

# Machine learning assisted Bayesian model comparison: learnt harmonic mean estimator

Jason D. McEwen<sup>1,2\*</sup>, Christopher G. R. Wallis<sup>1</sup>, Matthew A. Price<sup>1</sup>  
and Matthew M. Docherty<sup>1</sup>

<sup>1</sup>Mullard Space Science Laboratory (MSSL), University College  
London (UCL), Dorking, RH5 6NT, UK.

<sup>2</sup>Alan Turing Institute, London, NW1 2DB, UK.

\*Corresponding author(s). E-mail(s): [jason.mcewen@ucl.ac.uk](mailto:jason.mcewen@ucl.ac.uk);

## Abstract

We resurrect the infamous harmonic mean estimator for computing the marginal likelihood (Bayesian evidence) and solve its problematic large variance. The marginal likelihood is a key component of Bayesian model selection since it is required to evaluate model posterior probabilities; however, its computation is challenging. The original harmonic mean estimator, first proposed in 1994 by Newton and Raftery, involves computing the harmonic mean of the likelihood given samples from the posterior. It was immediately realised that the original estimator can fail catastrophically since its variance can become very large and may not be finite. A number of variants of the harmonic mean estimator have been proposed to address this issue although none have proven fully satisfactory. We present the *learnt harmonic mean estimator*, a variant of the original estimator that solves its large variance problem. This is achieved by interpreting the harmonic mean estimator as importance sampling and introducing a new target distribution. The new target distribution is learned to approximate the optimal but inaccessible target, while minimising the variance of the resulting estimator. Since the estimator requires samples of the posterior only it is agnostic to the strategy used to generate posterior samples. We validate the estimator on a variety of numerical experiments, including a number of pathological examples where the original harmonic mean estimator fails catastrophically. In all cases our learnt harmonic mean estimator is shown to be highly accurate. The estimator is computationally scalable and can be applied to problems of dimension  $\mathcal{O}(10^3)$  and beyond. Code implementing the learnt harmonic mean estimator is made publicly available.

# 1 Introduction

Model selection is a critical task in order to ascertain an appropriate statistical model to describe observational data. In the Bayesian formalism, model selection requires computing the *marginal likelihood*, the average likelihood of a model over its prior probability space, given observational data. The marginal likelihood (also called the *Bayesian evidence*) may then be used to compute model posterior odds and assign relative probabilities to different models. Computing the marginal likelihood is therefore a key ingredient in Bayesian inference. However, computing the marginal likelihood in practice requires the evaluation of a high-dimensional integral, which is computationally challenging.

The Bayesian formalism is one of the most common approaches to statistical inference. Consider the estimation of unknown parameters  $\theta \in \Theta$  (typically  $\Theta = \mathbb{R}^d$ ) from observed data  $y$  (typically  $y \in \mathbb{R}^n$ ), under a statistical model  $M$  relating the data to the parameters. Given the data  $y$  and model  $M$ , inferences of the parameters  $\theta$  are based on their posterior distribution through Bayes' theorem by

$$P(\theta | y, M) = \frac{P(y | \theta, M)P(\theta | M)}{P(y | M)} = \frac{\mathcal{L}(\theta)\pi(\theta)}{z}, \quad (1)$$

where for model  $M$  the likelihood  $P(y | \theta, M)$  specifies the probability of the data given the parameters and  $P(\theta | M)$  encodes prior information about the parameters. The denominator  $P(y | M)$ , termed the *marginal likelihood* or *Bayesian evidence*, measures the probability of the observed data under model  $M$ . For notational brevity we denote the likelihood by  $\mathcal{L}(\theta)$ , the prior by  $\pi(\theta)$  and the marginal likelihood by  $z$ . We drop the explicit dependence on the model  $M$ , except where explicitly required. For parameter inference the marginal likelihood can be ignored (since it simply normalises the posterior) and the shape of the posterior can be explored using Markov chain Monte Carlo (MCMC) sampling techniques, e.g. Metropolis-Hastings sampling ([Metropolis et al, 1953](#); [Hastings, 1970](#)).

For Bayesian model selection it is necessary to compute the marginal likelihood given by

$$z = P(y | M) = \int d\theta P(y | \theta, M)P(\theta | M) = \int d\theta \mathcal{L}(\theta)\pi(\theta). \quad (2)$$

The marginal likelihood is of critical importance for Bayesian model selection since it is required to compute the posterior probabilities of models. Noting Bayes' theorem, the relative posterior probability of competing models  $M_1$  and  $M_2$  is given by

$$\frac{P(M_1 | y)}{P(M_2 | y)} = \frac{P(y | M_1) P(M_1)}{P(y | M_2) P(M_2)}. \quad (3)$$

In the absence of prior information regarding model preferences it is reasonable to take the ratio of model prior probabilities  $P(M_1)/P(M_2)$  to be unity. In this case, the relative model posterior probability is given by the ratio of marginal likelihoods for the two competing models, which is also called the *Bayes factor*. In either case, computing marginal likelihoods is a critical component to evaluating model posterior odds, which can then be used to select the preferred model.

It is clear from (2) that evaluation of the marginal likelihood requires computation of an integral with dimension  $d$  given by the number of parameters of interest, which is typically high-dimensional. In principle, the marginal likelihood could be computed simply by Monte Carlo integration of the likelihood, given samples from the prior. While this estimator converges asymptotically to the true marginal likelihood as the number of Monte Carlo samples increases, in practice the accuracy of the estimator depends critically on its variance. Since in practice the prior is typically more diffuse than the likelihood this approach is inefficient, particularly in high and even moderate dimensional settings (Clyde et al, 2007). Consequently, this simple estimator is usually not effective in practice (see, e.g., Cai et al, 2021).

A variety of alternative methods have been proposed to compute the marginal likelihood. For excellent reviews see Clyde et al (2007) and Friel and Wyse (2012). The Savage-Dickey density ratio can be used for nested models (Trotta, 2007). For more general models, Laplace’s method is a widely used approach (Tierney and Kadane, 1986), which relies on the assumption that the posterior distribution can be adequately approximated by a Gaussian distribution. This assumption often does not hold and so marginal likelihood estimates computed by Laplace’s method may be inaccurate. Thermodynamic integration (e.g. O’Ruanaidh and Fitzgerald 1996), which is based on MCMC techniques, is a well-known, general approach for computing the marginal likelihood that has been applied successfully for low-dimensional problems (e.g. Marshall et al, 2003); however, it does require careful tuning. Annealed importance sampling (Neal, 2001), which approximates the target distribution using a tempering mechanism to adaptively define an importance sampling function, is another. Chib’s method (Chib, 1995; Chib and Jeliazkov, 2001) is based on the outputs of a Gibbs or Metropolis-Hasting (MH) sampler (Metropolis et al, 1953; Hastings, 1970), which poses some restrictions. Nested sampling (Skilling, 2006) was designed specifically with the computation of the marginal likelihood in mind, reparameterising the marginal likelihood into a one-dimensional integral of the likelihood with respect to the enclosed prior volume. The computational difficulty of nested sampling approaches is shifted to sampling of the prior distribution subject to a hard constraint defined by likelihood level-sets. Numerous nested sampling strategies have been proposed based on MCMC sampling (Skilling, 2006), ellipsoidal rejection sampling (Feroz and Hobson, 2008; Feroz et al, 2009), slice sampling (Handley et al, 2015), diffusive sampling (Brewer et al, 2011) and proximal sampling (Cai et al,

2021). In all of the above approaches, the sampling strategy is tightly coupled with the technique used to estimate the marginal likelihood. Furthermore, while nested sampling approaches have scaled to high-dimensional settings, notably proximal nested sampling to dimensions  $10^6$  and beyond (Cai et al, 2021), most techniques are limited to low-dimensional settings.

Ideally, the computation of the marginal likelihood would be agnostic to the sampling strategy. If the marginal likelihood estimator required samples from the posterior only, it could indeed then be decoupled from sampling. In this case, the most effective sampler for the problem at hand could be considered and the posterior samples recycled to estimate the marginal likelihood. While some techniques to compute the marginal likelihood from posterior samples have been proposed, they are generally not robust and limited to very low dimensions. The harmonic mean estimator (Newton and Raftery, 1994) involves computing the harmonic mean of the likelihood given samples of the posterior generated by any MCMC technique. However, it was immediately realised that the original estimator can fail catastrophically since its variance can become very large and may not be finite (a thorough review and inspection of the harmonic mean estimator and variants is presented in Sec. 2). In Heavens et al (2017) an approach based on  $k$ th nearest-neighbour distances is proposed to compute the marginal likelihood from posterior samples, although the technique is limited to low-dimensional settings.

In this article we present the *learnt harmonic mean estimator*, a variant of the original harmonic mean estimator that solves its large variance problem. This is achieved by interpreting the harmonic mean estimator as importance sampling and introducing a new target distribution. The new target distribution is learned to approximate the optimal but inaccessible target, while minimising the variance of the resulting estimator. The estimator requires samples of the posterior only and hence is agnostic to the strategy used to generate posterior samples. Posterior samples are split to first learn the target distribution and then to second infer the marginal likelihood using the learnt target. The resulting estimator is evaluated on a variety of numerical experiments, including a number of pathological examples where the original harmonic mean estimator has been shown to fail catastrophically. In all cases our learnt harmonic mean estimator is shown to be robust and highly accurate.

The remainder of this article is structured as follows. In Sec. 2 we review the harmonic mean estimator, its problematic source of large variance, and variants that have been introduced in an attempt to mitigate this issue. We present our learnt harmonic mean estimator in Sec. 3. In Sec. 4 we apply our estimator to numerous benchmark problems where ground truth marginal likelihood values are accessible, demonstrating in all cases that it is highly accurate. Particular attention has been paid to the design and implementation of the software code implementing our learnt harmonic mean estimator so that it can be easily applied by others to their problems of interest. We demonstrate the ease of use of the code in Sec. 5. Concluding remarks are made in Sec. 6.

## 2 Review of harmonic mean estimators

Harmonic mean estimators have been the focus of considerable discussion since first proposed by [Newton and Raftery \(1994\)](#). While the harmonic mean estimator is asymptotically consistent ([Newton and Raftery, 1994](#)), it was immediately realised that the original estimator can fail catastrophically ([Neal, 1994](#)) since its variance can become very large and may not be finite. A number of variants of the original estimator have been proposed to address its failings (e.g. [Raftery et al, 2006](#); [Robert and Wraith, 2009](#); [Lenk, 2009](#); [van Haasteren, 2014](#)), although harmonic mean estimators have generally been considered to be ineffective ([Clyde et al, 2007](#); [Friel and Wyse, 2012](#)). We review the original harmonic mean estimator and discuss why it is problematic. We then review variants of the original estimator and how they attempt to address this failing, which motivates the *learnt harmonic mean estimator* that we present in [Sec. 3](#).

