Efficient prior sensitivity analysis for Bayesian model comparison

Zixiao Hu[1,2], Jason D. McEwen[1,2]
[1]Mullard Space Science Laboratory (MSSL), University College London (UCL), Dorking RH5 6NT, UK
[2]The Alan Turing Institute, London NW1 2DB, UK

Abstract

Bayesian model comparison implements Occam's razor through its sensitivity to the prior. However, prior-dependence makes it important to assess the influence of plausible alternative priors. Such prior sensitivity analyses for the Bayesian evidence are expensive, either requiring repeated, costly model re-fits or specialised sampling schemes. By exploiting the learned harmonic mean estimator (LHME) for evidence calculation we decouple sampling and evidence calculation, allowing resampled posterior draws to be used directly to calculate the evidence without further likelihood evaluations. This provides an alternative approach to prior sensitivity analysis for Bayesian model comparison that dramatically alleviates the computational cost and is agnostic to the method used to generate posterior samples. We validate our method on toy problems and a cosmological case study, reproducing estimates obtained by full Markov chain Monte Carlo (MCMC) sampling and nested sampling re-fits. For the cosmological example considered our approach achieves up to $6000\times$ lower computational cost.

1. INTRODUCTION

While much of modern statistical practice focuses on predictive performance, scientific discovery often necessitates selecting between discrete, theoretically motivated hypotheses [1, 2]. Explicit model comparison consequently persists as a cornerstone of the scientific method. Various approaches exist, including frequentist hypothesis testing, information criteria [3], and cross-validation [4], but Bayesian model comparison is popular in many fields due to its evaluation of *relative plausibility* and principled *penalisation of model complexity* [5, 6, 7]. Central to this endeavour is the calculation of the marginal likelihood, or *Bayesian evidence*, which modern algorithms and software implementations have made feasible in many scenarios [8, 9, 10, 11, 12, 13].

The Bayesian evidence (or just *evidence*) embodies an automatic Occam's razor, naturally penalising complex models that do not sufficiently improve the fit to the data. The penalty arises from its definition as a prior predictive density. Models with vast, wasted parameter space, or those that place probability mass in physically forbidden regions, suffer a reduction in evidence, regardless of their maximum likelihood fit [5, 6].

However, this same mechanism imparts a direct sensitivity to the prior that warrants careful consideration. While the influence of the prior on parameter estimation typically "washes out" with increasing data, the evidence always remains sensitive to the prior volume [14]. Taken to the extreme, the evidence for any model can be made vanishingly small as the support of the prior goes to infinity (sometimes known as the Lindley-Bartlett paradox [15, 16]). Bayesian evidences are often criticised for this fragility; if priors are arbitrary, the resulting model comparison is also arbitrary.

To mitigate this, best practice dictates that priors should be well-motivated [17, 18]. These include physical priors (e.g. non-negative mass or flux; [19]), uninformative Jeffreys priors that are invariant to a parameter transformation [20], maximum entropy priors [21], or informative priors for example to regularise inverse problems (e.g. [22]). Data-driven specifications include those potentially specified by a generative model (e.g. [23, 24, 25, 26]), and data-informed priors, where the posterior of an *a priori* analysis is used as the prior for an analysis with new data (e.g. [27]). Alternative formulations such as fractional [28] or intrinsic Bayes factors [29] also attempt to reduce prior dependency by performing model comparison with priors constructed from subsets of the data. However, as no prior specification is truly devoid of influence, it is important to explore the sensitivity of the evidence to reasonable alternative choices [5, 30].

Despite its importance, such sensitivity analysis is difficult to perform in practice. Formal approaches to sensitivity analysis for posterior inference exist, ranging from global methods that consider entire classes of priors [31] to local methods based on infinitesimal perturbations [32], but these do not extend to Bayes factors. Conversely, the "informal" approach of checking discrete alternative priors is only tractable in simple limiting cases; one example is that under a highly informative likelihood, changing from one uniform prior to another leads the evidence to change by the ratio of the two constant prior densities (see Chapter 28.1 of [6] and e.g. [33, 34]). Some specific path-based sampling schemes for evidence calculation also allow sensitivity checks through re-weighting of a sequence of samples [35, 36]. Otherwise, in nearly all other cases prior dependence can only be assessed by expensive re-fitting. This latter case is the focus of this work.

To address this issue, we present an efficient approach to assess prior dependence by importance resampling existing posterior draws, which can be obtained from *any* sampling method (e.g. any Markov chain Monte Carlo (MCMC) sampling strategy, neural density estimators, and others). Although this is a common technique in sensitivity analysis for posterior inference ([37, 38]), its application to model comparison is underexplored since many effective techniques to compute the evidence are coupled to the sampling strategy, e.g. nested sampling [8, 9]. We leverage the learned harmonic mean estimator of the evidence [12, 13] to decouple sampling and evidence calculation, enabling the resampled draws to be used directly for evidence calculation without further likelihood evaluations.

The remainder of the paper is structured as follows. Section 2 describes our method in detail, including diagnostics to flag unreliable importance sampling and identify when unnecessary recalculation can be avoided. We validate the method on toy problems and a cosmological case study in Section 3, demonstrating up to $6000\times$ speed-ups compared to full MCMC and nested sampling re-runs. Finally, concluding remarks are given in Section 4.

2. METHODS

We first review the learned harmonic mean estimator for evidence calculation (Section 2.1), then describe our resampling approach for prior sensitivity analysis (Section 2.2).

2.1 Bayesian evidence calculation with the learned harmonic mean estimator

Consider a model M with parameters θ, prior $p(\theta \mid M)$, and likelihood $L(D \mid \theta, M)$ given data D, for which we have obtained N samples $\{\theta_i\}_{i=1}^N$ from the posterior $p(\theta \mid D, M)$ using any sampling method, such as MCMC or nested sampling.

