}{\Gamma((\delta+\nu_i)/2)} \times t_{ii}^{(\delta+\nu_i-2)} e^{-\frac{1}{2}t_{ii}^2} (2t_{ii}) \right] \times \left[\prod_{(r,s):r>s,(v_s,v_r)\in E} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t_{sr}^2} \right].$$

Since we are able to sample from the posterior distribution, $\text{HIW}_G(\delta + n, B + S)$, where $S = \sum_{i=1}^n x_i x_i'$, we are well-equipped to compute the Hybrid estimator. For this example, we take $\delta = 3$ and $B = I_5$, and in Figure 3, we present the errors for the different estimators when 25 MCMC samples are used for each approximation. Even with limited MCMC samples, the HybE retains its ability to produce reliable results that do not exhibit the high variance that we see in the WBSE. As the number of MCMC samples increases, the WBSE stabilizes and eventually beats the HybE.

3.4 Approximate Posterior Samples

Up until now, we have assumed that asymptotically exact samples from the posterior distribution are available to be used as input for the proposed approximation. In fact, for all previous numerical experiments, we have used samples drawn from the exact posterior distribution, ridding us of the need for burn-in or thinning. We now investigate how these algorithms perform when we only have approximate posterior samples. As a demonstration, we revisit the MVN-IG example in Section 3.1.1 and consider the case where $\beta \in \mathbb{R}^9$. We construct the following mean field approximation to the posterior distribution, $q(\beta, \sigma^2) = q(\beta)q(\sigma^2)$, where

$$q(\beta) \equiv \prod_{i=1}^{3} \mathcal{N}_3 \left(\mu_n^{(i)}, \sigma_0^2 V_n^{(i)} \right), \; q(\sigma^2) \equiv \mathcal{IG}\left(a_n, b_n\right).$$

Here, we have split the original 9-dimensional normal distribution into a product of 3-dimensional normal distributions, with the mean and covariance components extracted from the true posterior parameters. In particular, $\mu_n^{(1)} = (\mu_{n,1}, \mu_{n2}, \mu_{n3})'$, $\mu_n^{(2)} = (\mu_{n4}, \mu_{n5}, \mu_{n6})'$, $\mu_n^{(3)} = (\mu_{n7}, \mu_{n8}, \mu_{n9})'$. Each $V_n^{(i)}$ is defined as the corresponding 3×3 block matrix in V_n, and σ_0^2 is the posterior mean of σ^2.

In Figure 4 below, we observe that even with non-exact posterior samples, the Hybrid approximation produces accurate estimates, with an average error of 0.449 over 100 replications, compared to average errors of 0.698 and 1.035 for the CAME and BSE, respectively. While the latter two estimators have lower variance than the Hybrid approximations, neither covers the true marginal likelihood.

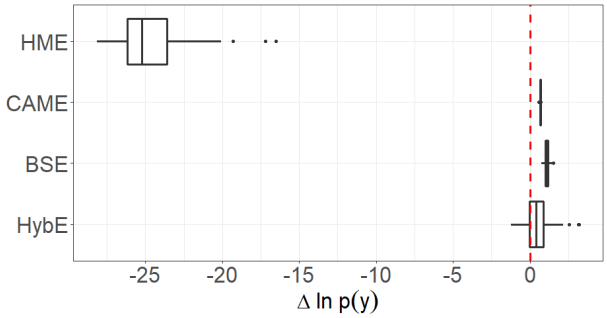

Figure 4: *Boxplots of the error (truth - estimate) for the MVN-IG example ($\beta \in \mathbb{R}^9$) with approximate posterior samples. For each of the 100 replications, we used 100 observations and 100 approximate posterior samples. The true log marginal likelihood is -147.3245.*

4 CONCLUSION

In this paper, we developed a novel algorithm that combines a variety of ideas to efficiently estimate the marginal likelihood. By first using a regression tree to identify high probability regions of the parameter space and then leveraging numerical integration ideas to obviate the need to trust the quality of the MCMC samples, we are able to construct an approximation that scales well with both the dimension and the complexity of the parameter space. From the simulation studies, we see that the Hybrid estimator is both accurate and reliable, providing robust approximations in situations when MCMC samples are either scarce or non-exact. Therefore, our contribution is multifaceted and bears practical value such that even in higher dimensions and in instances where generating MCMC samples is expensive and/or non-exact, the Hybrid estimator still delivers promising results.

Furthermore, the Hybrid approximation scheme outlined in this paper lays the groundwork for future work in a number of possible directions. One area of potential refinement is the construction of the partition of the parameter space. While we used CART for its convenience and interpretability, we found that the default objective function for CART was unsuitable for determining the representative point of each partition set, and we had to solve an additional optimization problem to obtain these points. Instead of this two-step roundabout approach, where we used CART to learn the partition and the objective function in Eq. (9) to identify representative points, we could modify the CART objective function to directly target the desired objective.