### 2.1 Original harmonic mean estimator

The harmonic mean estimator was first proposed by [Newton and Raftery \(1994\)](#), who showed that the marginal likelihood  $z$  can be estimated from the harmonic mean of the likelihood, given posterior samples. This follows by considering the expectation of the reciprocal of the likelihood with respect to the posterior distribution:

$$\rho = \mathbb{E}_{\mathbb{P}(\theta|y)} \left[ \frac{1}{\mathcal{L}(\theta)} \right] \quad (4)$$

$$= \int d\theta \frac{1}{\mathcal{L}(\theta)} \mathbb{P}(\theta|y) \quad (5)$$

$$= \int d\theta \frac{1}{\mathcal{L}(\theta)} \frac{\mathcal{L}(\theta)\pi(\theta)}{z} \quad (6)$$

$$= \frac{1}{z}, \quad (7)$$

where the final line follows since the prior  $\pi(\theta)$  is a normalised probability distribution. This relationship between the marginal likelihood and the harmonic mean motivates the *original harmonic mean estimator*:

$$\hat{\rho} = \frac{1}{N} \sum_{i=1}^N \frac{1}{\mathcal{L}(\theta_i)}, \quad \theta_i \sim \mathbb{P}(\theta|y), \quad (8)$$

where  $N$  specifies the number of samples  $\theta_i$  drawn from the posterior, and from which the marginal likelihood may naively be estimated by  $\hat{z} = 1/\hat{\rho}$ . For now we simply consider the estimation of the reciprocal of the marginal likelihood  $\hat{\rho}$  (we discuss estimation of the marginal likelihood itself and Bayes factors in more detail in [Sec. 3.3](#)).

As immediately realised by [Neal \(1994\)](#), this estimator can fail catastrophically since its variance can become very large and may not be finite. Review

articles that consider a variety of methods to estimate the marginal likelihood have also found that the harmonic mean estimator is not robust and can be highly inaccurate (Clyde et al, 2007; Friel and Wyse, 2012). To understand why the estimator can lead to extremely large variance we consider an importance sampling interpretation of the harmonic mean estimator.

### 2.1.1 Importance sampling interpretation

The harmonic mean estimator can be interpreted as importance sampling. Consider the reciprocal marginal likelihood, which may be expressed in terms of the prior and posterior by

$$\rho = \int d\theta \frac{1}{\mathcal{L}(\theta)} P(\theta | y) \quad (9)$$

$$= \int d\theta \frac{1}{z} \frac{\pi(\theta)}{P(\theta | y)} P(\theta | y). \quad (10)$$

It is clear the estimator has an importance sampling interpretation where the importance sampling target distribution is the prior  $\pi(\theta)$ , while the sampling density is the posterior  $P(\theta | y)$ , in contrast to typical importance sampling scenarios.

For importance sampling to be effective, one requires the sampling density to have fatter tails than the target distribution, i.e. to have greater probability mass in the tails of the distribution. Typically the prior has fatter tails than the posterior since the posterior updates our initial understanding of the underlying parameters  $\theta$  that are encoded in the prior, in the presence of new data  $y$ . For the harmonic mean estimator the importance sampling density (the posterior) typically does *not* have fatter tails than the target (the prior) and so importance sampling is not effective. This explains why the original harmonic mean estimator can be problematic. A number of variants of the original harmonic mean estimator have been introduced in an attempt to address this issue.

## 2.2 Adjusted harmonic mean estimator

Lenk (2009) show that while the original harmonic mean estimator is consistent, in practice it exhibits simulation pseudo-bias. Simulation pseudo-bias arises since the posterior simulation support is a subset of the prior support. Consequently, the prior is not sufficiently captured, which often results in an over-estimate of the marginal likelihood.

An *adjusted harmonic mean estimator* is introduced by Lenk (2009) to correct for simulation pseudo-bias:

$$\hat{\rho} = \frac{1}{P(\Lambda)} \frac{1}{N} \sum_{i=1}^N \frac{1}{\mathcal{L}(\theta_i)}, \quad \theta_i \sim P(\theta | y), \quad (11)$$

where  $P(\Lambda)$  is a pseudo-bias adjustment factor given by the prior probability of the posterior simulation support  $\Lambda \subset \Theta$ . Numerical methods to estimate  $P(\Lambda)$  are proposed, however, estimating the adjustment factor accurately is numerically challenging, particularly in high dimensions. Furthermore, while this adjusted estimator can mitigate simulation pseudo-bias it does not eliminate it (Pajor et al, 2017). Alternative approaches seek to eliminate the bias altogether.

### 2.3 Stabilised harmonic mean estimator

Raftery et al (2006) propose an alternative approach, a *stabilised harmonic mean estimator*, by introducing a variance stabilisation strategy that reduces the size of the parameter space. While this strategy can be applied to a variety of common hierarchical models it is not applicable in general, limiting its use.

### 2.4 Re-targeted harmonic mean estimator

The original harmonic mean estimator was revised by Gelfand and Dey (1994) by introducing an arbitrary density  $\varphi(\theta)$  to relate the reciprocal of the marginal likelihood to the likelihood through the following expectation:

$$\rho = \mathbb{E}_{P(\theta|y)} \left[ \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} \right] \quad (12)$$

$$= \int d\theta \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} P(\theta|y) \quad (13)$$

$$= \int d\theta \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} \frac{\mathcal{L}(\theta)\pi(\theta)}{z} \quad (14)$$

$$= \frac{1}{z}, \quad (15)$$

where the final line follows since the density  $\varphi(\theta)$  must be normalised. The above expression motivates the estimator:

$$\hat{\rho} = \frac{1}{N} \sum_{i=1}^N \frac{\varphi(\theta_i)}{\mathcal{L}(\theta_i)\pi(\theta_i)}, \quad \theta_i \sim P(\theta|y). \quad (16)$$

The normalised density  $\varphi(\theta)$  can be interpreted as an alternative importance sampling target distribution, as we will see, hence we refer to this approach as the *re-targeted harmonic mean estimator*. Note that the original harmonic mean estimator is recovered for the target distribution  $\varphi(\theta) = \pi(\theta)$ .

### 2.4.1 Importance sampling interpretation

With the introduction of the distribution  $\varphi(\theta)$ , the importance sampling interpretation of the harmonic mean estimator reads

$$\rho = \int d\theta \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} \mathbb{P}(\theta|y) \quad (17)$$

$$= \int d\theta \frac{1}{z} \frac{\varphi(\theta)}{\mathbb{P}(\theta|y)} \mathbb{P}(\theta|y). \quad (18)$$

It is clear that the distribution  $\varphi(\theta)$  now plays the role of the importance sampling target distribution. One is free to choose  $\varphi(\theta)$ , with the only constraint being that it is a normalised distribution. It is therefore possible to select the target distribution  $\varphi(\theta)$  such that it has narrower tails than the posterior, which we recall plays the role of the importance sampling density, thereby avoiding the problematic scenario of the original harmonic mean estimator. We therefore refer to  $\varphi(\theta)$  as the *target distribution* of the harmonic mean estimator.

The question of how to develop an effective strategy to select  $\varphi(\theta)$  for a given problem remains, which is particularly difficult in high-dimensional settings (Chib, 1995). Gelfand and Dey (1994) initially suggest using a multivariate Gaussian, although this approach is typically not effective since the tails of the distribution are generally not sufficiently narrow (Chib, 1995; Clyde et al, 2007).

### 2.4.2 Truncated harmonic mean estimator

A common strategy to select the target distribution is to set it to a normalised indicator function that is supported on a region  $\Omega$  of high posterior mass (so that the target has narrower tails than the posterior):

$$\varphi(\theta) = \frac{1}{V_\Omega} I_\Omega(\theta), \quad (19)$$

where  $V_\Omega$  represents the volume encapsulated in  $\Omega$  and the indicator function  $I_\Omega(\theta) = 1$  if  $\theta \in \Omega$  and zero otherwise. Since the indicator function effectively truncates the region of parameter space considered we refer to this approach as the *truncated harmonic mean estimator*.

Robert and Wraith (2009) propose a target distribution that corresponds to an indicator function with support  $\Omega$  determined from the convex hull of Monte Carlo samples within an  $\alpha\%$  highest posterior density (HPD) region. In practice they consider an ellipsoidal region defined by HPD samples, for which the volume can be computed analytically. van Haasteren (2014) take a similar approach and consider indicator functions defined over ellipsoidal regions.

While such approaches can be effective, in general the truncated harmonic mean estimator can be inaccurate and inefficient since each sample is



either used, with a uniform target density weight, or discarded. In scenarios that exhibit thin parameter degeneracies such approaches either capture large regions of low posterior mass, which is problematic (for reasons discussed above in Sec. 2.4.1), or can suffer prohibitive inefficiencies as the support of the target distribution  $\Omega$  can be a very small region of the full parameter space  $\Theta$  (resulting in very few samples being retained in the marginal likelihood computation).

The selection of appropriate target densities  $\varphi(\theta)$  for general problems remains an open question that is known to be difficult, particularly in high dimensions (Chib, 1995). One may gain insight into effective strategies to design the target density by considering the optimal target distribution.

### 2.4.3 Optimal importance sampling target

Consider the importance sampling target distribution given by the (normalised) posterior itself:

$$\varphi^{\text{optimal}}(\theta) = \frac{\mathcal{L}(\theta)\pi(\theta)}{z}. \quad (20)$$

This estimator is optimal in the sense of having zero variance, which is clearly apparent by substituting the target density into the re-targeted harmonic mean estimator of (16). Each term contributing to the summation is simply  $1/z$ , hence the estimator  $\hat{\rho}$  is unbiased, with zero variance.

Recall that the target density must be normalised. Hence, the optimal estimator given by the normalised posterior is not accessible in practice since it requires the marginal likelihood – the very term we are attempting to estimate – to be known. While the optimal estimator therefore cannot be used in practice, it can nevertheless be used to inform the construction of other estimators based on alternative importance sampling target distributions.

## 3 Learnt harmonic mean estimator

It is well-known that the original harmonic mean estimator can fail catastrophically since the variance of the estimator may become very large, as discussed in detail in Sec. 2. As also discussed in Sec. 2, however, this issue can be resolved by introducing an alternative (normalised) target distribution  $\varphi(\theta)$  (Gelfand and Dey, 1994), yielding what we term here the *re-targeted harmonic mean estimator*. From the importance sampling interpretation of the harmonic mean estimator, the re-targeted estimator follows by replacing the importance sampling target of the prior  $\pi(\theta)$  with the target  $\varphi(\theta)$ , where the posterior  $P(\theta | y)$  plays the role of the importance sampling density.

It remains to select a suitable target distribution  $\varphi(\theta)$ . On one hand, to ensure the variance of the resulting estimator is well-behaved, the target distribution should have narrower tails than the importance sampling density, i.e. the target  $\varphi(\theta)$  should have narrower tails than the posterior  $P(\theta | y)$  (as

discussed in Sec. 2). On the other hand, to ensure the resulting estimator is efficient and makes use of as many samples from the posterior as possible, the target distribution should not be too narrow. The optimal target distribution is the normalised posterior distribution since in this case the variance of the resulting estimator is zero (Sec. 2). However, the normalised posterior is not accessible since it requires knowledge of the marginal likelihood, which is precisely the term we are attempting to compute.

We propose learning the target distribution  $\varphi(\theta)$  from samples of the posterior. Samples from the posterior can be split into training and evaluation (cf. test) sets. Machine learning (ML) techniques can then be applied to learn an approximate model of the normalised posterior from the training samples, with the constraint that the tails of the learnt target are narrower than the posterior, i.e.

$$\varphi(\theta) \stackrel{\text{ML}}{\simeq} \varphi^{\text{optimal}}(\theta) = \frac{\mathcal{L}(\theta)\pi(\theta)}{z}. \quad (21)$$

We term this approach the *learnt harmonic mean estimator*.