The learned harmonic mean estimator (LHME) [12, 13], which builds on the approach of [39], computes the Bayesian evidence by introducing a learned target distribution $\varphi(\theta)$ to stabilise the standard harmonic mean estimator [40, 41]:

$$\hat{\rho} = \frac{1}{N} \sum_{i=1}^{N} \frac{\varphi(\theta_i)}{L(D \mid \theta_i, M)\, p(\theta_i \mid M)}, \quad \theta_i \sim p(\theta \mid D, M). \tag{1}$$

where $\hat{\rho}$ is an unbiased estimator of the reciprocal evidence $\rho = 1/p(D \mid M)$ under the usual Monte Carlo expectation. If $\varphi(\theta)$ is the normalised posterior, $\hat{\rho}$ has zero variance and is therefore optimal [12]. Since this is intractable, the LHME approximates the optimal $\varphi(\theta)$ by learning a density estimator, such as a normalising flow trained on posterior samples [42, 13]. We emphasise that the learned $\varphi(\theta)$ need not be perfectly accurate, but critically must have thinner tails than the posterior for the variance of the estimator to be finite. Various strategies can be performed to ensure this [12, 13]. Recent approaches use normalising flows [13], where the variance of the base distribution of the normalising flow is reduced after training (lowering the "temperature" $T < 1$). This concentrates $\varphi(\theta)$ within the bulk of the posterior samples, avoiding the exploding variance problem of the original estimator.

2.2 Prior sensitivity calculation

We leverage the decoupling of sampling and evidence calculation that the LHME provides to reuse existing posterior samples for prior sensitivity analysis.

Let us begin by describing the general procedure for reusing posterior draws for one model to approximate the evidence under an alternative model, before specialising to prior sensitivity. Consider an original model M_0 and

an alternative model M_1, which are each defined by two likelihoods and priors that may or may not be the same. We are interested in the case where M_0 and M_1 are indexed by the same parameters θ, but which have different distributional specifications. The posteriors for the models are:

$$p(\theta \mid D, M_k) = \frac{L(D \mid \theta, M_k)\, p(\theta \mid M_k)}{p(D \mid M_k)}, \quad k = 0, 1. \tag{2}$$

Given samples $\{\theta_i\}_{i=1}^N$ from the original posterior $p(\theta \mid D, M_0)$, we can obtain samples from the alternative posterior $p(\theta \mid D, M_1)$ via *sampling importance-resampling (SIR)* [43, 44], which proceeds as follows. Define the unnormalised importance weights as a ratio of the unnormalised posteriors:

$$w_i = \frac{L(D \mid \theta_i, M_1)\, p(\theta_i \mid M_1)}{L(D \mid \theta_i, M_0)\, p(\theta_i \mid M_0)}, \quad i = 1, \ldots, N. \tag{3}$$

Furthermore, let

$$\tilde{w}_i = \frac{w_i}{\sum_{j=1}^N w_j} \tag{4}$$

denote the self-normalised weights. We now obtain a new set of unweighted samples, which we will call $\{\theta_i^*\}_{i=1}^N$, by resampling from the original set $\{\theta_i\}_{i=1}^N$ with replacement, where the probability of selecting each θ_j is $\tilde{w}_j$. Equivalently, the resampled draws are i.i.d. from the empirical distribution $\sum_{j=1}^N \tilde{w}_j\, \delta_{\theta_j}(\theta)$, where the $\delta_{\theta_j}(\theta)$ are delta functions centred at θ_j. The samples θ_i^* are then distributed approximately according to the target distribution $p(\theta \mid D, M_1)$, since

$$p(\theta \mid D, M_1) \propto \frac{L(D \mid \theta, M_1)\, p(\theta \mid M_1)}{L(D \mid \theta, M_0)\, p(\theta \mid M_0)}\, p(\theta \mid D, M_0). \tag{5}$$

For valid importance sampling, the support of the new posterior (M_1) must be contained within the original posterior (M_0). Specifically, the mass of the new posterior should be well-represented by the original samples; we discuss diagnostics to assess these conditions below.

Armed with the resampled draws, we can compute the evidence for the alternative model M_1 using the LHME. We define the target density $\varphi_1(\theta)$ for the new model M_1, which can be constructed either by reusing the original learned target $\varphi_0(\theta)$, or by re-training it on the resampled draws. The reciprocal evidence for M_1 can then be estimated as:

$$\hat{\rho}_1 \approx \frac{1}{N} \sum_{i=1}^N \frac{\varphi_1(\theta_i^*)}{L(D \mid \theta_i^*, M_1)\, p(\theta_i^* \mid M_1)}, \quad \theta_i^* \sim \sum_{j=1}^N \tilde{w}_j\, \delta_{\theta_j}(\theta). \tag{6}$$

We now specialise to the case of prior sensitivity analysis, where the models M_0 and M_1 share the same likelihood function but differ in their prior specifications. The importance weights simplify to:

$$w_i = \frac{p(\theta_i \mid M_1)}{p(\theta_i \mid M_0)}, \quad i = 1, \ldots, N \tag{7}$$

since the likelihood terms cancel. The resampled draws $\{\theta_i^*\}$ are then approximately distributed according to the posterior under the new prior $p(\theta \mid D, M_1)$.