Another aspect of the current algorithm that can be further developed is the current formulation of the local approximation to Ψ in each of the partition sets. The piecewise constant approximation in Eq. (5), though providing encouraging results, is a rather simplistic way to approximate Ψ, particularly when moving to higher dimensions. A natural extension to the constant approximation is to use a local Taylor expansion to introduce higher order terms, giving piecewise linear and quadratic approximations.

Acknowledgements

The authors are grateful for the anonymous reviewers for providing valuable comments and suggestions. The authors also express special thanks to Donald Chung for insightful discussion about code optimizations.

References

Peter Binev, Albert Cohen, Wolfgang Dahmen, Ronald DeVore, and Vladimir Temlyakov. Universal algorithms for learning theory part i: Piecewise constant functions. *Journal of Machine Learning Research*, 6:1297–1321, 2005.

Z. I. Botev. The normal law under linear restrictions: simulation and estimation via minimax tilting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(1):125–148, 2016. doi: 10.1111/rssb.12162.

Zdravko Botev and Leo Belzile. *TruncatedNormal: Truncated Multivariate Normal and Student Distributions*, 2019. URL `https://CRAN.R-project.org/package=TruncatedNormal`. R package version 2.1.

Leo Breiman. *Classification and regression trees*. Wadsworth International Group, 1984.

Siddhartha Chib. Marginal likelihood from the gibbs output. *Journal of the American Statistical Association*, 90(432):1313–1321, 1995. doi: 10.1080/01621459.1995.10476635.

Siddhartha Chib and Ivan Jeliazkov. Marginal likelihood from the metropolis–hastings output. *Journal of the American Statistical Association*, 96(453): 270–281, 2001. doi: 10.1198/016214501750332848.

A Philip Dawid and Steffen L Lauritzen. Hyper markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, pages 1272–1317, 1993.

Nial Friel and Jason Wyse. Estimating the evidence - a review. *Statistica Neerlandica*, 66(3):288–308, 2012. doi: 10.1111/j.1467-9574.2011.00515.x.

Subhashis Ghosal and Aad Van Der Vaart. Convergence rates of posterior distributions for noniid observations. *The Annals of Statistics*, 35(1):192–223, 2007.

Quentin F. Gronau, Henrik Singmann, and Eric-Jan Wagenmakers. bridgesampling: An R package for estimating normalizing constants. *Journal of Statistical Software*, 92(10):1–29, 2020. doi: 10.18637/jss.v092.i10.

Dirk P. Kroese, Thomas Taimre, and Zdravko I. Botev. *Handbook of Monte Carlo methods*. Wiley-Blackwell, 2011.

Peter Lenk. Simulation pseudo-bias correction to the harmonic mean estimator of integrated likelihoods. *Journal of Computational and Graphical Statistics*, 18(4):941–960, 2009. doi: 10.1198/jcgs.2009.08022.

Xiao-Li Meng and Stephen Schilling. Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11(3):552–586, 2002. doi: 10.1198/106186002457.

Xiao-Li Meng and Wing Hung Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistia Sinica*, 6:831–860, 1996. URL `www.jstor.org/stable/24306045`.

Radford M. Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001. doi: 10.1023/a:1008923215028.

Michael A. Newton and Adrian E. Raftery. Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1):3–26, 1994. doi: 10.1111/j.2517-6161.1994.tb01956.x.

Anna Pajor. Estimating the marginal likelihood using the arithmetic mean identity. *Bayesian Analysis*, 12 (1):261–287, 2017. doi: 10.1214/16-ba1001.

Adrian E Raftery, Michael A Newton, Jaya M Satagopan, and Pavel N Krivitsky. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. *Bayesian Statistics*, 8: 1–45, 2007.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France, 07–09 Jul 2015. PMLR.

Alberto Roverato. Cholesky decomposition of a hyper inverse wishart matrix. *Biometrika*, 87(1):99–112, 2000.

Tim Salimans, Diederik Kingma, and Max Welling. Markov chain monte carlo and variational inference: Bridging the gap. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1218–1226, Lille, France, 07–09 Jul 2015. PMLR.

John Skilling. Nested sampling for general bayesian computation. *Bayesian Analysis*, 1(4):833–859, 2006. doi: 10.1214/06-ba127.

Terry Therneau and Beth Atkinson. *rpart: Recursive Partitioning and Regression Trees*, 2019. URL `https://CRAN.R-project.org/package=rpart`. R package version 4.1-15.

Luke Tierney and Joseph B. Kadane. Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81(393):82–86, 1986. doi: 10.1080/01621459.1986.10478240.