We are interested not only in an estimator for the marginal likelihood but also in an estimate of the variance of this estimator, and its variance. Such additional estimators are useful in their own right and can also provide valuable sanity checks that the resulting marginal likelihood estimator is well-behaved. We present corresponding estimators for the cases of uncorrelated and correlated samples. Harmonic mean estimators provide an estimation of the reciprocal of the marginal likelihood. We therefore also consider estimation of the marginal likelihood itself and its variance from the reciprocal estimators. Moreover, we present expressions to also estimate the Bayes factor, and its variance, to compare two models. Finally, we present models to learn the normalised target distribution  $\varphi(\theta)$  by approximating the posterior distribution, with the constraint that the target has narrower tails than the posterior, and discuss how to train such models. Training involves constructing objective functions that penalise models that would result in estimators with a large variance, with appropriate regularisation.

### 3.1 Uncorrelated samples

MCMC algorithms that are typically used to sample the posterior distribution result in correlated samples. By suitably thinning the MCMC chain (discarding all but every  $t$ th sample), however, samples that are uncorrelated can be obtained. In this subsection we present estimators for the reciprocal marginal likelihood and its variance under the assumption of uncorrelated samples from the posterior.

Consider the harmonic moments

$$\mu_n = \mathbb{E}_{P(\theta|y)} \left[ \left( \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} \right)^n \right], \quad (22)$$

and corresponding central moments

$$\mu'_n = \mathbb{E}_{\mathbb{P}(\theta|y)} \left[ \left( \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} - \mathbb{E}_{\mathbb{P}(\theta|y)} \left( \frac{\varphi(\theta)}{\mathcal{L}(\theta)\pi(\theta)} \right) \right)^n \right]. \quad (23)$$

We make use of the following harmonic moment estimators computed from samples of the posterior:

$$\hat{\mu}_n = \frac{1}{N} \sum_{i=1}^N \left( \frac{\varphi(\theta_i)}{\mathcal{L}(\theta_i)\pi(\theta_i)} \right)^n, \quad \theta_i \sim \mathbb{P}(\theta|y), \quad (24)$$

which are unbiased estimators of  $\mu_n$ , i.e.  $\mathbb{E}(\hat{\mu}_n) = \mu_n$ . The reciprocal marginal likelihood can then be estimated from samples of the posterior by

$$\hat{\rho} = \hat{\mu}_1 = \frac{1}{N} \sum_{i=1}^N \frac{\varphi(\theta_i)}{\mathcal{L}(\theta_i)\pi(\theta_i)}, \quad \theta_i \sim \mathbb{P}(\theta|y). \quad (25)$$

The mean and variance of the estimator read, respectively,

$$\mathbb{E}(\hat{\rho}) = \mathbb{E} \left[ \frac{1}{N} \sum_{i=1}^N \frac{\varphi(\theta_i)}{\mathcal{L}(\theta_i)\pi(\theta_i)} \right] = \mu_1 = \rho \quad (26)$$

and

$$\text{var}(\hat{\rho}) = \text{var} \left[ \frac{1}{N} \sum_{i=1}^N \frac{\varphi(\theta_i)}{\mathcal{L}(\theta_i)\pi(\theta_i)} \right] = \frac{1}{N} (\mu_2 - \mu_1^2). \quad (27)$$

Note that the estimator is unbiased.

Recall from Sec. 2 that the optimal target is given by the normalised posterior, i.e.  $\varphi^{\text{optimal}}(\theta) = \mathcal{L}(\theta)\pi(\theta)/z$ . It is straightforward to see that in this case

$$\mu_n = \hat{\mu}_n = \frac{1}{z^n}, \quad (28)$$

and thus the target distribution is optimal since

$$\text{var}(\hat{\rho}) = \frac{1}{N} (\mu_2 - \mu_1^2) = \frac{1}{N} (1/z^2 - (1/z)^2) = 0. \quad (29)$$

We are interested in not only an estimate of the reciprocal marginal likelihood but also its variance  $\text{var}(\hat{\rho})$ . It is clear from (27) that a suitable estimator of the variance is given by

$$\hat{\sigma}^2 = \frac{1}{N-1} (\hat{\mu}_2 - \hat{\mu}_1^2) = \frac{1}{N(N-1)} \sum_{i=1}^N \left( \frac{\varphi(\theta_i)}{\mathcal{L}(\theta_i)\pi(\theta_i)} \right)^2 - \frac{\hat{\rho}^2}{N-1}. \quad (30)$$

It follows that this estimator of the variance is unbiased since

$$\mathbb{E}(\hat{\sigma}^2) = \frac{1}{N}(\mu_2 - \mu_1^2) = \text{var}(\hat{\rho}). \quad (31)$$

The variance of the estimator  $\hat{\sigma}^2$  reads

$$\text{var}(\hat{\sigma}^2) = \frac{1}{(N-1)^2} \left[ \frac{(N-1)^2}{N^3} \mu'_4 - \frac{(N-1)(N-3)}{N^3} \mu'^2_2 \right], \quad (32)$$

where  $\mu'_n$  are central moments, which follows by a well-known result for the variance of a sample variance (e.g. [Rose and Smith, 2002](#), p. 264). An unbiased estimator of  $\text{var}(\hat{\sigma}^2)$  can be constructed from h-statistics (e.g. [Rose and Smith, 2002](#)), which provide unbiased estimators of central moments.

While we have presented general estimators for uncorrelated samples here, generating uncorrelated samples requires thinning the MCMC chain, which is highly inefficient. It is generally recognised that thinning should be avoided when possible since it reduces the precision with which summaries of the MCMC chain can be computed ([Link and Eaton, 2012](#)). Subsequently, we consider estimators that do not require uncorrelated samples and so can make use of considerably more MCMC samples of the posterior.

### 3.2 Correlated samples

We present an estimator of the reciprocal marginal likelihood, an estimate of the variance of this estimator, and its variance. These estimators make use of correlated samples in order to avoid the loss of efficiency that results from thinning an MCMC chain.

We propose running a number of independent MCMC chains and using all of the correlated samples within a given chain. A number of modern MCMC sampling techniques, such as affine invariance ensemble samplers ([Goodman and Weare, 2010](#)), naturally provide samples from multiple chains by their ensemble nature. Moreover, excellent software implementations are readily available, such as the `emcee` code<sup>1</sup> ([Foreman-Mackey et al, 2013](#)), which provides an implementation of the affine invariance ensemble samplers proposed by [Goodman and Weare \(2010\)](#). Alternatively, if only a single large chain is available then this can be broken into separate blocks, which are (approximately) independent for a suitably long block length. Subsequently, we use the terminology chains throughout to refer to both scenarios of running multiple MCMC chains or separating a single chain in blocks.

Consider  $C$  chains of samples, indexed by  $j = 1, 2, \dots, C$ , with chain  $j$  containing  $N_j$  samples. The  $i$ th sample of chain  $j$  is denoted  $\theta_{ij}$ . Since the chain of interest is typically clear from the context, for notational brevity we drop the chain index from the samples, i.e. we denote samples by  $\theta_i$  where the chain of interest is inferred from the context.

---

<sup>1</sup><https://emcee.readthedocs.io/en/stable/>

An estimator of the reciprocal marginal likelihood can be computed from each independent chain by

$$\hat{\rho}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} \frac{\varphi(\theta_i)}{\mathcal{L}(\theta_i)\pi(\theta_i)}, \quad \theta_i \sim P(\theta | y). \quad (33)$$

A single estimator of the reciprocal marginal likelihood can then be constructed from the estimator for each chain by

$$\hat{\rho} = \frac{\sum_{j=1}^C w_j \hat{\rho}_j}{\sum_{j=1}^C w_j}, \quad (34)$$

where the estimator  $\hat{\rho}_j$  of chain  $j$  is weighted by the number of samples in the chain, i.e.  $w_j = N_j$ . It is straightforward to see that the estimator of the reciprocal marginal likelihood is unbiased, i.e.  $\mathbb{E}(\hat{\rho}) = \rho$ , since  $\mathbb{E}(\hat{\rho}_j) = \rho$ .

The variance of the estimator  $\hat{\rho}$  is related to the population variance  $\sigma^2 = \mathbb{E}[(\hat{\rho}_i - \mathbb{E}(\hat{\rho}_i))^2]$  by

$$\text{var}(\hat{\rho}) = \frac{\sigma^2}{N_{\text{eff}}}, \quad (35)$$

where the effective sample size is given by

$$N_{\text{eff}} = \frac{(\sum_j^C w_j)^2}{\sum_j^C w_j^2}. \quad (36)$$

The estimator of the population variance, given by

$$\hat{s}^2 = \frac{N_{\text{eff}}}{N_{\text{eff}} - 1} \frac{\sum_{j=1}^C w_j (\hat{\rho}_j - \hat{\rho})^2}{\sum_j^C w_j}, \quad (37)$$

is unbiased, i.e.  $\mathbb{E}(\hat{s}^2) = \sigma^2$ . A suitable estimator for  $\text{var}(\hat{\rho})$  is thus

$$\hat{\sigma}^2 = \frac{\hat{s}^2}{N_{\text{eff}}} = \frac{1}{N_{\text{eff}} - 1} \frac{\sum_{j=1}^C w_j (\hat{\rho}_j - \hat{\rho})^2}{\sum_j^C w_j}, \quad (38)$$

which is unbiased, i.e.  $\mathbb{E}(\hat{\sigma}^2) = \text{var}(\hat{\rho})$ , since  $\hat{s}^2$  is unbiased.

The variance of the estimator  $\hat{\sigma}^2$  reads

$$\text{var}(\hat{\sigma}^2) = \frac{1}{N_{\text{eff}}^2} \text{var}(\hat{s}^2) = \frac{\sigma^4}{N_{\text{eff}}^3} \left( \kappa - 1 + \frac{2}{N_{\text{eff}} - 1} \right), \quad (39)$$

where in the second equality we have used a well-known result for the variance of the sample variance of independent and identically distributed (i.i.d.)

random variables (e.g. [Cho et al, 2005](#)). The kurtosis  $\kappa$  is defined by

$$\kappa = \text{kur}(\hat{\rho}_i) = \mathbb{E} \left[ \left( \frac{\hat{\rho}_i - \rho}{\sigma} \right)^4 \right]. \quad (40)$$

A suitable estimator for  $\text{var}(\hat{\sigma}^2)$  is thus

$$\hat{\nu}^4 = \frac{\hat{s}^4}{N_{\text{eff}}^3} \left( \hat{\kappa} - 1 + \frac{2}{N_{\text{eff}} - 1} \right) = \frac{\hat{\sigma}^4}{N_{\text{eff}}} \left( \hat{\kappa} - 1 + \frac{2}{N_{\text{eff}} - 1} \right), \quad (41)$$

where for the kurtosis we adopt the estimator

$$\hat{\kappa} = \frac{\sum_{j=1}^C w_j (\hat{\rho}_j - \hat{\rho})^4}{\hat{s}^4 \sum_{j=1}^C w_j} = \frac{\sum_{j=1}^C w_j (\hat{\rho}_j - \hat{\rho})^4}{N_{\text{eff}}^2 \hat{\sigma}^4 \sum_{j=1}^C w_j} \quad (42)$$

(although alternative estimators of the kurtosis may be considered).

The estimators  $\hat{\rho}$ ,  $\hat{\sigma}^2$  and  $\hat{\nu}^4$  provide a strategy to estimate the reciprocal marginal likelihood, its variance, and the variance of the variance, respectively. The variance estimators provide valuable measures of the accuracy of the estimated reciprocal marginal likelihood and provide useful sanity checks.