The new target $\varphi_1(\theta)$ is then constructed as follows. If the posterior is relatively insensitive to the prior change (guidance for determining this is given below), we reuse the original learned target as the new target, so that $\varphi_1(\theta) = \varphi_0(\theta)$. The evidence for the alternative prior is then computed immediately with Eq. 6. This is extremely cheap, requiring only evaluations of the new prior on existing samples. On the other hand, if the posterior changes significantly we re-train $\varphi_1(\theta)$ on the resampled draws. This can be fine-tuned from the original learned target for faster convergence, and typically takes only tens of seconds to a few minutes for low-dimensional problems. In computationally demanding settings, this is still orders of magnitude faster than a full model re-fit. For clarity, (model) *re-fit* refers to re-running the entire sampling procedure to obtain new posterior samples, while *re-train* refers to only re-training the normalising flow target $\varphi(\theta)$ on resampled draws.

We employ several diagnostics to assess the reliability of the estimates.

- First, we monitor the Pareto-$\hat{k}$ diagnostic, which estimates the shape parameter k of the generalised Pareto distribution fitted to the upper tail of the importance weights [45]. Importance sampling is typically empirically unreliable if $\hat{k}$ exceeds 0.7, so we adopt this as a heuristic to indicate that model-refit is necessary.

- Second, we use the effective sample size (ESS) of the importance weights,

$$\mathrm{ESS} = \frac{\left(\sum_{i=1}^{N} w_i\right)^2}{\sum_{i=1}^{N} w_i^2}, \tag{8}$$

to determine whether re-training of $\varphi(\theta)$ is required. Since the fractional ESS measures the overlap between the original and new posteriors, we recommend re-training if this drops below 0.95 (as a conservative default) to maintain low variance in the evidence estimate.

These checks are performed in addition to standard LHME diagnostics (see [12] for details). We also find that if the posterior is insensitive to the prior and weights are all equal, Pareto fitting diverges. In this case, when the maximum and minimum weights are equal to within some tolerance, we return $\hat{k} = -\infty$ directly as the $\hat{k}$ that reproduces a delta distribution for the weights in the limit.

We note that the above approach can be straightforwardly applied to other stabilised harmonic mean estimators (e.g. [46, 47, 48]), which differ in the construction of the target $\varphi(\theta)$, as well as other methods that require only posterior samples to compute the evidence [49, 50, 51, 52]. This work focuses on the learned variant due to its well-tested performance in real applications [53, 54, 55, 56, 57, 58, 59, 60, 61].

A schematic of the overall workflow is shown in Fig. 1.

3. RESULTS

We first present toy experiments to validate accuracy and diagnostics, then apply the method to a cosmological case study, comparing against full MCMC and nested sampling re-fits and report computational savings. In what follows, we report the natural logarithm of the evidence, $\log Z = \log p(D \mid M)$, whose estimate and uncertainty are obtained by inverting $\hat{\rho}$ and performing standard error propagation [12, 13].

3.1 Toy examples

We demonstrate the method on two toy examples where the ground truth evidence under alternative priors is available for comparison, and assess the performance of the diagnostics.

Gaussian in 10 dimensions. We first consider a 10-dimensional Gaussian problem to assess sensitivity to changes in prior strength. We use a Gaussian likelihood with zero mean and diagonal covariance (standard deviation $\sigma = 2 \times 10^{-4}$) and an original prior with zero mean and diagonal covariance ($\sigma = 1.0$). Posterior samples are taken with the No-U-Turn Sampler (NUTS) [62] for 16 chains of 1000 samples. We then evaluate the evidence under a set of alternative zero-mean Gaussian priors with widths ranging from $\sigma = 10^{-1.5}$ to 10^{-4}, as shown in Fig. 2a. The analytical evidence is available in closed form for comparison.

Fig. 3 shows the resampled evidence estimates and diagnostics. We report $\log Z$, the (non-reciprocal) log-evidence. The posterior is insensitive to the prior change for the four widest alternative priors, maintaining a fractional ESS close to 1.0. No re-training of $\varphi(\theta)$ is necessary. The second narrowest prior approaches the scale of the likelihood, so fractional ESS drops to around 0.65 and re-training of $\varphi(\theta)$ is required. Finally, the narrowest prior is more informative than the likelihood, causing the posterior to drastically change (ESS $\approx$ 0). Here, the importance sampled evidence estimate deviates significantly from the analytical value, which is correctly diagnosed by a Pareto-$\hat{k} > 0.7$.

Rosenbrock in two dimensions. We now assess the evidence for a sequence of informative priors against a broad baseline using the standard Rosenbrock likelihood ($a = 1$, $b = 100$). The baseline prior is uniform on $[-10, 10]^2$, and the original posterior was obtained with NUTS (32 chains of 2000 samples). For the alternative hypotheses we use narrow Gaussian priors ($\Sigma \approx 0.06$): the first is centred at the likelihood mode $(1, 1)$ and subsequent priors are shifted along the y-axis to emulate priors that increasingly conflict with the data. Contours of the posterior and alternative priors are shown in Fig. 2b.

Evidence estimates with diagnostics appear in Fig. 4. Here, ESS is low and re-training of $\varphi(\theta)$ is necessary for all priors. All evidences are consistent with ground truth (obtained by Riemann integration across a fixed grid) except for the prior with maximal shift, which is correctly flagged by Pareto-$\hat{k}$ as unreliable.