Additional sanity checks can also be considered. By the central limit theorem, for a large number of samples the distribution of  $\hat{\rho}_j$  approaches a Gaussian, with kurtosis  $\kappa = 3$ . If the estimated kurtosis  $\hat{\kappa} \gg 3$  it would indicate that the sampled distribution of  $\hat{\rho}_j$  has long tails, suggesting further samples need to be drawn. Similarly, the ratio of  $\hat{\nu}^2/\hat{\sigma}^2$  can be inspected to see if it is close to that expected for a Gaussian distribution with  $\kappa = 3$  of

$$\frac{\hat{\nu}^4}{\hat{\sigma}^4} = \frac{1}{N_{\text{eff}}} \left( 2 + \frac{2}{N_{\text{eff}} - 1} \right) = \frac{2}{N_{\text{eff}} - 1}, \quad (43)$$

or equivalently

$$\frac{\hat{\nu}^2}{\hat{\sigma}^2} = \sqrt{\frac{2}{N_{\text{eff}} - 1}}. \quad (44)$$

For the common setting where the number of samples per chain is constant, i.e.  $N_j = N$  for all  $j$ ,

$$N_{\text{eff}} = \frac{\left( \sum_j^C w_j \right)^2}{\sum_j^C w_j^2} = \frac{(NC)^2}{N^2 C} = C \quad (45)$$

and, say  $C = 100$ , we find

$$\frac{\hat{\nu}^2}{\hat{\sigma}^2} = 0.14. \quad (46)$$

In this setting significantly larger values of this ratio would suggest that further samples need to be drawn.

### 3.3 Bayes factors

We have so far considered the estimation of the reciprocal marginal likelihood and related variances only. However it is the marginal likelihood itself (not its reciprocal), or the Bayes factors computed to compare two models, that is typically of direct interest. We therefore consider how to compute these quantities of interest and a measure of their variance.

First, consider the mean and variance of the function  $f(X, Y) = X/Y$  of two uncorrelated random variables  $X$  and  $Y$ , which by Taylor expansion to second order are given by

$$\mathbb{E}\left(\frac{X}{Y}\right) \simeq \frac{\mathbb{E}(X)}{\mathbb{E}(Y)} + \frac{\mathbb{E}(X)}{\mathbb{E}(Y)^3}\sigma_Y^2 \quad (47)$$

and

$$\text{var}\left(\frac{X}{Y}\right) \simeq \frac{1}{\mathbb{E}(Y)^2}\sigma_X^2 + \frac{\mathbb{E}(X)^2}{\mathbb{E}(Y)^4}\sigma_Y^2, \quad (48)$$

respectively, where  $\sigma_X = \mathbb{E}[(X - \mathbb{E}(X))^2]$  and  $\sigma_Y = \mathbb{E}[(Y - \mathbb{E}(Y))^2]$ .

Using this result the marginal likelihood and its variance can be estimated from the reciprocal estimators by making use of the relations

$$\mathbb{E}(z) = \mathbb{E}\left(\frac{1}{\rho}\right) \simeq \frac{1}{\mathbb{E}(\rho)}\left(1 + \frac{\sigma_\rho^2}{\mathbb{E}(\rho)^2}\right) \quad (49)$$

and

$$\text{var}\left(\frac{1}{\rho}\right) \simeq \frac{\sigma_\rho^2}{\mathbb{E}(\rho)^4}, \quad (50)$$

respectively, by considering the case  $X = 1$  and  $Y = \rho$ .

Typically it is the Bayes factor given by the ratio of marginal likelihoods that is of most interest in order to compare models. Again using the expressions above for the mean and variance of the function  $f(X, Y) = X/Y$ , this time for the case  $X = \rho_2$  and  $Y = \rho_1$ , the Bayes factor and its variance can be estimated directly from the reciprocal marginal likelihood estimates and variances by making use of the relations

$$\mathbb{E}\left(\frac{z_1}{z_2}\right) = \mathbb{E}\left(\frac{\rho_2}{\rho_1}\right) \simeq \frac{\mathbb{E}(\rho_2)}{\mathbb{E}(\rho_1)}\left(1 + \frac{\sigma_{\rho_1}^2}{\mathbb{E}(\rho_1)^2}\right) \quad (51)$$

and

$$\text{var}\left(\frac{z_1}{z_2}\right) = \text{var}\left(\frac{\rho_2}{\rho_1}\right) \simeq \frac{\mathbb{E}(\rho_1)^2\sigma_{\rho_2}^2 + \mathbb{E}(\rho_2)^2\sigma_{\rho_1}^2}{\mathbb{E}(\rho_1)^4}, \quad (52)$$

respectively.

### 3.4 Learning the target density

While we have described estimators to compute the marginal likelihood and Bayes factors based on a learnt target distribution  $\varphi(\theta)$ , we have yet to consider the critical task of learning the target distribution. As discussed, the ideal target distribution is the posterior itself. However, since the target must be normalised, use of the posterior would require knowledge of the marginal likelihood – precisely the quantity that we are attempting to estimate. Instead, one can learn an approximation of the posterior that is normalised. The approximation itself does not need to be highly accurate. More critically, the learned target approximating the posterior must exhibit narrower tails than the posterior to avoid the problematic scenario of the original harmonic mean that can result in very large variance.

We present three examples of models that can be used to learn appropriate target distributions and discuss how to train them, although other models can of course be considered. Samples of the posterior are split into training and evaluation (cf. test) sets. The training set is used to learn the target distribution, after which the evaluation set, combined with the learnt target, is used to estimate the marginal likelihood. To train the models we typically construct and solve an optimisation problem to minimise the variance of the estimator, while ensuring it is unbiased. We typically solve the resulting optimisation problem by stochastic gradient descent. To set hyperparameters, we advocate cross-validation.

#### 3.4.1 Hypersphere

The simplest model one may wish to consider is a hypersphere, much like the truncated harmonic mean estimator. However, here we learn the optimal radius of the hypersphere, rather than setting the radius based on arbitrary level-sets of the posterior as considered previously.

Consider the target distribution defined by the normalised hypersphere

$$\varphi(\theta) = \frac{1}{V_S} I_S(\theta), \quad (53)$$

where the indicator function  $I_S(\theta)$  is unity if  $\theta$  is within a hypersphere of radius  $R$ , centred on  $\bar{\theta}$  with covariance  $\Sigma$ , i.e.

$$I_S(\theta) = \begin{cases} 1, & (\theta - \bar{\theta})^\top \Sigma^{-1} (\theta - \bar{\theta}) < R^2 \\ 0, & \text{otherwise} \end{cases}. \quad (54)$$

The values of  $\bar{\theta}$  and  $\Sigma$  can be computed directly from the training samples. Often, although not always, a diagonal approximation of  $\Sigma$  is considered for computational efficiency. The volume of the hypersphere required to normalise



the distribution is given by

$$V_S = \frac{\pi^{d/2}}{\Gamma(d/2 + 1)} R^d |\Sigma|^{1/2}. \quad (55)$$

Recall that  $d$  is the dimension of the parameter space, i.e.  $\theta \in \mathbb{R}^d$ , and note that  $\Gamma(\cdot)$  is the Gamma function.

To estimate the radius of the hypersphere we pose the following optimisation problem to minimise the variance of the learnt harmonic mean estimator, while also constraining it to be unbiased:

$$\min_R \hat{\sigma}^2 \quad \text{s.t.} \quad \hat{\rho} = \hat{\mu}_1. \quad (56)$$

By minimising the variance of the estimator we ensure, on one hand, that the tails of the learnt target are not so wide that they are broader than the posterior, and, on the other hand, that they are not so narrow that very few samples are effectively retained in the estimator. This optimisation problem is equivalent to minimising the estimator of the second harmonic moment:

$$\min_R \hat{\mu}_2. \quad (57)$$

Writing out the cost function explicitly in terms of the posterior samples, the optimisation problem reads

$$\min_R \sum_i C_i^2, \quad (58)$$

with costs for each sample given by

$$C_i = \frac{\varphi(\theta_i)}{\mathcal{L}(\theta_i)\pi(\theta_i)} \propto \begin{cases} \frac{1}{\mathcal{L}(\theta_i)\pi(\theta_i)R^d}, & (\theta - \bar{\theta})^\top \Sigma^{-1}(\theta - \bar{\theta}) < R^2 \\ 0, & \text{otherwise} \end{cases}. \quad (59)$$

This one-dimensional optimisation problem can be solved by straightforward techniques, such as the Brent hybrid root-finding algorithm.

While the learnt hypersphere model is very simple, it is good pedagogical illustration of the general procedure for learning target distributions. First, construct a normalised model. Second, train the model to learn its parameters by solving an optimisation problem to minimise the variance of the estimator while ensuring it is unbiased. If required, set hyperparameters or compare alternative models by cross-validation. While the simple learnt hypersphere model may be sufficient in some settings, it is not effective for multimodal posterior distributions or for posteriors with narrow curving degeneracies. For such scenarios we consider alternative learnt models.

### 3.4.2 Modified Gaussian mixture model

A modified Gaussian mixture model provides greater flexibility than the simple hypersphere model. In particular, it is much more effective for multimodal posterior distributions.

Consider the target distribution defined by the modified Gaussian mixture model

$$\varphi(\theta) = \sum_{k=1}^K \frac{w_k}{(2\pi)^{d/2} |\Sigma_k|^{1/2} s_k^d} \exp\left(\frac{-(\theta - \bar{\theta}_k)^T \Sigma_k^{-1} (\theta - \bar{\theta}_k)}{2s_k^2}\right), \quad (60)$$

for  $K$  components, with centres  $\bar{\theta}_k$  and covariances  $\Sigma_k$ , where the relative scale of each component is controlled by  $s_k$  and the weights are specified by

$$w_k = \frac{\exp(z_k)}{\sum_{k'=1}^K \exp(z_{k'})}, \quad (61)$$

which in turn depend on the weights  $z_k$ . Given  $K$ , the posterior training samples can be clustered by  $K$ -means. The values of  $\bar{\theta}_k$  and  $\Sigma_k$  can then be computed by the samples in cluster  $k$ . The model is modified relative to the usual Gaussian mixture model in that the cluster mean and covariance are estimated from the samples of each cluster, while the relative cluster scale and weights are fitted. Moreover, as before, a bespoke training approach is adopted tailored to the problem of learning an effective model for the learnt harmonic mean estimator.

To estimate the weights  $z_k$ , which in turn define the weights  $w_k$ , and the relative scales  $s_k$  we again construct an optimisation problem to minimise the variance of the learnt harmonic mean estimator, while also constraining it to be unbiased. We also regularise the relative scale parameters, resulting in the following optimisation problem:

$$\min_{\{z_k, s_k\}_{k=1}^K} \hat{\sigma}^2 + \frac{1}{2} \lambda \sum_{k=1}^K s_k^2 \quad \text{s.t.} \quad \hat{\rho} = \hat{\mu}_1, \quad (62)$$

for regularisation parameter  $\lambda$ . The problem may equivalently be written as

$$\min_{\{z_k, s_k\}_{k=1}^K} \hat{\mu}_2 + \frac{1}{2} \lambda \sum_{k=1}^K s_k^2, \quad (63)$$

or explicitly in terms of the posterior samples by

$$\min_{\{z_k, s_k\}_{k=1}^K} \sum_i C_i^2 + \frac{1}{2} \lambda \sum_{k=1}^K s_k^2. \quad (64)$$

The individual cost terms for each sample  $i$  are given by

$$C_i = \frac{\varphi(\theta_i)}{\mathcal{L}(\theta_i)\pi(\theta_i)} = \sum_{k=1}^K C_{ik}, \quad (65)$$

which include the following component from cluster  $k$ :

$$C_{ik} = \frac{w_k}{(2\pi)^{d/2}|\Sigma_k|^{1/2}s_k^d} \exp\left(\frac{-(\theta_i - \bar{\theta}_k)^\top \Sigma_k^{-1}(\theta_i - \bar{\theta}_k)}{2s_k^2}\right) \frac{1}{\mathcal{L}(\theta_i)\pi(\theta_i)}. \quad (66)$$

We solve this optimisation problem by stochastic gradient decent, which requires the gradients of the objective function. Denoting the total cost of the objective function by  $C = \sum_i C_i^2 + \frac{1}{2}\lambda \sum_{k=1}^K s_k^2$ , it is straightforward to show that the gradients of the cost function with respect to the weights  $z_k$  and relative scales  $s_k$  are given by

$$\frac{\partial C}{\partial z_k} = 2 \sum_i C_i (C_{ik} - w_k C_i) \quad (67)$$

and

$$\frac{\partial C}{\partial s_k} = 2 \sum_i \frac{C_i C_{ik}}{s_k^3} \left( (\theta_i - \bar{\theta}_k)^\top \Sigma_k^{-1} (\theta_i - \bar{\theta}_k) - ds_k^2 \right), \quad (68)$$

respectively.

The general procedure to learn the target distribution is the same as before: first, construct a normalised model; second, train the model by solving an optimisation problem to minimise the variance of the resulting learnt harmonic mean estimator. In this case we regularise the relative scale parameters and then solve by stochastic gradient descent. The number of clusters  $K$  can be determined by cross-validation (or other methods). While the modified Gaussian mixture model can effectively handle multimodal distributions, alternative models are better suited to narrow curving posterior degeneracies.

### 3.4.3 Kernel density estimation

Kernel density estimation (KDE) provides another alternative model to learn an effective target distribution. In particular, it can be used to effectively model narrow curving posterior degeneracies.

Consider the target distribution defined by the kernel density function

$$\varphi(\theta) = \frac{1}{N} \sum_i \frac{1}{V_K} K(\theta - \theta_i), \quad (69)$$

with kernel

$$K(\theta) = k\left(\frac{\theta^\top \Sigma_K^{-1} \theta}{R^2}\right), \quad (70)$$

where  $k(\theta) = 1$  if  $|\theta| < 1/2$  and 0 otherwise. The volume of the kernel is given by

$$V_K = \frac{\pi^{d/2}}{\Gamma(d/2 + 1)} R^d |\Sigma_K|^{1/2}. \quad (71)$$

The kernel covariance  $\Sigma_K$  can be computed directly from the training samples, for example by estimating the covariance or even simply by the separation between the lowest and highest samples in each dimension. A diagonal representation is often, although not always, considered for computational efficiency.

The kernel radius  $R$  can be estimated by following a similar procedure to those outlined above for the hypersphere and modified Gaussian mixture model to minimise the variance of the resulting estimator. Alternatively, since there is only a single parameter cross-validation is also effective.

## 4 Numerical experiments

We perform numerous numerical experiments to validate the learnt harmonic mean estimator by comparing to ground truth marginal likelihood values for a variety of example problems. The techniques presented in Sec. 3 are implemented in the `harmonic` software package<sup>2</sup>, which is discussed further in Sec. 5. Throughout we use `harmonic` with the `emcee` code<sup>3</sup> (Foreman-Mackey et al, 2013) to perform MCMC sampling. We consider problems with narrow curving posterior degeneracies, multimodal distributions, and scenarios where the original harmonic mean estimator has been shown to fail catastrophically, while applying all three of the strategies to learn the target density  $\varphi(\theta)$  that are discussed in Sec. 3.4. In all cases the learnt harmonic mean estimator is shown to be robust and highly accurate.

### 4.1 Rosenbrock

A common benchmark problem to test methods to compute the marginal likelihood is a likelihood specified by the Rosenbrock function. The Rosenbrock function exhibits a narrow curving degeneracy, which makes it challenging to explore the resulting posterior sufficiently to evaluate the marginal likelihood accurately.

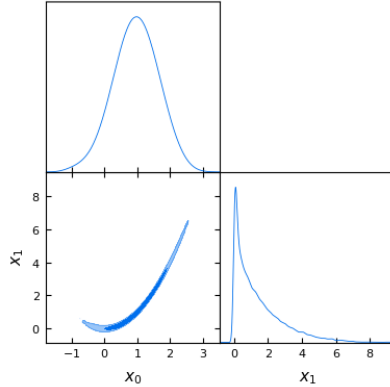
The Rosenbrock function is given by

$$f(x) = \sum_{i=1}^{d-1} \left[ 100(x_{i+1} - x_i^2)^2 + (x_i - 1)^2 \right], \quad (72)$$

where  $d$  denotes dimension. Due to its very narrow curving degeneracy, it can be difficult to numerically estimate the minimum of the Rosenbrock function, which can be seen analytically is given by  $f(x_{\min}) = 0$  at  $x_{\min} = (1, \dots, 1)$ .

<sup>2</sup><https://github.com/astro-informatics/harmonic>

<sup>3</sup><https://emcee.readthedocs.io/en/stable/>

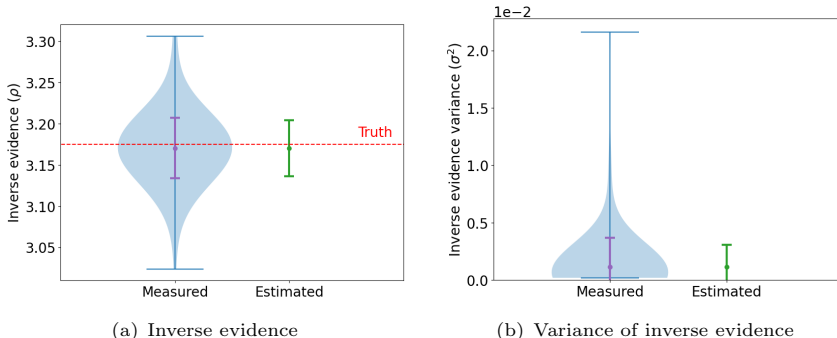


**Fig. 1** Rosenbrock posterior recovered by MCMC sampling using `emcee`. The Rosenbrock function exhibits a narrow curving degeneracy, which makes it challenging to explore the resulting posterior sufficiently to evaluate the marginal likelihood accurately.

We consider a log-likelihood given by  $\log \mathcal{L}(x) = -f(x)$  and consider a simple uniform prior with  $x_0 \in [-10, 10]$  and  $x_1 \in [-5, 15]$ .

We compute the marginal likelihood for dimension  $d = 2$  using the learnt harmonic mean estimator and by brute force to provide a ground truth for comparison, evaluating the marginal likelihood by numerical integration (which is possible in this low-dimensional setting). For our learnt harmonic mean estimator, computed using `harmonic`, we sample the resulting posterior distribution using `emcee`, drawing 5,000 samples for 200 chains, with burn in of 2,000 samples, yielding 3,000 posterior samples per chain. The recovered posterior distribution is illustrated in Fig. 1. We use 50% of the samples to fit a KDE model for the target distribution (recall that the KDE model is well-suited to problems with narrow curving degeneracies), using cross-validation to estimate the model hyperparameters. The remaining 50% of posterior samples are used to infer the marginal likelihood. Computation time is about one minute to compute on a standard laptop, including drawing all samples and performing cross-validation. We repeat this experiment 100 times in order to estimate the variance of the estimator and its variance, in order to compare to the variance and variance-of-variance estimators described in Sec. 3.2.

The distribution of marginal likelihood values compute by our learnt harmonic mean estimator for all 100 experiments are shown in Fig. 2. In addition, we show the values computed by the variance and variance-of-variance estimators (estimated) and compare them to the corresponding statistics measured from the 100 experiments (measured). Moreover, we plot the ground truth value computed by numerical integration. The marginal likelihood value computed is in close agreement with the ground truth and the variance and variance-of-variance estimators are in close agreement with the values computed from the experiments. It is clear that the learnt harmonic mean estimator is highly accurate, its variance is well-behaved and its error estimators are also highly accurate.



**Fig. 2** Marginal likelihood (evidence) computed by the learnt harmonic mean estimator (using `harmonic`) for the Rosenbrock benchmark problem. 100 experiments are repeated to recover empirical estimates of the statistics of the estimator. In panel (a) the distribution of marginal likelihood values computed by the learnt harmonic mean estimator are shown, with the mean and standard deviation of the distribution also shown (measured). For comparison the estimate of the standard deviation computed by the error estimator is also shown (estimated). The ground truth estimated by numerical integration is indicated by the red dashed line. In panel (b) the distribution of the variance of the estimator are shown, with its mean and standard deviation (estimated). For comparison the standard deviation computed by the variance-of-variance estimator is also shown (estimated). The learnt harmonic mean estimator and its error estimators are highly accurate.

## 4.2 Rastrigin

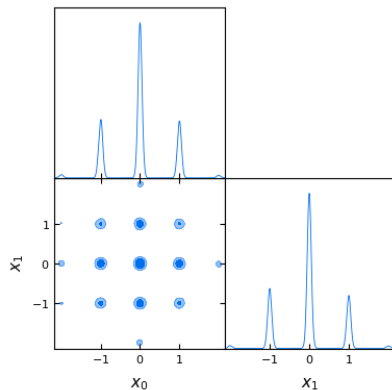
Another common benchmark problem to test marginal likelihood estimators is a likelihood specified by the Rastrigin function. The Rastrigin function exhibits multiple local peaks, which makes it challenging to explore the resulting posterior sufficiently to evaluate the marginal likelihood accurately.

The Rastrigin function is given by

$$f(x) = 10d + \sum_{i=1}^d [x_i^2 - 10 \cos(2\pi x_i)], \quad (73)$$

where  $d$  denotes dimension. Due to its highly multimodal behaviour, it can be difficult to numerically estimate the minimum of the Rastrigin function. Its local minima are given by integer coordinate values, with the global minimum at  $x_{\min} = 0$ . We consider a log-likelihood given by  $\log \mathcal{L}(x) = -f(x)$  and consider a simple uniform prior with  $x_i \in [-6, 6]$  for  $i = 1, \dots, d$ .

We compute the marginal likelihood for dimension  $d = 2$  in an identical manner as for the Rosenbrock example, that is, using the learnt harmonic mean estimator and by brute force to provide a ground truth for comparison, evaluating the marginal likelihood by numerical integration (which, again, is possible in this low-dimensional setting). For our learnt harmonic mean estimator, computed using `harmonic`, we sample the resulting posterior distribution using `emcee`, drawing 5,000 samples for 200 chains, with burn in of 2,000 samples, yielding 3,000 posterior samples per chain. The recovered posterior distribution



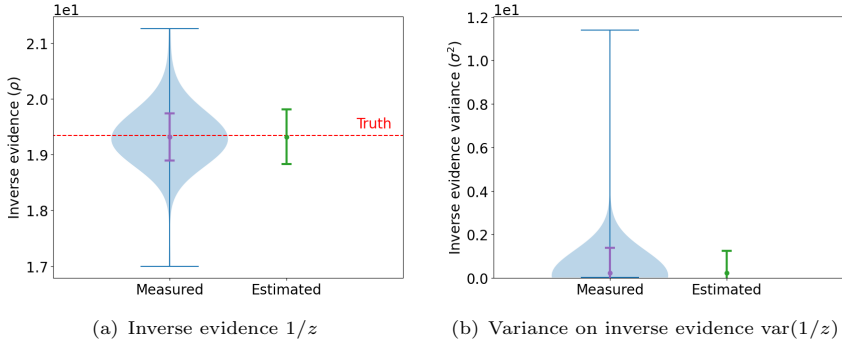
**Fig. 3** Rastrigin posterior recovered by MCMC sampling using `emcee`. The Rastrigin function exhibits multiple local peaks, which makes it challenging to explore the resulting posterior sufficiently to evaluate the marginal likelihood accurately.

is illustrated in Fig. 3. We again adopt a KDE model for the target distribution, avoiding the need to estimate the number of modes in the distribution, and use 50% of the samples to fit the model, using cross-validation to estimate the model hyperparameters. The remaining 50% of posterior samples are used to infer the marginal likelihood. Computation time is about one minute to compute on a standard laptop, including drawing all samples and performing cross-validation. We again repeat this experiment 100 times in order to estimate the variance of the estimator and its variance, in order to compare to the variance and variance-of-variance estimators.