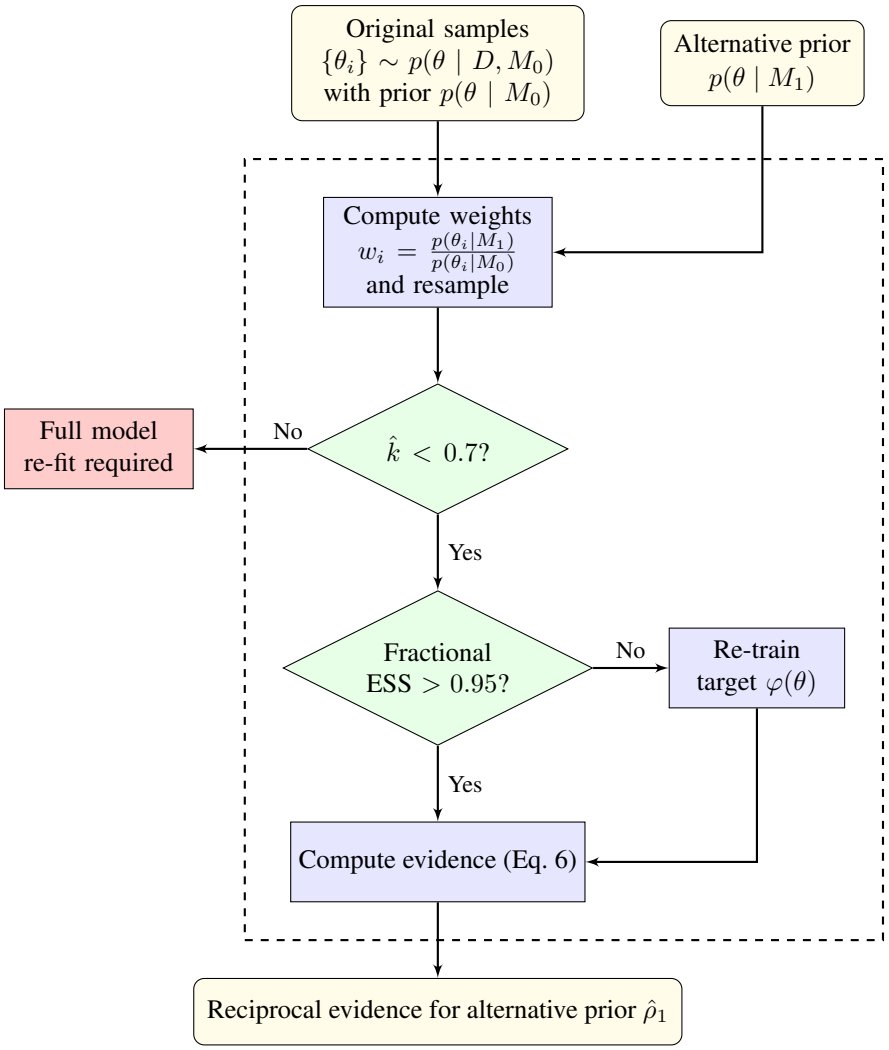


Figure 1: *Schematic of the prior sensitivity workflow. We first resample existing posterior draws to the posterior under the new prior. We then perform a two-stage diagnostic check: 1. if the Pareto-$\hat{k}$ diagnostic exceeds 0.7, the importance weights are unreliable and a full model re-fit is required; 2. if $\hat{k}$ is safe but fractional ESS is low (< 0.95), we re-train the learned target $\varphi(\theta)$ to maintain low variance (blue). After the two checks, the evidence for the alternative prior is computed by Eq. 6.*

3.2 Cosmological example

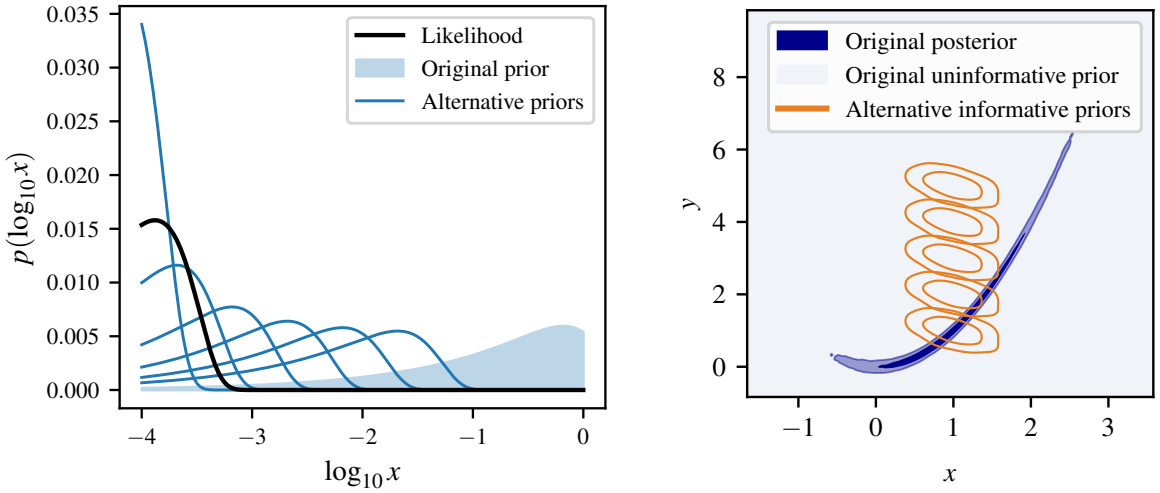
We apply our method to a cosmological case study comparing logarithmic and uniform priors for a scale parameter, which has invoked significant debate in the literature [63, 64, 65, 66, 67]. We consider the case of the tensor-to-scalar ratio of primordial perturbations r, reusing (Metropolis-Hastings) posterior samples provided by [65] and comparing against their reported nested sampling evidences.

The original analysis in [65] considers three different priors for r:

$$\log_{10} r \sim \mathcal{U}(-10, 0), \quad \log_{10} r \sim \mathcal{U}(-5, 0), \quad \text{and} \quad r \sim \mathcal{U}(0, 1). \tag{9}$$

We start with MCMC samples for the $\log_{10} r \sim \mathcal{U}(-10, 0)$ prior as the original posterior, and importance resample to the two alternative priors to obtain the evidences. We note that the error estimate of the LHME relies on the variance of the mean across from many MCMC chains, which can be noisy when the number of chains is low. A more reliable uncertainty can be obtained by bootstrap resampling the chains, taking at least 30 replicate chains, and computing the resulting standard deviation. The samples provided by [65] have 8 chains, so we adopt this approach. While this takes longer than the standard error estimate, computational costs remains on the order of minutes and orders of magnitude faster than a full model re-fit.