The distribution of marginal likelihood values computed by our learnt harmonic mean estimator for all 100 experiments are shown in Fig. 4. As before, we also show the values computed by the variance and variance-of-variance estimators (estimated) and compare them to the corresponding statistics measured from the 100 experiments (measured). Moreover, we plot the ground truth value computed by numerical integration. It is again clear that the learnt harmonic mean estimator is highly accurate, its variance is well-behaved and its error estimators are also highly accurate.

### 4.3 Normal-Gamma

An analytically tractable numerical example is considered in [Friel and Wyse \(2012\)](#) to assess the sensitivity of marginal likelihood estimators to changes in the prior. In this study [Friel and Wyse \(2012\)](#) found that the marginal likelihood values computed by the original harmonic mean estimator do not vary with the prior as the values computed analytically do, highlighting this example as a pathological failure of the original harmonic mean estimator. We consider the same pathological example here and demonstrate that our learnt harmonic mean estimator is highly accurate.



**Fig. 4** Marginal likelihood (evidence) computed by the learnt harmonic mean estimator (using `harmonic`) for the Rastrigin benchmark problem (as in Fig. 2). 100 experiments are repeated to recover empirical estimates of the statistics of the estimator. In panel (a) the distribution of marginal likelihood values computed by the learnt harmonic mean estimator are shown, with the mean and standard deviation of the distribution also shown (measured). For comparison the estimate of the standard deviation computed by the error estimator is also shown (estimated). The ground truth estimated by numerical integration is indicated by the red dashed line. In panel (b) the distribution of the variance estimator is shown, with its mean and standard deviation (estimated). For comparison the standard deviation computed by the variance-of-variance estimator is also shown (estimated). The learnt harmonic mean estimator and its error estimators are highly accurate.

We consider the *Normal-Gamma model* (Bernardo and Smith, 1994) with data

$$y_i \sim N(\mu, \tau^{-1}), \quad (74)$$

for  $i \in \{1, \dots, n\}$ , with mean  $\mu$  and precision (inverse variance)  $\tau$ . A normal prior is assumed for  $\mu$  and a Gamma prior for  $\tau$ :

$$\mu \sim N(\mu_0, (\tau_0 \tau)^{-1}), \quad (75)$$

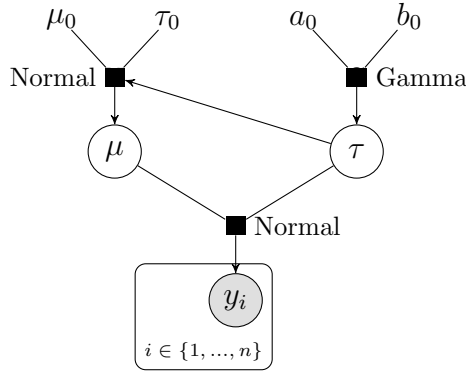
$$\tau \sim \text{Ga}(a_0, b_0), \quad (76)$$

with mean  $\mu_0 = 0$ , shape  $a_0 = 10^{-3}$  and rate  $b_0 = 10^{-3}$ . The precision scale factor  $\tau_0$  is varied to observe the impact of changing prior on the computed marginal likelihood. The joint prior for  $(\mu, \tau)$  then reads:

$$\pi(\mu, \tau) = \pi(\mu | \tau) \pi(\tau) \quad (77)$$

$$= \frac{b_0^{a_0} \sqrt{\tau_0}}{\Gamma(a_0) \sqrt{2\pi}} \tau^{a_0-1/2} \exp(-b_0 \tau) \exp(-\tau_0 \tau (\mu - \mu_0)^2 / 2). \quad (78)$$





**Fig. 5** Graphical representation of the Normal-Gamma model.

The likelihood is given by

$$\mathcal{L}(y) = \prod_{i=1}^n P(y_i | \mu, \tau) \tag{79}$$

$$= \prod_{i=1}^n \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{\tau}{2}(y_i - \mu)^2\right) \tag{80}$$

$$= \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left(-\frac{\tau}{2} \sum_{i=1}^n (y_i - \mu)^2\right) \tag{81}$$

$$= \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left(-\frac{\tau n}{2}(s^2 + (\bar{y} - \mu)^2)\right), \tag{82}$$

where  $y = (y_1, \dots, y_n)^T$ ,

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \tag{83}$$

and

$$s^2 = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2. \tag{84}$$

A graphical representation of the Normal-Gamma model is illustrated in Fig. 5.

For the Normal-Gamma model the marginal likelihood may be computed analytically by

$$z = (2\pi)^{-n/2} \frac{\Gamma(a_n) b_0^{a_0}}{\Gamma(a_0) b_n^{a_n}} \left(\frac{\tau_0}{\tau_n}\right)^{1/2}, \tag{85}$$

where

$$\tau_n = \tau_0 + n, \tag{86}$$

$$a_n = a_0 + n/2 \tag{87}$$

**Table 1** Marginal likelihood values computed analytically and by the learnt harmonic mean estimator for the Normal-Gamma example. While the original harmonic mean estimator fails catastrophically, our learnt harmonic mean estimator is highly accurate.

$\tau_0$	$10^{-4}$	$10^{-3}$	$10^{-2}$	$10^{-1}$	$10^0$
Analytic $\log(z)$	-144.5530	-143.4017	-142.2505	-141.0999	-139.9552
Estimated $\log(\hat{z})$	-144.5545	-143.3990	-142.2490	-141.1001	-139.9558
Error (learnt harmonic mean)	-0.0015	0.0027	0.0015	-0.0011	-0.0006
Error (original harmonic mean)	12.2100	—	9.7900	8.5000	7.1000

and

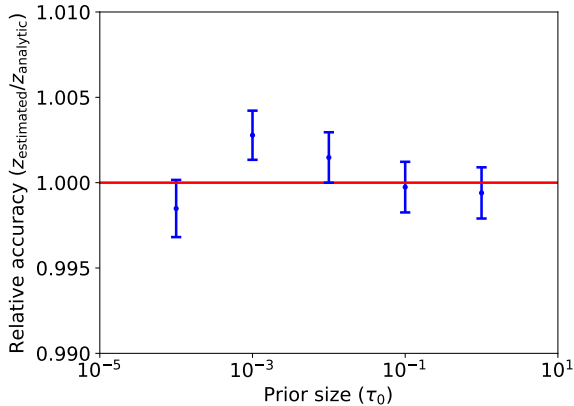
$$b_n = b_0 + \frac{1}{2} \sum_{i=1}^n (y_i - \bar{y})^2 + \frac{\tau_0 n (\bar{y} - \mu_0)^2}{2(\tau_0 + n)} \quad (88)$$

$$= b_0 + \frac{1}{2} n s^2 + \frac{\tau_0 n (\bar{y} - \mu_0)^2}{2(\tau_0 + n)}. \quad (89)$$

To assess the impact of altering the prior, we compute the marginal likelihood both analytically and using our learnt harmonic mean estimator for priors corresponding to  $\tau_0 \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^0\}$ . Data are simulated with underlying parameters  $(\mu, \tau) = (0, 1)$  to generate  $n = 100$  synthetic observations (the same experimental configuration considered by [Friel and Wyse 2012](#)).

For our learnt harmonic mean estimator we use `emcee` to draw 1,500 samples for 200 chains, with burn in of 500 samples, yielding 1,000 posterior samples per chain. We use 25% of the samples to learn the target model, using cross-validation to select between the hypersphere and modified Gaussian mixture model. In all cases the modified Gaussian mixture model is selected. The remaining 75% of posterior samples are used for inferring the marginal likelihood. Computation time is about one minute on a standard laptop for each experiment (i.e. each  $\tau_0$ ) considered, including drawing all samples.

The marginal likelihood values computed analytically and using our learnt harmonic mean estimator are shown in Table 1 for different priors as  $\tau_0$  is varied. For comparison, the errors between the analytic values and those estimated by our learnt harmonic mean estimator and the original harmonic mean estimator are also shown. Notice that the marginal likelihood values computed by the learnt harmonic mean estimator are highly accurate and do indeed vary with differing priors, in contrast to results computed by the original harmonic mean estimator ([Friel and Wyse, 2012](#)). The improvement in accuracy between the original and our learnt harmonic mean estimation is approximately four orders of magnitude in log space. In addition, to graphically compare the marginal likelihood values estimated by the learnt harmonic mean estimator to the analytic values, we plot in Fig. 6 the ratio of the estimated and analytic



**Fig. 6** Ratio of marginal likelihood values computed by the learnt harmonic mean estimator to those computed analytically. Errors bars corresponding to the estimated standard deviation of the learnt harmonic estimator are also shown. Notice that the marginal likelihood values computed by the learnt harmonic mean estimator are highly accurate and are indeed sensitive to changes in the prior.

values, with the uncertainties computed by our learnt harmonic mean estimator overlaid. It is clear that the marginal likelihood values computed by the learnt harmonic mean estimator closely estimate the analytic values and that the uncertainties are reasonable.

#### 4.4 Logistic regression models: Pima Indian example

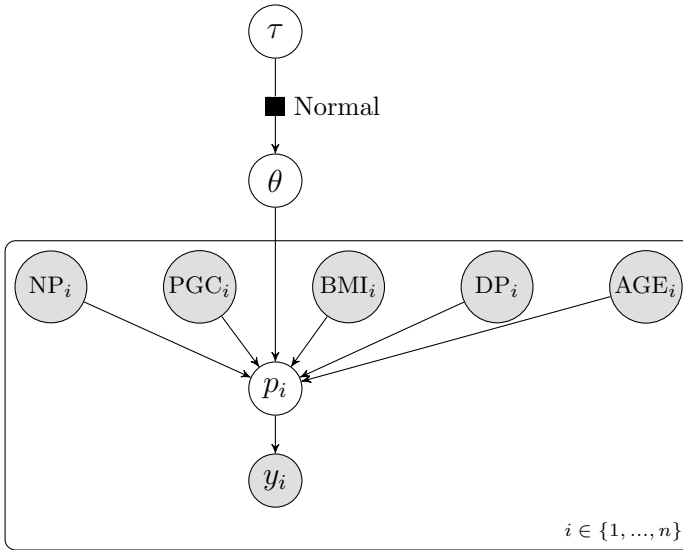
We consider the comparison of two logistic regression models using the *Pima Indians* data, which is another common benchmark problem for comparing estimators of the marginal likelihood. The original harmonic mean estimator has been shown to fail catastrophically for this example (Friel and Wyse, 2012), whereas we show here that our learnt harmonic mean estimator is highly accurate.

The Pima Indians data (Smith et al, 1988), originally from the National Institute of Diabetes and Digestive and Kidney Diseases, were compiled from a study of indicators of diabetes in  $n = 532$  Pima Indian women aged 21 or over. Seven primary predictors of diabetes were recorded, including: number of prior pregnancies (NP); plasma glucose concentration (PGC); diastolic blood pressure (BP); triceps skin fold thickness (TST); body mass index (BMI); diabetes pedigree function (DP); and age (AGE).