We recover evidences consistent with standard error to both the nested sampling re-fits reported in [65], and MCMC



(a) *Gaussian densities* ***(b)*** *Rosenbrock contours*

Figure 2: *(a) Illustration of prior widths versus the likelihood for the Gaussian toy example, in terms of the 1D marginal density (log-scale). (b) Rosenbrock posterior contours under an uninformative uniform prior (navy) and contours for alternative informative Gaussian priors (red) for different shifts along the y-axis. The original uninformative prior (light blue) covers the entire plotted region.*

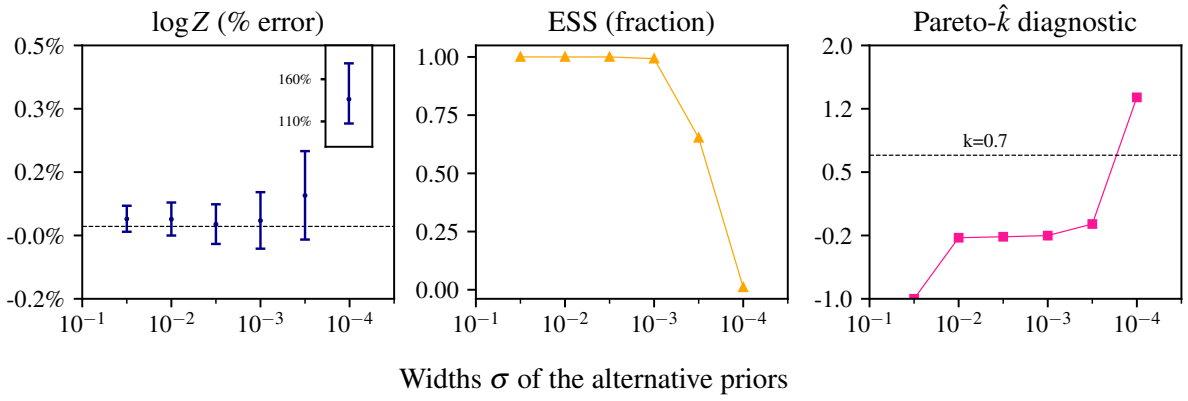


Figure 3: *Prior dependence for the Gaussian example. The width of the likelihood is 2×10^{-4}, which is between the narrowest and second narrowest priors considered. Left: Percent errors in the evidence computed with the LHME using importance resampled posterior draws, compared to the analytical solution as the strength of the prior is varied. Middle: Fractional ESS of the importance sampling. Right: Pareto-$\hat{k}$ diagnostic values for importance sampling. The evidence estimate is correctly flagged as unreliable when $\hat{k}$ is high (above 0.7).*

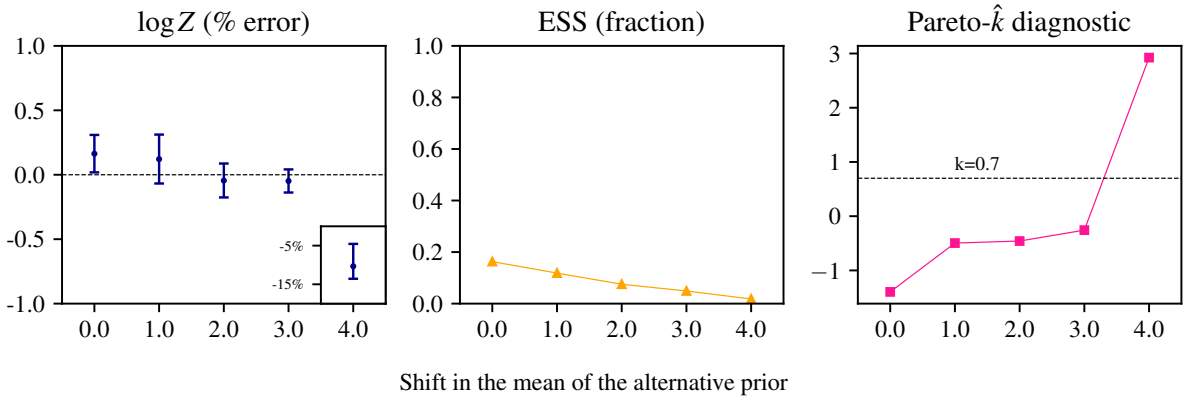


Figure 4: *Same as Fig. 3, but for the Rosenbrock example with informative priors at different locations. Here ESS is low throughout and re-training $\varphi(\theta)$ is necessary. Pareto-$\hat{k} > 0.7$ successfully identifies the case where the evidence estimate is unreliable.*

re-fits combined with LHME, as shown in Table 1. The variance is higher for the $r \sim \mathcal{U}(0,1)$ prior due to the larger change from the original posterior, but remains consistent and sufficient for practical use where order unity changes in log evidence are of interest. For the two alternative priors, the fractional ESS are 0.41 and 0.063 respectively, indicating moderate and significant changes in the posterior, so we re-train $\varphi(\theta)$ in both cases. Pareto-$\hat{k}$ diagnostics are $-\infty$ and 0.10 ± 0.01, both below the 0.7 threshold. $\hat{k} = -\infty$ for the $\log_{10} r \sim \mathcal{U}(-5,0)$ alternative prior indicates that the posterior is insensitive to the prior change, so the reduced ESS is due to the change in boundary of the prior support rather than a significant shift in the posterior mass.