The probability of diabetes  $p_i$  for person  $i \in \{1, \dots, n\}$  can be modelled by the standard logistic function

$$p_i = \frac{1}{1 + \exp(-\theta^T x_i)}, \quad (90)$$

with covariates  $x_i = (1, x_{i,1}, \dots, x_{i,d})^T$  and parameters  $\theta = (\theta_0, \dots, \theta_d)^T$ , where  $d$  is the total number of covariates considered. The likelihood function then



**Fig. 7** Graphical representation of logistic regression Model 2 for modelling diabetes in Pima Indians. Model 1 is similar but does not include the AGE covariate.

reads

$$\mathcal{L}(y|\theta) = \prod_{i=1}^n p_i^{y_i} (1 - p_i)^{1-y_i}, \quad (91)$$

where  $y = (y_1, \dots, y_n)^T$  is the diabetes incidence, *i.e.*  $y_i$  is unity if patient  $i$  has diabetes and zero otherwise. An independent multivariate Gaussian prior is assumed for the parameters  $\theta$ , given by

$$\pi(\theta) = \left(\frac{\tau}{2\pi}\right)^{d/2} \exp\left(-\frac{\tau}{2}\theta^T\theta\right), \quad (92)$$

with precision  $\tau$ .

Two different logistic regression models are considered, with different subsets of covariates:

Model  $M_1$  : covariates = {NP, PGC, BMI, DP} (and bias);

Model  $M_2$  : covariates = {NP, PGC, BMI, DP, AGE} (and bias).

A graphical representation of Model 2 is illustrated in Fig. 7 (Model 1 is similar but does not include the AGE covariate).

We compute the marginal likelihood for both Model 1 and Model 2 using our learnt harmonic mean estimator for  $\tau = 0.01$  and  $\tau = 1$ , as in Friel and Wyse (2012). A reversible jump algorithm (Green, 1995) is used by Friel and Wyse (2012) to compute benchmark Bayes factors  $\text{BF}_{12}$  of 13.96 and 1.30, respectively, for  $\tau = 0.01$  and  $\tau = 1$ , which are treated as ground truth values.

**Table 2** Marginal likelihood values computed by the learnt harmonic mean estimator for the Pima Indians logistic regression models for prior precision  $\tau = 0.01$ . While the original harmonic mean estimator fails catastrophically, our learnt harmonic mean estimator is highly accurate.

	Model $M_1$ $\log(z_1)$	Model $M_2$ $\log(z_2)$	$\log \text{BF}_{12}$ $= \log(z_1) - \log(z_2)$
Benchmark	–	–	2.63620
Estimated	-257.23656 $\pm 0.00264$	-259.86669 $\pm 0.00968$	2.63014 $\pm 0.01232$
Error (learnt harmonic mean)	–	–	0.00606
Error (original harmonic mean)	–	–	-2.67760

**Table 3** Marginal likelihood values computed by the learnt harmonic mean estimator for the Pima Indians logistic regression models for prior precision  $\tau = 1.0$ . While the original harmonic mean estimator fails catastrophically, our learnt harmonic mean estimator is highly accurate.

	Model $M_1$ $\log(z_1)$	Model $M_2$ $\log(z_2)$	$\log \text{BF}_{12}$ $= \log(z_1) - \log(z_2)$
Benchmark	–	–	0.26236
Estimated	-247.30633 $\pm 0.00239$	-247.56128 $\pm 0.00789$	0.25495 $\pm 0.01028$
Error (learnt harmonic mean)	–	–	0.00742
Error (original harmonic mean)	–	–	-0.44567

For our learnt harmonic mean estimator we use `emcee` to draw 5,000 samples for 200 chains, with burn in of 1,000 samples, yielding 4,000 posterior samples per chain. We use 25% of the samples to learn the target model, using cross-validation to select between the hypersphere and modified Gaussian mixture model. In all cases the modified Gaussian mixture model is selected. The remaining 75% of posterior samples are used for inferring the marginal likelihood. Computation time is typically a few minutes on a standard laptop, including drawing samples (note that fewer samples could likely be used to reduce computation time if required).

The marginal likelihood values computed by our learnt harmonic mean estimator are shown in Table 2 and Table 3 for the cases  $\tau = 0.01$  and  $\tau = 1$ , respectively. The benchmark values and errors of both the standard and learnt harmonic mean estimator are also shown for comparison. While the standard harmonic mean estimator fails catastrophically on this problem (Friel and Wyse, 2012), our learnt harmonic mean estimator is robust and highly accurate.

## 4.5 Non-nested linear regression models: Radiata pine example

We consider another example where the original harmonic mean estimator was shown to fail catastrophically (Friel and Wyse, 2012). In particular, we consider non-nested linear regression models for the *Radiata pine* data, which is another common benchmark data-set (Williams, 1959), and show that our learnt harmonic mean estimator is highly accurate.

For  $n = 42$  trees, the Radiata pine data-set includes measurements of the maximum compression strength parallel to the grain  $y_i$ , density  $x_i$  and resin-adjusted density  $z_i$ , for specimen  $i \in \{1, \dots, n\}$ . The question at hand is whether density or resin-adjusted density is a better predictor of compression strength. This motivates two Gaussian linear regression models:

$$\text{Model } M_1 : \quad y_i = \alpha + \beta(x_i - \bar{x}) + \epsilon_i, \quad \epsilon_i \sim \text{N}(0, \tau^{-1}); \quad (93)$$

$$\text{Model } M_2 : \quad y_i = \gamma + \delta(z_i - \bar{z}) + \eta_i, \quad \eta_i \sim \text{N}(0, \lambda^{-1}), \quad (94)$$

where  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ ,  $\bar{z} = \frac{1}{n} \sum_{i=1}^n z_i$ , and  $\tau$  and  $\lambda$  denote the precision (inverse variance) of the noise for the respective models.

For Model 1, Gaussian priors are assumed for the bias and linear terms:

$$\alpha \sim \text{N}(\mu_\alpha, (r_0\tau)^{-1}); \quad (95)$$

$$\beta \sim \text{N}(\mu_\beta, (s_0\tau)^{-1}), \quad (96)$$

with means  $\mu_\alpha = 3000$  and  $\mu_\beta = 185$ , and precision scales  $r_0 = 0.06$  and  $s_0 = 6$ . A gamma prior is assumed for the noise precision:

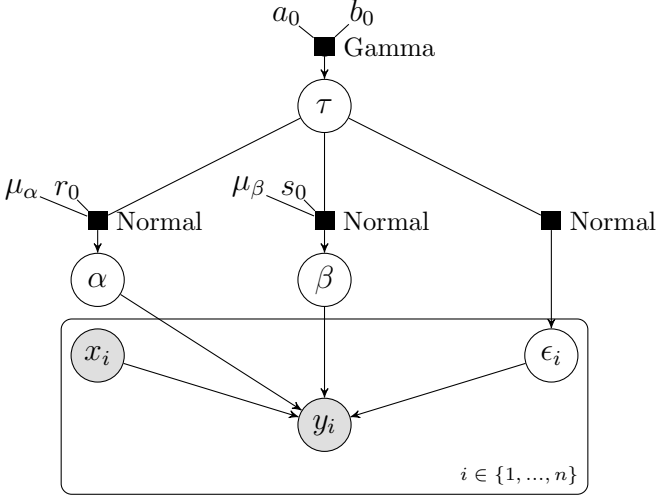
$$\tau \sim \text{Ga}(a_0, b_0), \quad (97)$$

with shape  $a_0 = 3$  and rate  $b_0 = 2 \times 300^2$ . The joint prior for  $(\alpha, \beta, \tau)$  then reads:

$$\pi(\alpha, \beta, \tau) = \pi(\alpha, \beta | \tau)\pi(\tau) \quad (98)$$

$$= \pi(\alpha | \tau)\pi(\beta | \tau)\pi(\tau) \quad (99)$$

$$= \frac{(b_0\tau_0)^{a_0}(r_0s_0)^{1/2}}{2\pi\Gamma(a_0)} \exp(-b_0\tau) \\ \times \exp\left(-\frac{\tau}{2}(r_0(\alpha - \mu_\alpha)^2 + s_0(\beta - \mu_\beta)^2)\right). \quad (100)$$



**Fig. 8** Graphical representation of the non-nested linear regression Model 1 for modelling maximum compression strength for Radiata pine. Model 2 is similar.

The likelihood for Model 1 is given by

$$\mathcal{L}(x, y) = \prod_{i=1}^n P(x_i, y_i | \alpha, \beta, \tau) \quad (101)$$

$$= \prod_{i=1}^n \sqrt{\frac{\tau}{2\pi}} \exp\left(-\frac{\tau}{2} (y_i - \alpha - \beta(x_i - \bar{x}))^2\right) \quad (102)$$

$$= \left(\frac{\tau}{2\pi}\right)^{n/2} \exp\left(-\frac{\tau}{2} \sum_{i=1}^n (y_i - \alpha - \beta(x_i - \bar{x}))^2\right), \quad (103)$$

where  $x = (x_1, \dots, x_n)^T$  and  $y = (y_1, \dots, y_n)^T$ . For Model 2, the priors adopted for  $(\gamma, \delta, \lambda)$  are the same as those adopted for  $(\alpha, \beta, \tau)$  of Model 1, respectively, with the same hyperparameters. The likelihood for Model 2 again takes an identical form to Model 1. A graphical representation of Model 1 is illustrated in Fig. 8 (Model 2 is similar).

One reason this problem has become a common benchmark for comparing marginal likelihood estimators is that the marginal likelihood of the two models can be computed analytically. The evidence for Model 1 is given by

$$z = \frac{(2b_0)^{a_0}}{\pi^{n/2}} \frac{\Gamma(a_0 + n/2)}{\Gamma(a_0)} \frac{|Q_0|^{1/2}}{|M|^{1/2}} (y^T y + \mu_0^T Q_0 \mu_0 - \nu_0^T M \nu_0 + 2b_0)^{-a_0 - n/2}, \quad (104)$$

where  $\mu_0 = (\mu_\alpha, \mu_\beta)^T$ ,  $Q_0 = \text{diag}(r_0, s_0)$ ,  $M = X^T X + Q_0$ , and  $\nu_0 = M^{-1}(X^T y + Q_0 \mu_0)$ , for the feature matrix  $X$  with row  $i$  containing  $(1, x_i - \bar{x})$ .

**Table 4** Marginal likelihood values computed analytically and by the learnt harmonic mean estimator for the Radiata pine non-nested linear regression models. While the original harmonic mean estimator fails catastrophically, our learnt harmonic mean estimator is highly accurate.

	Model $M_1$ $\log(z_1)$	Model $M_2$ $\log(z_2)$	$\log \text{BF}_{21}$ $= \log(z_2) - \log(z_1)$
Analytic	-310.12829	-301.70460	8.42368
Estimated	-310.12807	-301.70413	8.42394
	$\pm 0.00072$	$\pm 0.00074$	$\pm 0.00145$
Error (learnt harmonic mean)	0.00022	0.00047	0.00026
Error (original harmonic mean)	–	–	-0.17372

We compute the marginal likelihood analytically using (104) and also numerically using our learnt harmonic mean estimator. For the learnt harmonic mean estimator we use `emcee` to draw 20,000 samples for 400 chains, with burn in of 2,000 samples. We use 25% of the samples to learn the target model, adopting a simple hypersphere model, and use the remaining 75% for inferring the marginal likelihood. Computation time is a few minutes on a standard laptop, including drawing samples (note that fewer samples could likely be used to reduce computation time if required).