The computational savings for our method are listed in the rightmost column of Table 1. The original MCMC chains were run on 32 CPU cores, so we report CPU core-hours for consistency. Timings for the nested sampling runs were not recorded, but are typically within similar order of magnitude to MCMC. The resampled evidence estimates with bootstrapped uncertainties take 10 and 30 minutes on a single CPU core for the two alternative priors respectively, which is around $3000\times$ and $6000\times$ faster than the full re-fits.

A minor note is that the nested sampling evidences we report here differ slightly from [65], because cosmological theory codes typically reject certain unphysical parameter combinations at the level of the prior. The raw nested sampling output is instead the evidence for a prior restricted to the physically allowed region. Quantities shown in Table 1 are the corrected values for the actual prior used, which were reported in the code release of [65] but not the paper. The difference is not qualitatively significant, and one could even argue that the uncorrected evidence is more meaningful as a prior only over the physically allowed parameter space (following the discussion in Section 1). Nevertheless, we report the corrected version here for consistency between methods.

Table 1: *Log-evidence estimates ($\log Z$) for the cosmological model under different priors. Our resampling method (bottom row in each block) reproduces evidence calculations to within the quoted uncertainties, while avoiding computationally expensive full re-runs. Uncertainties for the resampled estimates are obtained via bootstrap resampling of the original MCMC samples. The rightmost column shows computational time in CPU core-hours for each method. Our resampled LHME approach provides savings in computational cost of around $3000\times$ and $6000\times$.*

Prior	**Method**	**log Z**	**Time (CPU-hours)**
Original	Nested sampling	-1432.43 ± 0.22	–
$\log_{10} r \sim \mathcal{U}(-10,0)$	MCMC + LHME	-1432.75 ± 0.01	1152
Alternative 1	Nested sampling	-1432.98 ± 0.22	–
$\log_{10} r \sim \mathcal{U}(-5,0)$	MCMC + LHME	-1432.98 ± 0.02	1440
	Resampled LHME (ours)	-1432.94 ± 0.06	0.5
Alternative 2	Nested sampling	-1435.44 ± 0.23	–
$r \sim \mathcal{U}(0,1)$	MCMC + LHME	-1435.57 ± 0.02	1056
	Resampled LHME (ours)	-1435.43 ± 0.33	0.16

4. CONCLUSION

We present a fast method to assess prior sensitivity of the Bayesian evidence, which is agnostic to the method used to generate samples, by resampling existing posterior draws. A two-stage diagnostic check for unreliable importance sampling is introduced to avoid unnecessary recalculation and check for failure. We validate on toy problems and a cosmological case study, reproducing Bayesian evidences obtained from full MCMC or nested sampling runs at several thousand times lower computational cost.

The learned harmonic mean estimator makes this efficient post-hoc approach possible by decoupling sampling and evidence calculation. Other post-hoc computational schemes that make use of these advantages are in development, such as to combine evidences across datasets, requiring only posterior samples from each dataset obtained in any manner [68].

Code for the method will be made publicly available in the `harmonic` [1] package. We hope that the increased accessibility brought by the method will encourage transparent prior dependence reporting in Bayesian model comparison applications, which is critical for robust scientific inferences.

References

[1] Leo Breiman. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, August 2001.

[2] Galit Shmueli. To Explain or to Predict? *Statistical Science*, 25(3):289–310, August 2010.

[3] Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*, 24(6):997–1016, November 2014.

[4] Juho Piironen and Aki Vehtari. Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, 27(3):711–735, May 2017.

[5] Robert E. Kass and Adrian E. Raftery. Bayes Factors. *J. Am. Statist. Assoc.*, 90(430):773–795, 1995.

[6] David J. C. MacKay. *Information Theory, Inference and Learning Algorithms*. Cambridge University Press, September 2003.

[7] Roberto Trotta. Bayes in the sky: Bayesian inference and model selection in cosmology. *Contemporary Physics*, 49(2):71–104, March 2008.

[8] John Skilling. Nested Sampling. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume 735, pages 395–405. AIP, November 2004.

[9] John Skilling. Nested sampling for general Bayesian computation. *Bayesian Analysis*, 1(4):833–859, December 2006.

[10] Xiao-Li Meng and Wing Hung Wong. Simulating Ratios of Normalizing Constants Via a Simple Identity: A Theoretical Exploration. *Statistica Sinica*, 6(4):831–860, 1996.

[11] Quentin F. Gronau, Henrik Singmann, and Eric-Jan Wagenmakers. Bridgesampling: An R Package for Estimating Normalizing Constants. *Journal of Statistical Software*, 92:1–29, February 2020.

[12] Jason D. McEwen, Christopher G. R. Wallis, Matthew A. Price, and Alessio Spurio Mancini. Machine learning assisted Bayesian model comparison: Learnt harmonic mean estimator, November 2023.

[13] Alicja Polanska, Matthew A. Price, Davide Piras, Alessio Spurio Mancini, and Jason D. McEwen. Learned harmonic mean estimation of the Bayesian evidence with normalizing flows. *The Open Journal of Astrophysics*, 8, October 2025.

[14] Sandip Sinharay and Hal S. Stern. On the Sensitivity of Bayes Factors to the Prior Distributions. *The American Statistician*, 56(3):196–201, 2002.

[1] `https://github.com/astro-informatics/harmonic`

[15] D. V. Lindley. A Statistical Paradox. *Biometrika*, 44(1-2):187–192, June 1957.

[16] M. S. Bartlett. A comment on D. V. Lindley's statistical paradox. *Biometrika*, 44(3-4):533–534, December 1957.

[17] Andrew Gelman, Daniel Simpson, and Michael Betancourt. The Prior Can Often Only Be Understood in the Context of the Likelihood. *Entropy*, 19(10):555, October 2017.