The analytic and estimated marginal likelihood values are shown in Table 4 for the two models, with the Bayes factor comparing the two models. For the values computed by our learnt harmonic mean estimator we also show the estimated uncertainty. The errors of both the standard and learnt harmonic mean estimator are shown for comparison. Note that the uncertainties estimated by the learnt harmonic mean estimator appear reasonable. While the standard harmonic mean estimator fails catastrophically on this problem (Friel and Wyse, 2012), our learnt harmonic mean estimator is highly accurate.

## 4.6 Gaussian in varying dimensions

Finally, we illustrate the application of our estimator beyond low-dimensional settings, considering experiments where the dimension of the parameter space increases. For simplicity we consider a Gaussian likelihood with a uniform prior, where the marginal likelihood can be computed analytically. We adopt the simple hypersphere model, which is effective for a Gaussian posterior. Results are illustrated in Table 5. Note that parameters were not optimised and accurate results could likely be obtained with fewer samples. Further note that the computation times recorded do not include the time accumulated during initial burn-in. It is apparent that the estimator is accurate in settings with dimensions  $\mathcal{O}(10^3)$  and potentially beyond.



**Table 5** Marginal likelihood values computed analytically and by the learnt harmonic mean estimator for a Gaussian posterior in varying dimensions. Note that parameters were not optimised and results could likely be computed to comparable accuracy with fewer samples (i.e. with lower computation time). Further note that the computation times recorded do not include the time accumulated during initial burn-in.

Dimension	Analytic $\log(z)$	Estimated $\log(z)$	Error (%)	Computation time
32	-29.406	-29.411	0.0180%	~10 sec
64	-58.812	-58.813	0.0008%	~20 sec
128	-117.62	-117.63	0.0026%	~3 min
256	-235.25	-235.25	0.0015%	~18 min
512	-470.50	-470.49	0.0006%	~20 min
1024	-940.99	-941.06	0.0073%	~3 hours

## 5 Software package

The learnt harmonic mean estimator is implemented in the `harmonic` software package<sup>4</sup>, which is open source and publicly available. Careful consideration has been given to the design and implementation of the code, following software engineer best practices (for example, at the time of release test coverage is over 96%).

Since the learnt harmonic mean estimator requires samples from the posterior distribution only, the `harmonic` code is agnostic to the method or code used to generate posterior samples. That said, `harmonic` works exceptionally well with MCMC sampling techniques that naturally provide samples from multiple chains by their ensemble nature, such as affine invariance ensemble samplers (Goodman and Weare, 2010). As discussed in Sec. 2, we advocate running a number of independent MCMC chains and using all of the correlated samples within a chain to avoid the loss of efficiency that otherwise results from thinning an MCMC chain. The `emcee` code<sup>5</sup> (Foreman-Mackey et al, 2013) provides an excellent implementation of the affine invariance ensemble samplers proposed by Goodman and Weare (2010). `emcee` is thus a natural choice for use with `harmonic` and we have specifically designed `harmonic` to ensure it works seamlessly with `emcee` (although of course other samplers can also be considered). In code Listing 1 we give an example of usage of `harmonic` with `emcee` to demonstrate how easy it is to use the combination for marginal likelihood estimation.

<sup>4</sup><https://github.com/astro-informatics/harmonic>

<sup>5</sup><https://emcee.readthedocs.io/en/stable/>

```

1 import numpy as np
2 import emcee
3 import harmonic
4
5 # Run sampler
6 sampler = emcee.EnsembleSampler(nchains, ndim, ln_posterior,
7                               args=[posterior_args])
8 (pos, prob, state) = sampler.run_mcmc(pos, samples_per_chain)
9 samples = np.ascontiguousarray(sampler.chain[:, nburn:, :])
10 lnprob = np.ascontiguousarray(sampler.lnprobability[:, nburn:])
11
12 # Set up chains
13 chains = harmonic.Chains(ndim)
14 chains.add_chains_3d(samples, lnprob)
15 chains_train, chains_infer = \
16     harmonic.utils.split_data(chains, training_prop)
17
18 # Fit model
19 model = harmonic.model.KernelDensityEstimate(ndim, domain,
20                                              hyper_parameters)
21 model.fit(chains_train.samples, chains_train.ln_posterior)
22
23 # Compute evidence
24 ev = harmonic.Evidence(chains_infer.nchains, model)
25 ev.add_chains(chains_infer)
26 ln_evidence, ln_evidence_std = ev.compute_ln_evidence()

```

**Listing 1** Example usage of `harmonic` to compute the marginal likelihood, using `emcee` to perform MCMC sampling.

## 6 Conclusions

We present the learnt harmonic mean estimator to solve the problematic large variance of the original estimator. The construction of our estimator follows by interpreting the harmonic mean estimator as importance sampling and introducing a new target distribution that is learned to approximate the optimal but inaccessible target (the normalised posterior), while minimising the variance of the resulting estimator. We discuss techniques to compute the variance of the estimator, its variance and to perform a number of additional computational sanity checks. The estimator is implemented in the publicly available `harmonic` software code. We demonstrate the application of our learnt harmonic mean estimator on numerous benchmark problems, including a number of pathological examples where the original harmonic mean estimator fails catastrophically. In all cases our estimator is robust and highly accurate. The current work opens up a number of avenues for future research. For example, similar approaches can be taken in MCMC sampling more generally where appropriate target, sampling densities or proposal distributions may be learned. Alternative more effective target models can be developed that better scale to higher dimensional settings. Since the learnt harmonic mean estimator is agnostic to the sampling strategy, it is also an ideal solution for computing the marginal likelihood for model comparison in simulation-based inference. We are already actively pursuing this avenue of research, with promising preliminary results.

**Acknowledgments.** This work was supported by the Leverhulme Trust and by EPSRC grant EP/W007673/1. For the purpose of open access, the authors

have applied a Creative Commons Attribution (CC BY) licence to any Author Accepted Manuscript version arising.

## References

- Bernardo J, Smith A (1994) Bayesian theory. New York
- Brewer BJ, Pártay LB, Csányi G (2011) Diffusive nested sampling. *Statistics and Computing* 21:649–656
- Cai X, McEwen JD, Pereyra M (2021) High-dimensional Bayesian model selection by proximal nested sampling. *SIAM Journal on Imaging Sciences*, submitted <https://arxiv.org/abs/arXiv:2106.03646>
- Chib S (1995) Marginal likelihood from the gibbs output. *Journal of the American Statistical Association* 90(432):1313–1321
- Chib S, Jeliazkov I (2001) Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association* 96:270–281
- Cho E, Cho MJ, Eltinge J (2005) The variance of sample variance from a finite population. *International Journal of Pure and Applied Mathematics* 21(3):389
- Clyde M, Berger J, Bullard F, et al (2007) Current challenges in bayesian model choice. In: *Statistical challenges in modern astronomy IV*, p 224
- Feroz F, Hobson MP (2008) Multimodal nested sampling: an efficient and robust alternative to Markov Chain Monte Carlo methods for astronomical data analyses. *Mon Not Roy Astron Soc* 384:449–463. <https://doi.org/10.1111/j.1365-2966.2007.12353.x>, <https://arxiv.org/abs/arXiv:0704.3704>
- Feroz F, Hobson MP, Bridges M (2009) MultiNest: an efficient and robust bayesian inference tool for cosmology and particle physics. *Mon Not Roy Astron Soc* 398:1601–1614. <https://doi.org/10.1111/j.1365-2966.2009.14548.x>, <https://arxiv.org/abs/arXiv:0809.3437>
- Foreman-Mackey D, Hogg DW, Lang D, et al (2013) emcee: The mcmc hammer. *PASP* 125:306–312. <https://doi.org/10.1086/670067>, <https://arxiv.org/abs/1202.3665>
- Friel N, Wyse J (2012) Estimating the evidence – a review. *Statistica Neerlandica* 66(3):288–308. <https://doi.org/10.1111/j.1467-9574.2011.00515.x>, URL <http://dx.doi.org/10.1111/j.1467-9574.2011.00515.x>, <https://arxiv.org/abs/arXiv:1111.1957>

- Gelfand AE, Dey DK (1994) Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)* 56(3):501–514
- Goodman J, Weare J (2010) Ensemble samplers with affine invariance. *Communications in applied mathematics and computational science* 5(1):65–80
- Green PJ (1995) Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika* 82(4):711–732
- van Haasteren R (2014) Marginal likelihood calculation with mcmc methods. In: *Gravitational Wave Detection and Data Analysis for Pulsar Timing Arrays*. Springer, p 99–120
- Handley WJ, Hobson MP, Lasenby AN (2015) POLYCHORD: nested sampling for cosmology. *Mon Not Roy Astron Soc* 450:L61–L65. <https://doi.org/10.1093/mnrasl/slv047>, <https://arxiv.org/abs/arXiv:1502.01856>
- Hastings WK (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 57:97–109
- Heavens A, Fantaye Y, Mootoovaloo A, et al (2017) Marginal Likelihoods from Monte Carlo Markov Chains. ArXiv <https://arxiv.org/abs/arXiv:1704.03472> [stat.CO]
- Lenk P (2009) Simulation pseudo-bias correction to the harmonic mean estimator of integrated likelihoods. *Journal of Computational and Graphical Statistics* 18(4):941–960
- Link WA, Eaton MJ (2012) On thinning of chains in mcmc. *Methods in ecology and evolution* 3(1):112–115
- Marshall PJ, Hobson MP, Slosar A (2003) Bayesian joint analysis of cluster weak lensing and Sunyaev-Zel’dovich effect data. *Mon Not Roy Astron Soc* 346:489–500
- Metropolis N, Rosenbluth AW, Rosenbluth MN, et al (1953) Equation of state by fast computing machines. *J Chemical Physics* 21:1087–1092
- Neal R (2001) Annealed importance sampling. *Statistics and Computing* 11:125–139
- Neal RM (1994) Contribution to the discussion of “approximate bayesian inference with the weighted likelihood bootstrap” by newton ma, raftery ae. *JR Stat Soc Ser A (Methodological)* 56:41–42
- Newton MA, Raftery AE (1994) Approximate bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society Series*

- B (Methodological) 56(1):3–48. URL <http://www.jstor.org/stable/2346025>
- O’Ruanaidh J, Fitzgerald WJ (1996) Numerical Bayesian methods applied to signal processing. Springer-Verlag New York
- Pajor A, et al (2017) Estimating the marginal likelihood using the arithmetic mean identity. *Bayesian Analysis* 12(1):261–287
- Raftery AE, Newton MA, Satagopan JM, et al (2006) Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. Preprint
- Robert CP, Wraith D (2009) Computational methods for bayesian model choice. In: Aip conference proceedings, American Institute of Physics, pp 251–262
- Rose C, Smith MD (2002) Mathstatica: mathematical statistics with mathematica. In: *Compstat*, Springer, pp 437–442
- Skilling J (2006) Nested sampling for general Bayesian computation. *Bayesian Analysis* 1:833–859
- Smith JW, Everhart JE, Dickson W, et al (1988) Using the adap learning algorithm to forecast the onset of diabetes mellitus. In: Proceedings of the annual symposium on computer application in medical care, American Medical Informatics Association, p 261
- Tierney L, Kadane JB (1986) Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* 81:82–86
- Trotta R (2007) Applications of Bayesian model selection to cosmological parameters. *Mon Not Roy Astron Soc* 378:72–82
- Williams EJ (1959) Regression analysis. wiley