[18] Fernando Llorente, Luca Martino, David Delgado, and Javier Lopez-Santiago. Marginal likelihood computation for model selection and hypothesis testing: An extensive review. *SIAM Review*, 65(1):3–58, February 2023.

[19] Rafael E Carrillo, Jason D McEwen, and Yves Wiaux. Purify: a new approach to radio-interferometric imaging. *Monthly Notices of the Royal Astronomical Society*, 439(4):3591–3604, 2014.

[20] Peter M Lee. *Bayesian statistics*. Oxford University Press London:, 1989.

[21] Edwin T. Jaynes. Prior Probabilities. *IEEE Transactions on Systems Science and Cybernetics*, 4(3):227–241, September 1968.

[22] Matthew A Price, Jason D McEwen, L Pratley, and Thomas D Kitching. Sparse bayesian mass-mapping with uncertainties: Full sky observations on the celestial sphere. *Monthly Notices of the Royal Astronomical Society*, 500(4):5436–5452, 2021.

[23] Jason D McEwen, Tobías I Liaudat, Matthew A Price, Xiaohao Cai, and Marcelo Pereyra. Proximal nested sampling with data-driven priors for physical scientists. In *Physical Sciences Forum*, volume 9, page 13. MDPI, 2023.

[24] Benjamin Remy, Francois Lanusse, Niall Jeffrey, Jia Liu, J-L Starck, Ken Osato, and Tim Schrabback. Probabilistic mass-mapping with neural score estimation. *Astronomy & Astrophysics*, 672:A51, 2023.

[25] Tobías I Liaudat, Matthijs Mars, Matthew A Price, Marcelo Pereyra, Marta M Betcke, and Jason D McEwen. Scalable bayesian uncertainty quantification with data-driven priors for radio interferometric imaging. *RAS Techniques and Instruments*, 3(1):505–534, 2024.

[26] Jessica J Whitney, Tobías I Liaudat, Matthew A Price, Matthijs Mars, and Jason D McEwen. Generative modelling for mass-mapping with fast uncertainty quantification. *Monthly Notices of the Royal Astronomical Society*, 542(3):2464–2479, September 2025.

[27] Justin Alsing and Will Handley. Nested sampling with any prior you like. *Monthly Notices of the Royal Astronomical Society: Letters*, 505(1):L95–L99, 2021.

[28] Anthony O'Hagan. Fractional bayes factors for model comparison. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):99–118, 1995.

[29] James O. Berger and Luis R. Pericchi. The Intrinsic Bayes Factor for Model Selection and Prediction. *Journal of the American Statistical Association*, 91(433):109–122, 1996.

[30] Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, New York, 3 edition, November 2013.

[31] James O. Berger. Robust Bayesian analysis: Sensitivity to the prior. *Journal of Statistical Planning and Inference*, 25(3):303–328, July 1990.

[32] Paul Gustafson. Local robustness in bayesian analysis. In David Ríos Insua and Fabrizio Ruggeri, editors, *Robust Bayesian Analysis*, pages 71–88. Springer New York, New York, NY, 2000.

[33] Harry Bevins, Will Handley, Pablo Lemos, Peter Sims, Eloy de Lera Acedo, and Anastasia Fialkov. Marginal Bayesian Statistics Using Masked Autoregressive Flows and Kernel Density Estimators with Examples in Cosmology. *Physical Sciences Forum*, 5(1):1, 2022.

[34] Vrund Patel, Amlan Chakraborty, and Luca Amendola. The prior dependence of the DESI results, October 2024.

[35] Nicolas Chopin and Christian P. Robert. Properties of nested sampling. *Biometrika*, 97(3):741–755, September 2010.

[36] Ewan Cameron and Anthony Pettitt. Recursive Pathways to Marginal Likelihood Estimation with Prior-Sensitivity Analysis. *Statistical Science*, 29(3):397–419, August 2014.

[37] James O. Berger, Elías Moreno, Luis Raul Pericchi, M. Jesús Bayarri, José M. Bernardo, Juan A. Cano, Julián De la Horra, Jacinto Martín, David Ríos-Insúa, Bruno Betrò, A. Dasgupta, Paul Gustafson, Larry Wasserman, Joseph B. Kadane, Cid Srinivasan, Michael Lavine, Anthony O'Hagan, Wolfgang Polasek, Christian P. Robert, Constantinos Goutis, Fabrizio Ruggeri, Gabriella Salinetti, and Siva Sivaganesan. An overview of robust Bayesian analysis. *Test*, 3(1):5–124, June 1994.

[38] Noa Kallioinen, Topi Paananen, Paul-Christian Bürkner, and Aki Vehtari. Detecting and diagnosing prior and likelihood sensitivity with power-scaling. *Statistics and Computing*, 34(1):57, December 2023.

[39] A. E. Gelfand and D. K. Dey. Bayesian Model Choice: Asymptotics and Exact Calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3):501–514, January 1994.

[40] 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):3–48, 1994.

[41] Radford M. Neal. 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, 1994.

[42] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing Flows for Probabilistic Modeling and Inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

[43] Donald B. Rubin. The Calculation of Posterior Distributions by Data Augmentation: Comment: A Noniterative Sampling/Importance Resampling Alternative to the Data Augmentation Algorithm for Creating a Few Imputations When Fractions of Missing Information Are Modest: The SIR Algorithm. *Journal of the American Statistical Association*, 82(398):543–546, 1987.

[44] A. F. M. Smith and A. E. Gelfand. Bayesian Statistics without Tears: A Sampling-Resampling Perspective. *The American Statistician*, 46(2):84–88, 1992.

[45] Aki Vehtari, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. Pareto Smoothed Importance Sampling. *Journal of Machine Learning Research*, 25(72):1–58, 2024.

[46] C. P. Robert and D. Wraith. Computational methods for Bayesian model choice. *AIP Conference Proceedings*, 1193(1):251–262, December 2009.

[47] Martin Metodiev, Marie Perrot-Dockès, Sarah Ouadah, Nicholas J. Irons, Pierre Latouche, and Adrian E. Raftery. Easily Computed Marginal Likelihoods from Posterior Simulation Using the THAMES Estimator. *Bayesian Analysis*, 20(3):1003–1030, September 2025.

[48] Dana Naderi, Christian P. Robert, Kaniav Kamary, and Darren Wraith§. Approximating evidence via bounded harmonic means, October 2025.

[49] Alan Heavens, Yabebal Fantaye, Arrykrishna Mootoovaloo, Hans Eggers, Zafiirah Hosenie, Steve Kroon, and Elena Sellentin. Marginal Likelihoods from Monte Carlo Markov Chains, April 2017.

[50] He Jia and Uros Seljak. Normalizing Constant Estimation with Gaussianized Bridge Sampling. In *Proceedings of The 2nd Symposium on Advances in Approximate Bayesian Inference*, pages 1–14. PMLR, February 2020.

[51] Rahul Srinivasan, Marco Crisostomi, Roberto Trotta, Enrico Barausse, and Matteo Breschi. Bayesian evidence estimation from posterior samples with normalizing flows. *Physical Review D*, 110(12):123007, December 2024.

[52] Stefano Rinaldi, Gabriele Demasi, Walter Del Pozzo, and Otto A. Hannuksela. Hierarchical inference of evidence using posterior samples, May 2024.

[53] A Spurio Mancini, M M Docherty, M A Price, and J D McEwen. Bayesian model comparison for simulation-based inference. *RAS Techniques and Instruments*, 2(1):710–722, January 2023.

[54] Davide Piras, Alicja Polanska, Alessio Spurio Mancini, Matthew A. Price, and Jason D. McEwen. The future of cosmological likelihood-based inference: Accelerated high-dimensional parameter estimation and model comparison. *The Open Journal of Astrophysics*, 7, September 2024.

[55] A. Polanska, T. Wouters, T. H. Pang, K. W. K. Wong, and J. D. McEwen. Accelerated bayesian parameter estimation and model selection for gravitational waves with normalizing flows. In *Proceedings of the Machine Learning and Physical Sciences Workshop as part of the 38th International Conference on Neural Information Processing Systems (NeurIPS)*, Sep 2024.

[56] A. Spurio Mancini, K. Lin, and J. D. McEwen. Field-level cosmological model selection: Field-level simulation-based inference for Stage IV cosmic shear can distinguish dynamical dark energy, October 2024.

[57] S. Paradiso, G. McGee, and W.J. Percival. Evaluating extensions to LCDM: An application of Bayesian model averaging and selection. *Journal of Cosmology and Astroparticle Physics*, 2024(10):021, October 2024.

[58] Richard Stiskalek and Harry Desmond. Symmetry in Hyper Suprime-Cam Galaxy Spin Directions. *Research Notes of the AAS*, 8(11):281, November 2024.

[59] Karim Carrion, Alessio Spurio Mancini, Davide Piras, and Juan Carlos Hidalgo. Testing interacting dark energy with Stage IV cosmic shear surveys through differentiable neural emulators. *Monthly Notices of the Royal Astronomical Society*, 539(4):3220–3228, June 2025.

[60] Guo-Hong Du, Peng-Ju Wu, Tian-Nuo Li, and Xin Zhang. Impacts of dark energy on weighing neutrinos after DESI BAO. *The European Physical Journal C*, 85(4):392, April 2025.

[61] Richard Stiskalek, Harry Desmond, Julien Devriendt, Adrianne Slyz, Guilhem Lavaux, Michael J. Hudson, Deaglan J. Bartlett, and Hélène M. Courtois. The Velocity Field Olympics: Assessing velocity field reconstructions with direct distance tracers. *Monthly Notices of the Royal Astronomical Society*, November 2025.

[62] Matthew D. Hoffman and Andrew Gelman. The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, January 2014.

[63] Fergus Simpson, Raul Jimenez, Carlos Pena-Garay, and Licia Verde. Strong Bayesian Evidence for the Normal Neutrino Hierarchy. *Journal of Cosmology and Astroparticle Physics*, 2017(06):029–029, June 2017.

[64] T. Schwetz, K. Freese, M. Gerbino, E. Giusarma, S. Hannestad, M. Lattanzi, O. Mena, and S. Vagnozzi. Comment on "Strong Evidence for the Normal Neutrino Hierarchy", March 2017.

[65] Lukas T. Hergt, Will J. Handley, Michael P. Hobson, and Anthony N. Lasenby. Bayesian evidence for the tensor-to-scalar ratio r and neutrino masses m_ν: Effects of uniform vs logarithmic priors. *Physical Review D*, 103(12):123511, June 2021.

[66] Raul Jimenez, Carlos Pena-Garay, Kathleen Short, Fergus Simpson, and Licia Verde. Neutrino Masses and Mass Hierarchy: Evidence for the Normal Hierarchy. *Journal of Cosmology and Astroparticle Physics*, 2022(09):006, September 2022.

[67] Stefano Gariazzo, Martina Gerbino, Thejs Brinckmann, Massimiliano Lattanzi, Olga Mena, Thomas Schwetz, Shouvik Roy Choudhury, Katherine Freese, Steen Hannestad, Christoph A. Ternes, and Mariam Tórtola. Neutrino mass and mass ordering: No conclusive evidence for normal ordering. *Journal of Cosmology and Astroparticle Physics*, 2022(10):010, October 2022.

[68] Zixiao Hu and Jason D. McEwen. Fast and flexible Bayesian model comparison with emulated likelihoods. In preparation, 2026.